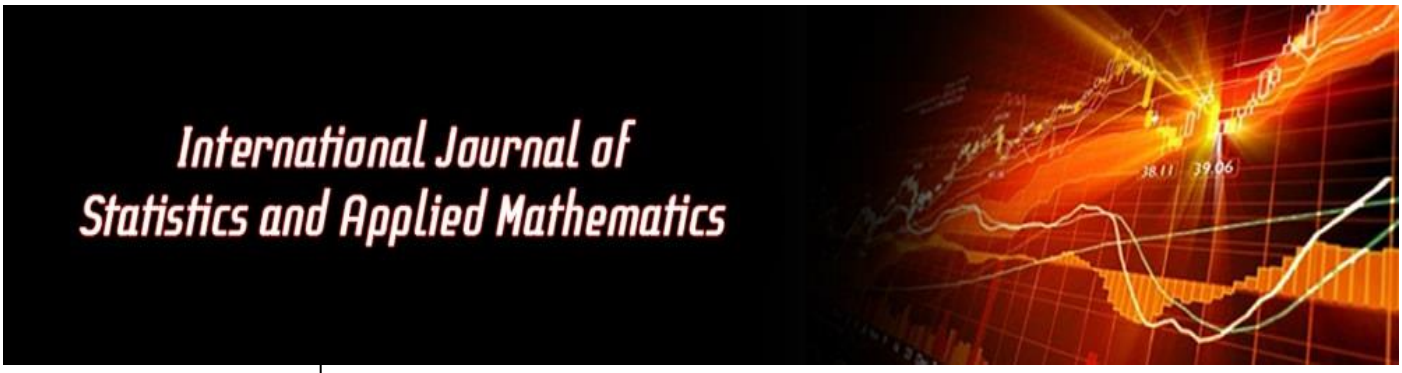
International Journal of Statistics and Applied Mathematics 2023; SP-8(4): 257-261



ISSN: 2456-1452
Maths 2023; SP-8(4): 257-261

https://www.mathsjournal.com
Received: 06-04-2023
Accepted: 11-05-2023

B Venkataviswateja
Agricultural College, Bapatla, ANGRAU, Andhra Pradesh, India

V Srinivasa Rao
Agricultural College, Bapatla, ANGRAU, Andhra Pradesh, India

A Dhandapani
Agricultural College, Bapatla, ANGRAU, Andhra Pradesh, India

G Raghunadha Reddy
Agricultural College, Bapatla, ANGRAU, Andhra Pradesh, India

D Ramesh
Agricultural College, Bapatla, ANGRAU, Andhra Pradesh, India

DVSLP Anand Kumar
Agricultural College, Bapatla, ANGRAU, Andhra Pradesh, India

Corresponding Author:
B Venkataviswateja
Agricultural College, Bapatla, ANGRAU, Andhra Pradesh, India

Modelling of leaf folder populations (*Cnaphalocrocis medinalis*) in Paddy: A count time series approach

B Venkataviswateja, V Srinivasa Rao, A Dhandapani, G Raghunadha Reddy, D Ramesh and ADVSLP Anand Kumar

DOI: https://doi.org/10.22271/maths.2023.v8.i4Sd.1078

Abstract
The present study was conducted to model the Leaf Folder pest population in Paddy at Agricultural Research Station, Bapatla. The secondary data between 2012-13 to 2020-21 was considered based on data availability. Correlation and stepwise regression were used to check the relationship between pest and weather parameters. The Minimum temperature had significant negative correlation and maximum temperature and maximum temperature and rainfall were significantly contributed, and having negative impact on Leaf Folder population. Count time series and machine learning models are used for fitting the Leaf Folder dataset. INGARCH-ANN model outperformed well than INGARCH, ZIPAR, ZINBAR, ANN models based on error comparison criteria (MSE and RMSE) and the statistical significance between the models utilized in the study were determined by Diebold- Marino test statistic (DM test). The order of prediction accuracy of the models under consideration is INGARCH-ANN>ANN >ZINBAR>ZIPAR>INGARCH.

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

1. Introduction
Agriculture is a major contributor to the Indian economy, producing around 28% of the GDP. Achieving self-sufficiency in food grain production has given high priority to the agricultural sector in development plans. However, pest and disease attacks can greatly affect agricultural production, resulting in losses of up to Rs 50,000 crores annually in India.

Among cereals, Rice (*Oryza sativa*, F: Graminae) is an important food crop and a major calorie source for over 3 billion people on the earth. Rice is one of the most consumed food crops in the world. Rice fields act as an 'environmental buffer' and have proven as a 'dynamic agro-based ecosystem' that helps to balance temperature and wind speed and to modulate the effect on the agro-surrounding. Globally, about 114 countries cultivate rice, assuring a primary source of income and employment for the rural population. About 100 million families are engaged in this cultivation in Asia. Asian countries have the largest share of global rice production. Almost 90% of rice is grown and consumed in Asia. The projected world's rice consumers is 8.27 billion in 2030.

Yield loss in rice due to insect pests is one among several factors that contribute to the gap between potential and actual rice yields across the nation. A total of twenty-one species of lepidopteran stem borers have been identified as rice pests across the world. The loss in yield of paddy due to leaf folder outbreak is 30-50%. So, there is a need to develop forewarning models to mitigate the losses.

This study aims to develop various models for forecasting pest populations in agriculture using weather parameters as exogenous variables. Recent advances in modelling have explored machine learning techniques for predicting agricultural fields, such as oil seed production, banana yield, rice yield and pests, 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 to predict pest populations by utilizing count data-driven approaches.

2. Methodology

The secondary data of Leaf Folder in Paddy was collected from the Agricultural Research Station (ARS), Bapatla under ANGRAU in Andhra Pradesh. Research station is in 15.9039°N 80.4671°E coordinates. It is situated at an elevation ranging from 8 m above MSL. The secondary data of Leaf Folder on Paddy available from 2012-13 to 2020-21. The data available in standard meteorological weeks (SMW). The pest data is counts of Leaf Folder collected in light traps arranged in the field. The weather parameters Maximum Temperature, Minimum Temperature, Rainfall, Relative Humidity Morning, relative Humidity Evening was also collected from meteorological station. The total data is divided into training data and testing data (last 10 observations).

2.1 Statistical models

2.1.1. Correlation analysis

Simple correlations were carried out to determine the degree of relationship between two variables. In the present study, the degree of relationships between pest population and each of the weather parameters *viz.*, minimum temperature, maximum temperature, morning relative humidity, evening relative humidity, rainfall, and sunshine hours were determined using Karl Pearson's correlation coefficient which can be measured using.

$$r_{xy} = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

2.1.2. Stepwise Regression

The stepwise regression procedure is a statistical method used to identify the most significant variables that contribute to the variation in a dependent variable. The procedure involves a series of steps that are repeated until the most significant variables are identified. The steps involved in the procedure are:

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

2.1.3. INGARCH (Integer valued generalized auto-regressive conditional Heteroscedastic) model

The integer-valued generalized autoregressive conditional Heteroscedastic (INGARCH) model is special case of generalized linear model where it follows poison and negative binomial distribution. The integer-valued generalized autoregressive conditional Heteroscedastic (INGARCH) models are the class of GLM in which the conditional distribution of dependent variable or observed count is assumed to follow popular discrete distributions like Poisson negative binomial, generalized Poisson and double Poisson distributions by Rathod *et al*. (2021) [9]. For the estimation of INGARCH model, conditional likelihood estimation was used.

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.

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

2.1.4 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* Regressors variables. 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 $\beta_1, \beta_2 \ldots. \beta_k$ are unknown parameters that are estimated from a set of data and their estimates are symbolized as $b_1, b_2 \ldots. b_k$. 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* repressors variables.

2.1.5 Zero-Inflated Negative Binomial Autoregressive (ZINBAR) Model

The zero-inflated negative binomial regression is used for count data that exhibit over dispersion and excess zeros. The data distribution combines the negative binomial distribution and the logit distribution by Kim *et al*. (2021). The possible values of y are the non-negative integers: 0, 1, 2.

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

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 component can include an exposure time *t* and a set of *k* regressors variables. The expression related these quantities is

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

Often, x_1 = 1, in which case β_1 is called the intercept. The regression coefficients β_1, $\beta_2 \ldots. \beta_k$ are known parameters that are estimated from a set of data. Their estimates are symbolized as b_1, b_2, …. b_k.

2.1.6 Artificial neural network model (ANN)

Artificial Neural Network (ANN) is the most widely used machine learning technique in recent years. In the area of time

series modelling, the ANN is commonly referred as the autoregressive neural network as it considers time lags as inputs. The time series framework for ANN can be mathematically modelled using a neural network with implicit functional representation of time. The general expression for the final output Y_t of a multi-layer feed-forward autoregressive neural network is expressed as follows:

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

ANNX is an Artificial Neural Network model with exogenous variables. Where X denotes the exogenous variables i.e., independent variables.

3. Results

The time series plot of Leaf Folder population of Bapatla centre was plotted and depicted in Fig1. The range of Leaf Folder is with Minimum count of 0 and Maximum count of 765. The Mean is 49.68 and standard deviation of the data is 106.25.

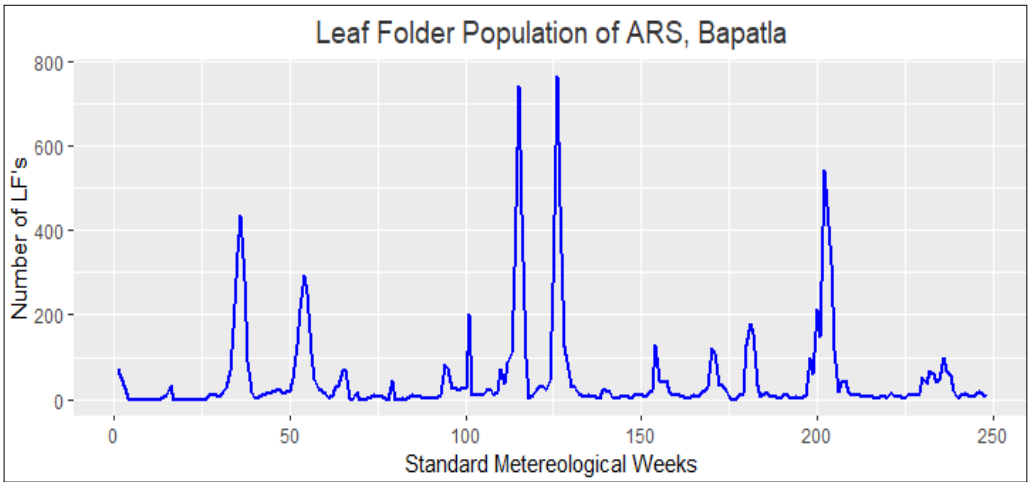


Fig 1: Leaf Folder population of Bapatla Research Station

Pearson correlation was made between Leaf Folder population and considered climatological variables, as to understand were depicted in Table 1. The Minimum temperature had significant negative correlation and maximum temperature, relative humidity evening and rainfall is showing non-significant negative correlation and relative humidity morning is showing non-significant positive correlation with Leaf folder data. The bivariate correlation between weather variables in Table 1 were self-explanatory.

Table 1: Results of correlation analysis for Leaf Folder population

Parameters	Leaf Folder	TMAX	TMIN	RHM	RHE
TMAX	-0.01369				
	0.8301				
TMIN	-0.15357	-0.25243			
	0.0155	<.0001			
RHM	0.02495	-0.39747	-0.29190		
	0.6958	< .0001	< .0001		
RHE	-0.06011	-0.07804	0.22117	0.44423	
	0.3458	0.2207	0.0005	< .0001	
RF	-0.11408	0.06279	0.11163	0.13587	0.38606
	0.0729	0.3247	0.0793	0.0324	< .0001

The step-wise linear regression analysis was carried out to identify the factors which influence the incidence of Leaf Folder population. It was revealed that; the results of step-wise regression analysis were depicted in Table 2. For the response variable Leaf Folder population, explanatory variables like maximum temperature and rainfall were significantly contributing, and having negative impact on Leaf Folder population for the data under consideration. Though the listed variables had a significant influence on the Leaf Folder populations but the model R^2 value for the fitted regression in the Bapatla station for Leaf Folder population was very low, which indicated that the model was not a strong fit, due to non-linearity and presence of high heterogeneity in dependent variable.

Table 2: Results of Stepwise Regression analysis for Leaf Folder population and weather variables

Variable	Estimate	S.E.	F-Value	Probability	R^2	Model R^2
Intercept	137.92882	37.29	13.68	0.00		
TMAX	-3.59914	1.59	5.94	0.01	0.02	0.02
RF	-0.26588	0.17124	2.41	0.12	0.01	0.03

3.1 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 arise 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.

Before starting the modelling, autocorrelation was tested using Box-Pierce non-correlation test. It was proved that autocorrelation is present in the data as the χ^2 value is 128.9 and the probability value is < 0.0001.

Count time series models like INGARCH, ZIPAR, ZINBAR and Machine learning model ANN and hybrid model INGARCH-ANN was fitted for the data. All the model's residuals show non-significant autocorrelation except INGARCH model. So, hybrid model was developed for INGARCH with ANN. 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 238.32 and 15.43 respectively. The similar scenario was evident in the testing dataset too as shown in the Table 4. The comparison of the models in this study was based on the observed differences between the predicted values of the models for the Leaf Folder dataset, using MSE and RMSE criteria. However, to determine the statistical significance between the models, the Diebold-Marino test statistic (DM test) was used. The results showed that compared to the INGARCH, ZIPAR, ZINBAR and ANN models, the INGARCH-ANN model was significantly better for the Leaf Folder data. This indicates that the INGARCH-ANN model had superior performance due to its greater capacity and ability to handle the non-linear nature of the Leaf Folder population. Table 5 provides further details on the performance of the models in both the testing and training datasets.

Table 3: Model performance comparison for training data set

Particulars	Training set	INGARCH	ZIPAR	ZINBAR	ANN	Ingarch ANN
Comparison Criteria	MSE	10176.5	9228.99	7301.99	326.20	238.32
	RMSE	100.87	96.06	85.45	18.06	15.43

Table 4: Model performance comparison for testing data set

SMW	Leaf Folder Testing data	INGARCH	ZIPAR	ZINBAR	ANN	Ingarch ANN
1	20	54.49	42.57	24.02	5.01	7.01
2	4	50.37	43.43	20.00	2.26	5.26
3	11	46.45	42.80	21.72	3.52	8.52
4	15	44.75	41.38	23.29	1.28	2.58
5	10	43.14	40.60	22.69	3.25	5.25
6	7	42.44	40.60	22.69	3.29	3.99
7	12	41.77	39.60	22.69	3.25	4.25
8	17	41.48	42.60	22.69	3.26	6.26
9	14	41.21	41.60	28.69	3.25	8.25
10	6	41.093	40.68	32.72	4.19	5.69
MSE		1129.58	922.07	208.56	91.43	57.01
RMSE		33.60	30.36	14.44	9.56	7.55

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

Models	DM Statistic	Probability
INGARCH vs ZIPAR	-1.43	0.15
INGARCH vs ZINBAR	-1.31	0.19
INGARCH vs ANN	-1.11	< 0.0001
ZIPAR vs ZINBAR	-1.28	0.20
ZIPAR vs ANN	2.426	< 0.0001
ZINBAR vs ANN	-1.12	0.26
INGARCH-ANN vs ANN	0.710	< 0.0001
INGARCH-ANN vs ZIPAR	1.81	< 0.0001
INGARCH-ANN vs ZIBAR	-1.84	< 0.0001
INGARCH-ANN vs INGARCH	-4.92	< 0.0001

3.2. Structure of best-fitted INGARCH-ANN Model for Leaf Folder population

Table 6: Inagrch-ANN model parameter specification for Leaf Folder population

Parameter	Specification
Input lag	4
Output variable	1
Hidden nodes	8
Hidden layer	1
Exogenous variables	5
Model	9:8S:1L
Network type	Feedforward
Activation function (I: H)	Sigmoidal
Activation function (H: O)	Identity
Box Test for Non-Correlation	$\chi^2 = 0.001$(p =0.89)

In this study, a sigmoidal activation function was implemented in the input to hidden layer, while a linear activation function was used in the hidden to output layer. In this, weather variables such as maximum temperature, minimum temperature, morning relative humidity, evening relative humidity, and rainfall were considered in input layer which were exogenous variables. Candidate models were evaluated based on their mean squared error (MSE) and root mean squared error (RMSE) values. The best model being selected as the NNAR (4, 8) model with 9 tapped delays and 8 hidden nodes (9:8S:1L). This model consisted of an average of 50 networks, each with a 9:8S:1L network structure and 89 weights. Additionally, a Box-Pierce non-autocorrelation test was conducted on the residuals, which indicated that the residuals were non-auto-correlated (probability value - 0.89).

4. Conclusion

The study was carried out with an objective to establish an efficient forewarning service to forecast Leaf Folder population for designing and implementing of effective location-specific pest management strategies to avoid Paddy yield losses. The INGARCH-ANN model outperformed among the count time series models. The order of prediction accuracy of models under consideration is INGARCH-ANN>ANN>ZINBAR>ZIPAR>INGARCH as per the error criteria.

5. References

1. Assefa E, Tadesse M. Factors related to the use of antenatal care services in Ethiopia: Application of the

zero-inflated negative binomial model. Women and Health. 2017;57(7):804-821.
2. Barajas LG, Egerstedt MB, Kamen EW, Goldstein A. Stencil printing process modelling and control using statistical neural networks. IEEE transactions on electronics packaging manufacturing. 2008;31(1):9-18.
3. Khedhiri S. Statistical modelling of COVID-19 deaths with excess zero counts. Epidemiologic Methods. 2021;10(1):5-4.
4. Kim H, Shoji Y, Tsuge T, Aikoh T, Kuriyama K. Understanding recreation demands and visitor characteristics of urban green spaces: A use of the zero-Inflated negative binomial model. Urban Forestry and Urban Greening. 2021;65:127332.
5. Kim JY, Kim HY, Park D, Chung Y. Modelling of fault in RPM using the GLARMA and INGARCH model. Electronics Letters. 2018;54(5):297-299.
6. Lee Y, Lee S. 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. 2019;48(6):1901-1911.
7. Majo MC, Soest A. The fixed-effects zero-inflated Poisson model with an application to health care utilization. 2011;4(1):5-7.
8. Raihan MA, Alluri P, Wu W, Gan A. Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero-inflated negative binomial models. Accident Analysis & Prevention. 2018;123:303-313.
9. Rathod S, Yerram S, Arya P, Katti G, Rani J, Padmakumari AP. Climate-Based modelling and prediction of rice gall midge populations using count time series and machine learning approaches. Agronomy. 2021;12:1.