Modelling of Jassids (*Amrasca biguttula*) in Cotton- A Count Time Series Approach

ABSTRACT

This study was aimed to model Jassids population in cotton at Regional Agricultural Research Station(RARS), Nandyal. The secondary standard meteorological weekwise(SMW) data between 2008-2021 was considered based on data availability in the research station. Count time series models and machine learning models are used for modelling the Jassids population dataset Among the models evaluated in the study, the INGARCH-ANN model performed better than the INGARCH, ZIPAR, ZINBAR, and ANN models, according to error comparison metrics (MSE and RMSE). The statistical significance between the models was assessed using the Diebold-Mariano (DM) test. The order of prediction accuracy of the models under consideration is INGARCH-ANN>ANN> ZIPAR >ZINBAR>INGARCH. Overall, the study suggests that employing the Hybrid model could effectively model the jassids population in cotton at RARS, Nandyal.

Keywords: Modelling, ANN, ZIPAR, ZINBAR, INGARCH, MSE, RMSE.

1) Introduction

Agriculture plays a significant role in India's economy, contributing approximately 28% to the GDP. Ensuring self-sufficiency in food grain production has been a top priority in development strategies for the sector. Nevertheless, pest and disease outbreaks can severely impact agricultural output, leading to annual losses of up to Rs 50,000 crore in India.

Cotton holds immense significance in India's agricultural landscape and economy. It is one of the country's principal cash crops, contributing significantly to the livelihoods of millions of farmers and workers. India is among the world's largest cotton producers and exporters, with cotton cultivation spread across various states. The textile industry, which heavily relies on cotton, is a crucial driver of economic growth and employment in the country. Cotton's versatility makes it a key resource not only for clothing but also for various industrial products like oil, animal feed, and cosmetics. The success of the cotton crop has a cascading effect on multiple sectors, making it a vital component of India's agricultural and economic stability. Around 30% of yield loss is occurring in cotton due to pest attacks. One of the major pests in cotton crop is Jassids.

Jassids, small sap-sucking insects, pose a significant threat to cotton crops in India. These pests belong to the family of leafhoppers and have the capacity to cause considerable damage to cotton plants. Jassids feed on the sap of the cotton leaves, leading to the weakening of plants, reduced photosynthesis, and boll shedding. The damage caused by jassids can result in stunted growth, decreased cotton yield, and compromised fiber quality. In certain cases, severe infestations can lead to a phenomenon known as "hopper burn," where leaves turn yellow and brown due to excessive sap removal. The damage percentage inflicted by jassids on cotton crops can vary greatly depending on factors like climate conditions, pest management practices, and the availability of natural predators. So, there is a need to develop forewarning models to mitigate the losses.

So, this study focuses on advancing pest population forecasting by examining the effectiveness of modern statistical and machine learning models in addressing the limitations of traditional methods. Traditional approaches like Multiple Linear Regression (MLR) and Auto-Regressive Integrated Moving Average (ARIMA) often struggle with nonlinear data and excess zeroes in pest counts, resulting in less accurate forecasts.

Hence, this study aims to develop models for forecasting jassids populations in cotton by incorporating weather parameters as exogenous variables. Recent advances in modeling have investigated machine learning approaches for predicting agricultural fields, such as oil seed production, rice yield and pests, banana yield, tomato crop blight severity, and Paddy borer disease. The study focuses on developing generalized linear model (INGARCH Model), zero-inflated models, and machine learning models and Hybrid models to predict pest populations by utilizing count data driven approaches.

2) Methodology

The secondary data of Jassids in cotton was collected form the Regional Agricultural Research station (RARS), Nandyal under ANGRAU in Andhra Pradesh. Research station is in 15.4777° N, 78.4873° E co-ordinates. It is situated at an elevation ranging from 203 m above MSL. The secondary data of Jassids on cotton available from 2008-2021. The data available in standard meteorological weeks (SMW). The pest data is counts of Jassids collected in light trap arranged in the field. The weather parameters Maximum Temperature, Minimum

Temperature, Rainfall, Relative Humidity Morning, Relative Humidity Evening was also collected from Nandyal meteorological station. The total data is divided into training data and testing data (last 10 observations).

2.1 Statistical models

2.1.1 Stepwise Regression

The stepwise regression procedure is a statistical technique designed to identify the most important variables that influence the variation in a dependent variable. This method involves a series of iterative steps, each aimed at selecting the most significant variables. The procedure includes the following steps:

1. Variable Selection
2. Forward Selection
3. Backward Elimination
4. Stepwise Selection
5. Significance Testing

2.1.2 INGARCH (Integer Valued Generalized Autoregressive Conditional Heteroscedastic) model

The integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) model is a specific type of generalized linear model that adheres to Poisson and negative binomial distributions. INGARCH models belong to the GLM class where the conditional distribution of the dependent variable or observed count is typically assumed to follow discrete distributions such as Poisson, negative binomial, generalized Poisson, and double Poisson, as outlined by Rathod et al. (2021). Conditional likelihood estimation was employed for the estimation of the INGARCH model.

Let us denote the count time series by $\{Y_t: t \in N\}$ and time varying r-dimensional covariate vector say $\{X_t: t \in N\}\, i.e. X_t = (X_{t,1} \ldots, X_{t,r})^T$. The conditional mean becomes $E(\frac{Y_t}{F_{t-1}}) = \lambda_t$ and F_t is historical data. The generalized model form is expressed as follows;

$$g(\lambda_t) = \beta_o + \sum_{k=1}^{p} \alpha_k \tilde{g}\ (Y_{t-i_k}) + \sum_{l=1}^{q} \beta_1 g\ (\lambda_{t-jl}) + \eta^T$$

2.1.3 Zero Inflated Poisson Autoregressive (ZIPAR) Model

Poisson regression is used to predict a dependent variable that consists of count data given one or more independent variables. The zero inflated poison autoregressive (ZIPAR) model is expressed as follows

$$pr(Y_i = j) = \pi + (1 - \pi)exp(-\mu), if\ j = 0$$

The poison distribution is described as follows

$$(1 - \pi)\frac{\mu^y exp(-\mu)}{y_i}, if\ j>0$$

Where y_i is the logistic link function defined below.

The Poisson component can include an exposure time *t* and a set of *k* regressor variable. the expression relating these quantities is

$$\mu_i = exp(ln(t_i) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Often, $x_1 = 1$, in which case β_1 is called the intercept, the regression coefficients β_1, β_2 …. β_k are unknown parameters that are estimated from a set of data and their estimates are symbolized as b_1, b_2 …. b_k this logistic link function π_i is given by

$$\pi_i = \frac{\lambda_i}{1+\lambda_i}$$

were $\lambda_i = exp(in\ t_i) + y_1 z_{1i} + y_2 z_{2i} + \ldots + y_m z_{mi})$

The logistic component includes time *t* and a set of *m* regressor variables.

2.1.4 Zero Inflated Negative Binomial Autoregressive (ZINBAR) Model

The zero-inflated negative binomial regression model is employed for count data that display both overdispersion and an abundance of zeros. This model integrates the negative binomial distribution with the logit distribution, as described by Kim et al. (2021). The possible values for the variable y are non-negative integers: 0, 1, 2, and so on.

$$(y_i = j) = \begin{cases} \Pi_i + (1 - \Pi_i) g(y_i = 0) if j = 0 \\ (1 - \Pi_i) g(y_i) if j > 0 \end{cases}$$

Where π_i is the logistic link function defined below and $g(y_i)$ is the negative binomial distribution given by

$$g(y_i) = pr(Y = \frac{y_i}{\mu_i, \alpha}) = \frac{\tau(y_i + \alpha^{-1})}{\tau(\alpha^{-1})\,(y_i + 1)} (\frac{1}{1 + \alpha\mu_i}) \alpha^{-1} (\frac{\alpha\mu_i}{1 + \alpha\mu_i}) y_i$$

The negative binomial model can incorporate an exposure time ttt along with a set of kkk regressor variables. The corresponding expression is

$$\mu_i = \exp in\,(t_i) + \beta_1\ {}_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}\ )$$

Often, $x_1 = 1$, in which case β_1 is called the intercept. The regression coefficients β_1, β_2 …. β_k are known parameters that are estimated from a set of data. Their estimates are symbolized as b_1, b_2, …. b_k. Y_t

2.1.5 Artificial neural network model (ANN)

Artificial Neural Network (ANN) is the most widely used machine learning technique in recent years. In time series modeling, the Artificial Neural Network (ANN) is often termed the autoregressive neural network because it utilizes time lags as inputs. This approach can be mathematically represented within a time series framework using a neural network that implicitly captures the function of time. The general formula for the final output Yt of a multi-layer feedforward autoregressive neural network is given as:

$$Y_t = \alpha_0 + \sum_{j=1}^{q} \alpha_j\, g\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij}\, Y_{t-p}\right) + \epsilon_t$$

ANNX refers to an Artificial Neural Network model that includes exogenous variables, where "X" represents these external or independent variables.

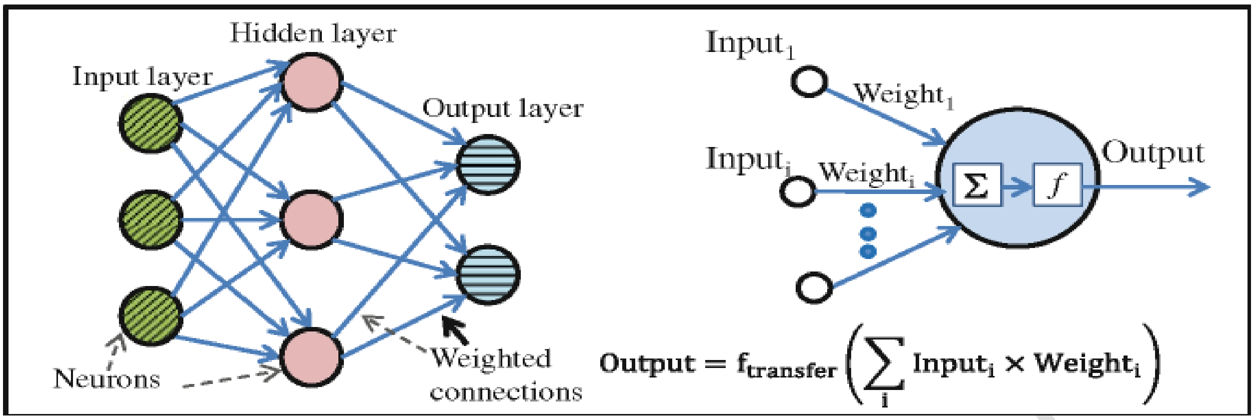

Fig 1. General form of Artificial Neural Network model

2.1.6 Two stage modelling

The proposed two stage modelling in this work considers the time series Y_t as a combination of both autocorrelated original time series and significant residuals of the model. This approach follows the Zhang's (Zhang 2003) hybrid approach, accordingly the relationship between autocorrelated count time series and significant residuals were considered.

In this work, the autocorrelated count time series were modelled using INGARCH, ZIPAR and ZINBAR models (Stage-I) and significant residuals were modelled using ANN model (Stage II).

The proposed methodology consists of two steps; Firstly, an INGARCH, ZIPAR and ZINBAR models are employed to model the count time series data. In the second step, if the residuals obtained from INGARCH, ZIPAR and ZINBAR models were found (Stage II) to be significant by Box pierce test and confirm the nonlinearity by BDS test, then they were modelled and predicted using ANN model. Finally, the forecasted values from stage 1 and stage 2 components were combined to generate aggregate the forecasted values.

$$\hat{Y}_t = \widehat{S_1} + \widehat{S_2}$$

Where, $\widehat{S_1}$and $\widehat{S_2}$ represents the predicted count time series and predicted significant residual components respectively. The graphical representation of two stage methodology is expressed in following Figure 2.

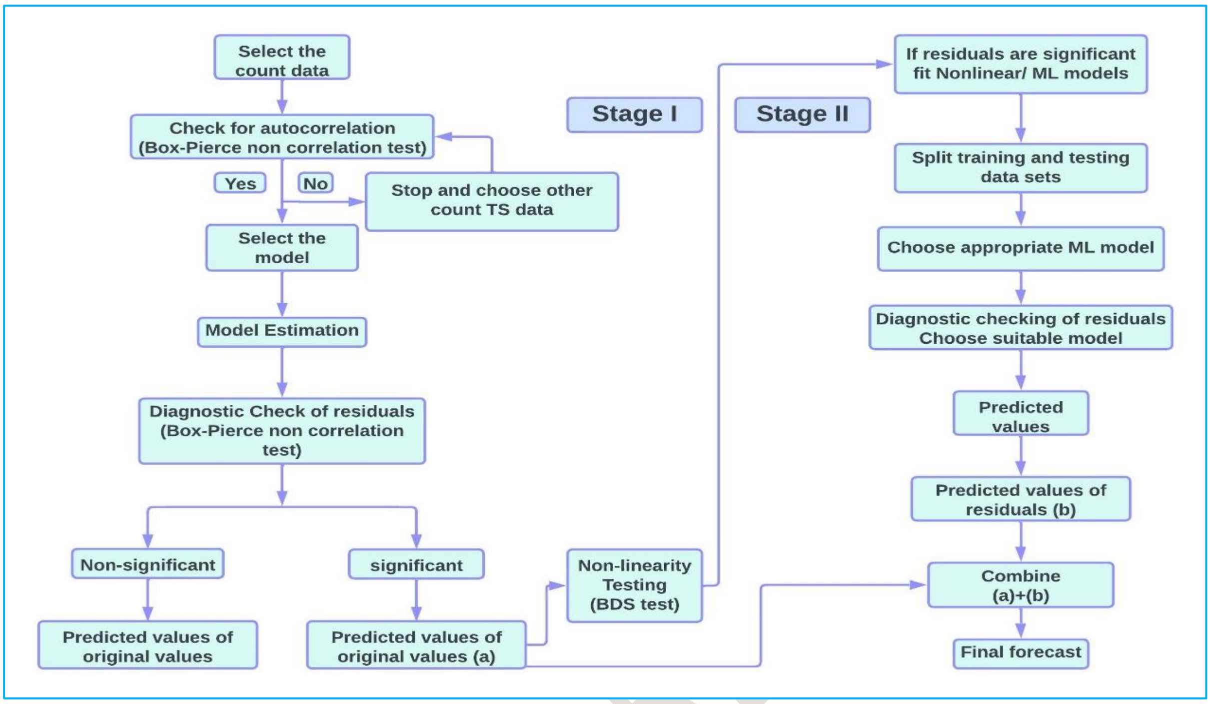

Fig. 2. Schematic representation of two stage methodology

3) Results

The time series plot showing the Jassid population at the Nandyal research station is illustrated in Fig. 3. The Jassid population ranges from a minimum count of 0 to a maximum count of 22.56. The Mean catches is 3.91 and standard deviation of the data is 3.82 with Coefficient of variation 98.07%.

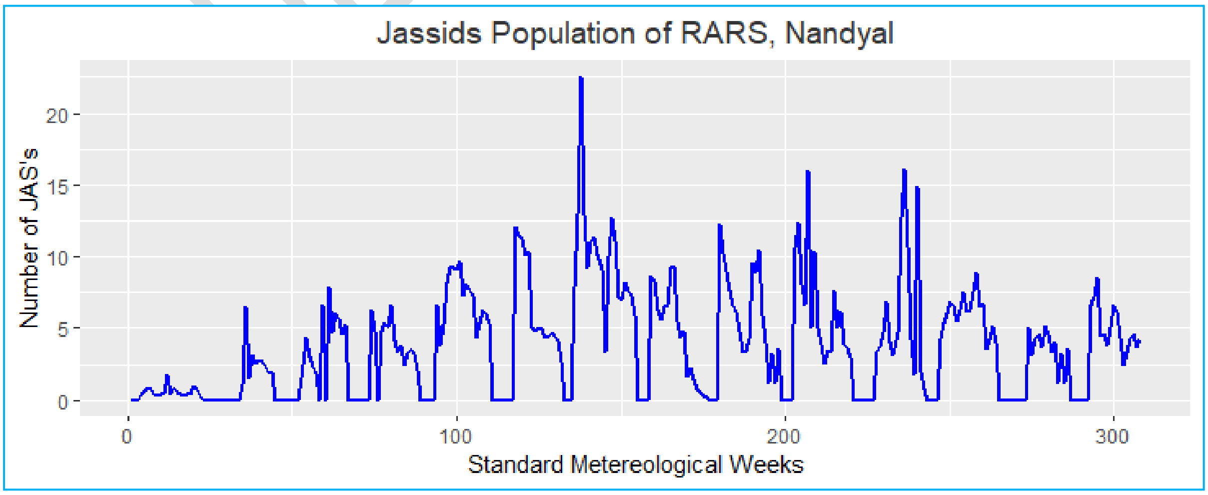

Fig. 3. Jassids population of RARS, Nandyal

A stepwise linear regression analysis was conducted to identify the factors influencing the incidence of the Jassid population. But there was no single explanatory variable which showed significant relationship with Jassids population. So, not able draw conclusions from the stepwise regression. The non significance (p>0.05) of explanatory variables were checked using multiple linear regression presented in the Table 1.

Table 1. Results of Multiple Linear Regression analysis for Jassid population with weather variables

Variables	**Coefficient**	***p*-value**
Intercept	9.93	0.19
TMAX	-0.14	0.51
TMIN	-0.04	0.73
RHM	-0.00	0.98
RHE	-0.00	0.94
RF	-0.004	0.52

Developing various count time series models

In count data, only non-negative integer values can be used for observations, which can exhibit discreteness, skewness, excess zeros, and unusual events. Count data arises from counting rather than ranking. Time series of count data are made up of tallies of observed events over a specific period, and a count time series model must consider the dependence between observations and the over-dispersion comparable with the mean. Count time series analysis has rapidly developed in various fields and can be used to estimate the effects of pest and disease dynamics in agriculture, health implications of environmental pollutants, and environmental science for daily rainfall, among others.

Prior to beginning the modelling process, autocorrelation was tested using the Box-Pierce test for non-correlation. The results indicated the presence of autocorrelation in the data, with a χ^2 value of 152.13 and a p-value of less than 0.0001.

Count time series models such as INGARCH, ZIPAR, ZINBAR, along with the machine learning model ANN and the hybrid model INGARCH-ANN, were applied to the data. The residuals from all models, except for the INGARCH model, showed non-significant autocorrelation. So, hybrid model was developed for INGARCH with ANN as shown in Table 2. All the five models were compared based on error criteria known as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). From the Table 3, looking into error criteria it was evident that INGARCH-ANN model outperforms best with lowest MSE and RMSE values 25.46 and 5.05 respectively. The similar scenario was evident in the testing dataset too as shown in the Table 4.

The models in this study were compared based on the observed differences between their predicted values for the Jassids dataset, using MSE and RMSE as evaluation criteria. To assess the statistical significance between the models, the Diebold-Mariano (DM) test was employed. The results, presented in Table 5, indicated that the INGARCH-ANN model significantly differed from the INGARCH, ZIPAR, ZINBAR, and ANN models. This suggests that the INGARCH-ANN model outperformed the others, likely due to its enhanced capacity to handle the non-linear nature of the Jassid population.

Table 2. Box-pierce test for residuals of models of Jassids population

Particulars	**INGARCH ANN**	**ANN**	**ZIPAR**	**ZINBAR**	**INGARCH**
χ^2	0.001	0.03	0.007	0.003	0.14
p-value	0.96	0.84	0.93	0.95	<0.001

Table 3. Model performance comparison for training data set of Jassids population

Particulars	**Training set**	**INGARCH ANN**	**ANN**	**ZIPAR**	**ZINBAR**	**INGARCH**
Comparison Criteria	MSE	25.46	27.56	28.49	29.02	33.29
	RMSE	5.05	5.25	5.34	5.39	5.77

Table 4. Model performance comparison for testing data set

SMW	Jassid testing dataset	INGARCH ANN	ANN	ZIPAR	ZINBAR	INGARCH
1	5.1	2.69	3.15	1.72	3.74	2.4
2	6.6	3.41	3.75	2.68	3.73	1.89
3	6.2	2.61	2.51	3.09	2.73	2.04
4	4.4	3.62	3.95	1.23	2.72	1.95
5	2.5	2.33	0.33	1.64	2.64	1.54
6	3.4	2.08	1	1.28	2.74	1.37
7	4.2	1.8	1.21	2.69	1.74	1.42
8	4.6	1.79	3.01	3.44	1.53	1.39
9	3.7	2.85	3.88	3.3	1.82	1.25
10	4.2	2.07	1.57	2.99	0.75	1.2
MSE		5.02	5.46	5.70	5.78	9.09
RMSE		2.24	2.34	2.39	2.41	3.01

Table 5. Diebold Mariano test for significance comparison of model performance

Models	DM Statistic	Probability
INGARCH Vs ZIPAR	3.21	0.001
INGARCH Vs ZINBAR	3.16	0.001
INGARCH Vs ANN	7.57	<0.001
ZIPAR Vs ZINBAR	-2.13	0.03
ZIPAR Vs ANN	2.42	0.001
ZINBAR Vs ANN	2.47	0.01
INGARCH-ANN Vs ANN	**0.65**	**<0.0001**
INGARCH-ANN Vs ZIPAR	**1.68**	**<0.0001**
INGARCH-ANN Vs ZINBAR	**-1.89**	**<0.0001**
INGARCH-ANN Vs INGARCH	**-3.52**	**<0.0001**

Structure of best fitted INGARCH-ANN Model for Jassids population:

In this study, a sigmoidal activation function was applied between the input and hidden layers, while a linear activation function was used between the hidden and output layers.

Weather variables, including maximum temperature, minimum temperature, morning relative humidity, evening relative humidity, and rainfall, were incorporated into the input layer as exogenous variables. The ANN models were assessed using mean squared error (MSE) and root mean squared error (RMSE) metrics. The best model being selected as the NNAR (3,8) model with 8 tapped delays and 6 hidden nodes (8:6S:1L). This model consisted of an average of 50 networks, each with an 8:6S:1L network structure and 92 weights. Additionally, a Box-Pierce non-autocorrelation test was conducted on the residuals, which indicated that the residuals were non-autocorrelated (p-value = 0.96) presented in Table 6.

Table 6. INAGRCH-ANN model parameter specification for Jassids population

Parameter	**Specification**
Input lag	3
Output variable	1
Hidden nodes	6
Hidden layer	1
Exogenous variables	5
Model	8:6S:1L
Network type	feed forward
Activation function(I:H)	Sigmoidal
Activation function(H:O)	Identity
Box Test for Non-Correlation	$\chi^2 = 0.001$ (p=0.96)

Conclusion

The objective of this study was to develop an effective forecasting service for predicting Jassid populations, which would aid in designing and implementing location-specific pest management strategies to prevent cotton yield losses. Among the count time series models tested, the INGARCH-ANN model demonstrated superior performance. The order of prediction accuracy of models under consideration is INGARCH-ANN>ANN>ZIPAR>ZINBAR>INGARCH as per the error criteria.

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

Option 2:

Author(s) hereby declare that generative AI technologies such as Large Language Models, etc have been used during writing or editing of manuscripts. This explanation will include the name, version, model, and source of the generative AI technology and as well as all input prompts provided to the generative AI technology

Details of the AI usage are given below:

1.

2.

3.

References:

Agrawal, R and Mehta, S. C. 2007. Weather based forecasting of crop yields, pests and diseases-IASRI models. *Journal of Indian Society of Agricultural Statistics.* 61(2): 255-263.

Assefa, E and Tadesse, M. 2017. Factors related to the use of antenatal care services in Ethiopia: application of the zero-inflated negative binomial model. *Women and health*. 57(7): 804-821.

Barajas, L.G., Egerstedt, M.B., Kamen, E.W and Goldstein, A. 2008. Stencil printing process modeling and control using statistical neural networks. *IEEE transactions on electronics packaging manufacturing*. 31(1): 9-18.

Khedhiri, S. 2021. Statistical modeling of COVID-19 deaths with excess zero counts. *Epidemiologic Methods*. 10(1): 5-4.

Kim, H., Shoji, Y., Tsuge, T., Aikoh, T and Kuriyama, K. 2021. Understanding recreation demands and visitor characteristics of urban green spaces: A use of the zero- Inflated negative binomial model. *Urban Forestry and Urban Greening*. 65:127332.

Kim, J.Y., Kim, H.Y., Park, D and Chung, Y. 2018. Modelling of fault in RPM using the GLARMA and INGARCH model. *Electronics Letters*. 54(5): 297-299.

Lee, Y and Lee, S. 2019. On causality test for time series of counts based on Poisson INGARCH models with application to crime and temperature data. *Communications in Statistics-Simulation and Computation*. *48*(6): 1901-1911.

Majo, M.C and Soest, A. 2011. The fixed-effects zero-inflated Poisson model with an application to health care utilization. 4(1): 5-7.

Raihan, M.A., Alluri, P., Wu, W and Gan, A. 2018. Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. *Accident Analysis & Prevention*. 123: 303-313.

Rathod, S., Yerram, S., Arya, P., Katti, G., Rani, J., Padmakumari, A.P., Somasekhar, N., Padmavathi, C., Ondrasek, G., Amudan, S and Malathi, S. 2021. Climate-Based Modeling and Prediction of Rice Gall Midge Populations Using Count Time Series and Machine Learning Approaches. *Agronomy*. 12(1).

Reddy, B.N.K., Rathod, S., Kallakuri, S., Sridhar, Y., Admala, M., Malathi, S., Pandit, P and Jyostna, B., 2022. Modelling the Relationship between Weather Variables and Rice Yellow Stem Borer Population: A Count Data Modelling Approach. *International Journal of Environment and Climate Change*, 12(11): 3623-3632.

Zhang, G.P. 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing. 50: 159-175.

Zhu, F. 2012. Modeling time series of counts with COM-Poisson INGARCH models. *Mathematical and Computer Modelling*. 56(9-10): 191-20.

Zulkifli, M., Ismail, N and Razali, A.M. 2011. Zero-inflated Poisson versus zero-inflated negative binomial: Application to theft insurance data. *The 7th IMT-GT International Conference on Mathematics, Statistics and its Applications*.