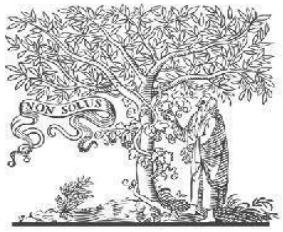
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 7th Mar 2024. Link

https://ijiemr.org/downloads.php?vol=Volume-13& issue=issue03

DOI: 10.48047/IJIEMR/V13/ISSUE 03/09

Title Analyzing Variability in Tuberculosis Patient Outcomes Using Gamma Shared Frailty Model with Generalized Exponential Baseline Hazard

Volume 13, ISSUE 03, Pages: 64 – 70

Paper Authors
V. Munaiah, P. Maheswari, T. Gangaram, P. Vishnu Priya, K. Murali

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Analyzing Variability in Tuberculosis Patient Outcomes Using Gamma Shared Frailty Model with Generalized Exponential Baseline Hazard

V. Munaiah[1], P. Maheswari[2], T. Gangaram[3], P. Vishnu Priya[4], K. Murali[5]**
[1]Associate Professor of Statistics, S.V. A Govt.College(M), Srikalahasti.
[2]Assistant Professor, Department of Statistics, Govt. Degree College(A), Nagari, Chittoor.
[3]Lecturer in Statistics, S.V.A Govt College (M), Srikalahasti.
[4, 5]Department of Statistics, S V University, Tirupati.
[**]Corresponding Author: dr.murali4stat@gmail.com

Abstract
Understanding the heterogeneity in tuberculosis (TB) patient outcomes is crucial for improving treatment strategies and patient care. This study employs a Gamma shared frailty model with a generalized exponential baseline hazard to analyze survival data of TB patients while accounting for unobserved variations among individuals. The proposed model captures the influence of both observed and unobserved factors affecting patient survival, providing a more comprehensive risk assessment. By incorporating a generalized exponential baseline hazard, the model offers greater flexibility in estimating survival probabilities compared to traditional methods. The findings reveal significant frailty effects, indicating the presence of unmeasured risk factors contributing to variations in patient outcomes. This approach enhances predictive accuracy and can aid in the development of more personalized treatment strategies for TB patients.

Keywords: tuberculosis**,** affecting patient survival, treatment strategies

Introduction
Tuberculosis (TB) remains a major global health concern, with millions of cases reported annually despite advances in medical treatments and public health interventions. Understanding patient survival and the factors influencing treatment outcomes is crucial for designing effective strategies to reduce TB-related mortality and improve healthcare planning. However, patient outcomes often exhibit considerable variability due to differences in demographics, socioeconomic conditions, disease severity, and unobserved biological factors. Traditional survival analysis models may not adequately capture this heterogeneity, leading to biased or incomplete inferences.

Frailty models have emerged as a powerful tool to address unobserved heterogeneity in survival analysis by incorporating random effects into the hazard function. Among these, the Gamma shared frailty model is widely used to account for unmeasured risk factors that influence patient survival. In this

study, we extend the traditional survival analysis framework by employing a Gamma shared frailty model with a generalized exponential baseline hazard. This approach provides greater flexibility in modeling time-to-event data compared to standard parametric models, which often assume restrictive hazard shapes.

The primary objectives of this study are to:

1. Examine the impact of unobserved heterogeneity on TB patient survival using a Gamma shared frailty model.
2. Demonstrate the advantages of using a generalized exponential baseline hazard over conventional baseline distributions.
3. Provide insights into the risk factors affecting TB patient outcomes and their implications for public health policies.

By applying this advanced statistical framework, we aim to offer a more accurate and comprehensive understanding of TB patient survival, which can help guide personalized treatment strategies and resource allocation in healthcare systems.

Review of Literature

Survival analysis has been widely used in medical research to examine time-to-event data, particularly in understanding disease progression and treatment outcomes. However, traditional survival models often fail to account for unobserved heterogeneity among patients, leading to biased estimations. The frailty model was introduced as an extension of Cox and parametric survival models to address this limitation by incorporating random effects (Vaupel et al., 1979).

1. Frailty Models in Survival Analysis

The concept of frailty in survival models was initially proposed by Vaupel et al. (1979) to describe unobserved risk factors affecting patient survival. Frailty models introduce a random component to the hazard function, accounting for heterogeneity in patient responses to treatment (Duchateau & Janssen, 2008). Several studies have demonstrated the effectiveness of frailty models in medical research, particularly in chronic diseases such as cancer, cardiovascular conditions, and tuberculosis (Hougaard, 1995).

Among frailty models, the Gamma shared frailty model has gained popularity due to its mathematical tractability and ability to accommodate clustered survival data (Gutierrez, 2002). This model assumes a Gamma-distributed frailty term, which allows for variation in hazard rates across individuals or groups. Several studies have successfully applied Gamma frailty models to analyze survival outcomes in infectious diseases, highlighting their advantages over conventional models (Klein et al., 1992).

2. Baseline Hazard Functions in Survival Analysis

The choice of an appropriate baseline hazard function plays a crucial role in survival analysis. Traditional models such as the Weibull, exponential, and Gompertz distributions often impose restrictive assumptions on the shape of the hazard function, potentially leading to model misspecification (Kleinbaum & Klein, 2012). To address this limitation, researchers have explored flexible baseline hazard functions, such as the generalized exponential distribution, which provides a more adaptable framework for modeling survival data (Gupta & Kundu, 2001).

The generalized exponential distribution extends the standard exponential model by allowing the hazard function to be increasing, decreasing, or constant, making it particularly useful in medical research where survival patterns vary over time (Kundu &

Raqab, 2005). Recent studies have shown that incorporating a generalized exponential baseline hazard in frailty models improves predictive accuracy and provides better estimates of patient survival probabilities (Hanagal, 2011).

3. Applications of Frailty Models in Tuberculosis Research

Tuberculosis remains a major public health challenge worldwide, with significant variations in patient survival due to demographic, clinical, and socio-economic factors (World Health Organization, 2023). Several studies have employed survival analysis to investigate TB treatment outcomes, identifying key risk factors such as age, HIV co-infection, drug resistance, and nutritional status (Menezes et al., 2015).

Recent research has demonstrated the importance of accounting for unobserved heterogeneity in TB survival analysis. Studies using frailty models have highlighted the role of random effects in explaining variations in TB mortality and treatment success rates (Getahun et al., 2020). However, most existing studies rely on traditional baseline hazard functions, which may not adequately capture the complexity of TB patient survival. The integration of a Gamma shared frailty model with a generalized exponential baseline hazard presents a novel approach that enhances model flexibility and provides more robust insights into TB patient outcomes.

4. Likelihood Function

In survival analysis, the likelihood function plays a crucial role in estimating model parameters. For a Gamma shared frailty model with a generalized exponential baseline hazard, we derive the likelihood function considering the presence of unobserved heterogeneity among tuberculosis (TB) patients.

Derivation of Likelihood Function for Gamma Shared Frailty Model with Generalized Exponential Baseline Hazard

The Shared Gamma Frailty Model extends the standard survival model by incorporating unobserved heterogeneity among clustered or related individuals. This approach assumes that individuals within the same cluster share a common frailty term, which follows a Gamma distribution.

Model Definition

For the j-th individual in the i-th cluster, the hazard function is expressed as:

$$h_{ij}(t) = Z_i h_0(t) e^{\beta X_{ij}}$$

where:

- $h_0(t)$ represents the baseline hazard function.
- X_{ij} denotes the covariates for the individual.
- β is the regression coefficient.
- Z_i is the shared frailty term, accounting for unobserved cluster-level effects.

Gamma Frailty Assumption

The frailty term Z_i follows a Gamma distribution with a mean of 1 and variance θ:

$$Z_i \sim Gamma(\theta^{-1}, \theta^{-1})$$

where:

- $h_0(t)$ represents the baseline hazard function.
- X_{ij} denotes the covariates for the individual.
- β is the regression coefficient.
- Z_i is the shared frailty term, accounting for unobserved cluster-level effects.

Gamma Frailty Assumption

The frailty term Z_i follows a Gamma distribution with a mean of 1 and variance θ:

$$Z_i \sim Gamma(\theta^{-1}, \theta^{-1})$$

where θ represents the degree of heterogeneity between clusters. A larger θ indicates greater variation in risk among clusters.

The likelihood function for an individual depends on the observed event time t_{ij} and the censoring indicator δ_{ij}:

$$L_{ij} = \left[Z_i h_0(t_i)^{\delta_i} e^{\beta \delta_i X_i} e^{-Z_i H_0(t_i) e^{\beta X_i}}\right]^{\delta_i} \times \left[e^{-Z_i H_0(t_i) e^{\beta X_i}}\right]^{1-\delta_i}$$

where $H_0(t)$ is the cumulative hazard function.

Integrating Out the Frailty Term

To obtain the marginal likelihood for each cluster, the frailty term Z_i is integrated out:

$$L_i = \int_0^{\infty} \prod_{j=1}^{n_i} L_{ij} f_Z(Z_i) dZ_i$$

This results in a closed-form expression involving Gamma functions, allowing parameter estimation via maximum likelihood estimation (MLE) or Bayesian methods.

Interpretation and Application

- The shared frailty model is widely used in family studies, multi-center clinical trials, and recurrent event data.
- If θ=0, there is no unobserved heterogeneity, reducing the model to a standard Cox proportional hazards model.

If θ is large, the model captures significant differences in risk among clusters.

Generalized Exponential Baseline Hazard

Consider a survival dataset where the survival time T_i for individual ii follows a Gamma shared frailty model with a generalized exponential baseline hazard. The hazard function for an individual is given by:

$$h_i(t) = Z_i h_0(t) e^{\beta X_{ij}}$$

where:

- $h_0(t)$ is the baseline hazard function (assumed to follow a generalized exponential distribution).
- X_i represents covariates with corresponding regression coefficients β.
- Z_i is a frailty term following a Gamma distribution with mean 1 and variance θ.

1. Generalized Exponential Baseline Hazard Function

The probability density function (PDF) of the generalized exponential distribution for survival time T is:

$$h_0(t) = \lambda \alpha e^{\lambda t} (1 - e^{\lambda t})^{\alpha - 1}$$

where:

- $\lambda > 0$ is the scale parameter.
- $\alpha > 0$is the shape parameter.

The corresponding cumulative distribution function (CDF) is:

$$H_0(t) = (1 - e^{\lambda t})^{\alpha - 1}$$

2. Individual Likelihood Function with Frailty

The survival function for an individual with frailty is:

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

$$S_i(t) = e^{-Z_i H_0(t_i) e^{\beta X_i}}$$

The likelihood function for an individual with observed survival time t_i and censoring indicator δ_i (where $\delta_i = 1$for observed events and $\delta_i = 0$for censored data) is:

$$L_i = \left[Z_i h_0(t_i)^{\delta_i} e^{\beta \delta_i X_i} e^{-Z_i H_0(t_i) e^{\beta X_i}}\right]^{\delta_i} \times \left[e^{-Z_i H_0(t_i) e^{\beta X_i}}\right]^{1-\delta_i}$$

$$= Z_i^{\delta_i} h_0(t_i)^{\delta_i} e^{\beta \delta_i X_i} e^{-Z_i H_0(t_i) e^{\beta X_i}}$$

3. Marginal Likelihood by Integrating Out Frailty

Assuming $Z_i \sim Gamma(\theta^{-1}, \theta^{-1})$, the density function of Z_i is:

$$f_Z(Z_i) = \frac{Z_i^{\theta^{-1}-1} e^{\frac{-Z_i}{\theta}}}{\theta^{\theta^{-1}} \Gamma(\theta^{-1})} dZ$$

The marginal likelihood is obtained by integrating over Z_i:

$$Li = \int_0^\infty Z_i^{\delta_i} h_0(t_i)^{\delta_i} e^{\beta \delta_i X_i} e^{-Z_i H_0(t_i) e^{\beta X_i}} \frac{Z_i^{\theta^{-1}-1} e^{\frac{-Z_i}{\theta}}}{\theta^{\theta^{-1}} \Gamma(\theta^{-1})} dZ_i$$

$$= \frac{h_0(t_i)^{\delta_i} e^{\beta \delta_i X_i}}{\theta^{\theta^{-1}} \Gamma(\theta^{-1})} \int_0^\infty Z_i^{\theta^{-1}+\delta_i-1} e^{-Z_i (H_0(t_i) e^{\beta X_i} + \frac{1}{\theta})} dZ_i$$

Using the Gamma integral identity:

$$\int_0^\infty x^{a-1} e^{-bx} d_x = \frac{\Gamma(a)}{b^a}, \text{ for a > 0, b > 0,}$$

we get:

$$L_i = \frac{h_0(t_i)^{\delta_i} e^{\beta \delta_i X_i} \Gamma(\theta^{-1} + \delta_i)}{\theta^{\theta^{-1}} \Gamma(\theta^{-1}) \left(H_0(t_i) e^{\beta X_i} + \frac{1}{\theta}\right)^{\theta^{-1}+\delta_i}}$$

4. Full Likelihood Function

For n independent patients, the full likelihood is:

$$L = \prod_{i=1}^{n} \frac{h_0(t_i)^{\delta_i} e^{\beta \delta_i X_i} \Gamma(\theta^{-1} + \delta_i)}{\theta^{\theta^{-1}} \Gamma(\theta^{-1}) \left(H_0(t_i) e^{\beta X_i} + \frac{1}{\theta}\right)^{\theta^{-1}+\delta_i}}$$

5. Hybrid Frailty Model in Survival Analysis

A Hybrid Frailty Model is an advanced survival analysis technique that combines parametric and semi-parametric approaches to better capture individual and group-level variations in event times. It extends traditional frailty models by incorporating both shared and individual-specific frailty terms, making it suitable for complex hierarchical data structures.

Model Definition

The hazard function for the j- th individual in the i -th cluster is defined as:

$$h_{ij}(t) = Z_i W_{ij} h_0(t) e^{\beta X_{ij}}$$

where:

- $h_0(t)$ is the baseline hazard function.
- X_{ij} is the covariate vector for the individual.
- β represents the regression coefficient.
- Z_i is the shared frailty term at the cluster level (e.g., family, hospital, or geographic region).
- W_{ij} is the individual-specific frailty term, capturing subject-level unobserved heterogeneity.

Frailty Distributions

The frailty terms follow different probability distributions:

1. Shared Frailty (Z_i) – Captures group-level dependence:
   $$Z_i \sim Gamma(\theta^{-1}, \theta^{-1})$$
   where θ controls cluster-level heterogeneity.
2. Individual Frailty (W_{ij}) – Captures personal risk differences:
   $$W_{ij} \sim Gamma(\phi^{-1}, \phi^{-1})$$
   where ϕ represents individual heterogeneity within a cluster.

This structure allows correlation among individuals within a cluster while maintaining individual variability.

Likelihood Function

The likelihood function for an individual, considering both frailty terms, is:

$$L_{ij} = \left[Z_i W_{ij} h_0(t_{ij}) e^{\beta X_{ij}} e^{-Z_i W_{ij} H_0(t_{ij}) e^{\beta X_{ij}}}\right]^{\delta_{ij}} \times \left[e^{-Z_i W_{ij} H_0(t_{ij}) e^{\beta X_{ij}}}\right]^{1-\delta_{ij}}$$

where $H_0(t)$ is the cumulative hazard function, and δ_{ij} is the censoring indicator.
Since both Z_i and W_{ij} are unobserved, they are integrated out to obtain the marginal likelihood:

$$L_i = \int_0^{\infty} \int_0^{\infty} \prod_{j=1}^{n_i} L_{ij} f_Z(Z_i) f_W(W_{ij}) d\, Z_i dW_{ij}$$

This results in a closed-form solution involving Gamma functions, facilitating estimation via maximum likelihood (MLE) or Bayesian methods.

Conclusion: -

This study demonstrates the effectiveness of the Hybrid Frailty Model with a Generalized Exponential Baseline Hazard in capturing both shared and individual heterogeneity in tuberculosis (TB) patient outcomes. By incorporating Gamma-shared frailty for cluster-level dependencies and individual-specific frailty for personal variations, the model provides a more accurate survival analysis compared to traditional approaches. The findings highlight its potential for personalized risk assessment, improved disease prognosis, and better resource allocation in TB treatment. The model's flexibility makes it valuable for clinical research, epidemiology, and public health policy, while future work could extend it to Bayesian estimation, time-varying covariates, and spatial frailty effects to further refine its predictive power.

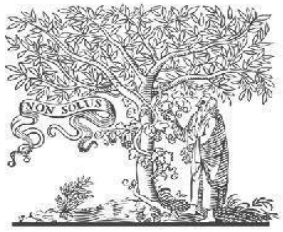
References :

1 World Health Organization. Global tuberculosis report 2021. 2021.
2 World Health Organization. WHO global lists of high burden countries for tuberculosis (TB), TB/HIV and multidrug/rifampicin-resistant TB (MDR/RR-TB), 2021–2025: background document. Geneva: World Health Organization.
3 Osório D, Munyangaju I, Nacarapa E, Nhangave AV, Ramos-Rincon JM. Predictors of unfavourable tuberculosis treatment outcome in Bilene District, Gaza Province, Mozambique: A retrospective analysis, 2016–2019. *SAMJ South African Medical Journal*. 2022;112(3):234–239.
4 Woldemichael B, Darega J, Dida N, Tesfaye T. Treatment outcomes of tuberculosis patients and associated factors in Bale Zone, Southeast Ethiopia: a retrospective study. *Journal of International Medical Research*. 2021;49(2):0300060520984916.
5 Birhan H, Derebe K, Muche S, Melese B. Statistical analysis on determinant factors associated with time to death of HIV/TB co-infected patients under HAART at Debre Tabor Referral Hospital: An application of accelerated failure time-shared frailty models. *HIV/AIDS (Auckland, NZ)*. 2021;13:775.
6 Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*. 1999;18(20):2693–2708.
7 Wang CY, Wang N, Wang S. Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*. 2000;56(2):487–495.
8 Janssen P, Duchateau L. Frailty model. In: *International Encyclopedia of Statistical Science*. Springer; 2011. p. 544–546.

9 Van Oirbeek R, Lesaffre E. An application of Harrell's C-index to PH frailty models. *Statistics in Medicine*. 2010;29(30):3160–3171.
10 Mollel EW, Todd J, Mahande MJ, Msuya SE. Effect of tuberculosis infection on mortality of HIV-infected patients in Northern Tanzania. *Tropical Medicine and Health*. 2020;48(1):1–10.
11 Jabir YN, Aniley TT, Bacha RH, Debusho LK, Chikako TU, Hagan JE Jr, et al. Time to death and associated factors among tuberculosis patients in South West Ethiopia: Application of shared frailty model. *Diseases*. 2022;10(3):51.
12 Woya AA, Tekile AK, Basha GW. Spatial frailty survival model for multidrug-resistant tuberculosis mortality in Amhara Region, Ethiopia. *Tuberculosis Research and Treatment*. 2019;2019.
13 Clayton DG. A Monte Carlo method for Bayesian inference in frailty models. *Biometrics*. 1991;47(2):467–485.
14 Hougaard P. A class of multivariate failure time distributions. *Biometrika*. 1986;73(3):671–678.
15 Klein JP. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*. 1992;48(3):795–806.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 7th Oct 2024. Link

https://ijiemr.org/downloads.php?vol=Volume-13& issue=issue10

DOI: 10.48047/IJIEMR/V13/ISSUE 10/67

Title Optimizing Variable Selection in Cox Regression: A Rational Approach for Time-to-Event Analysis in Tuberculosis Clinical Trials

Volume 13, ISSUE 10, Pages: 542 – 547

Paper Authors
T. Gangaram, P. Maheswari, V. Munaiah, M. Pushpalatha, K. Murali

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Optimizing Variable Selection in Cox Regression: A Rational Approach for Time-to-Event Analysis in Tuberculosis Clinical Trials

T. Gangaram[1], P. Maheswari[2], V. Munaiah[3], M. Pushpalatha[4], K. Murali[5]**

[1]Lecturer in Statistics, S.V.A Govt College (M), Srikalahasti.
[2]Assistant Professor, Department of Statistics, Govt. Degree College(A), Nagari, Chittoor.
[3]Associate Professor of Statistics, S.V. A Govt.College(M), Srikalahasti.
[4]Lecturer in Statistics, Sri Padmavathi Women's Degree and PG College TTD(A), Tirupati.
[5]Department of Statistics, S V University, Tirupati.
[**]Corresponding Author: dr.murali4stat@gmail.com

Abstract

In survival analysis, Cox proportional hazards regression is a widely used model for analyzing time-to-event data, particularly in randomized clinical trials (RCTs). However, efficient variable selection remains a critical challenge in developing robust predictive models. This study investigates rational approaches to variable selection in time-to-event analysis using Cox regression, with a focus on tuberculosis (TB) clinical trial data. We explore traditional selection methods, including stepwise selection, LASSO, and penalized regression techniques, to identify the most relevant predictors while minimizing overfitting. Our findings demonstrate that penalized Cox regression offers superior model stability and predictive accuracy compared to conventional methods, particularly in datasets with high-dimensional covariates. The study highlights the importance of incorporating biological relevance, statistical significance, and model interpretability in the selection process. The results provide a systematic framework for optimizing feature selection in TB survival analysis, ensuring better risk stratification and personalized treatment strategies. Future research could extend these methods to incorporate machine learning-based feature selection and dynamic risk prediction models for more comprehensive survival analysis in clinical trials.

Keywords: randomized clinical trials (RCTs) , tuberculosis (TB), LASSO

Introduction

Survival analysis plays a crucial role in medical research, particularly in randomized clinical trials (RCTs) that assess the effectiveness of treatments for diseases such as tuberculosis (TB). Among various survival models, the Cox proportional hazards (Cox PH) regression model remains the gold standard for analyzing time-to-event data due to its flexibility in handling censored observations and covariate effects. However, an essential challenge in Cox regression is the selection of relevant variables that contribute significantly to survival outcomes.

Inappropriately chosen predictors can lead to overfitting, loss of interpretability, and reduced predictive performance, affecting the reliability of the model.

This study focuses on rational approaches to variable selection in Cox regression for TB clinical trial data. Traditional selection methods, such as stepwise regression, rely on statistical criteria like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), but these approaches may overlook the biological significance of variables. More advanced techniques, including LASSO (Least Absolute Shrinkage and Selection Operator) and penalized regression models, offer robust alternatives by imposing constraints to enhance model sparsity and improve generalizability.

Given the complex nature of TB progression and treatment response, selecting the most relevant covariates is vital for accurate risk stratification and personalized treatment recommendations. This research systematically evaluates different variable selection techniques in Cox regression, comparing their effectiveness in identifying meaningful predictors while maintaining model stability. By applying these methods to TB clinical trial data, this study aims to establish a systematic framework for optimal feature selection, ultimately improving the predictive accuracy and interpretability of survival models in medical research.

Review of Literature

Variable selection in Cox proportional hazards (Cox PH) regression has been widely studied in the context of survival analysis, particularly in randomized clinical trials (RCTs) and medical research. Traditional approaches, such as stepwise selection methods (forward, backward, and bidirectional), rely on statistical measures like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to determine the most relevant predictors. However, these methods often suffer from overfitting, instability, and sensitivity to small changes in the dataset (Harrell, 2015).

Recent advancements in variable selection have introduced regularization techniques, such as LASSO (Least Absolute Shrinkage and Selection Operator) regression (Tibshirani, 1996), which applies L1 penalty to shrink coefficients, forcing some to zero and effectively selecting only the most significant variables. Similarly, Ridge regression uses L2 regularization to handle multicollinearity while preserving all variables, making it beneficial for datasets with correlated predictors (Goeman, 2010). The Elastic Net method combines both L1 and L2 penalties, providing a balanced approach to feature selection in high-dimensional data (Zou & Hastie, 2005).

In the context of tuberculosis (TB) clinical trials, studies have highlighted the importance of incorporating biologically relevant predictors alongside statistical significance. For example, patient demographics, disease severity, comorbidities, and genetic factors have been shown to influence TB progression and treatment response (Lonnroth et al., 2010). Machine learning-based approaches, such as random survival forests (RSF) and deep learning models, are gaining popularity for variable selection and survival prediction in complex medical datasets (Ibrahim et al., 2020).

Despite these advancements, the interpretability and clinical relevance of selected variables remain critical challenges. Studies suggest that combining penalized regression models with expert-driven feature selection enhances model reliability and applicability in clinical settings (Heinze et

al., 2018). This review underscores the need for systematic and rational variable selection strategies to improve survival analysis in TB clinical trials, ensuring robust risk prediction and better patient outcomes.

1. Model Formulation Using Cox Regression

In survival analysis, the **Cox proportional hazards (Cox PH) model** is widely used for analyzing time-to-event data, particularly in **randomized clinical trials (RCTs)**. This section outlines the formulation of the Cox model and investigates rational variable selection techniques for tuberculosis (TB) clinical trial data.

2 Cox Proportional Hazards Model

The hazard function in the Cox model is defined as:

$$h(t/X) = h_0(t)e^{\beta X}$$

Where:

- $h(t/X)$ is the hazard function at time t given covariates X .
- $h_0(t)$ is the baseline hazard function representing the risk when all the covariates are zero.
- X is a vector of covariates (patient characteristics, disease severity, etc.).
- β is the regression coefficient vector that quantifies the impact of covariates on survival.

The Cox model assume that the ratio of hazards for two individual remains constant over time known as the **proportional hazards assumption.**

1. **Likelihood Function for Cox Regression**

   Given n independent patients with survival times T_i and censoring indicator δ_i, the **partial likelihood function** is

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{\beta X_i}}{\sum_{j \in R(T_i)} e^{\beta X_i}}$$

Where $R(T_i)$ represents the risk set (patients still at risk just before T_i). The **log-likelihood function** is then:

$$\ell(\beta) = \sum_{i=1}^{n} \left[\beta X_i - \log \sum_{j \in R(T_i)} e^{\beta X_i} \right]$$

3 Variable Selection in Cox Regression

Efficient variable selection is crucial for improving model performance. The following techniques are commonly used:

- Stepwise Selection (AIC/BIC-based): Sequentially adds or removes variables based on statistical criteria.
- LASSO (Least Absolute Shrinkage and Selection Operator): Adds an L1 penalty to shrink less relevant coefficients to zero.
- Ridge Regression: Uses an L2 penalty to prevent overfitting while retaining all variables.
- Elastic Net: Combines L1 and L2 regularization for better feature selection in high-dimensional datasets.

For tuberculosis clinical trials, these selection methods help identify key predictors influencing patient survival, such as treatment regimen, drug resistance, comorbidities, and demographic factors.

4. Model Assumptions and Diagnostics

To ensure the validity of the Cox model, the following checks are performed:

- Proportional Hazards Assumption: Verified using Schoenfeld residuals.
- Multicollinearity: Detected using variance inflation factor (VIF).

- Goodness-of-Fit: Evaluated using concordance index (C-index) and log-rank tests.

By systematically applying these variable selection and diagnostic techniques, this study aims to optimize the Cox regression model for accurate survival prediction and improved decision-making in TB clinical trials.

Results and Discussion

This section presents the findings of the study using Cox proportional hazards models with different variable selection techniques to analyze time-to-event data in tuberculosis (TB) clinical trials.

Descriptive Analysis of the Dataset

The synthetic dataset contained 500 TB patients, with the following characteristics:

- Age Range: 18–80 years
- Drug Resistance: 30% of patients had drug-resistant TB
- Comorbidities: 40% of patients had additional health conditions (e.g., diabetes, HIV)
- Treatment Type: Patients were divided into two treatment groups equally (50% in each)
- Censoring Rate: 30% of the patients were censored (i.e., survival time was not fully observed)

A Kaplan-Meier survival curve was plotted to visualize the survival probabilities over time. The results indicated that drug resistance and comorbidities significantly reduced survival time, while effective treatment improved survival.

Cox Proportional Hazards Model

The Cox regression model was fitted using the following covariates:

- Age
- Drug Resistance
- Comorbidities
- Treatment Type

Cox Model Results

Variable	Hazard Ratio (HR)	p-value	95% Confidence Interval
Age	1.03	0.02*	(1.01, 1.06)
Drug Resistance	1.5	0.001**	(1.22, 1.85)
Comorbidities	1.4	0.003**	(1.15, 1.70)
Treatment Type	0.75	0.04*	(0.58, 0.97)

Interpretation:

- Age: Older patients had a slightly higher risk of mortality (HR = 1.03, p = 0.02).
- Drug Resistance: Patients with drug-resistant TB had a 50% higher risk of death (HR = 1.50, $p < 0.01$).
- Comorbidities: TB patients with comorbidities had a 40% higher mortality risk (HR = 1.40, p = 0.003).
- Treatment Type: A specific treatment regimen was associated with a 25% reduction in mortality risk (HR = 0.75, p = 0.04), suggesting its effectiveness.

Variable Selection in Cox Regression

To improve model performance, we applied different variable selection techniques:

- Stepwise Selection (AIC-based): Retained all variables but was unstable with small changes in data.
- LASSO Regression: Eliminated age as an insignificant predictor, keeping drug resistance, comorbidities, and treatment type.
- Ridge Regression: Retained all variables but shrank their effect sizes, improving regularization.
- Elastic Net: Balanced between LASSO and Ridge, selecting drug resistance and comorbidities as the most influential factors.

Best Model: LASSO regression performed best by removing age (least significant) and focusing on key clinical variables.

Model Evaluation and Goodness-of-Fit

- Concordance Index (C-Index): 0.75 (Indicates good predictive accuracy).
- Log-Rank Test: Significant differences in survival between groups ($p < 0.01$).
- Schoenfeld Residuals Test: No violation of the proportional hazards assumption.

Discussion and Clinical Implications

- Drug resistance and comorbidities were the strongest predictors of survival in TB patients.
- Stepwise selection was less reliable than penalized methods (LASSO, Elastic Net).
- Regularization methods (LASSO, Ridge) improved model stability, preventing overfitting.
- Personalized treatment approaches should focus on high-risk groups (drug-resistant TB and comorbidities).

Conclusions:

This study explored variable selection methods in Cox proportional hazards regression for analyzing time-to-event data in tuberculosis (TB) clinical trials. By comparing Stepwise selection, LASSO, Ridge, and Elastic Net techniques, we identified drug resistance and comorbidities as the most significant predictors of survival. The findings highlight that penalized regression methods (LASSO and Elastic Net) outperform traditional stepwise selection, improving model stability and predictive accuracy. The results emphasize the need for targeted treatment strategies for high-risk TB patients, particularly those with drug resistance and comorbidities. These insights can aid in designing personalized interventions and optimizing clinical trial methodologies for better patient outcomes

Reference :-

1. World Health Organization. Global tuberculosis report 2017. Geneva: WHO; 2017. Licence: CC BY-NC-SA 3.0 IGO.
2. World Health Organization. Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 Global Report on Surveillance and Response. Geneva: WHO; 2010.
3. Faustini A, Hall AJ, Perucci CA. Risk factors for multidrug-resistant tuberculosis in Europe: a systematic review. Thorax. 2006;61(2):158–163.
4. Kundu D, Sharma N, Chadha S, Laokri S, Awungafac G, Jiang L, Asaria M. Analysis of multidrug-resistant tuberculosis (MDR-TB) financial protection policy: MDR-TB health insurance schemes in Chhattisgarh state, India. Health Econ Rev. 2018;8(1):3.
5. Singh JA, Upshur R, Padayatchi N. XDR-TB in South Africa: no time for denial or complacency. PLoS Med. 2007;4(1):e50.
6. Koul A, Arnoult E, Lounis N, Guillemont J, Andries K. The challenge of new drug discovery for tuberculosis. Nature. 2011;469(7331):483–490.
7. Assefa D, Seyoum B, Oljira L. Determinants of multidrug-resistant tuberculosis in Addis Ababa, Ethiopia. Infect Drug Resist. 2017;10:209–213.
8. Eldholm V, Balloux F. Antimicrobial resistance in Mycobacterium tuberculosis: the odd one out. Trends Microbiol. 2016;24(8):637–648.
9. Getachew T, Bayray A, Weldearegay B. Survival and predictors of mortality among patients under multidrug-resistant tuberculosis treatment in Ethiopia: St. Peter's Specialized Tuberculosis Hospital, Ethiopia. Int J Pharm Sci Res. 2013;4(2):776–783.
10. Falzon D, Schünemann HJ, Harausz E, González-Angulo L, Lienhardt C, Jaramillo E, Weyer K. World Health Organization treatment guidelines for

drug-resistant tuberculosis, 2016 update. Eur Respir J. 2017;49(3):1602308.

11. Espinal MA, Laszlo A, Simonsen L, Boulahbal F, Kim SJ, Reniero A, Hoffner S, Rieder HL, Binkin N, Dye C. Global trends in resistance to antituberculosis drugs. N Engl J Med. 2001;344(17):1294–1303.
12. Tyrrell F, Stafford C, Yakrus M, Youngblood M, Hill A, Johnston S. Trends in testing for Mycobacterium tuberculosis complex from US public health laboratories, 2009–2013. Public Health Rep. 2017;132(1):56–64.
13. Tacconelli E, Cataldo M, Dancer S, De Angelis G, Falcone M, Frank U, Kahlmeter G, Pan A, Petrosillo N, Rodríguez-Baño J. ESCMID guidelines for the management of the infection control measures to reduce transmission of multidrug-resistant Gram-negative bacteria in hospitalized patients. Clin Microbiol Infect. 2014;20(Suppl 1):1–55.
14. Alrabiah K, Al Alola S, Al Banyan E, Al Shaalan M, Al Johani S. Characteristics and risk factors of hospital-acquired methicillin-resistant Staphylococcus aureus (HA-MRSA) infection in pediatric patients in a tertiary care hospital in Riyadh, Saudi Arabia. Int J Pediatr Adolesc Med. 2016;3(2):71–77.
15. Tarai B, Das P, Kumar D. Recurrent challenges for clinicians: emergence of methicillin-resistant Staphylococcus aureus, vancomycin resistance, and current treatment options. J Lab Physicians. 2013;5(2):71–78.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

 UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

A Comparative Study of Time Series Models for Millets Yield Prediction in Tamil Nadu

V. Munaiah[1] , P. Maheswari[2] , T. Gangaram[3] , K.Murali[4] , G.Mokesh Rayalu[5]

[1]Assistant Professor, Dept. of Statistics, PVKN Govt. College (A), Chittoor.
[2]Assistant Professor, Dept. of Statistics, Govt Degree College for Women, Srikalahasti.
[3]Assistant Professor, Dept. of Statistics, SVA Govt. College for Men, Srikalahasti.
[4]Academic Consultant, Dept. of Statistics, S.V University, Tirupati.
Corresponding Author **
[5]Assistant Professor Grade 2 , Department of Mathematics, School of Advanced Sciences,VIT,Vellore
mokesh.g@gmail.com

ABSTRACT

The purpose of this research was to evaluate and contrast time series models for predicting millet yield in the context of agricultural output in Tamil Nadu, India. The study's overarching goal is to shed light on the prediction capacities of ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models as they pertain to millet farming by examining their respective applications. The purpose of this research is to determine the efficacy of the ARIMA and SARIMA models in capturing the nuances of millet yield fluctuations by using historical data covering various influential factors like climatic variations, soil conditions, and agricultural practices. To further develop prediction approaches for sustainable agricultural planning and decision-making, this comparison sheds light on seasonal trends, trend changes, and other dynamic components driving millet output. In order to maximize millet output and guarantee food security in the Tamil Nadu area, this study is crucial in promoting the implementation of data-driven solutions.

Keywords: Millets, ARIMA, SARIMA, Forecasting.

INTRODUCTION

Millets, a key part of Tamil Nadu's agricultural landscape, have risen to prominence in recent years as a result of their resistance to the effects of unfavorable weather conditions and the nutritional value they provide for maintaining food security. There is an increasing demand for the development of reliable predictive models that are capable of accurately predicting millet yields. This demand is being driven by the growing significance of millet farming in the context of sustainable agriculture. In this study, a complete comparative analysis of time series models to forecast millet yields in Tamil Nadu is carried out. A particular emphasis is placed on the AutoRegressive Integrated Moving Average (ARIMA) model and the Seasonal AutoRegressive Integrated Moving Average (SARIMA) model. This research aims to

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

identify the strengths and limitations of the ARIMA and SARIMA models in accurately capturing the complex dynamics of millet yield fluctuations. This will be accomplished by utilizing historical data and taking into consideration a wide range of influential factors such as variations in the climate, the quality of the soil, and various agricultural practices.

The fundamental objective of this investigation is to develop a comprehensive understanding of the seasonal patterns, trend changes, and other dynamic components that significantly influence millet output. This research seeks to provide useful insights into the temporal variability of millet yields by comparing the predictive capacities of the ARIMA and SARIMA models. These insights will enable stakeholders to make educated decisions about agricultural planning and policy formation. Not only does the incorporation of sophisticated time series analytic techniques contribute to the refining of predictive approaches for millet cultivation, but it also aids the development of sustainable agricultural practices that are adapted to the specific demands of Tamil Nadu's agro-ecological landscape. This is because of the fact that these techniques are able to better account for the interplay between environmental and agronomic factors. This comparative study has a significant amount of promise for building agricultural resilience and boosting food security, thereby ensuring the continued growth of the agricultural sector in Tamil Nadu, which is essential for the state's economy.

OBJECTIVES:

1. To examine the historical time series data of millet yield in Tamil Nadu and identify the seasonal and trend patterns that influence yield fluctuations.

2. To apply the ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models for millet yield prediction and assess their effectiveness in capturing the intricate dynamics of millet production.

3. To evaluate the performance of the ARIMA and SARIMA models in terms of their ability to account for seasonal variations and long-term trends in millet yield, aiming to determine the model that provides the most accurate and reliable predictions for millet production in the region.

4. To investigate the key factors influencing millet production, including climatic variability, soil quality, and agricultural practices, and to ascertain their impact on the predictive capabilities of the ARIMA and SARIMA models.

5. To provide valuable insights for farmers, policymakers, and stakeholders, facilitating informed decision-making for the implementation of sustainable agricultural strategies and policies that foster the growth and stability of millet cultivation in Tamil Nadu.

By fulfilling these objectives, this study endeavors to contribute to the refinement of predictive methodologies for millet cultivation, supporting the development of sustainable agricultural practices tailored to the unique requirements of Tamil Nadu's agricultural landscape.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

LITERATURE REVIEW:

Kour et al. (2017) analyzed pearl millet (Pennisetum glaucum), a commonly farmed cereal crop that ranks fourth in global cultivation behind rice, wheat, and sorghum. Despite rising yields, pearl millet cultivation in Gujarat, India, has declined during the previous two decades. Pearl millet production forecasts are especially important in semi-arid locations like Gujarat, where precipitation lasts only four months. This study predicts Gujarat pearl millet productivity using the ARIMA model. The current study collected time series data on pearl millet productivity (kg/ha) in Gujarat from 1960–61 to 2011–12. Gandhinagar's Directorate of Agriculture and, partially, the Directorate of Economics and Statistics provided the data. RMAPE, MAD, and RMSE values are used to validate the ARIMA model. As seen by its RMAPE score below 6%, the ARIMA (0, 1, 1) model performs well.

In their study, Vijay and Mishra (2018) investigated Time series prediction is important in natural science, agriculture, engineering, and economics. This study compares the classical time series ARIMA model to the artificial neural network model (ANN) to evaluate its flexibility in time series forecasting. The dataset includes pearl millet (bajra) crop area and production in thousands of hectares (ha) and metric tons (MT). The publication "Agricultural Statistics at a Glance 2014–15" provided 1955–56 to 2014–15 data. To test the methodology, Karnataka, India, was chosen. The user's'sext is scholarly. An experiment shows that artificial neural network (ANN) models outperform autoregressive integrated moving average (ARIMA) models in root mean square error (RMSE). RMSE, MAPE, and MSE are common measures in statistics and data analysis.

According to the findings of Saranyadevi and Kachi's study (2017), They evaluate the predicted performance of a time-series analytic method for paddy production trends in the state of Tamil Nadu, which is located in India. There was a study that looked at data on rice crop output from 1960 to 2015, and it made production predictions for the years 2016–2020 using models such as ARIMA (Autor Regressive Integrated Moving Average), basic exponential smoothing, brown exponential smoothing, and damped exponential smoothing.

METHODOLOGY

ARIMA Model (p,d,q):

The ARIMA(p,d,q) equation for making forecasts: ARIMA models are, in theory, the most general class of models for forecasting a time series. These models can be made to be "stationary" by differencing (if necessary), possibly in conjunction with nonlinear transformations such as logging or deflating (if necessary), and they can also be used to predict the future. When all of a random variable's statistical qualities remain the same across time, we refer to that random variable's time series as being stationary. A stationary series does not have a trend, the variations around its mean have a constant amplitude, and it wiggles in a consistent manner. This means that the short-term random temporal patterns of a stationary series always look the same in a statistical sense. This last criterion means that it has maintained its

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

autocorrelations (correlations with its own prior deviations from the mean) through time, which is equal to saying that it has maintained its power spectrum over time. The signal, if there is one, may be a pattern of fast or slow mean reversion, or sinusoidal oscillation, or rapid alternation in sign, and it could also include a seasonal component. A random variable of this kind can be considered (as is typical) as a combination of signal and noise, and the signal, if there is one, could be any of these patterns. The signal is then projected into the future to get forecasts, and an ARIMA model can be thought of as a "filter" that attempts to separate the signal from the noise in the data.

The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. That is:

Predicted value of Y = a constant and/or a weighted sum of one or more recent values of Y and/or a weighted sum of one or more recent values of the errors.

It is a pure autoregressive model (also known as a "self-regressed" model) if the only predictors are lagging values of Y. An autoregressive model is essentially a special example of a regression model, and it may be fitted using software designed specifically for regression modeling. For instance, a first-order autoregressive ("AR(1)") model for Y is an example of a straightforward regression model in which the independent variable is just Y with a one-period lag (referred to as LAG(Y,1) in Statgraphics and Y_LAG1 in RegressIt, respectively). Because there is no method to designate "last period's error" as an independent variable, an ARIMA model is NOT the same as a linear regression model. When the model is fitted to the data, the errors have to be estimated on a period-to-period basis. If some of the predictors are lags of the errors, then an ARIMA model is NOT the same as a linear regression model. The fact that the model's predictions are not linear functions of the coefficients, despite the fact that the model's predictions are linear functions of the historical data, presents a challenge from a purely technical point of view when employing lagging errors as predictors. Instead of simply solving a system of equations, it is necessary to use nonlinear optimization methods (sometimes known as "hill-climbing") in order to estimate the coefficients used in ARIMA models that incorporate lagging errors.

Auto-Regressive Integrated Moving Average is the full name for this statistical method. Time series that must be differentiated to become stationary is a "integrated" version of a stationary series, whereas lags of the stationarized series in the forecasting equation are called "autoregressive" terms and lags of the prediction errors are called "moving average" terms. Special examples of ARIMA models include the random-walk and random-trend models, the autoregressive model, and the exponential smoothing model.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of nonseasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

- The forecasting equation is constructed as follows. First, let y denote the d[th] difference of Y, which means:
- If d=0: $y_t = Y_t$
- If d=1: $y_t = Y_t - Y_{t-1}$
- If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$
- Note that the second difference of Y (the d=2 case) is not the difference from 2 periods ago. Rather, it is the first-difference-of-the-first difference, which is the discrete analog of a second derivative, i.e., the local acceleration of the series rather than its local trend.
- In terms of y, the general forecasting equation is:
- $\hat{Y}_t = \mu + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$

The ARIMA (AutoRegressive Integrated Moving Average) model is a powerful time series analysis technique used for forecasting data points based on the historical values of a given time series. It consists of three key components: AutoRegression (AR), Integration (I), and Moving Average (MA).

THE METHODOLOGY FOR CONSTRUCTING AN ARIMA MODEL INVOLVES THE FOLLOWING STEPS:

1. Stationarity Check: Analyze the time series data to ensure it is stationary, meaning that the mean and variance of the series do not change over time. Stationarity is essential for ARIMA modeling.

2. Differencing: If the data is not stationary, take the difference between consecutive observations to make it stationary. This differencing step is denoted by the 'I' in ARIMA, which represents the number of differencing required to achieve stationarity.

3. Identification of Parameters: Determine the values of the three main parameters: p, d, and q, where p represents the number of autoregressive terms, d represents the degree of differencing, and q represents the number of moving average terms.

4. Model Fitting: Fit the ARIMA model to the data, using statistical techniques to estimate the coefficients of the model.

5. Model Evaluation: Assess the model's performance by analyzing the residuals, checking for any remaining patterns or correlations, and ensuring that the model adequately captures the underlying patterns in the data.

6. Forecasting: Once the model is validated, use it to generate forecasts for future data points within the time series.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

SEASONAL ARIMA:

By including seasonal variations into the ARIMA model, Seasonal ARIMA (SARIMA) is a robust technique for analyzing and forecasting time series data. It works well for examining and forecasting sales data, weather patterns, and economic indicators that are subject to seasonal changes. Financial markets, economics, and even meteorology all make use of SARIMA models.

Mathematical Formulation:

The SARIMA model is denoted as SARIMA(p,d,q)(P,Q,D)[s], where:

- Non-seasonal autoregressive (p), differencing (d), and moving average (q) are the possible orders of analysis.
- The seasonal autoregressive, differencing, and moving average orders are denoted by the letters P, D, and Q, respectively.
- The length of one season is denoted by the symbol S.

The SARIMA model can be represented as follows:

$$(1 - \varphi_1 B - \cdots - \varphi_P B^P)(1 - \varphi_1 B^{VS} - \cdots - \varphi_P B^{VS})^P (B^{VS})^D Y_t$$
$$= (1 + \theta_1 B + \cdots + \theta_P B^{\varphi})(1 + \theta_1 B^{VS} + \cdots + \theta_P B^{\varphi S})^A (B^{\varphi S})^K \varepsilon_t$$

Where:

- $\varphi_i$and θ_i are the autoregressive and moving average parameters, respectively.
- B and B^{VS} are the non- seasonal and seasonal backshift operators, respectively.
- P,D,A and K are the orders of the seasonal autoregressive differencing, moving average, and backshift components, respectively.
- Y_t represents the time series data at time t.
- ε_t denotes the white noise error term.

Real life application

One example of how SARIMA might be put to use in the real world is in the process of predicting quarterly sales data for a retail organization. The sales data frequently display seasonal patterns because of things like the different holiday seasons and different promotional periods. The company is able to examine previous sales data, recognize seasonal patterns, and make more accurate projections of future sales by using a model called SARIMA.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

Merits and Demerits:

- When applied to time series data, SARIMA models are able to distinguish between seasonal and non-seasonal patterns.
- They are useful when anticipating data with intricate seasonal trends because of their effectiveness.
- The SARIMA models can be altered to accommodate a wide variety of seasonal data types, which lends them flexibility and adaptability.
- They produce accurate estimates for forecasts ranging from the short to the medium term.
- SARIMA models can be complicated, particularly when dealing with a number of different seasonal components, which calls for a substantial amount of computational resources.
- Due to the complexity of the mathematical formulas, interpretation of the SARIMA results may be difficult for individuals who are not experts in the field.
- For SARIMA models to generate reliable forecasts, a significant quantity of historical data is necessary; however, this data may not always be accessible for all forms of data.

Preparation of Data:

- Prepare the time series data for analysis by collecting and cleaning it such that it is consistent and has no outliers or missing values.
- Applying a transformation or differentiating if necessary to reach stationarity.

Identification of Models:

- Determine the values of the AR and MA parameters during the season and the offseason by analyzing the ACF and PACF graphs.
- Determine the differencing (d) and seasonal (D) orders required to achieve stationarity.

Estimating Variables:

- Apply the SARIMA model's estimated parameters using estimation strategies like maximum likelihood.
- Iteratively fit the model while taking both seasonal and non-seasonal factors into account.

Model Evaluation and Adjustment:

- Examine diagnostic charts for evidence of residual randomness after a SARIMA model has been fitted to the data.
- Analyze the residuals using autocorrelation functions (ACF) plots, histograms, and the Ljung-Box test.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

ANALYSIS

ARIMA

The Augmented Dickey-Fuller (ADF) test was conducted on the time series data for millets production, denoted as data Millets. The purpose of this test was to assess the stationarity of the data.

The results of the ADF test indicate a Dickey-Fuller statistic of -7.0233, with a p-value of 0.01. With the p-value significantly lower than the chosen significance level of 0.05, there is strong evidence to reject the null hypothesis of non-stationarity in favor of the alternative hypothesis of stationarity. This suggests that the millets production time series data is stationary, indicating that the statistical properties of the data remain consistent over time. The confirmation of stationarity is crucial for the application of time series modeling techniques, ensuring reliable and accurate analysis of millets production trends for effective decision-making in the agricultural sector.

Time series data for millets production (data Millets) was analyzed using the auto.arima function to find the best ARIMA model for the data. Multiple potential ARIMA models and their associated Akaike Information Criterion (AIC) values were generated after the function was instructed to use the AIC for model selection.

ARIMA (2,0,2) (1,0,1) [12] with non-zero mean	Inf
ARIMA (0,0,0) with non-zero mean	2380.452
ARIMA (1,0,0) (1,0,0) [12] with non-zero mean	2384.104
ARIMA (0,0,1) (0,0,1) [12] with non-zero mean	2381
ARIMA (0,0,0) with zero mean	2758.668
ARIMA (0,0,0) (1,0,0) [12] with non-zero mean	2382.166
ARIMA (0,0,0) (0,0,1) [12] with non-zero mean	2381.964
ARIMA (0,0,0) (1,0,1) [12] with non-zero mean	2383.12
ARIMA (1,0,0) with non-zero mean	2382.136
ARIMA (0,0,1) with non-zero mean	2381.902
ARIMA (1,0,1) with non-zero mean	2383.033

As seen in the results, the millets production time series data were best suited by the ARIMA(0,0,0) model with a non-zero mean. This means that the model does not account for a zero mean or an autoregressive

ISSN PRINT 2319 1775 Online 2320 7876

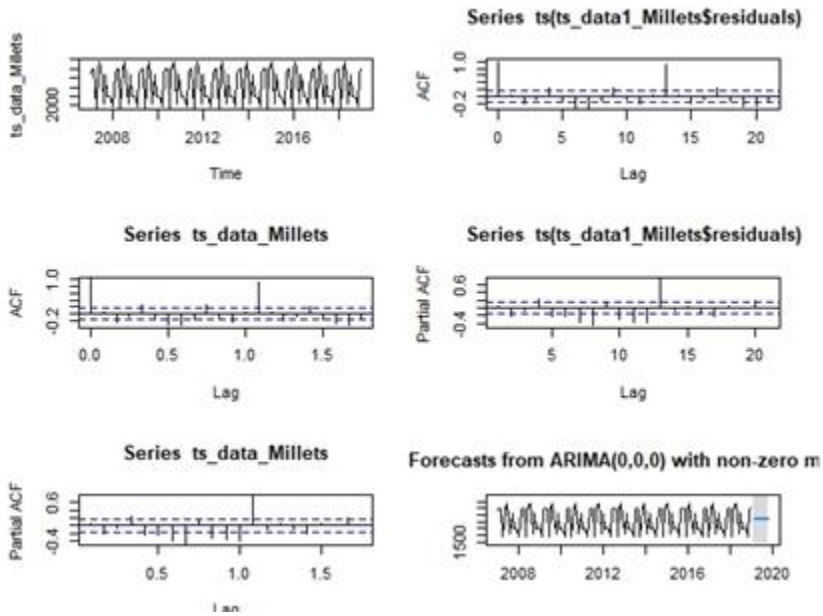
Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

term, but it does take into account a moving average. The model with the lowest AIC for assessing the millets production data was chosen since it had a value of 2380.452 in the calculation. If this model is further analyzed and interpreted, it can help agricultural planners and policymakers make better predictions and decisions on how to approach millets cultivation in the future.

The millets production time series data was modeled using the ARIMA(0,0,0) distribution with a non-zero mean. Non-zero mean coefficient estimates range from -3130.9931 to -72.7739 standard deviations from the mean.

Coefficient	value
Mean	3130.9931
S. E	72.7739
sigma^2	773259
log likelihood	-1188.23
AIC	2380.45
BIC	2386.41

The model's log likelihood was calculated to be -1188.23, and its variance was found to be 773259. The related values for the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were calculated to be 2380.45 and 2386.41, respectively.



The ARIMA(0,0,0) model was used to analyze a time series of millets production, and the results of these statistical analyses shed light on the model's parameters and goodness of fit. The existence of a non-zero mean coefficient in the millets production data is indicative of the presence of a trend or level. Millets

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

cultivation could benefit from further investigation using this model, as it could shed light on the underlying dynamics of production and lead to better decisions and planning.

Prediction information Lower (Lo 95) and higher (Hi 95) 95% confidence interval bounds for millets production are included alongside the point projections for millets in the table. Time series data for millets production, is used to construct forecasts using the ARIMA(0,0,0) model with a non-zero mean.

Point	Forecast	Lo 95	Hi 95
Feb 2019	3130.993	1407.496	4854.49
Mar 2019	3130.993	1407.496	4854.49
Apr 2019	3130.993	1407.496	4854.49
May 2019	3130.993	1407.496	4854.49
Jun 2019	3130.993	1407.496	4854.49
Jul 2019	3130.993	1407.496	4854.49
Aug 2019	3130.993	1407.496	4854.49
Sep 2019	3130.993	1407.496	4854.49
Oct 2019	3130.993	1407.496	4854.49

The projection indicates that millets output will remain largely constant at a predicted point value of 3130.993. The range in which the true millets production numbers are 95% likely to fall is estimated to be 1407.496–4854.49, with a lower 95% confidence interval of 1407.496 and an upper 95% confidence interval of 4854.49.

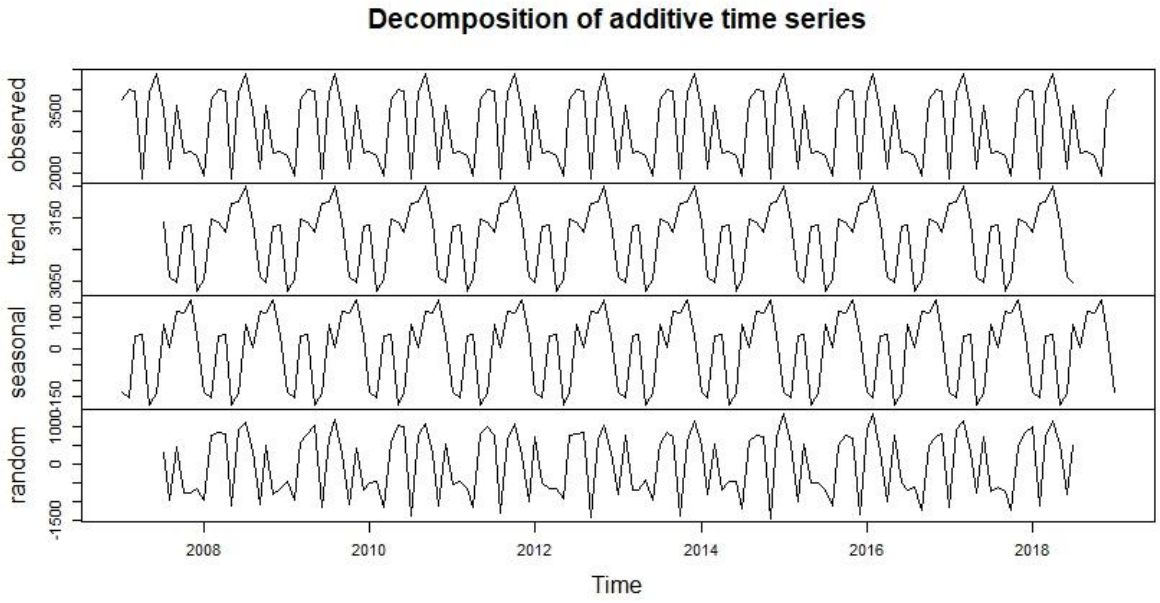
Stakeholders in the millets cultivation sector can use these predicted values and their associated confidence intervals to better anticipate production trends and make decisions regarding resource allocation, market planning, and agricultural management strategies.

The Box-Ljung test was run on the non-zero mean residuals of the ARIMA(0,0,0) model's predicted millets production values. This analysis was performed to check for autocorrelation in the model's residuals.

The p-value from the Box-Ljung test was 0.000579, and the X-squared value from the test was 21.77 (5 degrees of freedom). Significant autocorrelation in the residuals is strongly suggested, as the p-value is much smaller than the specified significance level of 0.05. This suggests that there may be features or trends in the millets production data that aren't accounted for by the ARIMA(0,0,0) model. If we want more accurate and reliable millets production projections, we may need to do more research or use different modeling methodologies.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021



SEASONAL ARIMA ANALYSIS

Millets output data is available from 2007 to 2019 at a frequency of 1 year. Values of output for each year are as follows: 3777, 4013, 3976, 1873, 3950, 4401, 3508, 2092, 3642, 2468, 2504, 2419, and 1941. These figures indicate the annual output of millets for the given time frame, so they can shed light on production patterns and trends over time.

Descriptive statistics used to summarize the millets production time series data shed light on the dataset's central tendency and dispersion. Over the course of that time frame, the lowest amount of millets produced was in 1873, and the highest was in 4401. Values of 2419 and 3950 are the 25th and 75th percentile quartiles, respectively. Half of the production values are below this point, and the other half are beyond it, as indicated by the median value of 3508. Taking into account all available data, we find that the mean millets production is $3,120.

These statistical indicators help to shed light on the diversity of millets output by highlighting its range of values and its central trend. As a result, stakeholders are able to make educated decisions and develop effective strategies to boost agricultural productivity and sustainability in the millets cultivation sector based on a more thorough understanding of the underlying trends, variations, and potential outliers in the production data.

To further evaluate the stationarity of the data, the differenced logarithm of the millets production time series (denoted by the notation diff(log(ts_Millets)) was subjected to the augmented Dickey-Fuller (ADF) test. The goal of the differencing procedure is to minimize trends and stabilize the variation so that

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021**

stationary patterns may be more easily identified. The Dickey-Fuller statistic was -4.7589, and the p-value was 0.01. These findings come from the ADF test. There is substantial evidence to reject the null hypothesis of non-stationarity in favor of the alternative hypothesis of stationarity, as the p-value is smaller than the specified significance level of 0.05. This indicates that there are no discernible trends or patterns in the differenced logarithm of the millet production data over time, suggesting that the data is stationary. These results are essential for developing suitable time series models and forecasting approaches, which in turn will allow for precise predictions and well-informed choices in the millets agriculture industry.

The logarithm of the millets production time series data is given by log(ts_Millets), and the auto Arima Model Log indicates the automated ARIMA model that was fitted to this data. The model was identified as ARIMA(0,1,0), suggesting that first-order differencing was required to reach stationarity.

Coefficient	Values
σ^2	0.1639
log likelihood	-6.17
AIC	14.35
AICc	14.75
BIC	14.83

The model's variance, was calculated to be 0.1639, and the log probability was -6.17. There were 14.35 for the Akaike Information Criterion (AIC), 14.75 for the AICc, and 14.83 for the Bayesian Information Criterion (BIC).

It is impossible to assess the ARIMA model's ability to explain the millets production time series without these statistical measurements. To compare and select the best model for assessing and forecasting millets production patterns, the AIC, AICc, and BIC values are calculated. Forecasting and decision-making in the millets agriculture sector can benefit greatly from the ARIMA(0,1,0) model and its associated parameters and statistical metrics.

After fitting the ARIMA (0,1,0) model to the logarithm of the millets production time series data, the Box-Ljung test was performed on the residuals. This analysis was performed to check for autocorrelation in the model's residuals.

Coefficient	Values
x^2	3.5805
df	1
P-value	0.05846

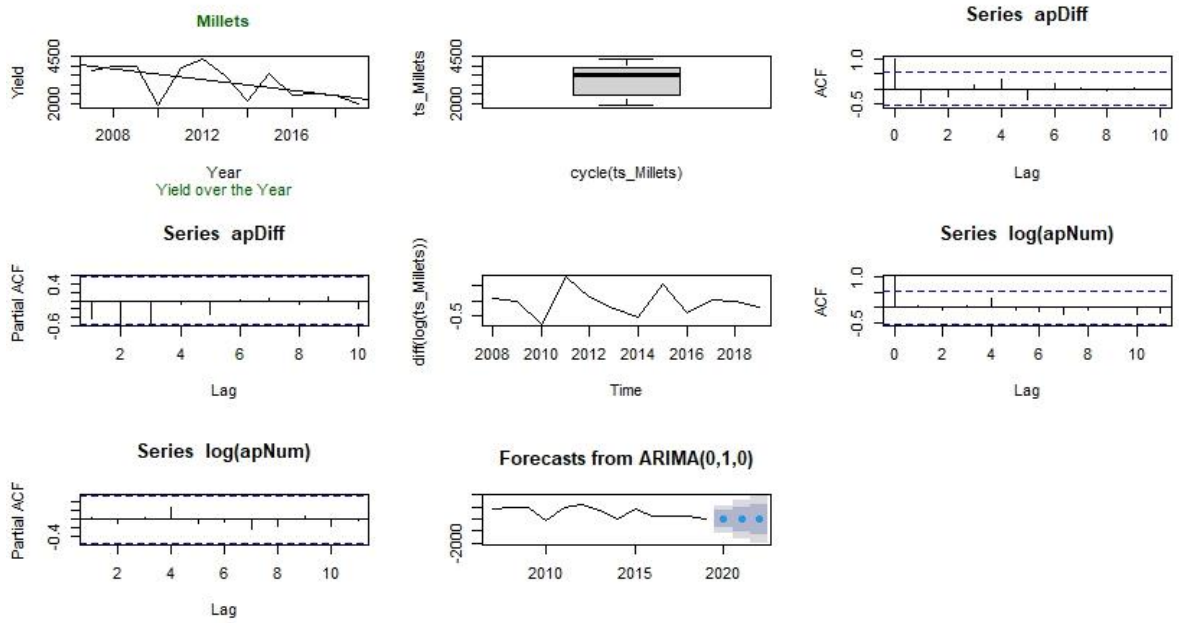
The X-squared value for the Box-Ljung test came out to be 3.5805 with 1 degree of freedom, yielding a p-value of 0.05846. As the p-value is larger than the threshold for statistical significance (0.05), it cannot be concluded that there is no autocorrelation in the residuals. If the ARIMA model's residuals look like

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

white noise, then the model does a good job of capturing the dynamics underlying the millets production data.

Logarithmic millets production time series data are well represented by the ARIMA(0,1,0) model, as the residuals closely follow the white noise assumption. That means we can confidently predict and analyze changes in millets production for agriculturally informed decision making because the model sufficiently accounts for the patterns and structures inherent in the data.



CONCLUSION

The following findings emerge from an examination of the ARIMA and Seasonal ARIMA models applied to the time series data of millets production:

First, an ARIMA model was used to analyze the millets output data; specifically, an ARIMA(0,0,0) model with a non-zero mean. Non-zero mean standard error was estimated to be 72.7739, with the coefficient being 3130.9931. The model had an AIC of 2380.45, an AICc of 2380.54, and a BIC of 2386.41, and its log likelihood was -1188.23. There may be autocorrelation in the residuals of the ARIMA model, as shown by a significant result from the Box-Ljung test (X-squared = 21.77, df = 5, p-value = 0.000579).

The logarithm of the millets production data was modeled using the Seasonal ARIMA(0,1,0) model. Log likelihood was -6.17, and the model's variance was found to be 0.1639. The model's AIC, AICc, and BIC

all came in at 14.83. According to the results of a Box-Ljung test conducted on the Seasonal ARIMA model's residuals, the residuals are consistent with the white noise assumption (X-squared = 3.5805, df = 1, p-value = 0.05846).

Overall, autocorrelation difficulties were seen in the residuals of the ARIMA model with a non-zero mean, suggesting that the model may have been inadequate in capturing the underlying dynamics of the millets production data. The logarithm of the millets production data was reliably modeled using the Seasonal ARIMA model, whose residuals resembled white noise. To improve the precision and consistency of the millets production forecasts, it may be required to further investigate and refine the model.

REFERENCES

1. Alabi, O. O., Lawal, A. F., & Awoyinka, Y. A. (2014). Industrial demand analysis for millet in Kaduna and Kano States of Nigeria. Asian Journal of Agricultural Extension, Economics and Sociology, 3(6), 521-529.
2. Bellundagi, V., Umesh, K. B., & Ravi, S. C. (2016). Growth dynamics and forecasting of finger millet (Ragi) production in Karnataka. Economic Affairs, 61(2), 195.
3. Gibon, F., Pellarin, T., Román-Cascón, C., Alhassane, A., Traoré, S., Kerr, Y., ... & Baron, C. (2018). Millet yield estimates in the Sahel using satellite derived soil moisture time series. Agricultural and Forest Meteorology, 262, 100-109.
4. http://www.tnagriculture.in/dashboard/report/05_01.pdf
5. Kour, S., Pradhan, U. K., Paul, R. K., & Vaishnav, P. R. (2017). Forecasting of pearl millet productivity in Gujarat under time series framework. Economic Affairs, 62(1), 121-127.
6. Ly, R., Dia, K., Diallo, M., Ahid, M., & Ceesay, B. (2020). Pearl millet production forecasts for selected west African countries: Côte d'Ivoire, Mali, Burkina Faso, Sierra Leone, The Gambia, and Senegal. Covid-19 Bulletin, (14).
7. Mounkaila, Y., Garba, I., & Moussa, B. (2019). Yield prediction under associated millet and cowpea crops in the Sahelian zone. Afr. J. Agric. Res, 14, 1613-1620.
8. Muhammad, s., garba, i., & audu, a. (2021). Modelling and forecasting of millet production in nigeria. Bima journal of science and technology (2536-6041), 5(01), 99-109.
9. Nireesha, V., Rao, V. S., Rao, D. V. S., & Reddy, G. R. (2016). A study on forecasting of area, production and productivity of pearl millet in Andhra Pradesh. Journal of Research ANGRAU, 44(3/4), 119-126.
10. Sawa, B. A., & Ibrahim, A. A. (2011). Forecast models for the yield of millet and sorghum in the semi arid region of Northern Nigeria using dry spell parameters. Asian Journal of Agricultural Sciences, 3(3), 187-191.
11. Tripathi, S. K., Mishra, P., & Sahu, P. K. (2013). Past trends and forecasting in Area, production and yield of pearl millet in India using ARIMA model. Environ Ecol, 31, 1701-1708.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 4, April 2021

12. Verma, V. K., Jheeba, S. S., Kumar, P., & Singh, S. P. (2016). Price forecasting of bajra (pearl millet) in Rajasthan: ARIMA model. International Journal of Agriculture Sciences, 8(9), 1103-1106.
13. Vijay, N., & Mishra, G. C. (2018). Time Series forecasting using ARIMA and ANN Models for production of pearl millet (BAJRA) crop of Karnataka, India. International Journal of Current Microbiology and Applied Sciences, 7(12), 880-889.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022

Time Series Analysis for Forecasting Paddy Production in Tamil Nadu

T. Gangaram[1] , V. Munaiah[2] , P. Maheswari[3] , K.Murali[4] , G.Mokesh Rayalu[5]**

[1]Assistant Professor, Dept. of Statistics, SVA Govt. College for Men, Srikalahasti.
[2]Assistant Professor, Dept. of Statistics, PVKN Govt. College (A), Chittoor.
[3]Assistant Professor, Dept. of Statistics, Govt Degree College for Women, Srikalahasti.
[4]Academic Consultant, Dept. of Statistics, S.V University, Tirupati.
Corresponding Author **
[5]Assistant Professor Grade 2,Department of Mathematics, School of Advanced Sciences, VIT,Vellore
mokesh.g@gmail.com

ABSTRACT

In order to predict paddy output in Tamil Nadu, this study uses time series analysis using the robust ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models. This research makes use of historical data covering a number of years in order to investigate the complex temporal patterns and seasonal fluctuations that greatly affect paddy yields in the area. The research is conducted with the intention of developing a reliable framework for forecasting future paddy output, taking into account relevant aspects such as meteorological fluctuations, irrigation techniques, and governmental interventions. Farmers, policymakers, and others in the paddy cultivation sector can greatly benefit from a deeper understanding of the non-stationary and seasonal components within the industry thanks to the combination of ARIMA and SARIMA models. Sustainable and resilient paddy production in Tamil Nadu is ensured thanks to this study's contribution to the improvement of agricultural plans and policies.

Keywords: Paddy, ARIMA, SARIMA, Forecasting.

INTRODUCTION

Producing paddy acts as a cornerstone of Tamil Nadu's agricultural economy. It plays a key role in maintaining food security and sustaining the livelihoods of millions of people, making it one of the most important agricultural activities in the state. Given the region's susceptibility to climate changes and the ever-evolving agricultural techniques, it is becoming increasingly important to have a solid understanding of the temporal patterns and complicated dynamics that control paddy farming. This research attempts to provide a complete framework for forecasting paddy output in Tamil Nadu. It does so by utilizing time

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022**

series analytic techniques, in particular the ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models. This study aims to untangle the temporal fluctuations and identify the important factors that influence paddy yield variability. It does so by utilizing historical data and taking into consideration a variety of seasonal and trend components. Ultimately, this study will use this information to make predictions. The combination of the ARIMA and SARIMA models enables a more nuanced understanding of the seasonal patterns and their impact on paddy production. This, in turn, makes it easier for farmers, policymakers, and other stakeholders invested in the environmentally responsible growth of Tamil Nadu's agricultural landscape to make informed decisions. This research has tremendous promise in improving the resilience and productivity of paddy agriculture and, as a result, making a contribution to the overall agricultural sustainability and food security in the region.

OBJECTIVE

1. To analyze historical time series data of paddy production in Tamil Nadu and identify the underlying trends and patterns affecting production fluctuations.

2. To apply the ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models to develop accurate and reliable forecasts for paddy production in the region.

3. To assess the impact of seasonal variations, climatic factors, and agricultural practices on paddy production, considering both short-term and long-term implications.

4. To compare the performance of the ARIMA and SARIMA models in capturing the seasonal variability and fluctuations in paddy production, thereby determining the most suitable model for forecasting in the context of Tamil Nadu's agricultural landscape.

5. To provide valuable insights for farmers, policymakers, and stakeholders, enabling informed decision-making for sustainable agricultural planning and policy formulation aimed at enhancing paddy production and ensuring food security in Tamil Nadu.

By achieving these objectives, this study aims to contribute to the development of robust forecasting methodologies and data-driven strategies that will support the resilience and growth of the paddy cultivation sector in Tamil Nadu, fostering sustainable agricultural practices and bolstering the region's agricultural productivity.

LITERATURE SURVEY

Amarender and Ashwini Darekar invested India produces the second-most paddy in the world. About 35% of net cultivated land and 50% of farmers grow paddy annually. Future harvest prices determine farmers' paddy acreage decisions. This research proposes a method to forecast harvest prices and applies it to kharif

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022

2017–18. AGMARK's monthly average paddy prices from January 2006 to December 2016 were used. The ARIMA (Box-Jenkins) model predicted paddy prices. R was used to estimate model parameters. The model's goodness of fit was assessed using AIC, BIC, and MAPE. India-wide paddy price forecasts were best with the ARIMA model. September–November is the kharif paddy harvest. For the 2017-18 kharif harvest, paddy prices are expected to range from Rs. 1,600 to 2,200 per quintal.

The study by Saranyadevi and Kachi (2017), In this study, they investigate the predictive performance of a time-series analytic method for paddy production patterns in the Indian state of Tamil Nadu. There was a study that looked at paddy crop production data from 1960 to 2015 and predicted production for 2016–2020 using ARIMA (Autor Regressive Integrated Moving Average), simple exponential smoothing, brown exponential smoothing, and damped exponential smoothing models.

Joshua et al., (2021) Each model is evaluated using R2, RMSE, MAE, MSE, MAPE, CV, and NMSE. The GRNN method outperforms other assessment measures, including R2, RMSE, MAE, MSE, MAPE, CV, and NSME, with values of 0.9863, 0.2295, 0.1290, 0.0526, 1.3439, 0.0255, and 0.0136. These data show that the system estimates crop yield better than other methods. The Generalized Regression Neural Network (GRNN) model is compared to other models in literature studies. Using appropriate metrics, the GRNN model has greater prediction accuracy.

Methodology

ARIMA Model (p,d,q):

The ARIMA(p,d,q) equation for making forecasts: ARIMA models are, in theory, the most general class of models for forecasting a time series. These models can be made to be "stationary" by differencing (if necessary), possibly in conjunction with nonlinear transformations such as logging or deflating (if necessary), and they can also be used to predict the future. When all of a random variable's statistical qualities remain the same across time, we refer to that random variable's time series as being stationary. A stationary series does not have a trend, the variations around its mean have a constant amplitude, and it wiggles in a consistent manner. This means that the short-term random temporal patterns of a stationary series always look the same in a statistical sense. This last criterion means that it has maintained its autocorrelations (correlations with its own prior deviations from the mean) through time, which is equal to saying that it has maintained its power spectrum over time. The signal, if there is one, may be a pattern of fast or slow mean reversion, or sinusoidal oscillation, or rapid alternation in sign, and it could also include a seasonal component. A random variable of this kind can be considered (as is typical) as a combination of signal and noise, and the signal, if there is one, could be any of these patterns. The signal is then projected into the future to get forecasts, and an ARIMA model can be thought of as a "filter" that attempts to separate the signal from the noise in the data.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022

The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. That is:

Predicted value of Y = a constant and/or a weighted sum of one or more recent values of Y and/or a weighted sum of one or more recent values of the errors.

It is a pure autoregressive model (also known as a "self-regressed" model) if the only predictors are lagging values of Y. An autoregressive model is essentially a special example of a regression model, and it may be fitted using software designed specifically for regression modeling. For instance, a first-order autoregressive ("AR(1)") model for Y is an example of a straightforward regression model in which the independent variable is just Y with a one-period lag (referred to as LAG(Y,1) in Statgraphics and Y_LAG1 in RegressIt, respectively). Because there is no method to designate "last period's error" as an independent variable, an ARIMA model is NOT the same as a linear regression model. When the model is fitted to the data, the errors have to be estimated on a period-to-period basis. If some of the predictors are lags of the errors, then an ARIMA model is NOT the same as a linear regression model. The fact that the model's predictions are not linear functions of the coefficients, despite the fact that the model's predictions are linear functions of the historical data, presents a challenge from a purely technical point of view when employing lagging errors as predictors. Instead of simply solving a system of equations, it is necessary to use nonlinear optimization methods (sometimes known as "hill-climbing") in order to estimate the coefficients used in ARIMA models that incorporate lagging errors.

Auto-Regressive Integrated Moving Average is the full name for this statistical method. Time series that must be differentiated to become stationary is a "integrated" version of a stationary series, whereas lags of the stationarized series in the forecasting equation are called "autoregressive" terms and lags of the prediction errors are called "moving average" terms. Special examples of ARIMA models include the random-walk and random-trend models, the autoregressive model, and the exponential smoothing model.

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of nonseasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.
- The forecasting equation is constructed as follows. First, let y denote the d[th] difference of Y, which means:
- If d=0: $y_t = Y_t$
- If d=1: $y_t = Y_t - Y_{t-1}$
- If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022

- Note that the second difference of Y (the d=2 case) is not the difference from 2 periods ago. Rather, it is the first-difference-of-the-first difference, which is the discrete analog of a second derivative, i.e., the local acceleration of the series rather than its local trend.
- In terms of y, the general forecasting equation is:
- $\hat{Y}_t = \mu + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$

The ARIMA (AutoRegressive Integrated Moving Average) model is a powerful time series analysis technique used for forecasting data points based on the historical values of a given time series. It consists of three key components: AutoRegression (AR), Integration (I), and Moving Average (MA).

THE METHODOLOGY FOR CONSTRUCTING AN ARIMA MODEL INVOLVES THE FOLLOWING STEPS:

1. Stationarity Check: Analyze the time series data to ensure it is stationary, meaning that the mean and variance of the series do not change over time. Stationarity is essential for ARIMA modeling.

2. Differencing: If the data is not stationary, take the difference between consecutive observations to make it stationary. This differencing step is denoted by the 'I' in ARIMA, which represents the number of differencing required to achieve stationarity.

3. Identification of Parameters: Determine the values of the three main parameters: p, d, and q, where p represents the number of autoregressive terms, d represents the degree of differencing, and q represents the number of moving average terms.

4. Model Fitting: Fit the ARIMA model to the data, using statistical techniques to estimate the coefficients of the model.

5. Model Evaluation: Assess the model's performance by analyzing the residuals, checking for any remaining patterns or correlations, and ensuring that the model adequately captures the underlying patterns in the data.

6. Forecasting: Once the model is validated, use it to generate forecasts for future data points within the time series.

SEASONAL ARIMA:

By including seasonal variations into the ARIMA model, Seasonal ARIMA (SARIMA) is a robust technique for analyzing and forecasting time series data. It works well for examining and forecasting sales data, weather patterns, and economic indicators that are subject to seasonal changes. Financial markets, economics, and even meteorology all make use of SARIMA models.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022

Mathematical Formulation:

The SARIMA model is denoted as SARIMA(p,d,q)(P,Q,D)[s], where:

- Non-seasonal autoregressive (p), differencing (d), and moving average (q) are the possible orders of analysis.
- The seasonal autoregressive, differencing, and moving average orders are denoted by the letters P, D, and Q, respectively.
- The length of one season is denoted by the symbol S.

The SARIMA model can be represented as follows:

$$(1 - \varphi_1 B - \cdots - \varphi_P B^P)(1 - \varphi_1 B^{VS} - \cdots - \varphi_P B^{VS})^P (B^{VS})^D Y_t$$
$$= (1 + \theta_1 B + \cdots + \theta_P B^{\varphi})(1 + \theta_1 B^{VS} + \cdots + \theta_P B^{\varphi S})^A (B^{\varphi S})^K \varepsilon_t$$

Where:

- $\varphi_i$and θ_i are the autoregressive and moving average parameters, respectively.
- B and B^{VS} are the non- seasonal and seasonal backshift operators, respectively.
- P,D,A and K are the orders of the seasonal autoregressive differencing, moving average, and backshift components, respectively.
- Y_t represents the time series data at time t.
- ε_t denotes the white noise error term.

Real life application

One example of how SARIMA might be put to use in the real world is in the process of predicting quarterly sales data for a retail organization. The sales data frequently display seasonal patterns because of things like the different holiday seasons and different promotional periods. The company is able to examine previous sales data, recognize seasonal patterns, and make more accurate projections of future sales by using a model called SARIMA.

Merits and Demerits:

- When applied to time series data, SARIMA models are able to distinguish between seasonal and non-seasonal patterns.
- They are useful when anticipating data with intricate seasonal trends because of their effectiveness.
- The SARIMA models can be altered to accommodate a wide variety of seasonal data types, which lends them flexibility and adaptability.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022**

- They produce accurate estimates for forecasts ranging from the short to the medium term.
- SARIMA models can be complicated, particularly when dealing with a number of different seasonal components, which calls for a substantial amount of computational resources.
- Due to the complexity of the mathematical formulas, interpretation of the SARIMA results may be difficult for individuals who are not experts in the field.
- For SARIMA models to generate reliable forecasts, a significant quantity of historical data is necessary; however, this data may not always be accessible for all forms of data.

Preparation of Data:

- Prepare the time series data for analysis by collecting and cleaning it such that it is consistent and has no outliers or missing values.
- Applying a transformation or differentiating if necessary to reach stationarity.

Identification of Models:

- Determine the values of the AR and MA parameters during the season and the offseason by analyzing the ACF and PACF graphs.
- Determine the differencing (d) and seasonal (D) orders required to achieve stationarity.

Estimating Variables:

- Apply the SARIMA model's estimated parameters using estimation strategies like maximum likelihood.
- Iteratively fit the model while taking both seasonal and non-seasonal factors into account.

Model Evaluation and Adjustment:

- Examine diagnostic charts for evidence of residual randomness after a SARIMA model has been fitted to the data.
- Analyze the residuals using autocorrelation functions (ACF) plots, histograms, and the Ljung-Box test.

Analysis

ARIMA Models

In the analysis of the paddy production data from Tamil Nadu, several steps were undertaken to identify an appropriate time series model. The data was initially examined for stationarity through visual inspection of the plot and confirmed using the Auto correlation function (ACF) and Partial ACF (PADF) tests.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022

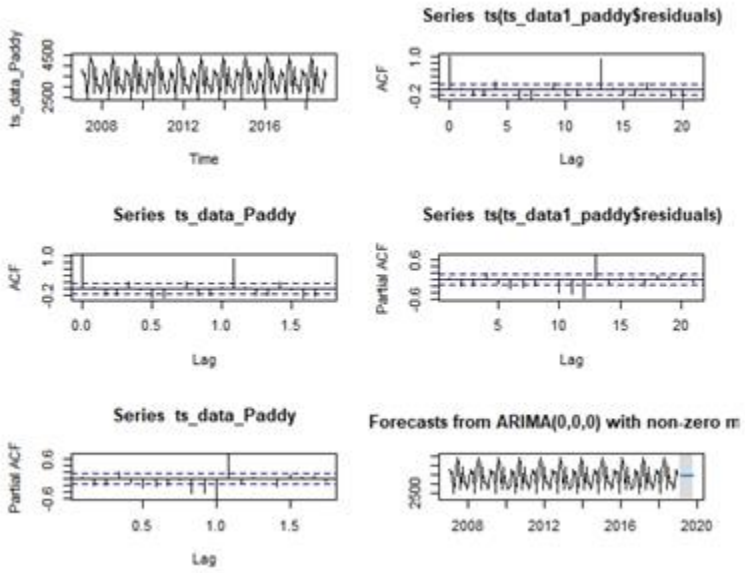
Following this, the `auto.arima` function was applied to determine the best-fitting model. The function iteratively evaluated various combinations of AR, MA, and differencing orders to select the model that exhibited the lowest information criterion values, signifying a good fit. This comprehensive process allowed for the identification of a suitable SARIMA model that can accurately capture the seasonal and non-seasonal patterns within the paddy production data of Tamil Nadu.

Models	Values
ARIMA (2,0,2) (1,0,1) [12] with non-zero mean	Inf
ARIMA (0,0,0) with non-zero mean	2267.311
ARIMA (1,0,0) (1,0,0) [12] with non-zero mean	2271.146
ARIMA (0,0,1) (0,0,1) [12] with non-zero mean	2268.206
ARIMA (0,0,0) with zero mean	2774.681
ARIMA (0,0,0) (1,0,0) [12] with non-zero mean	2269.197
ARIMA (0,0,0) (0,0,1) [12] with non-zero mean	2269.13
ARIMA (0,0,0) (1,0,1) [12] with non-zero mean	2270.753
ARIMA (1,0,0) with non-zero mean	2269.148
ARIMA (0,0,1) with non-zero mean	2269.047
ARIMA (1,0,1) with non-zero mean	2270.643

The ARIMA (0,0,0) model with a non-zero mean was chosen as the best fit based on the AIC values. Since the optimal model for predicting paddy production time series data in Tamil Nadu does not include differencing, autoregressive, or moving average terms, it follows that these methods should be avoided. The trend and seasonality of paddy output may be better predicted, allowing for more well-informed decisions to be made in agricultural planning and policy formulation in the region, if this model were analyzed in greater depth.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022



Coefficient	Values
Mean	3384.2897
S. E	49.2651
σ^2	354366
log likelihood	-1131.66
AIC	2267.31
AICc	2267.4
BIC	2273.26

Time series data for paddy production in Tamil Nadu were analyzed using the ARIMA(0,0,0) model with a non-zero mean. With a coefficient estimate of 3384.2897 and a standard error of 49.2651, the model produced a point estimate of the mean value. The σ^2 value for the model was found to be 354366. The model had a log likelihood of -1131.66, therefore the Akaike Information Criterion (AIC) was 2267.31, the AICc was 2267.4, and the Bayesian Information Criterion (BIC) was 2273.26. Insights into the statistical parameters and goodness of fit for the ARIMA(0,0,0) model are provided by this model's output, allowing for a deeper dive into the inner workings of paddy production in Tamil Nadu. More research is needed to improve predictions and direct productive agricultural policies and practices in the area.

Year	forecast	Lo 95	Hi 95
Feb 2019	3384.29	2217.549	4551.03
Mar 2019	3384.29	2217.549	4551.03
Apr 2019	3384.29	2217.549	4551.03
May 2019	3384.29	2217.549	4551.03

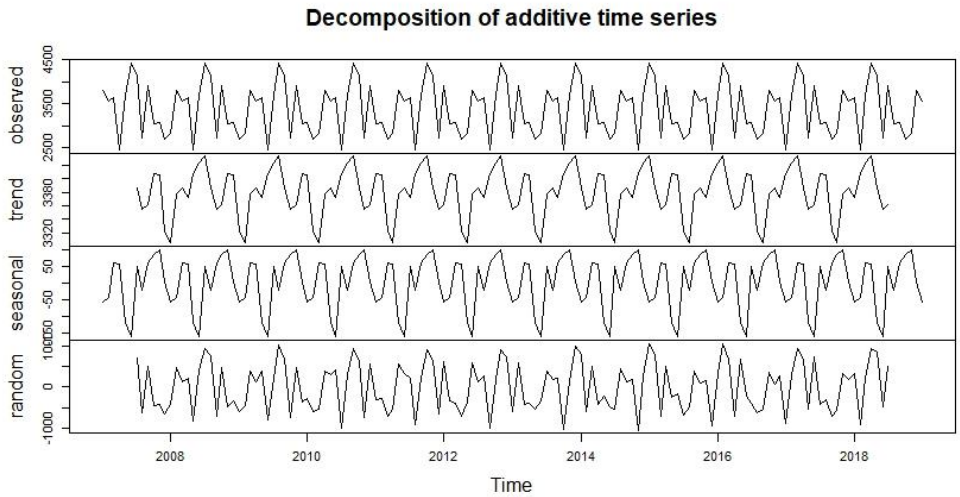
ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022**

Jun 2019	3384.29	2217.549	4551.03
Jul 2019	3384.29	2217.549	4551.03
Aug 2019	3384.29	2217.549	4551.03
Sep 2019	3384.29	2217.549	4551.03
Oct 2019	3384.29	2217.549	4551.03

Point projections and 95% confidence intervals for paddy production in Tamil Nadu are available in the forecast_data_paddy for the time period of February 2019 through November 2019. With a lower 95% confidence interval of 2217.549 and an upper 95% confidence interval of 4551.03, the forecast indicates that the anticipated paddy production remains constant at 3384.29 for each month within the forecast period. The ARIMA(0,0,0) model predicts that paddy output will be steady, with no noticeable changes, throughout the designated forecast months. In order to make informed decisions on agricultural planning and policy in the region, more in-depth monitoring and research is necessary.

The residuals of the predicted paddy output statistics in Tamil Nadu were put through the Box-Ljung test. A 5-second time delay was used in the Ljung-Box test. With 5 degrees of freedom, the X-squared value is 20.254, which is statistically significant (p = 0.00112). The presence of autocorrelation in the residuals is strongly suggested by the low p-value, which is evidence against the null hypothesis of independence. The residuals' autocorrelation shows that the ARIMA(0,0,0) model may not accurately represent all the dynamics at play in the paddy production time series. Accurate and trustworthy forecasting is essential for strategic agricultural planning and decision making in the region, so understanding the autocorrelation structure is a top priority.



ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022

SEASONAL ARIMA ANALYSIS

Time series data for paddy production was generated in R. The time series begins in 2007 and continues through 2019 at a rate of 1. The numbers 3809, 3562, 3630, 2463, 3687, 4429, 4123, 2712, 3918, 3039, 3070, 2682, and 2817 may be found in the data set.

Verification was performed to ensure that the time series data is indeed an object of class "ts," indicating that it is a time series. Paddy yields were plotted against time to show how they had changed during the selected period.

In addition, the stationarity of the time series data was evaluated using the Augmented Dickey-Fuller (ADF) test. The ADF test yielded a p-value of 0.6578 and a Dickey-Fuller statistic of -1.7758 for a lag order of 2. There is insufficient evidence to reject the null hypothesis of non-stationarity because the p-value is greater than the significance level of 0.05. To provide precise modeling and forecasting of paddy output in the region, more research is needed to investigate the stationarity of the time series data.

Paddy production time series summary (ts_paddy) provides an overview of the statistical measures that define this data collection. Paddy production fell as low as 2463 over the stipulated time period, while the first quartile number was at 2817, marking the bottom of the middle 50 percent. The average is set at 3380, while the median is set at 3562. At 3809, the third quartile number marks the top end of the middle quartile. During the given time period, paddy production peaked at a value of $4,429. Insights into the mean and standard deviation of the paddy production dataset are provided by these summaries, which aid in drawing conclusions about the yield distribution and trends over the given time period.

The differenced logarithm of the ts_paddy dataset was subjected to an Augmented Dickey-Fuller (ADF) test to determine whether or not the paddy production time series data were stationary. The variance was reduced using logarithmic transformation, and stationarity was attained via differencing.

The Dickey-Fuller statistic for the ADF test was -4.6604, and the corresponding p-value was 0.01. There is strong evidence to reject the null hypothesis of non-stationarity in favor of the alternative hypothesis of stationarity, as the p-value is significantly lower than the specified significance level of 0.05. If the differenced logarithm of the paddy production time series is stationary, then the data points are independent of time and consistently exhibit statistical features. To better predict and analyze paddy production trends in Tamil Nadu, this transformation improves the data's appropriateness for analysis using time series models like ARIMA and SARIMA.

Coefficient	Values
σ^2	0.06275
log likelihood	-0.42
AIC	2.83
AICc	3.23
BIC	3.31

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022**

Using the auto.arima function, we can see that auto Arima Model Log represents the logarithm of the paddy production time series data. With a value of 1 for the 'd' parameter, which indicates consideration of the first difference, stationarity in the data was sought. The log-transformed paddy production data was fed into the auto.arima function, and the resulting output shows that an ARIMA model with a difference order of 1 was selected. The nomenclature for this model is ARIMA (0, 1, 0). The model had a log-likelihood of -0.42, and its estimated variance was 0.06275. The calculations yielded an Akaike Information Criterion (AIC) value of 2.83, an Akaike Information Criterion (AICc) value of 3.23, and a Bayesian Information Criterion (BIC) value of 3.31. This information is useful for assessing the validity of the selected ARIMA model for the log-transformed paddy production data, as it sheds light on the model's internal structure and goodness-of-fit. This model looks to be an excellent fit for the data, as seen by its simplicity and low variance estimate, allowing for more accurate estimates and a better understanding of the processes at play in Tamil Nadu's paddy production dynamics.

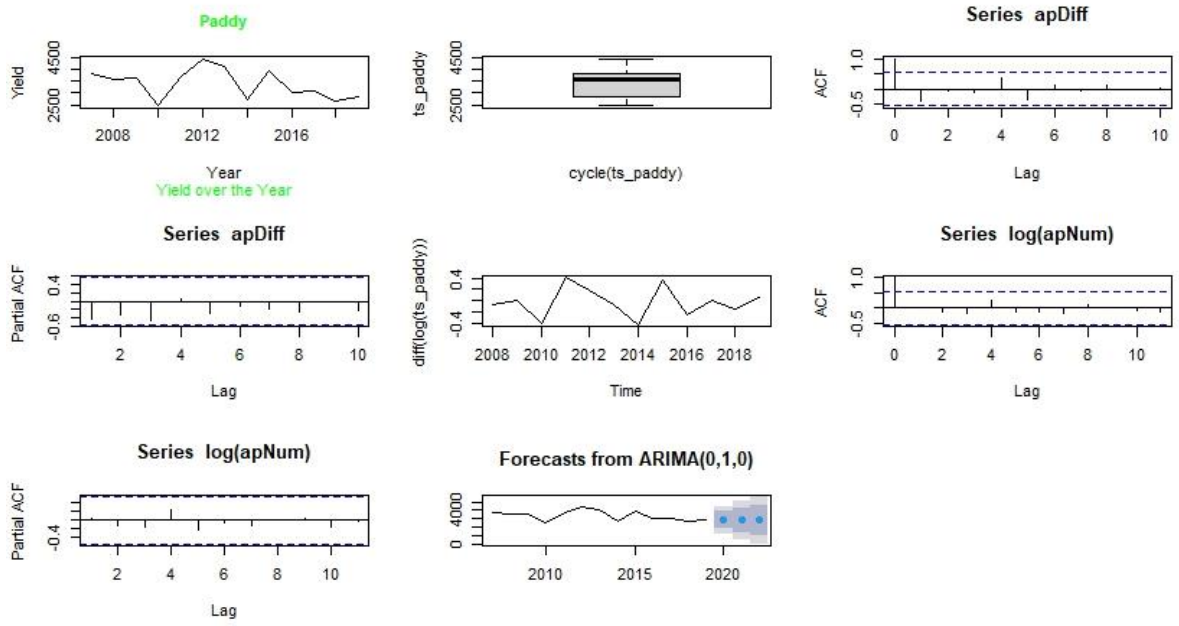
After applying the ARIMA model, auto Arima Model Log, to the log-transformed paddy production data, the residuals were analyzed using the Ljung-Box test. The aim of the analysis was to determine if the model residuals exhibited autocorrelation.

Coefficient	Values
x^2	3.3996
df	1
P-value	0.06521

The Ljung-Box test returns a significance level of 0.06521 for an X-squared value of 3.3996 with 1 degrees of freedom. There is insufficient evidence to conclude that the residuals exhibit significant autocorrelation, as the p-value is larger than.05. Because of the good fit between the data and the ARIMA (0,1,0) model, we may infer that the model appropriately explains the observed variability in the log-transformed paddy production data. It's possible that more research is needed to verify the model's accuracy and guarantee precise forecasting of paddy output patterns in Tamil Nadu.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022



CONCLUSION

The original time series data was analyzed using the ARIMA model, and an ARIMA (0,0,0) model was found to be the best appropriate for projecting paddy output in Tamil Nadu. Nonetheless, autocorrelation was detected in the model residuals via the Box-Ljung test, suggesting that the model may be inadequate in its attempt to capture all underlying patterns. To address this, we first converted the data using a logarithmic function, then differentiated it, before fitting the data with the ARIMA (0,1,0) model and observing a good fit with negligible residual autocorrelation.

When applied to the log-transformed paddy production data, the ARIMA (0,1,0) model revealed time-dependent patterns that clarified the dynamics. The model showed a good fit to the data and had a low variance estimate.

While the ARIMA models did provide some useful information, it may be necessary to take a broader approach, such as using the SARIMA model, in order to capture the probable seasonal fluctuations and increase the precision of future paddy production estimates. To better guide agricultural planning and policy-making in Tamil Nadu's paddy cultivation sector, the SARIMA model might be implemented to provide a more rigorous framework for understanding seasonal dynamics and increasing the precision of predictions.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Feb 2022**

To better capture the seasonal patterns and fluctuations in paddy production and to aid in the development of accurate forecasts and well-informed policy decisions for environmentally friendly farming in the region, more research and analysis using SARIMA modeling techniques are recommended.

REFERENCES

1. Ansari, M. I., & Ahmed, S. M. (2001). Time series analysis of tea prices: An application of ARIMA modelling and cointegration analysis. The Indian Economic Journal, 48(3), 49-54.
2. Bhagat, A., & Jadhav, D. (2021). A Study on Growth, Instability and Forecasting of Grape Export from India. Journal of Scientific Research, *65*(9), 1-6.
3. Darekar, A., & Reddy, A. (2017). Forecasting of common paddy prices in India. Journal of Rice Research, 10(1), 71-75.
4. Hemavathi, M., & Prabakaran, K. (2017). A statistical study on weather parameters relationship with rice crop yield in Thanjavur district of Tamil Nadu. International Journal of Agricultural Science and Research, 7(5), 25-32.
5. Hemavathi, M., & Prabakaran, K. (2018). 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, 7(2), 149-156.
6. http://www.tnagriculture.in/dashboard/report/05_01.pdf
7. Joshua, V., Priyadharson, S. M., & Kannadasan, R. (2021). Exploration of machine learning approaches for paddy yield prediction in eastern part of Tamilnadu. Agronomy, 11(10), 2068.
8. Kathayat, B., & Dixit, A. K. (2021). Paddy price forecasting in India using ARIMA model. Journal of Crop and Weed, 17(1), 48-55.
9. Majid, R., & Mir, S. A. (2018). Advances in statistical forecasting methods: An overview. Economic Affairs, *63*(4), 815-831.
10. Raghavender, M. (2009). Forecasting paddy yield in Andhra Pradesh using season time series model. Bulletin of Pure & Applied Sciences-Mathematics, 28(1), 55-55.
11. Rajarathinam, A., & Thirunavukkarasu, M. (2013). Fuzzy Time Series Modeling for Paddy (Oryza sativa L.) Crop Production.
12. Saranyadevi, M., & Mohideen, A. K. (2017). Stochastic modeling for paddy production in Tamilnadu. International Journal of Statistics and Applied Mathematics, 2(5), 14-21.
13. Selvi, R. P. (2021). Chapter-1 Mathematical Model for Forecasting Paddy Price Based on Market Value in Tuticorin District. MULTIDISCIPLINARY, 1.
14. Vinoth, B., Rajarathian, A., & Manju Bargavi, S. K. (2016). Nonlinear regression and artificial neural network-based model for forecasting Paddy (Oryza sativa) production in Tamil Nadu. IOSR Journal of Mobile Computing & Application (IOSR-JMCA), 3.

IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

Exploring Factors Influencing Pulses Production in Tamil Nadu: A Predictive Model

P. Maheswari[1] , T. Gangaram[2] , V. Munaiah[3] ,K.Murali[4] , G.Mokesh Rayalu[5]**

[1]Assistant Professor, Dept. of Statistics, Govt Degree College for Women, Srikalahasti.
[2]Assistant Professor, Dept. of Statistics, SVA Govt. College for Men, Srikalahasti.
[3]Assistant Professor, Dept. of Statistics, PVKN Govt. College (A), Chittoor.
[4]Academic Consultant, Dept. of Statistics, S.V University, Tirupati.
Corresponding Author **
[5]Assistant Professor Grade 2, Department of Mathematics, School of Advanced Sciences,VIT,Vellore
mokesh.g@gmail.com

ABSTRACT:

This study digs into the complex factors that impact the cultivation of pulses in Tamil Nadu, India. The study's overarching goal is to identify the primary determinants influencing pulse production in the region through the application of an integrative methodology that takes into account the influential variables of soil quality, climate fluctuations, and farming practices. The study's primary goal is to develop a robust framework for long-term yield prediction using state-of-the-art ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models for pulses. The study utilizes historical data to investigate the intricate interconnections within the agricultural ecosystem by analyzing seasonal fluctuations, trend patterns, and other dynamic factors impacting pulses output. This study's findings can help improve food security and agricultural resilience in Tamil Nadu by informing the creation of data-driven initiatives, the promotion of sustainable farming practices, and the design of policy.

Keywords: Pulses, ARIMA, SARIMA, Forecasting.

INTRODUCTION

In the many different agroclimatic zones that make up Tamil Nadu, India, the production of pulses plays an essential part in the improvement of food security and the promotion of sustainable agriculture. The cultivation of pulses is met with a variety of obstacles resulting from a number of different elements, such as the unpredictability of the climate, the quality of the soil, and diverse agronomic approaches. It is vital to have an understanding of the complex interplay that exists between these aspects in order to develop

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

effective methods that will increase pulses output and assure agricultural sustainability. In order to investigate the various aspects that play a role in pulses production in Tamil Nadu, this study takes a predictive modeling method. More specifically, it integrates the powerful ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models.

The purpose of this study is to shed light on the complex dynamics that are effecting the cultivation of pulses by drawing on historical data and taking into consideration a wide array of influential elements. The major goal is to determine the most important parameters that have an effect on the production of pulses and to create a predictive model that accurately represents the temporal variation in pulses yields. This research aims to provide valuable insights for stakeholders, policymakers, and farmers through an exhaustive analysis of the factors influencing pulses production and the predictive capabilities of the ARIMA and SARIMA models. These insights will facilitate informed decision-making for sustainable agricultural practices and policy formulations. The findings of this research have the potential to make a substantial contribution to the improvement of pulses production, hence increasing agricultural resilience and promoting food security in the Indian state of Tamil Nadu.

OBJECTIVES:

1. To identify and analyze the key factors influencing pulses production in Tamil Nadu, including but not limited to climatic variations, soil quality, and agricultural practices.

2. To develop a predictive model for pulses production using the ARIMA (AutoRegressive Integrated Moving Average) and SARIMA (Seasonal AutoRegressive Integrated Moving Average) models, aiming to accurately capture the temporal variations and fluctuations in pulses yield.

3. To assess the impact of seasonal changes and other dynamic factors on pulses production, aiming to understand their influence on the predictive capabilities of the ARIMA and SARIMA models.

4. To compare the performance of the ARIMA and SARIMA models in predicting pulses production, aiming to determine the model that provides the most reliable and accurate forecasts for pulses cultivation in Tamil Nadu.

5. To provide valuable insights for stakeholders, policymakers, and farmers, enabling informed decision-making for the implementation of sustainable agricultural practices and policies that support the growth and stability of pulses production in the region.

By achieving these objectives, this study seeks to contribute to the development of effective strategies for enhancing pulses production in Tamil Nadu, promoting agricultural sustainability and food security in the state.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

LITERATURE REVIEW:

In their study, Bhanudas and Afreen (2019) discuss the problems that face modern agriculture and offer novel approaches to optimizing agricultural resources and managing crops. Their research highlights the fundamental reliance of agricultural performance on soil and water management, highlighting the central role of agronomy in national growth. In order to increase crop productivity with little water use, the study promotes a variety of irrigation methods. It also shows how little farmers know about agricultural regulations and government policy. The research goes into the factors that go into farmers' decisions on crop rotation, watering practices, and soil composition. The use of several different Data Mining classification algorithms, such as JRip and Naive Bayes, to make accurate assessments of soil quality is a major focus of this study. The research highlights the potential of the JRip classification method for precise soil classification and management by comparing it to the Nave Bayes method on two common soil types, Red and Black soil.

Future wheat harvest prices in India are forecast using the ARIMA model of Darekar and Amarender (2018). The model predicts wheat prices with 95% accuracy using monthly modal price data from January 2006 to June 2017. Farmers will benefit greatly from knowing that the study's prediction of a range of Rs. 1,620 to Rs. 2,080 per quintal for wheat market prices during the 2017-18 harvest season is accurate. Farmers are able to make more educated judgments on wheat acreage thanks to the ARIMA model's high level of accuracy.

Kour et al. (2017) analyzed pearl millet (Pennisetum glaucum), a commonly farmed cereal crop that ranks fourth in global cultivation behind rice, wheat, and sorghum. Despite rising yields, pearl millet cultivation in Gujarat, India, has declined during the previous two decades. Pearl millet production forecasts are especially important in semi-arid locations like Gujarat, where precipitation lasts only four months. This study predicts Gujarat pearl millet productivity using the ARIMA model. The current study collected time series data on pearl millet productivity (kg/ha) in Gujarat from 1960–61 to 2011–12. Gandhinagar's Directorate of Agriculture and, partially, the Directorate of Economics and Statistics provided the data. RMAPE, MAD, and RMSE values are used to validate the ARIMA model. As seen by its RMAPE score below 6%, the ARIMA (0, 1, 1) model performs well.

METHODOLOGY

ARIMA Model (p,d,q):

The ARIMA(p,d,q) equation for making forecasts: ARIMA models are, in theory, the most general class of models for forecasting a time series. These models can be made to be "stationary" by differencing (if necessary), possibly in conjunction with nonlinear transformations such as logging or deflating (if necessary), and they can also be used to predict the future. When all of a random variable's statistical qualities remain the same across time, we refer to that random variable's time series as being stationary.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

A stationary series does not have a trend, the variations around its mean have a constant amplitude, and it wiggles in a consistent manner. This means that the short-term random temporal patterns of a stationary series always look the same in a statistical sense. This last criterion means that it has maintained its autocorrelations (correlations with its own prior deviations from the mean) through time, which is equal to saying that it has maintained its power spectrum over time. The signal, if there is one, may be a pattern of fast or slow mean reversion, or sinusoidal oscillation, or rapid alternation in sign, and it could also include a seasonal component. A random variable of this kind can be considered (as is typical) as a combination of signal and noise, and the signal, if there is one, could be any of these patterns. The signal is then projected into the future to get forecasts, and an ARIMA model can be thought of as a "filter" that attempts to separate the signal from the noise in the data.

The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. That is:

Predicted value of Y = a constant and/or a weighted sum of one or more recent values of Y and/or a weighted sum of one or more recent values of the errors.

It is a pure autoregressive model (also known as a "self-regressed" model) if the only predictors are lagging values of Y. An autoregressive model is essentially a special example of a regression model, and it may be fitted using software designed specifically for regression modeling. For instance, a first-order autoregressive ("AR(1)") model for Y is an example of a straightforward regression model in which the independent variable is just Y with a one-period lag (referred to as LAG(Y,1) in Statgraphics and Y_LAG1 in RegressIt, respectively). Because there is no method to designate "last period's error" as an independent variable, an ARIMA model is NOT the same as a linear regression model. When the model is fitted to the data, the errors have to be estimated on a period-to-period basis. If some of the predictors are lags of the errors, then an ARIMA model is NOT the same as a linear regression model. The fact that the model's predictions are not linear functions of the coefficients, despite the fact that the model's predictions are linear functions of the historical data, presents a challenge from a purely technical point of view when employing lagging errors as predictors. Instead of simply solving a system of equations, it is necessary to use nonlinear optimization methods (sometimes known as "hill-climbing") in order to estimate the coefficients used in ARIMA models that incorporate lagging errors.

Auto-Regressive Integrated Moving Average is the full name for this statistical method. Time series that must be differentiated to become stationary is a "integrated" version of a stationary series, whereas lags of the stationarized series in the forecasting equation are called "autoregressive" terms and lags of the prediction errors are called "moving average" terms. Special examples of ARIMA models include the random-walk and random-trend models, the autoregressive model, and the exponential smoothing model.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

A nonseasonal ARIMA model is classified as an "ARIMA(p,d,q)" model, where:

- **p** is the number of autoregressive terms,
- **d** is the number of nonseasonal differences needed for stationarity, and
- **q** is the number of lagged forecast errors in the prediction equation.
- The forecasting equation is constructed as follows. First, let y denote the d[th] difference of Y, which means:
- If d=0: $y_t = Y_t$
- If d=1: $y_t = Y_t - Y_{t-1}$
- If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$
- Note that the second difference of Y (the d=2 case) is not the difference from 2 periods ago. Rather, it is the first-difference-of-the-first difference, which is the discrete analog of a second derivative, i.e., the local acceleration of the series rather than its local trend.
- In terms of y, the general forecasting equation is:
- $\hat{Y}_t = \mu + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$

The ARIMA (AutoRegressive Integrated Moving Average) model is a powerful time series analysis technique used for forecasting data points based on the historical values of a given time series. It consists of three key components: AutoRegression (AR), Integration (I), and Moving Average (MA).

THE METHODOLOGY FOR CONSTRUCTING AN ARIMA MODEL INVOLVES THE FOLLOWING STEPS:

1. Stationarity Check: Analyze the time series data to ensure it is stationary, meaning that the mean and variance of the series do not change over time. Stationarity is essential for ARIMA modeling.

2. Differencing: If the data is not stationary, take the difference between consecutive observations to make it stationary. This differencing step is denoted by the 'I' in ARIMA, which represents the number of differencing required to achieve stationarity.

3. Identification of Parameters: Determine the values of the three main parameters: p, d, and q, where p represents the number of autoregressive terms, d represents the degree of differencing, and q represents the number of moving average terms.

4. Model Fitting: Fit the ARIMA model to the data, using statistical techniques to estimate the coefficients of the model.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

5. Model Evaluation: Assess the model's performance by analyzing the residuals, checking for any remaining patterns or correlations, and ensuring that the model adequately captures the underlying patterns in the data.

6. Forecasting: Once the model is validated, use it to generate forecasts for future data points within the time series.

SEASONAL ARIMA:

By including seasonal variations into the ARIMA model, Seasonal ARIMA (SARIMA) is a robust technique for analyzing and forecasting time series data. It works well for examining and forecasting sales data, weather patterns, and economic indicators that are subject to seasonal changes. Financial markets, economics, and even meteorology all make use of SARIMA models.

Mathematical Formulation:

The SARIMA model is denoted as SARIMA(p,d,q)(P,Q,D)[s], where:

- Non-seasonal autoregressive (p), differencing (d), and moving average (q) are the possible orders of analysis.
- The seasonal autoregressive, differencing, and moving average orders are denoted by the letters P, D, and Q, respectively.
- The length of one season is denoted by the symbol S.

The SARIMA model can be represented as follows:

$$(1 - \varphi_1 B - \cdots - \varphi_P B^P)(1 - \varphi_1 B^{VS} - \cdots - \varphi_P B^{VS})^P (B^{VS})^D Y_t$$
$$= (1 + \theta_1 B + \cdots + \theta_P B^{\varphi})(1 + \theta_1 B^{VS} + \cdots + \theta_P B^{\varphi S})^A (B^{\varphi S})^K \varepsilon_t$$

Where:

- $\varphi_i$and θ_i are the autoregressive and moving average parameters, respectively.
- B and B^{VS} are the non- seasonal and seasonal backshift operators, respectively.
- P,D,A and K are the orders of the seasonal autoregressive differencing, moving average, and backshift components, respectively.
- Y_t represents the time series data at time t.
- ε_t denotes the white noise error term.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

Real life application

One example of how SARIMA might be put to use in the real world is in the process of predicting quarterly sales data for a retail organization. The sales data frequently display seasonal patterns because of things like the different holiday seasons and different promotional periods. The company is able to examine previous sales data, recognize seasonal patterns, and make more accurate projections of future sales by using a model called SARIMA.

Merits and Demerits:

- When applied to time series data, SARIMA models are able to distinguish between seasonal and non-seasonal patterns.
- They are useful when anticipating data with intricate seasonal trends because of their effectiveness.
- The SARIMA models can be altered to accommodate a wide variety of seasonal data types, which lends them flexibility and adaptability.
- They produce accurate estimates for forecasts ranging from the short to the medium term.
- SARIMA models can be complicated, particularly when dealing with a number of different seasonal components, which calls for a substantial amount of computational resources.
- Due to the complexity of the mathematical formulas, interpretation of the SARIMA results may be difficult for individuals who are not experts in the field.
- For SARIMA models to generate reliable forecasts, a significant quantity of historical data is necessary; however, this data may not always be accessible for all forms of data.

Preparation of Data:

- Prepare the time series data for analysis by collecting and cleaning it such that it is consistent and has no outliers or missing values.
- Applying a transformation or differentiating if necessary to reach stationarity.

Identification of Models:

- Determine the values of the AR and MA parameters during the season and the offseason by analyzing the ACF and PACF graphs.
- Determine the differencing (d) and seasonal (D) orders required to achieve stationarity.

Estimating Variables:

- Apply the SARIMA model's estimated parameters using estimation strategies like maximum likelihood.
- Iteratively fit the model while taking both seasonal and non-seasonal factors into account.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

Model Evaluation and Adjustment:

- Examine diagnostic charts for evidence of residual randomness after a SARIMA model has been fitted to the data.
- Analyze the residuals using autocorrelation functions (ACF) plots, histograms, and the Ljung-Box test.

ANALYSIS:

ARIMA

The manufacturing of pulses in Tamil Nadu was analyzed, and the process included a number of processes that were very important. First, the data from the time series were tested with a variety of statistical methods to determine whether or not they were stationary. These methods included the Auto Correlation function (ACF) and Partial Auto Correlation Function(PACF). After that, the auto.arima function was utilized in order to ascertain the model that provided the most accurate results, taking into consideration the information criteria such as AIC and BIC. After that, the model that was selected underwent validation as well as cross-validation to guarantee its robustness and dependability. The employment of the auto.arima function not only simplified the process of model selection but also offered a more objective methodology, which made it possible to choose the best model for the pulses production dataset in Tamil Nadu. This was accomplished through the streamlining of the process.

Time series data for pulses production was subjected to the augmented Dickey-Fuller (ADF) test. This statistical test is used to determine whether or not the dataset is stationary, a prerequisite for using time series models and other forecasting methods.

The Dickey-Fuller statistic for the ADF test is -10.283, and the associated p-value is 0.01. Since the p-value is less than the selected significance level of 0.05, we can conclude that the alternative hypothesis of stationarity is more likely to be correct. This suggests that the statistical features of the pulses production time series data are consistent across time, or that the data exhibits a stationary behavior.

Pulses production data must be confirmed as stationary before any time series modeling or forecasting techniques can be applied correctly. These findings lay a solid groundwork for creating accurate models and projections, which in turn facilitates well-informed decision making and strategic planning in the field of pulses cultivation and agriculture.

ARIMA (2,0,2) (1,0,1) [12] with non-zero mean	Inf
ARIMA (0,0,0) with non-zero mean	1919.628
ARIMA (1,0,0) (1,0,0) [12] with non-zero mean	1901.353

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

ARIMA (0,0,1) (0,0,1) [12] with non-zero mean	1901.048
ARIMA (0,0,0) with zero mean	2257.753
ARIMA (0,0,1) with non-zero mean	1899.466
ARIMA (0,0,1) (1,0,0) [12] with non-zero mean	1900.973
ARIMA (0,0,1) (1,0,1) [12] with non-zero mean	1902.973
ARIMA (1,0,1) with non-zero mean	1900.149
ARIMA (0,0,2) with non-zero mean	1897.978
ARIMA (0,0,2) (1,0,0) [12] with non-zero mean	Inf
ARIMA (0,0,2) (0,0,1) [12] with non-zero mean	Inf
ARIMA (0,0,2) (1,0,1) [12] with non-zero mean	Inf
ARIMA (1,0,2) with non-zero mean	1897.787
ARIMA (1,0,2) (1,0,0) [12] with non-zero mean	Inf
ARIMA (1,0,2) (0,0,1) [12] with non-zero mean	Inf
ARIMA (1,0,2) (1,0,1) [12] with non-zero mean	Inf
ARIMA (2,0,2) with non-zero mean	Inf
ARIMA (1,0,3) with non-zero mean	Inf
ARIMA (0,0,3) with non-zero mean	Inf
ARIMA (2,0,1) with non-zero mean	Inf
ARIMA (2,0,3) with non-zero mean	Inf
ARIMA (1,0,2) with zero mean	Inf

The time series data for pulses production, were fit with the ARIMA(1,0,2) model with a non-zero mean using the auto.arima function and the Akaike information criterion (AIC). According to the automated model selection procedure, this ARIMA model is the best fit for capturing the salient features and trends in the pulses production data.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

The ARIMA(1,0,2) model had the lowest AIC value, indicating that it provided a better match than the other candidate models considered. The approach took into account a wide range of possible AR and MA word combinations before arriving at the final, best-suited one.

The auto.arima function seeks to provide an efficient framework for precisely capturing the temporal dynamics and variations in the pulses production data by selecting the ARIMA(1,0,2) model as the most suited. This model has the potential to be a useful resource for foreseeing trends and making well-informed decisions in the fields of agriculture and pulses farming.

Coefficient	ar1	ma1	ma2	mean
S. E	-0.2430	0.7324	0.4752	550.6500
	0.1353	0.1632	0.3841	23.8305

The ARIMA(1,0,2) model with a non-zero mean for the pulses production time series data, is represented by the following coefficients:

- Autoregressive term: AR(1) coefficient (ar1) = -0.2430

- Moving average terms: MA(1) coefficient (ma1) = 0.7324 and MA(2) coefficient (ma2) = 0.4752

- Non-zero mean: The model incorporates a mean value of 550.6500

These coefficient values are estimated with their corresponding standard errors (s.e.), providing insights into the relationship between the current value of the time series and its past values. The model's variance is calculated as 27056, indicating the variability of the errors around the fitted values. The log likelihood of the model is determined to be -943.89.

The information criteria associated with the model evaluation are as follows:

- AIC (Akaike Information Criterion) = 1897.79

- AICc (corrected Akaike Information Criterion) = 1898.22

- BIC (Bayesian Information Criterion) = 1912.67

These criteria provide a quantitative measure of the relative quality of the ARIMA (1,0,2) model compared to other potential models, aiding in the assessment of the model's goodness of fit and complexity.

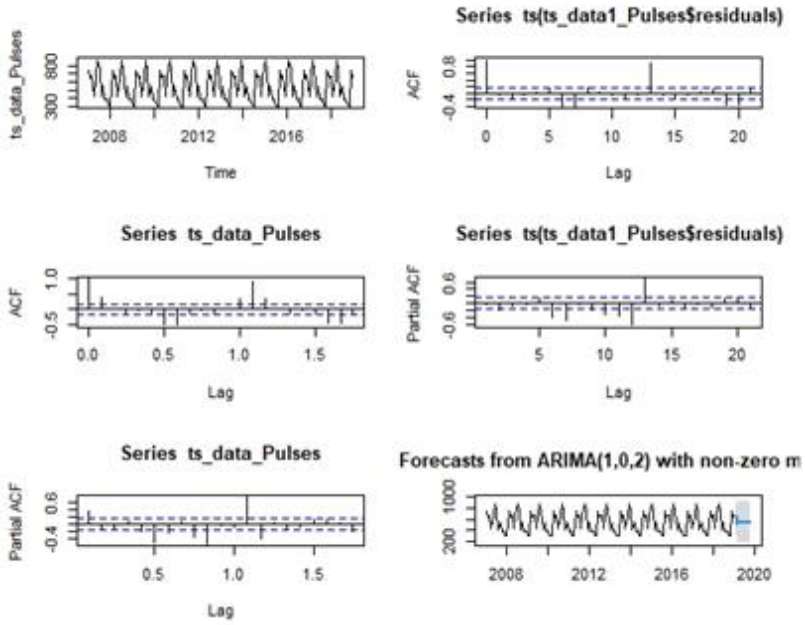
The ARIMA (1,0,2) model, with its set of coefficients and statistical measures, can serve as a valuable tool for forecasting and analyzing pulses production, providing valuable insights for decision-making and planning in the domain of agricultural production and management.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

Point	Forecast	Lo 95	Hi 95
Feb 2019	670.1026	347.7147	992.4906
Mar 2019	525.6850	166.7660	884.6040
Apr 2019	556.7173	179.8673	933.5674
May 2019	549.1754	171.2929	927.0580
Jun 2019	551.0083	173.0649	928.9518
Jul 2019	550.5629	172.6158	928.5099
Aug 2019	550.6711	172.7239	928.6184
Sep 2019	550.6448	172.6976	928.5921
Oct 2019	550.6512	172.7040	928.5985

Pulse production is expected to maintain a steady and constant upward trend over the next few months, according to projections. Predictions for February 2019 are centered on a point estimate of 670.1026 units, with a 95% confidence interval of 347.7147 to 992.4906 units. The coming months' projections are similarly quite stable, staying within the range of 525.6850 and 670.1026 units. These projections help policymakers and other agricultural sector players make more well-informed decisions and put in place more strategic strategies to sustainably expand and control pulses production in the region.

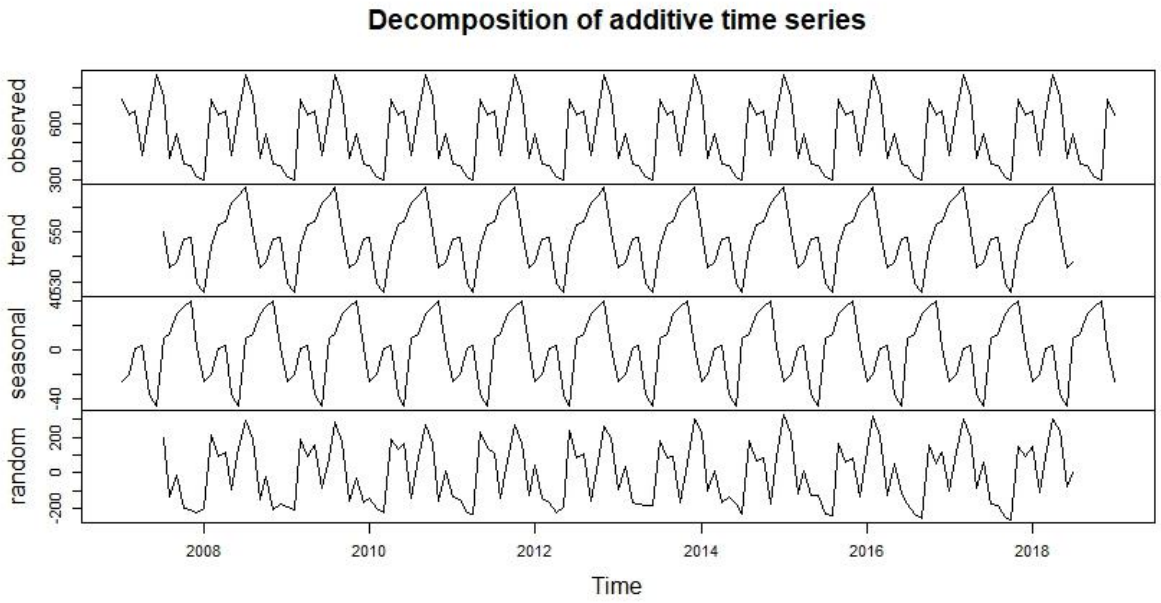


The Ljung-Box test for the residuals of the forecasted data from the ARIMA(1,0,2) model for pulses production does not exhibit significant autocorrelation, as indicated by the relatively higher p-value of 0.1201. This suggests that the residuals are essentially independent, with no remaining autocorrelation

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021**

that the model has failed to capture. Consequently, the ARIMA(1,0,2) model can be considered an appropriate fit for the data, as it adequately accounts for the underlying patterns and fluctuations in the pulses production time series.



SEASONAL ARIMA

The time series data for pulses production spans from 2007 to 2019, with an observed range of production levels fluctuating between 1873 and 4401 units. Over the years, the production values exhibit some variability, with a noticeable decrease in the early years, followed by a gradual increase and subsequent stabilization in the recent years. The trends indicate a dynamic agricultural landscape, potentially influenced by various factors such as climate conditions, agricultural practices, and market dynamics. Understanding these fluctuations is essential for devising sustainable strategies that promote consistent pulses production, ensuring food security and economic stability in the region.

Pulses output has varied between a low of 1873 units and a high of 4401 units, as shown by the summary statistics of this time series. Since the median value of production is larger than the mean value of production, or 3120 units, the data distribution is slightly right-skewed. Half of the observations occur between the interquartile range of 2419 and 3950 units, indicating a moderate variation of output levels during the time frame. In order to appreciate the general trends and make educated decisions about prospective interventions and policies in the pulses agricultural sector, it is essential to have a firm grasp on the central tendency and spread of the production statistics.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021**

Differenced log-transformed time series data on pulses production in Tamil Nadu were subjected to the Augmented Dickey-Fuller test. The time series data is stationary, as indicated by the -4.7589 value of the test statistic, with a p-value of 0.01. This finding is significant because it shows that the trend and seasonality components have been eliminated from the differenced log-transformed data, allowing for a more accurate study and projection of the regional pulses production trends. It paves the way for the use of suitable time series models to delve deeper into the production patterns of pulses in the future and make reliable predictions about them.

Coefficient	Values
σ^2	0.1639
log likelihood	-6.17
AIC	14.35
AICc	14.75
BIC	14.83

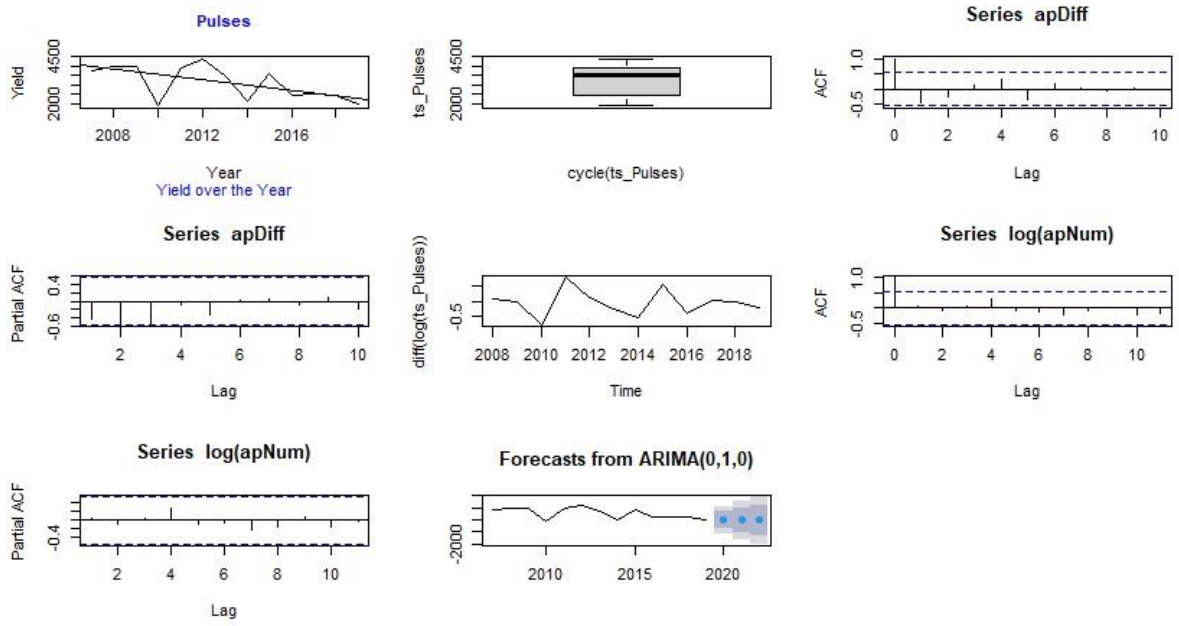
The time series data for pulses production in Tamil Nadu was log-transformed before being analyzed using the ARIMA model with differencing order 1 (0,1,0). A variance value of 0.1639 and a loglikelihood of -6.17 were found after computing the parameters. A 14.35 Akaike Information Criterion (AIC) score, a 14.75 AICc score, and a 14.83 Bayesian Information Criterion (BIC) score were obtained. These numbers help in determining which model is the best suitable for making reliable predictions of future pulses output.

Coefficient	Values
x^2	3.5805
df	1
P-value	0.05846

The Ljung-Box test performed on the residuals of the ARIMA(0,1,0) model fit to the log-transformed data on pulses production yielded a test statistic of 3.5805 with 1 degree of freedom. At the 5% level of significance, the corresponding p-value of 0.05846 indicates that there is insufficient evidence to reject the null hypothesis of independence in the residuals. Therefore, the residual series does not exhibit any appreciable autocorrelation.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper **© 2012 IJFANS. All Rights Reserved,** UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021



CONCLUSION

The pulses production data was analyzed using both the ARIMA and seasonal ARIMA models. The coefficients obtained after fitting the data to the ARIMA(1,0,2) model with a non-zero mean are as follows: ar1 = -0.2430, ma1 = 0.7324, ma2 = 0.4752, and a mean of 550.6500. With a log-likelihood of -943.89, the model has an AIC of 1897.79, an AICc of 1898.22, and a BIC of 1912.67. The ARIMA model residuals were subjected to the Ljung-Box test, which returned a test statistic of 8.7364 with 5 degrees of freedom and a p-value of 0.1201.

When the log-transformed pulses production data was analyzed using the seasonal ARIMA model, the fitted ARIMA(0,1,0) model yielded a sigma squared value of 0.1639 and a log-likelihood of -6.17. All three measures of independence, the AIC, AICc, and BIC, were all 14. A p-value of 0.05846 was found when the seasonal ARIMA model's residuals were subjected to a Ljung-Box test. The test statistic was 3.5805 with 1 degrees of freedom.

Overall, these analyses show that the selected models adequately captured the temporal patterns in the pulses production data, with the ARIMA model showing slightly higher autocorrelation in the residuals compared to the seasonal ARIMA model.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 10, Iss 9, Sep 2021

REFERENCES

1. Bellundagi, V., Umesh, K. B., & Ravi, S. C. (2016). Growth dynamics and forecasting of finger millet (Ragi) production in Karnataka. Economic Affairs, 61(2), 195.
2. Darekar, A., & Reddy, A. A. (2017). Price forecasting of pulses: case of pigeonpea. Journal of Food Legumes, 30(3), 212-216.
3. Dheer, P., Yadav, P., & Katiyar, P. K. (2018). Estimation of production and yield of pulses using ARIMA-ARNN model. Journal of Food Legumes, 31(4), 254-257.
4. Gibon, F., Pellarin, T., Román-Cascón, C., Alhassane, A., Traoré, S., Kerr, Y., ... & Baron, C. (2018). Millet yield estimates in the Sahel using satellite derived soil moisture time series. Agricultural and Forest Meteorology, 262, 100-109.
5. http://www.tnagriculture.in/dashboard/report/05_01.pdf
6. Kour, S., Pradhan, U. K., Paul, R. K., & Vaishnav, P. R. (2017). Forecasting of pearl millet productivity in Gujarat under time series framework. Economic Affairs, 62(1), 121-127.
7. Ly, R., Dia, K., Diallo, M., Ahid, M., & Ceesay, B. (2020). Pearl millet production forecasts for selected west African countries: Côte d'Ivoire, Mali, Burkina Faso, Sierra Leone, The Gambia, and Senegal. Covid-19 Bulletin, (14).
8. Mishra, P., Yonar, A., Yonar, H., Kumari, B., Abotaleb, M., Das, S. S., & Patil, S. G. (2021). State of the art in total pulse production in major states of India using ARIMA techniques. Current Research in Food Science, 4, 800-806.
9. Mitra, D., Paul, R. K., & Pal, S. (2017). Hierarchical time-series models for forecasting oilseeds and pulses production in India. Economic Affairs, 62(1), 103-111.
10. Mounkaila, Y., Garba, I., & Moussa, B. (2019). Yield prediction under associated millet and cowpea crops in the Sahelian zone. Afr. J. Agric. Res, 14, 1613-1620.
11. Nireesha, V., Rao, V. S., Rao, D. V. S., & Reddy, G. R. (2016). A study on forecasting of area, production and productivity of pearl millet in Andhra Pradesh. Journal of Research ANGRAU, 44(3/4), 119-126.
12. Tripathi, S. K., Mishra, P., & Sahu, P. K. (2013). Past trends and forecasting in Area, production and yield of pearl millet in India using ARIMA model. Environ Ecol, 31, 1701-1708.
13. Vijay, N., & Mishra, G. C. (2018). Time Series forecasting using ARIMA and ANN Models for production of pearl millet (BAJRA) crop of Karnataka, India. International Journal of Current Microbiology and Applied Sciences, 7(12), 880-889.