Journal of Computer Science

Research Article

Time Series Analysis for Predicting Tea Harvest Yield: A SARIMAX-Based Approach

Pallavi Nagpal, Deepika Chaudhary and Jaiteg Singh

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

Article history
Received: 20-02-2025
Revised: 02-09-2025
Accepted: 04-10-2025

Corresponding Author:
Deepika Chaudhar
Chitkara University Institute of Engineering and Technology,
Chitkara University, Punjab, India
Email: deepika.chaudhary@chitkara.edu.in

Abstract: Precise prediction of tea yield is crucial for both agricultural planning and economic forecasting. Projection of the future trends, which rely on present and historical data, is the process termed 'forecasting'. Yield prediction is fundamental for research and development. By gathering the yield data over historic times, the researchers can make their work more valuable by predicting the yield patterns. This can help in analyzing the impact of changing environmental variables that can lead to a change in the yield prediction. To predict tea yield on the basis of historical yield and meteorological variables, this study proposed the optimal use of the Seasonal Autoregressive Integrated Moving Average with Exogenous Variable (SARIMAX) model, which was applied to the historical data from the years 1985-2022. The model integrates both the seasonality of tea production and external factors that influence the crop growth, such as rainfall, temperature, etc., which are known to influence crop growth. Two of the competing models, SARIMAX (1,1,1) and SARIMAX (1,0,0) (0,0,1,12), were applied and validated on statistical parameters log-likelihood, AIC, BIC, residual diagnostics, the Ljung-Box test, and the Q-statistic. The results showed that the hyperparameter-tuned model SARIMAX (1,0,0) (0,0,1,12) successfully captured both temporal and seasonal patterns. This model yielded a lower AIC (-553.70) and exhibited consistent residuals, normally distributed and free from autocorrelation. The results indicate the robustness of SARIMAX in yield predictions and also highlight its role in the planning and framing of agricultural policies.

Keywords: Forecasting, Tea Crop, Prediction, Algorithms, SARIMAX, Time Series

Introduction

Situated in the breathtaking Himalayan territory of Himachal Pradesh in India, Kangra is known worldwide for its outstanding tea, which continues to serve as a crucial element of its economy and heritage. In addition to its rainfall, elevation, and soil conditions, Kangra's natural and picturesque surroundings provide all the ideal circumstances for growing tea. The projection of the future trends, which rely on present and historical data, is the process termed as forecasting. The term forecasting is a game-changing tool while taking important decisions in a number of fields, from pricing to weather and from weather to yield (Dharmaraja et al., 2020). Forecasting leads to the curtailing of doubts. For forecasting, a number of methods are being used. This can range from the use of easy speculation to the statistical models. Inspecting a large number of variables in the study will help to identify the type of technique to be used. Commonly applied approaches are Time series analysis, Regression techniques, and various hybrid models. It is always a challenging task to predict yield. Accurate predictions of yield make many decisions better in terms of weather-related impacts, natural disasters, resource utilization, economy enhancements, and other important tasks like planning for policy making (Van Klompenburg et al., 2020). This practice emphasizes the use of accurate tools and methods so that accurate predictions can be made. Precision agriculture is valuable for optimizing agricultural practices and improving efficiency. Well-informed predictions of crop yield make the planting and harvesting schedules better.

Yield prediction is fundamental for research and development. By gathering the yield data over historic times, the researchers can make their work more valuable. Such data enables the analysis of changing environmental variables and their influence on crop yields, thereby improving prediction accuracy. Moreover, yield forecasting

plays a crucial role in shaping global trade (Bali and Singla, 2022). International trade also depends on the future forecasts, which directly help in reducing costs and balancing the supply chain. In agriculture, better tea yield prediction has a lot more issues to check for Figure 1 illustrates this in a graphical manner Time series models are advantageous over traditional machine learning models for several reasons. The time series models have the capability to read the patterns and the effect of the previous data on the new data, i.e., lag effects and the self-similarity, i.e., autocorrelation, whereas in traditional machine learning algorithms, the lag features need to be manually engineered. The time series models are well capable of understanding the seasonal patterns along with the long-term trends, whereas the other traditional machine learning models are not capable of handling long-term trends as well as the seasonal patterns, as they have to be explicitly feature-engineered for this. Minimum efforts are required for feature engineering in time series models. Time series models have strong exploration capabilities, whereas in traditional machine learning models, to predict the future values, we need to do special settings (specialized pre-processing), i.e., to make the lag features, or make trend or seasonal settings. Time series models are directly concerned with the domain-specific effects, such as how strong the trend is or what the seasonality magnitude is, whereas in traditional machine learning models, the interpretation is not time-based. This study helps in exploring the field of forecasting tea crop yields in the Kangra Valley by using time series machine learning models. The objective of our study is to estimate tea yields precisely in the Kangra region. The data set used in the underlying study is secondary data, which is collected from two organizations, i.e., the Tea Board of Palampur, from where tea production data was collected, and the other, weather data was collected from the Agriculture University Palampur, like Maximum temperature, rainfall, soil temperature, etc., from 1985 to 2022.

The dataset spans 38 years. For model development, we used an 80:20 split. The training set comprises 30 years, and for testing, the next 8 years ensured sufficient historical depth to capture both the short-term dynamics and the long-term cycles.

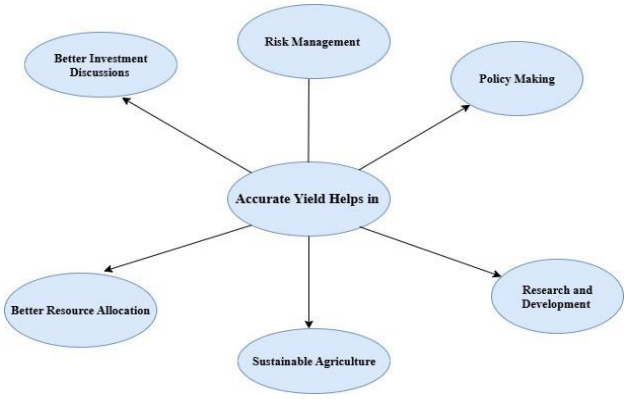


Fig. 1: Multi-Facets of Yield Prediction

In order to overcome the constraints of the past studies, which focus on a short time span and univariate forecasting, this study focuses on the SARIMAX framework with exogenous variables. The time series models perform well with smaller data sets as they are based on the statistical properties of time-dependent data and can reasonably forecast even if the data is limited by identifying patterns present in the historical data. SARIMA incorporates seasonal variations, making it a candidate for agricultural yield forecasting. This study seeks to:

(a) Examine how Kangra's tea output has been impacted by past weather trends and agricultural techniques
(b) Predict tea crop yields using time series-based models
(c) Offer practical advice and insights to the Kangra tea growers and help them make better decisions.

Literature Review

In this section, an in-depth literature review on time series models is presented. The papers for the purpose of review were selected from Scopus-indexed journals and were of high impact. The estimation of the production of any crop by using historical data and other influential factors is of utmost importance. Better forecasting helps in optimizing the resource planning and decreasing the losses related to post- harvest. Better forecasting also helps in stabilizing the market trends. Thus, the forecasting techniques play an important role in food security (Sam et al., 2019). Traditional forecasting methods like linear regression have been applied in past research, and these methods are not appropriate for modelling the nonlinear dynamics due to multiple interacting variables. Recently speaking, ARIMA has become evident as a leading time series forecasting technique because of its ability to capture complex temporal patterns. Specifically, the datasets that exhibit trends and seasonality, for them, ARIMA is particularly well-suited (Ensafi et al., 2022). The advanced approaches like ARIMA and its extended versions have helped a lot in the field of forecasting, which helps in capturing the complex hidden patterns in the historical data. The evaluation of the models is being done on the basis of reliable metrics like AIC, BIC, RMSE, etc. (Ospina et al., 2023). This study seeks to evaluate the role of advanced ARIMAX models in forecasting crop production, including the influence of exogenous variables. The effectiveness of the model is evaluated using statistical metrics like AIC and BIC (Alawsi et al., 2022). A study by Nipa et al. applied the SARIMAX model to predict tea yield in Bangladesh. The historical data was collected from 2005 to 2017.AIC was the primary criterion used to choose the best models. These models may enhance the prediction of tea results in the years to come, more accurately forecasting future climatic conditions.

The investigation of the variance of meteorological variables in Moulvibazar and Sylhet, including relative rainfall, humidity, wind speed, light duration, and both maximum and minimum temperatures, was done by using a seasonal time series dataset (Begum et al., 2024). Deka et al. (2022) described the use of a time series model for India's yearly tea production, called the ARIMA model. R programming was employed in the study's analysis. The Ministry of Commerce's Tea Board of India provided the secondary data for the analysis, which spans 70 years from 1947 to 2016. Using several diagnostic and assessment standards, ARIMA (1,1,2) was the best-fitted model. Then, Raj et al. (2019) suggested that the tea yield can be predicted with the help of the SARIMAX models. The data was collected from the years 1981- 2005. The relation between the climatic variables and the tea yield was measured. The other models were also employed, such as SMLR and ANN. Reddy and Sureshbabu (2023) proposed that their study forecast the rice yield for the place named Ananthapura in Andhra Pradesh. The data was collected from 2008 to 2014, and the prediction was done for the next four years according to the seasons. The current study elaborated the fact that the use of the SARIMA model outperforms other existing models (Reddy et al., 2020).

Further, Mishra et al. (2024) investigated and forecasted the production of potatoes by using a Time Series model, and the training of the model was done on the data for the year 1961- 2009 in 8 South Asian countries. Respective countries and their predictions for potato yield were thoroughly discussed along with the choice of the model. The analysis of the study proposed that the ARIMA model gave the least and the lowest forecast errors. Further, (Goyal et al., 2024) proposed that the ARIMAX model outperforms the best while comparing it with the ARIMA model for predicting the wheat yield with respect to climate change. The lower error metrics proved to be the turning points in proving the above statement. Later on, Mirpulatov et al. (2023), in their proposed findings, contrast the Neural Prophet, the SARIMAX, and the Prophet model for the crop yield of Sugar beet and Soybean. The favorable values of R2 were given by SARIMAX for the crop yield of soybean, and on the other hand, Neural Prophet gave the suitable metrics for the sugar beet. Further, Airlangga, (2024) suggested that the SARIMA model helped in proving the dynamics of prediction for the maize crop to a remarkable extent. The SARIMA (1, 1, 2) x (2, 2, 2, 12) model has shown the best fit predictive metrics. The advanced machine learning model can be combined with SARIMA to deal with the nonlinear and complex agricultural data. (Patrick, 2023) used different models like SARIMAX, LSTM, and state space for the banana tea yield by considering the effect of climate change. The data was collected for the years 1961-2020. The prediction results proved that time series models, along with the ensemble models, outperformed the others. For the country of Tanzania, it was proven that climate change plays a vital role in the predictive measures. 0.99 was the calculated R^2 for the ensemble model for banana yield. This study helps in evaluating the outcomes of time series models combined with other machine learning models. To forecast the Rabi as a major crop of India, the outcomes revealed that the ARIMAX hybrid model outperforms the other models. The time series models are capable of handling the nonlinear and complex patterns (Pandit et al., 2023). In this study, the SARIMAX model has been employed to forecast important meteorological parameters such as temperature, rainfall, and humidity. SARIMAX was chosen to generate short-term forecasts. These forecasts are then integrated into a decision support system for major Bangladeshi crops like rice, maize, jute, wheat, and potato. By combining soil conditions, weather forecasts, and disease prediction, the system helped in identifying the most suitable crops (Ahmed et al., 2024).

In another study, for maize yield prediction, the authors applied the SARIMAX model. The best-performing model, SARIMA (1,1,2) × (2,2,2,12), achieved a strong fit with historical maize production data, with an AIC of 339914.85 and a BIC of 339950.64. The model's short-term forecasts showed less than a 2% deviation from actual production, demonstrating high accuracy, while long-term forecasts exhibited greater variability due to external environmental and economic uncertainties (Airlangga, 2024). The authors Mirpulatov et al. (2023) applied a time series model including SARIMAX, Prophet, and Neural Prophet to generate the weather inputs for MONICA crop simulation in the Kursk region of Russia. The objective was to evaluate how different weather forecast approaches affect the yield predictions. The study simulated the crop rotation of soybean and sugar beet. Among the tested methods, SARIMAX provided the best performance for forecasting soybean yield. The next section highlights the methodology for the present study.

Methods

This section highlights the methodology adopted for the study. The reason for adopting time series models has already been discussed in the introduction. One of the standard time series forecasting models is ARIMA. To predict the expected trends and gain meaningful wisdom, historical time series data is used by statistical models that are based on the past values, and then the trend pattern is followed. For short-term forecasting, ARIMA models give outstanding results and help in modeling the non- stationary data. In ARIMA AR stands for Auto Regressive (AR) where AR means that they use the past values to depict the future values, Integrated (I) is by applying the differencing d times, the integrated part in ARIMA model helps in converting the non-stationary series to the stationary series Moving Average (MA) to smoothen out fluctuations in time series data, the MA helps to depict the impact on

current values due to past forecasts.

The notation of ARIMA is with the symbols p, d, and q. These standard notations are filled with the integer parameters to signify the type of the ARIMA model being used for the analysis. The parameter specifies the number of past values (lags) used to predict the current values, and d is the degree of differencing, which is a technique used to check how many times differencing is applied to make the particular series stationary. q: To calculate the moving average, the number of data points used is also known as the size of the moving average window. The elongated version of the ARIMA model, which incorporates the seasonal variable with it, is the SARIMA model. To handle the strong seasonal patterns, the SARIMA model is used. Adding the seasonal patterns to this model makes it more capable of focusing on the repeating patterns in the data. The wise selection of the parameters is of utmost importance. The SARIMA model is well capable of handling the irregularities along with the changing pattern. After the determination of all the required parameters, data can be used to fit into the model. In this study, where the data set captures the seasonal variation, a time series SARIMAX model has been used for yield prediction, and the results obtained were compared with the results obtained from other machine learning models. SARIMA is an influential time series forecasting model that is widely applicable in crop yield prediction. When seasonal components were added to the ARIMA model, it became a SARIMA model for repeating patterns or at regular intervals, like a monthly basis or a yearly basis, etc. In the agricultural field, yield prediction is a noteworthy implementation of machine learning. The SARIMA model is widely used for yield prediction. The problem definition for crop yield prediction by using the SARIMA model can be discussed as: The time series-based dataset was taken, where historical yield data was taken on a yearly basis, and as the yield is predicted on the change of environment variables, e.g., Max temperature, Rainfall, wind speed, Humidity etc.

Therefore these features were selected on the basis of intervals. Non-seasonal, seasonal, and seasonal parameters were quantified to express the horizon of the model. It includes the pre-processing of the data so that proper parameters can be selected for the model. SARIMAX is designed for time series forecasting as it accounts for trend, seasonality, and exogenous variables. In summary, the SARIMA model helps in predicting the crop yield in a very accurate and précised way. For the said work, a better understanding of the data is required. As our data is real-time data, the data is pre-processed, the missing values are filled, and the data is scaled, so that accuracy can be achieved. The yield values were also standardized so that all the values could be on a comparable scale, along with the steps that are necessary for the model selection and its evaluation (refer to Figure 2). All the experiments were conducted in an Anaconda environment (Python 3.9), on a system equipped with an Intel Core i7 processor and 16 GB RAM, and Windows 10. The implementation made use of the Python libraries statsmodels for SARIMAX, pandas and numpy for data processing, and matplotlib for data visualization. The exogenous variables were selected based on the domain knowledge and the statistical feature selection methods. Initially, correlation analysis was conducted to examine the degree of association between each variable and crop yield. Variables showing a significant correlation were retained for further study. Although PCA was tested, it was not adopted in the final model, in order to maintain the interpretability. After the clear problem statement, the SARIMAX model can be set as the benchmark among the time series models for model training and its evaluation.

The methodology of the present study is depicted in Figures 2 and 3. First of all, load the CSV, then split the dataset into endogenous and exogenous variables. The data was divided into different parts for training and testing, followed by the generation of diagnostic reports. The future forecast was generated on the dashboard.

Hyperparameters Selection

The notation SARIMA (p, d, q) (P, D, Q, m) refers to a SARIMA model, which extends the ARIMA model to account for seasonality in time series data. It is a powerful model used for forecasting time series data that exhibits both non-seasonal and seasonal patterns. Here's an explanation of each parameter: Non-seasonal Parameters (p, d, q): here, p (AR – Autoregressive term) represents the number of lag observations included in the model.

Fig. 2: Steps in the SARIMAX Model

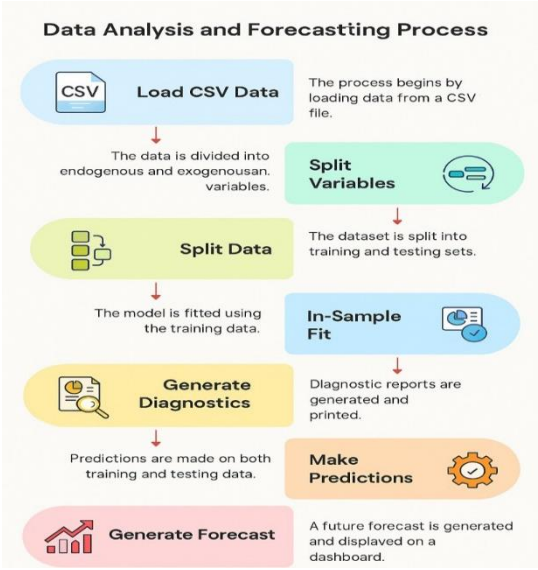


Fig. 3: Research Methodology

In simpler terms, it's the number of previous time steps (or lags) that are used to predict the current value. d (Differencing term): d is the number of times the time series is differenced to make it stationary, i.e., to remove trends and seasonality in the data. q (MA - Moving Average term) q represents the number of lagged forecast errors in the prediction model. It refers to the number of previous error terms that are included to predict the current value. The P, D, and Q are the Seasonal Parameters, where P is similar to p but applies to the seasonal part of the time series. It represents the number of seasonal lags that influence the current value. D is the Seasonal differencing term, which is the seasonal counterpart to d and represents the number of seasonal differences needed to make the seasonal part of the series stationary. Q represents Seasonal MA - Seasonal Moving Average term corresponds to q, but for the seasonal component. It represents the number of lagged seasonal errors to be used in the model. To capture the seasonal periodicity, m is used, which is the number of periods in a full seasonal cycle (also called the seasonality length). It defines the frequency of the seasonal component. In our study, (p, d, q) handles the seasonality, i.e., yearly. The exogenous variables are rainfall and temperature. The above factors make our model more powerful because our data has cycles as well as external variables. To access the quality of fit for the proposed model, BIC and AIC values are calculated, where AIC helps in penalizing the model's complexity. Similarly, BIC helps in evaluating the model's efficiency with a strong penalty. To check how well the model fits the data, it is being assessed with the help of the Log Likelihood. A model with the lower AIC is preferred to check for the optimal balance in the model. Consequently, minimized AIC values ensure the adequate predictive power of the models. This model has been used to generate future yield values. To evaluate the best fit of the chosen forecasting models, several widely accepted metrics were used to check the performance. Here, Mean MAE measures forecast error as the average magnitude, MSE, and RMSE place greater emphasis on larger errors, making them more sensitive to deviations and helping identify more precise models. MSE represents the average squared error, and RMSE, being the square root of MSE, offers a more interpretable measure in the same units as the original data, facilitating easier comparison. To ensure robustness and generalization, we employ a rolling window cross-validation approach. This involves dividing the dataset into overlapping windows, training the model on each, and testing it on the following window. Through repeated iterations, we assess how well the model adapts to changing data patterns while avoiding overfitting. This method further guarantees the model's reliability in forecasting across different subsets of the dataset. SARIMAX extends its applications in a number of fields, predicting the stock prices or forecasting upcoming weather reports, etc.

The quality of any model depends upon the relevance of the data taken for the study. The steps involved in the yield prediction using the SARIMA model. In our study, we have applied the SARIMAX model. For this, we have used a number of iterations, and the number of hyperparameters has been changed vice versa, for the seasonal and the non-seasonal notation of the particular model. Two models have been compared in the present study: Model 1 and Model 2. Model 1 is SARIMAX (1,0,1) x (1,1,1,12) and Model 2 is SARIMA (1,0, 0) x (0, 0, [1], 12). The notation for the Model 1 for the non-seasonal part is described as SARIMA (1,0,1), which has been used here (p = 1, autoregressive lag 1), as the model predicts the value on the basis of the last month or year. Differencing d=0, our data needed no differencing to make our data trend-free, as our data was already stationary. (q = 1, moving average) means the model remembers the last forecast error. The notation for the seasonal part (P, D, Q, m), i.e., (1,1,1,12). Here, P is taken as 1, i.e., lag taken as 12 months, so that a prediction could be made. D is taken as 1, and seasonality is being removed to make the model more stable by taking seasonal differencing as 1. Seasonal Moving Average is taken as 1. M =12, seasonal cycle length. It means that our data has a seasonal yearly seasonality. In the context of the hyperparameters taken into account, Non-seasonal is capturing short-term lag and forecast errors. Seasonal order (1,1,1,12) is capturing agricultural and weather data. After hyperparameter tuning, the SARIMAX model with order (1, 0, 0) and seasonal order (0, 0, 1, 12) was selected.

Results

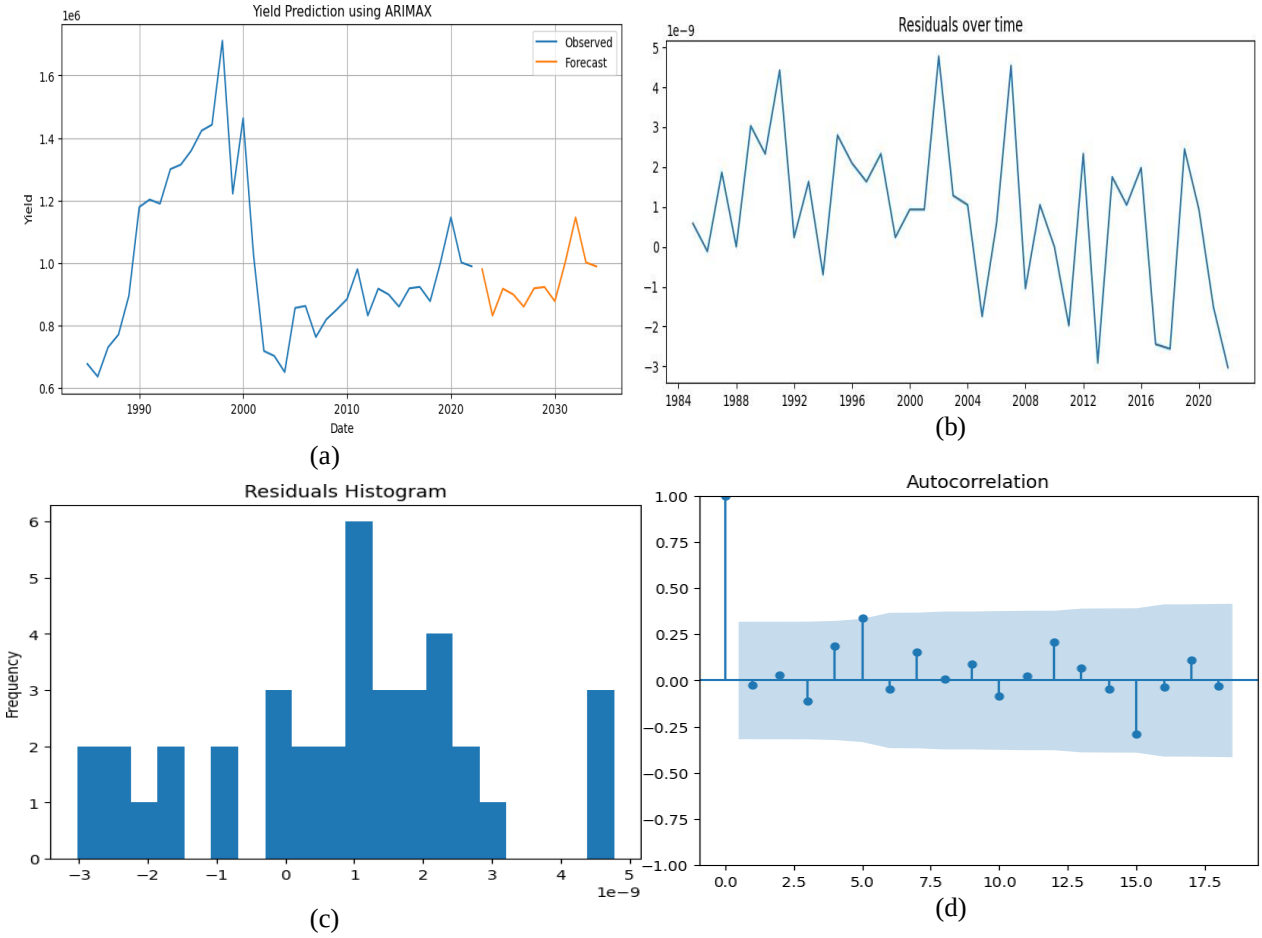
The results with respect to the future predictions are presented in the form of a graphs (Fig. 4). In this, observed and forecasted values are given. The forecast is done for the next 12 years.

Table 2 presents various metrics of the results obtained through the SARIMAX Model 1 and Model 2. All the parameters indicate that Model 1 is an average fit, and Model 2 is a good fit for this data and captures the variations very well.

The histogram depicts that the residuals are clustered near 0. This means that our model's predictions are unbiased. The shape of the distribution is roughly symmetric, with small deviations. It matches the Jarque-Bera test, which already confirmed that the residuals are approximately normal. The residual ACF, PACF, and QQ –plot is shown in Figure 4. The ACF (Auto Correlation Function) Plot of the residuals is depicted in Figure 4. The shaded blue area is 95 % confidence interval. The residual autocorrelation bars inside the band mean that they are not statistically significant. Nearly all lags fall inside. This is a clear sign that no autocorrelation is left.

Table 2: Performance Metrics of SARIMAX (Model 1 and Model 2)

C	Formula	Results Obtained (Model I)	Results Obtained (Model II)	Good Fit
AIC	AIC = 2k - 2ln (L) *k Where K is the parameters, and L is the Likelihood function, and 2ln (L) represents the goodness of the fit.	-512.767	-559.138	Model II
BIC	BIC = k * ln (n) - 2 * ln (L). K acts as the penalty for the complexity of the model, n represents the number of data points, and L represents the goodness of the fit in the model (likelihood function)	-311.402	-357.715	Model II



(a) (b) (c) (d)

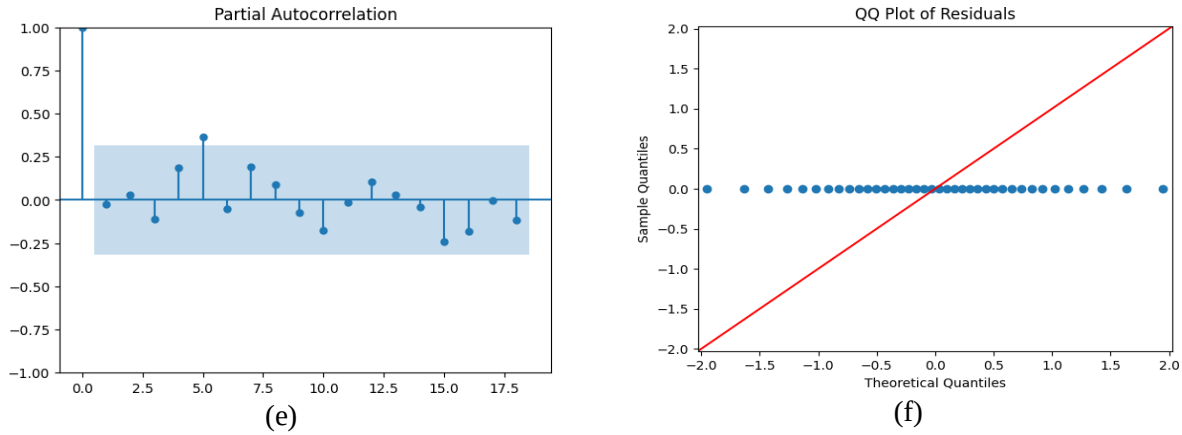


Fig. 4: Statistical Parameters -Model II

The next possible check is the QQ plot, which shows the visual confirmation of the normality of the residuals. This section discusses the comparative evaluation and the factors that make the SARIMAX the most reliable method for this task. At first glance, this study appears similar to the prior research that has applied ARIMA or SARIMAX models.

Table 2 Performance Metrics of SARIMAX (Model 1 and Model 2) However, it differs in several important aspects. Unlike earlier univariate approaches, the present study employs the SARIMAX framework, which explicitly incorporates the exogenous variables. This multivariate perspective enables the model to not only capture the temporal dependencies but also the direct influence of the weather variables on yield data. Furthermore, the dataset comprises the different regions, i.e., Himachal Pradesh (1985-2022), which is longer than the past analyzed studies. Together, these factors distinguish the present study by providing a more comprehensive and interpretable work and context-specific forecasting framework. The SARIMAX model used in the study balances the predictive performance with the interpretability, allowing for clearer insights into the role of the exogenous factors in determining yield outcomes. Time series models provide the special capabilities related to temporal dynamics, which makes them more suitable for sequential data forecasting. For tea yield prediction, this model helps in predicting accurate results, which can help society in a number of ways, like planning and preparing for the disastrous conditions in the near future. SARIMAX can be combined well with the traditional machine learning models as well as deep learning models to give the best prediction for crop yield. This helps in increasing the accuracy of the forecasting in the desired agricultural field of yield prediction. The improved technologies have helped in implementing the SARIMAX models, which can give fast results. The enhanced libraries of Python have helped a lot in defining the outputs of the SARIMAX model. SARIMAX being efficient in handling the real-time stream of data, i.e., the Big Data environments, solutions are best predicted with SARIMAX. Overall, on studying the parameters obtained through SARIMAX, Model 2 understands the data very well and is best suited in situations where the data set is small and has seasonal variations. The SARIMAX model 2 consistently helped in demonstrating better and superior results. In a nutshell, for many days, SARIMAX has been used for predicting yield forecasts, but the integration of exogenous variables along with the dealing of big data environments along with the handling of capturing the linearity and the non -linearity of data in combining it with the machine learning and deep learning models have made SARIMAX an efficient model in the field of yield forecasting.

The validation of the present study has been carried out in two distinct phases in order to establish the robustness of the SARIMAX models for forecasting tea yield. In the first phase, several studies have been examined where SARIMAX and its variations have been effectively applied for forecasting crop yields by incorporating climatic and other exogenous factors. For instance, one study focused on classifying the annual rainfall of the Ceará (1901 to 2020) into six categories. Also helped in building the SARIMAX models (1947-2020) to forecast grain area, yield, and the production value and prices using rainfall as an exogenous variable in comparison to the ARIMA model (Lemos et al., 2024). Furthermore, another study emphasized that the irrigation % is a valuable predictor. ARIMAX is superior to ARIMA, and hybrid ARIMAX models are the best options when capturing non-linearity in crop yields (Pandit et al., 2024). Similarly, research conducted using the data from 1948-2023 demonstrated that integrating Impulse Indicator Saturation into the ARIMAX models enhances the wheat yield forecasting accuracy, making it a stronger tool for early warning systems and production planning (Zulfiqar et al., 2024). These studies highlight a consistent pattern and seasonality combined with the relevant climate drivers/ exogenous variables, which yield

the best performance. Thus, SARIMAX models are robust in comparison to ARIMA or SARIMA models, which aligns with the present study, echoing our model II with seasonal MA (1) and exogenous factor delivered the best performance. In the second phase, the focus shifts specifically to tea yield forecasting, where prior studies have predominantly applied ARIMA and the SARIMAX approaches. (Chandra, 2023) identified that the ARIMA (0,1,1) model is the most appropriate for predicting tea yield in India. For selecting the best-fit model, various statistical methods were used. The acceptability of the model was further verified using the Ljung-Box test, ACF, and PACF criteria and other residual diagnostics. The AIC was 11.546, and the BIC was 11.659, highlighting its suitability for the dataset under consideration. Similarly, another study on tea production in Bangladesh found that the ARIMA (0,2,1) model was the most suitable. AIC was 774.68, and BIC was 778.06 (Hossain and Abdulla, 2015). Complementing these findings, Esack et al assessed the effect of agrometeorological variables on tea yield. Their study applied the SARIMAX model and emphasized the importance of the BIC values in model diagnostics, indicating that the lower BIC values are a sign of a stronger fit. Model adequacy was also confirmed using the Ljung-Box test and other metrics. SARIMAX (0,1,0) (2,0,0) $_4$ model was determined to be the best for a particular region with AIC of 1561.21. However, the study also noted that while SARIMAX yielded good results regionally, ANN outperformed SARIMAX overall in terms of predictive accuracy (Raj et al., 2019).

Conclusion

The SARIMAX model will help a lot in this domain to explore and identify pertinent trends and patterns by using the historical data sets in constructing the stepping stone for forecasting models. This study will help in analyzing and expressing the complex relationship between tea yield and climate, and help in unfolding the hidden patterns of crop yield with the climate. All over the globe, recent advancements are being used in the field of agriculture, which directly or indirectly help the farmers in well planning of the resources, time saving, cost saving, and this will lead to the maximization of profits. In the future, this model can be used in various spheres. We need to address more challenges related to the complexity of the data and the adaptability of the changing environmental variables. So that the SARIMAX models can become more capable of producing reliable forecasts. Though SARIMAX models are well capable of drawing their attention with the software tool named Python, there is always a chance for improvement and creating more user-friendly platforms in the near future. More focus should be given on the enhancement of the model so that extreme weather changing trends can also be identified in the near future, in case of extreme rainfall, extreme temperature, drought, or other conditions. To capture complex data trends, continuous exploration should be done for timely yield prediction along with neural networks or other deep learning techniques.

Acknowledgment

The authors sincerely thank the publisher for providing the opportunity and the support to publish this research work. We appreciate the platform and resources offered, which made it possible to disseminate our findings to a broader research community. We also extend our gratitude to the editorial team for their valuable time, constructive feedback, and assistance throughout the review of the publication process. This support has greatly contributed to the successful completion of this publication.

Funding Information

The authors declare that no financial support was received for the conduct of this study.

Author's Contributions

Pallavi Nagpal: Conceptualization of the study, methodology design, SARIMAX model development, data analysis and interpretation, writing the original draft, and revision of the manuscript.

Deepika Chaudhary: Exploratory data analysis and assistance in model validation.

Jaiteg Singh: Contributed to the Investigation process and supervised the research work.

Ethics

The submitted manuscript is original and has not been previously published. The corresponding author confirms that all authors have approved the final manuscript, and the study does not involve any ethical issues.

References

Ahmed, F. U., Das, A., & Zubair, Md. (2024). A Machine Learning Approach for Crop Yield and Disease Prediction Integrating Soil Nutrition and Weather Factors. *Proceeding of the International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (ICACCESS)*, 1–6. https://doi.org/10.1109/icaccess61735.2024.10499459

Airlangga, G. (2024). Advanced Forecasting of Maize Production using SARIMAX Models: An Analytical Approach. *Jurnal Media Informatika Budidarma*, *8*(1), 361–370. https://doi.org/10.30865/mib.v8i1.7268

Alawsi, M. A., Zubaidi, S. L., Al-Bdairi, N. S. S., Al-Ansari, N., & Hashim, K. (2022). Drought Forecasting: A Review and Assessment of the Hybrid Techniques and Data Pre-Processing. *Hydrology*, *9*(7), 115. https://doi.org/10.3390/hydrology9070115

Bali, N., & Singla, A. (2022). Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey. *Archives of Computational Methods in Engineering*, *29*(1), 95–112. https://doi.org/10.1007/s11831-021-09569-8

Begum, N., Alam, M., Mazumde, Md. S. J., Mim, M. R., & Monshi, M. H. (2024). Modeling of Climate Change Prediction and its Impact on the Tea Production in Sylhet District, Bangladesh. *Journal of Tropical Crop Science*, *11*(02), 105–119. https://doi.org/10.29244/jtcs.11.02.105-119

Chandra, R. P. (2023). Econometric Modeling for High Impact Sustainable Organic Tea Production: The Box-Jenkins Approach. *Asian Journal of Economics, Business and Accounting*, *23*(24), 141–154. https://doi.org/10.9734/ajeba/2023/v23i241193

Deka, S., Hazarika, P. J., Goswami, K., & Patowary, A. N. (2022). Forecasting tea production in India: a time series approach. *International Journal of Agricultural & Statistical Sciences*, *18*(1), 105.

Dharmaraja, S., Jain, V., Anjoy, P., & Chandra, H. (2020). Empirical Analysis for Crop Yield Forecasting in India. *Agricultural Research*, *9*(1), 132–138. https://doi.org/10.1007/s40003-019-00413-x

Ensafi, Y., Amin, S. H., Zhang, G., & Shah, B. (2022). Time-series forecasting of seasonal items sales using machine learning – A comparative analysis. *International Journal of Information Management Data Insights*, *2*(1), 100058. https://doi.org/10.1016/j.jjimei.2022.100058

Goyal, M., Agarwal, S., Ghalawat, S., & Malik, J. S. (2024). ARIMA and ARIMAX Analysis on the Effect of Variability of Rainfall, Temperature on Wheat Yield in Haryana. *Indian Journal of Extension Education*, *60*(1), 95–99. https://doi.org/10.48165/ijee.2024.60118

Hossain, Md. M., & Abdulla, Faruq. (2015). Forecasting the tea production of Bangladesh: Application of ARIMA model. *Jordan Journal of Mathematics and Statistics*, *8*(3), 257–270.

Lemos, J. de J. S., & Bezerra, F. N. R. (2024). ARIMAX Model to Forecast Grain Production under Rainfall Instabilities in Brazilian Semi-Arid Region. *Global Journal of Human-Social Science: E Economics*, *24*(1), 1–15.

Mirpulatov, I., Gasanov, M., & Matveev, S. (2023). Soil Dynamics and Crop Yield Modeling Using the MONICA Crop Simulation Model and Time Series Forecasting Methods. *Agronomy*, *13*(8), 2185. https://doi.org/10.3390/agronomy13082185

Mishra, P., Al khatib, A. M. G., Mohamad Alshaib, B., Kuamri, B., Tiwari, S., Singh, A. P., Yadav, S., Sharma, D., & Kumari, P. (2024). Forecasting Potato Production in Major South Asian Countries: a Comparative Study of Machine Learning and Time Series Models. *Potato Research*, *67*(3), 1065–1083. https://doi.org/10.1007/s11540-023-09683-z

Ospina, R., Gondim, J. A. M., Leiva, V., & Castro, C. (2023). An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil. *Mathematics*, *11*(14), 3069. https://doi.org/10.3390/math11143069

Pandit, P., Sagar, A., Ghose, B., Dey, P., Paul, M., Alqadhi, S., Mallick, J., Almohamad, H., & Abdo, H. G. (2023). Hybrid time series models with exogenous variable for improved yield forecasting of major Rabi crops in India. *Scientific Reports*, *13*(1), 22240. https://doi.org/10.1038/s41598-023-49544-w

Patrick, S., Mirau, S., Mbalawata, I., & Leo, J. (2023). Time series and ensemble models to forecast banana crop yield in Tanzania, considering the effects of climate change. *Resources, Environment and Sustainability*, *14*, 100138. https://doi.org/10.1016/j.resenv.2023.100138

Raj, E. E., Ramesh, K. V., & Rajkumar, R. (2019). Modelling the impact of agrometeorological variables on regional tea yield variability in South Indian tea-growing regions: 1981-2015. *Cogent Food & Agriculture*, *5*(1), 1581457. https://doi.org/10.1080/23311932.2019.1581457

Reddy, P. C. S., & Sureshbabu, A. (2020). *An Applied Time Series Forecasting Model for Yield Prediction of Agricultural Crop*. *1118*. https://doi.org/10.1007/978-981-15-2475-2_16

Sam, A. S., Abbas, A., Surendran Padmaja, S., Kaechele, H., Kumar, R., & Müller, K. (2019). Linking Food Security with Household's Adaptive Capacity and Drought Risk: Implications for Sustainable Rural Development. *Social Indicators Research*, *142*(1), 363–385. https://doi.org/10.1007/s11205-018-1925-0

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Zulfiqar, H., Ahmad, R., & Shahzad, U. (2024). Hybrid ARIMA-IIS Approach for Wheat Yield Forecasting: An Integrated Approach. *IRASD Journal of Economics*, *6*(1), 109–127. https://doi.org/10.52131/joe.2024.0601.0197