The Andhra Agric. J 69 (2): 295-302, 2022

Forewarning models for Early Shoot Borer populations (*Chilo infuscatellus*) in Sugarcane - A Count Time Series Approach

B Venkataviswateja, V Srinivasa Rao, A Dhandapani, G Raghunadha Reddy, D Ramesh, A D V S L P Anand Kumar and M Visalakshi

Department of Statistics and Computer Applications, Agricultural College, Bapatla, A. P.

ABSTRACT

The present study was conducted to model the early shoot borer pest population in sugarcane at Regional Agricultural Research Station, Anakapalle. The secondary data between 2013-2020 (308 observations) was considered based on data availability. Correlation and stepwise regression were used to check the relationship between pest and weather parameters. The maximum temperature, minimum temperature and relative humidity morning showed significant positive correlation and maximum temperature and rainfall were significantly contributing, and had positive impact on ESB population. Count time series and machine learning models were used for fitting the ESB dataset. ANN model was outperformed well than INGARCH, ZIPAR, ZINBAR models based on error comparison criteria (MSE and RMSE) and the statistical significance between the models was verified in the study by using Diebold- Marino test statistic (DM test). The order of prediction accuracy of the models under consideration was identified as ANN>ZINBAR>ZIPAR>INGARCH.

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

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 crore annually in India.

Sugarcane crop losses due to pest damage are a major problem faced by farmers in India every year. The most common pests that attack sugarcane crops in India include the sugarcane borers, shoot borer, termites, and armyworms. These pests cause significant damage to the crop, leading to reduced yields and economic losses for farmers. Sugarcane borers, particularly early shoot borers, are highly destructive pests that feed on sugarcane stalks. They can cause significant damage to sugarcane crops, leading to reduced growth and yield. The percentage of crop loss due to shoot borers in sugarcane crops in India depends on various factors such as infestation severity, crop management practices, and environmental conditions. However, studies indicate that shoot borers can cause considerable damage to sugarcane crops, resulting in substantial yield losses.

According to a study conducted by the Indian Council of Agricultural Research (ICAR), the average yield loss due to shoot borers in sugarcane crops in India is estimated to be around 10-15%. However, in severe infestations, yield losses can reach up to

40-50%. In addition to yield losses, shoot borers can also cause quality deterioration in sugarcane crops by increasing the fiber content and reducing sugar recovery. This can lead to further economic losses for farmers. So, Early warning systems based on pest modelling can provide a tool for predicting and investigating pest and disease status, enabling timely control measures and loss assessment.

This study aims to develop various models for forecasting pest populations in agriculture using weather parameters as exogenous variables. Recent advances in modeling 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 sugarcane 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.

MATERIAL AND METHODS

The secondary data of Early shoot borer (ESB) in sugarcane was collected from the Regional Agricultural Research station (RARS), Anakapalle under ANGRAU in Andhra Pradesh. Research station is situated in17.6913°N 83.0039°E co-ordinates and an elevation ranging from 29 m above MSL. The secondary data of ESB on sugarcane was collected from 2013 to 2020 (308 observations) as standard meteorological weeks (SMW). In this, the pest data of ESB count was 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 meteorological station. In the study the total data was divided into training data and testing data (last 10 observations) for fitting the model.

Statistical models

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

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

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

INGARCH (Integer Valued Generalized Autoregressive 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). 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;

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

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

$$\mu_i = exp(ln(\grave{t}_i) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Often, $x_1 = 1$, in which case β_1 is called the intercept, the regression coefficients $\beta_1, \beta_2 \ldots \beta_\kappa$ are unknown parameters that are estimated from a set of data and their estimates are symbolized as $b_1, b_2 \ldots\ldots b_k$ this logistic link function Π 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.

Zero Inflated Negative Binomial Autoregressive (ZINBAR) Model

The zero inflated negative binomial regression is used for count data that exhibit overdispersion 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* regressor variable. The expression related these quantities is

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

Often, $X_1 = 1$, in which case β_1 is called the intercept. The regression coefficients $\beta_1, \beta_2 \ldots. \beta_k$ are known parameters that are estimated from a set of data. Their estimates are symbolized as $b_1, b_2 \ldots\ldots b_k$

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.

RESULTS AND DISCUSSION

The time series plot of ESB population of Anakapalle centre was plotted and depicted in Fig1. The range of ESB was with Minimum count of 0 (in 22 SMW's) and Maximum count of 62 (in 2 SMW's). The Mean and standard deviation of ESB during the period was recorded as 7.93 and 8.11 respectively.

Pearson correlation coefficients between ESB population and considered climatological variables were depicted in Table1. The maximum temperature and Minimum temperature showing significant positive correlation and relative humidity morning showing significant negative correlation with Early shoot borer data. The relative humidity evening and Rainfall had non-significant negative and positive correlation with the ESB data respectively. The bivariate correlation between weather variables in Table 1 were self-explanatory.

The step-wise linear regression analysis was carried out to identify the factors influencing the incidence of ESB population and weather variables. The results of step-wise regression analysis were depicted in Table 2. As per the explanatory variables like maximum temperature and rainfall are significantly contributing on the response variable ESB population, and had positive impact. Though the listed variables had significant influence on the ESB populations, the model R^2 value for the fitted regression in the

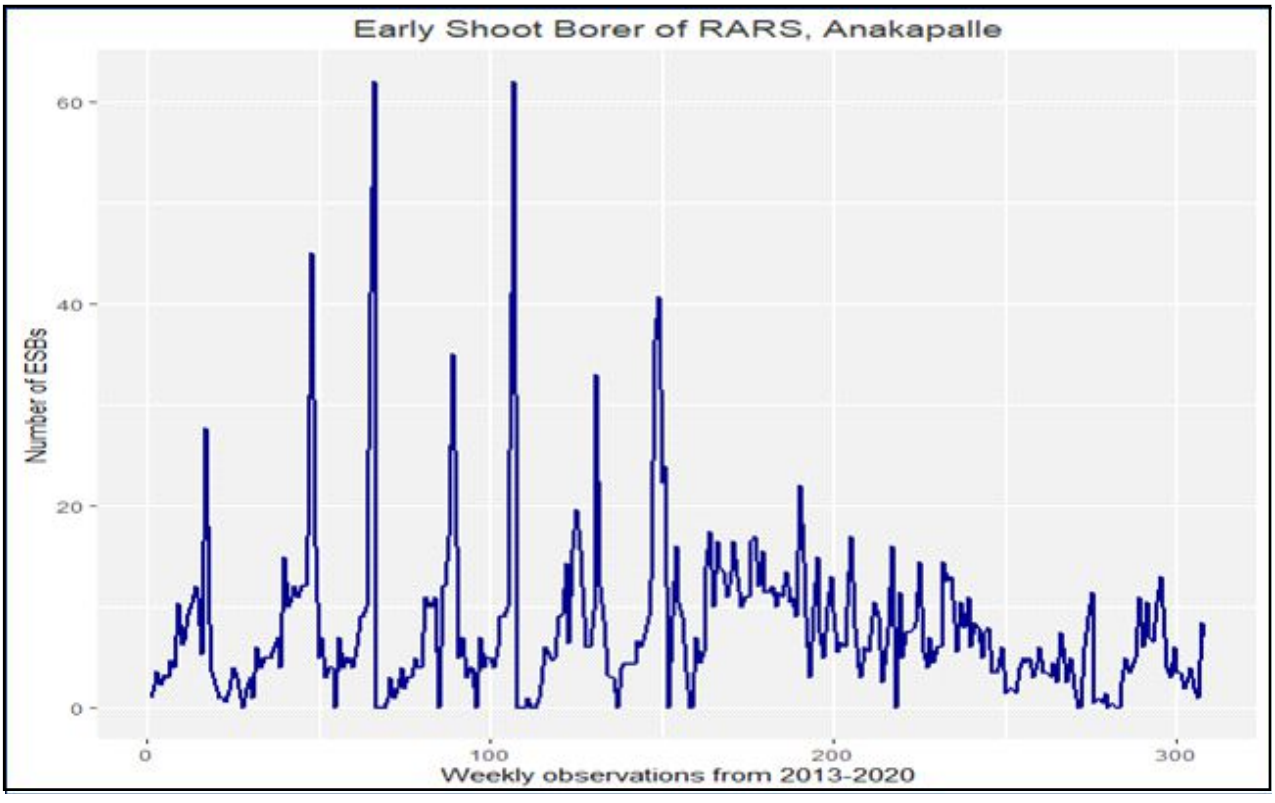

Fig 1. ESB population of Anakapalle Research Station

Table 1. Results of correlation analysis for ESB population

Parameters	ESB	TMAX	TMIN	RHM	RHE
TMAX	**0.18**				
	***p* = 0.001**				
TMIN	**0.11**	0.47			
	***p* = 0.04**	*p* = <0.0001			
RHM	**-0.12**	-0.3	-0.19		
	***p* = 0.02**	*p* = <0.0001	*p* = 0.004		
RHE	-0.01	-0.14	0.55	0.18	
	p = 0.76	*p* = 0.01	*p* = <0.0001	*p* = 0.001	
RF	0.1	-0.16	0.21	0.01	0.41
	p = 0.05	*p* = 0.004	*p* = 0.002	*p* = 0.83	*p* = <0.0001

Table 2. Results of Stepwise Regression analysis for ESB population and weather variables

Variable	Estimate	S.E.	F -Value	Probability	R^2	Model R^2
Intercept	-14.5	6.04	5.77	0.02		
TMAX	0.65	0.18	13.32	0	0.03	0.03
RF	0.03	0.01	6.29	0.01	0.02	0.05

Anakapalle station for ESB population was low, which indicated that the model was not appropriate due to existence of non-linearity and presence of high heterogeneity in dependent variable.

Developing various count time series models

Before starting the modelling, auto correlation was tested using Box-Pierce non correlation test. It was proved that autocorrelation was present in the data as per χ^2 value is 74.29 and the probability value is < 0.0001.

Count time series models like INGARCH, ZIPAR, ZINBAR and Machine learning model ANN were tried to fit for the data. All the model's residuals shown non-significant autocorrelation as shown in the Table 3. All the four models were compared based on error criteria known as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). From the Table 3, as per error criteria it was evident that ANN model outperformed with lowest MSE and RMSE values i.e., 6.71 and 2.59 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 ESB 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, and ZINBAR models, the ANN model was significantly better for the ESB data. This indicated that the ANN model had superior performance due to its greater capacity and ability to handle the non-linear nature of the ESB population. The performance of the models in both the testing and training datasets were depicted in Table 5.

Table 3. Model performance comparison for training data set

Particulars	Training set	ANN	ZINBAR	ZIPAR	INGARCH
Comparison Criteria	MSE	6.71	50.42	69.43	115.72
	RMSE	2.59	7.1	8.33	10.75
Box -Pierce Non correlation test	χ^2 -squared	0.01	0.00	2.32	1.99
	p-value	0.93	0.97	0.13	0.66

Table 4. Model performance comparison for testing data set

SMW	ESB Testing data	ANN	ZINBAR	ZIPAR	INGARCH
1	6.00	4.04	5.85	4.72	2.03
2	3.50	6.66	5.73	4.68	1.48
3	3.50	4.17	6.2	5.09	1.17
4	2.00	4.87	5.1	5.13	0.99
5	3.00	3.88	8.46	7.64	0.89
6	4.00	6.57	5.72	6.28	0.83
7	2.00	4.07	5.55	6.69	0.8
8	1.00	3.66	7.92	7.44	0.78
9	8.50	7.14	8.78	8.3	0.77
10	7.00	8.08	9.92	9.99	0.77
MSE		4.43	12.38	13.95	14.08
RMSE		2.1	3.52	3.73	3.75

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

Models	DM Statistic	Probability
INGARCH *Vs* ZIPAR	1.06	0.28
INGARCH *Vs* ZINBAR	0.34	0.76
INGARCH *Vs* ANN	3.22	<0.001
ZIPAR *Vs* ZINBAR	-0.839	0.40
ZIPAR *Vs* ANN	1.919	<0.005
ZINBAR *Vs* ANN	2.24	<0.005

Table 6. ANN model parameter specification for ESB population

Parameter	Specification
Input lag	3
Output variable	1
Hidden nodes	8
Hidden layer	1
Exogenous variables	5
Model	8:8S:1L
Network type	Feed forward
Activation function(I:H)	Sigmoidal
Activation function(H:O)	Identity
Box Test for Non-Correlation	χ^2 = 0.12 (p=0.74)

Structure of best fitted ANN Model for ESB population

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. The input layer included weather variables such as maximum temperature, minimum temperature, morning relative humidity, evening relative humidity, and rainfall as exogenous variables. Candidate models were evaluated based on their mean squared error (MSE) and root mean squared error (RMSE) values, with the best model being selected as the NNAR (3,8) model with 8 tapped delays and 8 hidden nodes (8:8S:1L). This model consisted of an average of 50 networks, each with an 8-8-1 network structure and 81 weights. Additionally, a Box-Pierce non-correlation test was conducted on the residuals, indicating that they were non-correlated as per probability value of 0.74.

CONCLUSION

The study was carried out with an objective to establish an efficient forewarning service to forecast ESB population for designing and implementation of effective location specific pest management strategies to avoid sugarcane yield losses. The maximum temperature and Minimum temperature showing significant positive correlation and relative humidity morning showing significant negative correlation with Early shoot borer data. The relative humidity evening and Rainfall had non-significant negative and positive correlation with the ESB data respectively. The ANN model outperformed among the count time series models. The order of prediction accuracy of models under consideration is ANN>INGARCH >ZINBAR>ZIPAR as per the error criteria.

LITERATURE CITED

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-Simulationand 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).

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*.