The Pharma Innovation Journal 2023; 12(4): 2319-2326

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277-7695
ISSN (P): 2349-8242
NAAS Rating: 5.23
TPI 2023; 12(4): 2319-2326

www.thepharmajournal.com
Received: 09-02-2023
Accepted: 13-03-2023

Baswaraj Biradar
Agricultural Research Station, Bidar, University of Agricultural Sciences, Raichur, Karnataka, India

Sunil A Kulkarni
Agricultural Research Station, Bidar, University of Agricultural Sciences, Raichur, Karnataka, India

Shobharani M
Agricultural Research Station, Bidar, University of Agricultural Sciences, Raichur, Karnataka, India

Sidramappa
Agricultural Research Station, Bidar, University of Agricultural Sciences, Raichur, Karnataka, India

Gnyanadev Bulla
Agricultural Research Station, Bidar, University of Agricultural Sciences, Raichur, Karnataka, India

Santosh Rathod
Indian Institute of Rice Research (ICAR-IIRR), Hyderabad, Telangana, India

Naveena K
Centre for Water Resources Development and Management Kunnamangalam, Kozhikode, Kerala. India

Gayathri Chitikela
Professor, Jayashankar Telangana State Agricultural University, Hyderabad, Telangana, India

Fakeerappa Arabhanvi
Krishi Vigyan Kendra, University of Agricultural Sciences, Raichur, Karnataka, India

Corresponding Author:
Baswaraj Biradar
Agricultural Research Station, Bidar, University of Agricultural Sciences, Raichur, Karnataka, India

Enhancing legume crop protection: Machine learning approach for accurate prediction of lepidopteran pest populations in Kalyan Karnataka

Baswaraj Biradar, Sunil A Kulkarni, Shobharani M, Sidramappa, Gnyanadev Bulla, Santosh Rathod, Naveena K, Gayathri Chitikela and Fakeerappa Arabhanvi

Abstract
Legumes are a vital source of high-protein food and play a crucial role in nitrogen fixation in the atmosphere. However, their productivity is threatened by various lepidopteran pests. In this study, we aimed to develop a robust statistical model for predicting the pest population of soybean, pigeonpea, and chickpea, using climatological input parameters as influencing variables. To achieve this, we employed improved statistical and machine learning models, such as INGARCH, Random Forest, Support Vector Regression (SVR), and neural network (ANN), to predict pest populations in the North Eastern Transitional belts of Kalyan Karnataka. We conducted the study at ARS, examining soybean tobacco caterpillar incidence, pod borer incidence in pigeon pea and chickpea crops, using various crop varieties, including Soybean: JS 335, Pigeonpea: BSMR-736, and Chickpea-JG-11, over a 15-year period (2006 to 2020), incorporating historical weather data comprising rainfall, maximum and minimum temperature, and relative humidity (morning and evening). Our results demonstrate that the ANN model is a highly viable and effective alternative for modeling and predicting the incidence of lepidopteron pests based on time-series data. Moreover, the Diebold-Mariano test statistics confirm the superiority of the ANN models over INGARCH, SVM, and Random Forest models. It is expected that machine learning techniques will be extensively used in the future to model the count time series of various crop pests in other crops.

Keywords: Artificial neural network, INGARCH, support vector regression, random forest, soybean tobacco caterpillar and pod borer in grams

Introduction
Legume crops are highly valuable as they are rich in protein and help improve soil fertility through nitrogen fixation. However, these crops are susceptible to damage by pests such as the tobacco caterpillar and pod borer, which can significantly reduce crop yields. In order to minimize these losses, it is crucial for farmers to have access to reliable and sustainable pest forecasting models. These models can help farmers anticipate pest outbreaks and take necessary measures to manage them effectively. By doing so, farmers can protect their crops from damage and maximize their yields, contributing to food security and economic stability.

These crops are primarily cultivated in the North Eastern Transitional region of Kalyan Karnataka, as well as other districts such as Belgavi, Dharwad, Haveri, Vijayapur, and Bagalkote. In India, soybean covered an area of 10.80 million hectares in 2018-19, producing 12.10 million tonnes with a productivity of 1120 kg per ha. Pigeon pea covered an area of 4.5 lakh ha, producing 3.3 lakh tonnes with a productivity of 728 kg per ha, while Chickpea covered 8.95 million hectares, producing 7.06 million tonnes with a productivity of 801 kg per ha.

In Karnataka, soybean covered 3.4 lakh ha, producing 3.39 million tonnes with a productivity of 1000 kg per ha. Pigeon pea and chickpea covered 8.8 and 12.6 lakh ha respectively, producing 8.1 and 9.4 million tonnes with an average productivity of 967 and 784 kg per ha respectively. In Bidar district, soybean, pigeon pea, and chickpea covered an area of 182448, 87952, and 66350 ha respectively, with a production of 273672, 105542, and 89573 metric tonnes and productivity of 1500, 1200, and 1350 kg per ha respectively (Anonymous, 2018, 2019, 2020, 2022).

Tobacco caterpillar and Spodoptera litura larvae defoliate soybean leaves, reducing the weight and number of pods and grain, with severity linked to delayed sowing and heavy rainfall from June to mid-August (Prasad *et al*., 2013; Sasvihalli, *et al*., 2017) [17, 21]. Climatic factors significantly influence the pest's intensity, dynamics and infestation period. A pest-weather forewarning model can predict and prevent pest infestations, helping soybean farmers make timely management decisions.

In India, pigeon pea productivity has been a concern due to damage caused by insect pests, with nearly 250 species known to infest the crop. Among these, Helicoverpa armigera is a major pest causing 60 to 90 percent loss in grain yield under favorable conditions (Sujithra and Chander, 2014) [22]. Similarly, chickpea yields are affected by the gram pod borer Helicoverpa armigera, with biotic stress being the major factor responsible for low yields (Dhingra *et al.*, 2003) [10]. The population density of insect pests is influenced by changes in weather conditions such as temperature, rainfall, relative humidity, sunshine hours, and wind velocity.

Count time series modeling is a popular statistical approach in which integer auto-correlated discrete count observations are considered as inputs, and the observations are assumed to be derived from Poisson and negative Binomial distributions. Kim, 2014 examined machine learning- and regression-based crop pest prediction techniques. Hybrid time series and machine learning models for agricultural yield projections were created by Alam *et al*. (2018) [1] and (2019) [2], while Rathod and Paramesha (2022) [18] explained concepts of various ML models and their applications in agriculture. Gorlapalli *et al*. (2022) [12] developed ML-based models to forecast drought in Hyderabad region of India. The severity of early tomato blight was predicted (Paul *et al*., 2019) [16] as well as the sugarcane borer disease (Huang, 2018) [13].

Predicting lepidopteran pest populations can help farmers take preventive measures, but past models were limited to classical methods like regression analysis and time series models. These methods may not be effective for non-Gaussian, heterogeneous, and nonlinear data. However, machine learning models like SVR and ANN are data-driven and can be effective. This study aims to develop a robust statistical model to predict pest populations in soybean, chickpea, and redgram using climatological variables as influencing factors. The model can aid in decision-making and crop management planning.

Previous studies in India have used count time series models to predict pest and disease populations in agriculture. Arya *et al*. (2015) [8] used the Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) model to forecast pests in Guntur and Faridkot Districts, while Rathod *et al*. (2021) [19] compared machine learning and count time series models to compute gal midge population in rice crops in Hyderabad. This study aims to develop a robust statistical model for predicting lepidopteron pest populations in soybean, chickpea, and redgram crops in the north eastern transitional belt of Kalyana Karnataka region by analyzing the causal relationships between lepidopteron caterpillar populations and weather parameters using Karl Pearson correlation and comparing the performance of models like INGARCH, Random Forest, SVR, and ANN.

Materials and Methods

Data Collection

The study was conducted at Agricultural Research Station, Bidar on Soybean tobacco caterpillar incidence in soybean crop, Pod borer incidence in pigeon pea and Chick pea crops respectively (Supplementary fig. 1, 2, 3). Various varieties such as (soybean JS-335), (pigeon pea BSMR-736) and (chick pea JG-11) and 15 years [Historical Climatic data] comprising of Rainfall, Max (T), Min (T), RHM, RHE & Wind speed] i.e from 2006 to 2020 week wise was collected from Surface meteorological observatory, Agricultural Research Station, Bidar. Weekly Observations belonging to 2019 to 2020 were used as testing/validation sets, and remaining observations were used as the training data set.

Totally 11-week observation was taken in soybean crop from (29th SMW to 43th SMW) on tobacco caterpillar, 15-week observation (34th SMW to 48th SMW on pigeon pea) and 17-week observation (45th SMW to 9th SMW) for a period of 15 years that is from 2006 to 2020.

Statistical and Machine Learning Models

Descriptive statistics, time series plots, and Pearson's correlation analysis were used to describe and understand the data. In addition, machine learning models including Artificial Neural Network with explanatory variable (ANNX), Support Vector Machine with explanatory variable (SVMX), and Random Forest model with explanatory variable were compared with count time series models such as INGARCHX model.

Integer-valued generalized autoregressive conditional heteroscedastic (INGARCH) model

The class of generalized linear model (GLM) known as INGARCH models assumes that the conditional distribution of the dependent variable would follow well-known discrete distributions, such as the Poisson, and negative binomial distributions.

Let's imagine that the count time series is Yt: t ∈N and that the time- varying r- dimensional covariate vector is Xt: t ∈N, which is expressed as Xt = (Xt, 1, …,Xt. r)T. When the Ftrepresents historical data, the conditional mean is thus E(Yt|Ft-1) = λt. This is how the generalized model form is stated:

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \alpha_k \tilde{g}\left(Y_{t-i_k}\right) + \sum_{l=1}^{q} \beta_1 g\left(\lambda_{t-j_l}\right) + \eta^T$$

If g and$\tilde{g}$ are same, i.e., g(x) = (x) = x. Additionally, Yt follows the (Poisson) INGARCH (p, q) model with p > 1and q ≥ 0

If (a) Yt is Poisson distributed when conditioned on Yt-1, Yt-2,

(b) the conditional mean t = E[Yt| Yt-1, Yt-2,….] fulfils

$\lambda_t = \beta_0 + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \beta_j \lambda_{t-j}$with $\beta_0 > 0$ $and \alpha_{1,\ldots}$, $\alpha_p, \beta_1, \ldots, \beta_q \geq 0$

If we assume that the distribution of Yt| Yt-1 is Poisson, we then have an INGARCH model of order p and q also known as INGARCH (p, q) model. The INGARCH (p) model is the one that exists if q =0. The term Autoregressive Conditional Poisson (ACP) models is also used to describe these models.

Support Vector Regression (SVR)

The main concept of SVR is to convert the original input space into a high-dimensional variable space before creating the regression or time series model in the newly created high-dimensional feature space. A data set vector has the form Z = {Xi, Yi} i=1, N where Xi ∈Rn denotes the input vector, Yi is the scalar output, and N denotes the size of the data set. This is how the generic equation SVR is expressed as:

$$f(x) = WT\Phi(x) + b$$

Where W stands for weight vector, b for bias term, and superscript T for transportation. By minimizing the regularized risk function shown below, the coefficients W and b are calculated from data:

$$R(\theta) = \frac{1}{2}\|w\|^2 + C\left[\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon\left(y_i, f(x_i)\right)\right]$$

In order to prevent both underfitting and overfitting of the model, this regularized risk function concurrently minimizes the empirical error and regularized term. The first part in the in above equation $\frac{1}{2}\|w\|^2$ is referred to as the "regularized term", which assess flatness of the function. A function will be as flat as possible if $\frac{1}{2}\|w\|^2$ is minimized. The second term, denoted as $\frac{1}{N}\sum_{i=1}^{N} L_\varepsilon\left(y_i, f(x_i)\right)$ is known as the "empirical error" and it was calculated using the Vapnik-insensitive loss function as follows:

$$L_\varepsilon\left(y_i, f(x_i)\right) = f(x) = \begin{cases} |y_i, f(x_i) - \varepsilon|; & |y_i - f(x_i)| \geq \varepsilon, \\ 0 & |y_i - f(x_i)| < \varepsilon, \end{cases}$$

Where f(xi) is an estimate value and yi is the actual value. The radial basis function (RBF), which is denoted as follows, is the kernel function that is most frequently utilized.

$$k(x, xi) = \exp(-\gamma \| x - xi\|2)$$

Two hyper-parameters must be optimized in order for the RBF kernel function to work at its best: the regularization parameter C, which balances the model complexity and approximation accuracy, and the kernel bandwidth parameter, which represents the RBF kernel function's variance. Similar to the INGARCH model, exogeneous variables are also used in SVR and ANN for modelling and forecasting.

Artificial Neural Network (ANN)

ANN is a popular machine learning method that has gained significant usage in recent decades. In time series modeling, ANN is also known as the autoregressive neural network as it considers time lags as inputs. To model the time series framework using ANN, a neural network with an implicit functional representation of time is used. The final output Yt of a multi-layer feed forward autoregressive neural network can be represented by the general expression:

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

The model parameters also known as the synopsis weights, are α_j (j = 0, 1, 2, …, q) and β_{ij} (i= 0, 1, 2, …, p; j = 0, 1, 2, …, q); p is the number of input nodes, q is the number of hidden nodes and g is the activation function. The error function between real and anticipated values is minimized during the training phase of an ANN. The following is how the error function of an autoregressive ANN is expressed as:

$$E = \frac{1}{N}\sum_{t=1}^{N}(e_t)^2 = \frac{1}{N}\sum_{t=1}^{N}\left\{X_t - \left(w_0 + \left(\sum_{J=1}^{Q} w_J\, g\left(w_{oj} + \sum_{i=1}^{P} w_{ij}\, X_{t-i}\right)\right)\right)\right\}^2$$

Where N represents the overall count of error phrases. The neural network Wij parameters are altered by the amount of changes in ΔWij, where ΔWij = $-\eta\frac{\partial E}{\partial W_{ij}}$, where η is the learning rate (Rathod and Mishra, 2018; Zhang, 2003) [20, 23]. The exogeneous variable will be utilized to stimulate the pest count, just like in the INGARCHX and SVRX models, making the ANNX model.

Random Forest

The Random forest is a machine learning algorithm that combines predictions from multiple decision trees with different depths (Liu *et al.*, 2012) [12]. Each decision tree is trained on a bootstrapped dataset, and the algorithm uses a bootstrap sampling technique to randomly collect a certain number of samples for each tree. The algorithm creates multiple trees during training, and it grows them as much as possible without trimming. Because of its randomness, random forest is less prone to overfitting. Variable significance can be recorded in the model and determined from the permissible out-of-bag data (Liaw and Wiener, 2002). The final prediction is the mean of the outputs of all trees in the forest. The random forest model was developed using the Random Forest package in R (Liaw and Wiener, 2002).

Comparison criteria

RMSE and MAE are widely used accuracy measures in time series modeling to evaluate the performance of a forecasting model. The root mean square error (RMSE) is a commonly used metric that measures the difference between the predicted and actual values, expressed as the square root of the average of the squared differences. The equation for RMSE is,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

Where N is the number of observations, Yi is the actual value and $\hat{y}_i$ is the predicted value.

Another commonly used metric is the mean absolute error (MAE), which measures the average magnitude of the forecasting errors without considering their direction. The equation for MAE is:

$$MAE = \frac{\sum|y_i - x_i|}{n}$$

Diebold-Marino Test

In order to compare the statistical significance of the residuals from the various models, the Diebold-Mariano (DM) test is performed [14]. Consider two models residuals as r1, and r2, where di is the absolute difference between the residuals and

di = | r1|-| r2|. The autocovariance function γk is written as follows:

$$\gamma k = \frac{1}{n}\sum_{i=k+1}^{n}(d_i - \bar{d})(d_{j-k} - \bar{d})$$

According to the Diebold-Mariano test statistic

$$DM = \frac{\bar{d}}{\sqrt{[\gamma_0 + 2\sum_{k=1}^{h-1}\gamma_k]/n}}$$

Where, h = n1/3 + 1, the null hypothesis(H0), which that the prediction accuracy for two models is equal to zero, and the alternative hypothesis (H1), which states that the forecast accuracy for two models is different, are used in the experiment.

Results and Discussion

Summary statistics were computed to understand the nature of the data for tobacco caterpillars in soybean, pod borer caterpillar in red gram and chickpea, as well as the weather variables (Table 1). The incidence of spodoptera caterpillars in soybean ranged from 0 to 9, while in red gram and chickpea crops, pod borer incidence was observed between 0 to 3. During soybean growth, the daily average rainfall was 44.5mm, while in red gram and chickpea, 26.1 mm and 1.21mm of rainfall were recorded respectively. The summary statistics of weather variables indicated high heterogeneity.

Figure 1 revealed that the association between spodoptera caterpillars in soybean and weather variables such as Tmax, Tmin, and wind speed was weak and non-significant, while relative humidity (morning), relative humidity (evening), and rainfall parameters showed a significant effect, albeit with a weak relationship. In the case of pigeon pea pod borer, their association with weather elements, particularly Tmax and Tmin, showed positive significance, while relative humidity (morning), relative humidity (evening), rainfall, and wind speed were non-significant. Similarly, in the case of chickpea pod borer, the association with abiotic factors, mainly RHm, showed a significant relationship, while other parameters were non-significant. Due to the heterogeneity in weather data from 2006 to 2020, abiotic factors such as RHm and RHe were found to be significant.

Before fitting the count time series models, such as the INGARCHX model, it was necessary to check for autocorrelation in the data. The fitted Ljung-Box test statistic showed the presence of autocorrelation in all the series. However, most of the parameters for the INGARCHX model exhibited a non-significant connection with the pest population, which may reduce the accuracy of prediction. (Table 2).

Table 2 suggests that there was no significant difference in the association of weather parameters with different lepidopteran caterpillars in soybean, pigeon pea, and chickpea crops. This indicates that the model's performance was not satisfactory and it was not a good fit. The P values for all three lepidopteran pests incidences exceeded 0.05%, which further supports the lack of significant difference. The INGARCH model was used to fit the association between various crop caterpillars of different crops belonging to the leguminaceae family, but the climatic variables were found to be highly heterogeneous, which resulted in the failure of the INGARCH models to prove successful.

Table 3 provides results for the trained artificial neural network with explanatory variables (ANNX) and its application to the testing sample. The ANNX model was selected based on the lowest training error values for RMSE and MAE. The selected model configurations for chickpea pod borer, Red gram pod borer, and Soybean tobacco caterpillar were 15 input nodes and 10 hidden nodes, 10 input nodes and 2 hidden nodes, and 9 input nodes and 8 hidden nodes, respectively. The model used a feed-forward network with a sigmoidal function in the input to the hidden layer and an identity function in the hidden layer to output node. The nonlinear support vector regression model with explanatory variables is a powerful machine learning algorithm used for regression analysis. In this case, the Radial kernel function was used as it is suitable for non-linear problems and can handle complex data distributions. Finally, the Random Forest algorithm was utilized for the development of the random forest model.

Table 4 compares the performance of different models used for predicting the incidence of lepidopteran pests in three different crops belonging to the family Fabaceae (soybean, pigeon pea, and chickpea). The models used for comparison include Poison, ANNX, SVMX, and Random Forest. The ANNX model outperformed the other models in all three crops, as indicated by the lower values of MAE and RMSE. The P-values for the ANNX model were also found to be higher compared to other models, indicating a better fit.

The lepidopteran pest population data for each crop was analyzed using a training and testing set before being subjected to the BOX-PIERCE test. The results showed that the climatic parameters were highly heterogeneous, suggesting the need for a more sophisticated model for prediction.

In summary, the ANNX model was found to be superior in predicting the incidence of lepidopteran pests in different leguminous crops belonging to the family Fabaceae. The comparison criteria (MAE & RMSE) highlighted the observed differences between the predicted values of the models.

The study conducted Diebold-Mariano tests to compare the accuracy of different models used in predicting lepidopteran pest incidence in soybean, pigeon pea, and chickpea crops. The results showed that the ANNX model outperformed other models in all three crops. The ANNX model had a significant difference ($p<0.05$) compared to SVMX, Random Forest, and INGARCH models in most cases. However, the SVMX Vs Random Forest model did not show a significant difference ($p>0.05$) in soybean tobacco caterpillar incidence.

The INGARCH model was not considered a good fit for predicting lepidopteran pest incidence in these crops because of the heterogeneous nature of the weather data over time. ANNX was found to be the most effective model for predicting lepidopteran pest incidence due to its ability to capture the complex and nonlinear nature of the data.

Table 1: Summary statistics for incidence of chickpeapod borer, Red gram pod borerand Soybean tobacco caterpillarwith climatological variables

	Tmax	Tmin	RH1	RH2	WS	Rainfall	JG-11
Chickpea Pod borer							
Mean	29.57	14.46	72.96	42.83	4.41	1.21	1.18
Standard Error	2.19	2.43	11.87	13.98	2.88	6.06	1.01
Kurtosis	0.47	-0.51	-0.50	-0.85	0.42	49.02	-1.39
Skewness	0.41	-0.08	-0.32	0.32	1.15	6.68	0.40
Range	12.70	10.80	58.00	60.00	11.50	53.80	2.77
Minimum	24.00	9.30	35.00	16.00	0.60	0.00	0.00
Maximum	36.70	20.10	93.00	76.00	12.10	53.80	2.77
Red gram Pod Borer							
Mean	30.03	18.90	80.58	58.31	5.63	26.10	1.16
Standard Error	1.46	2.35	9.35	14.38	3.90	57.88	0.85
Kurtosis	-0.44	1.07	-0.60	-0.53	-0.34	54.32	-1.29
Skewness	-0.03	-1.10	-0.41	-0.32	0.86	6.18	0.24
Range	7.00	12.10	41.00	66.00	15.90	590.00	2.56
Minimum	26.10	10.60	57.00	22.00	0.70	0.000	0.00
Maximum	33.10	22.70	98.00	88.00	16.60	590.00	2.56
Soybean Tobacco Caterpillar							
Mean	29.51	20.74	86.72	67.49	8.04	44.51	1.33
Standard Error	1.61	0.95	8.67	10.13	5.25	61.90	1.44
Kurtosis	-0.15	2.12	1.62	0.89	1.13	35.69	25.92
Skewness	-0.30	-1.08	-1.46	-0.27	1.04	4.62	5.11
Range	7.30	6.20	38.00	65.00	25.60	590.00	8.59
Minimum	25.40	16.50	60.00	30.00	0.80	0.00	0.71
Maximum	32.70	22.70	98.00	95.00	26.40	590.00	9.30

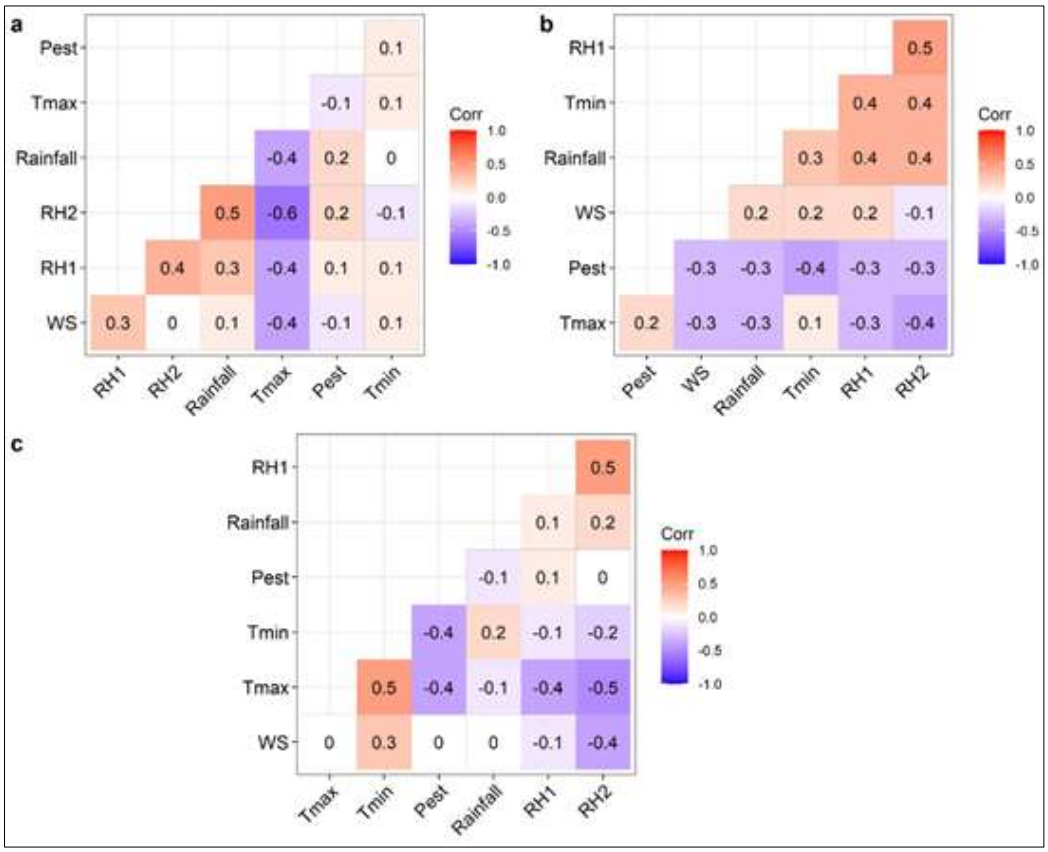

Fig 1: Pearson correlation coefficients between Incidence of Soybean Tobacco Caterpillar (a) Red gram Pod Borer (b) and Pod borer in chickpea (Figure1.c) with climatologicalvariables

Table 2: Parameter estimation of the INGARCHX model for Incidence of lepidopteran species in leguminous crops

	Estimate	Std. Error	z value	Pr(>|z|)
Chickpea pod borer				
Intercept	8.8216e-07	1.5227e00	0.0000	1.000
beta_1	6.6948e-01	8.7542e02	7.6476	2.048e-14
eta_1	2.1952e-12	4.3455e00	0.0000	1.000
eta_2	1.0080e-09	3.2976e02	0.0000	1.000
eta_3	3.6136e-03	6.4952e03	0.5563	0.578
eta_4	2.5593e-09	7.1027e03	0.0000	1.000
eta_5	1.6200e-07	1.9481e02	0.0000	1.000
eta_6	7.6808e-11	9.0529e03	0.0000	1.000
Red gram pod borer				
Intercept	5.3754e-01	2.3028e00	0.2334	0.8154
beta_1	2.0604e01	1.2905e01	1.5966	0.1103
eta_1	6.8589e03	7.1724e02	0.0956	0.9238
eta_2	3.7324e04	6.1807e02	0.0060	0.9952
eta_3	2.8741e06	1.0461e02	0.0003	0.9998
eta_4	2.3693e07	9.4940e03	0.0000	1.0000
eta_5	5.3927e04	2.2113e02	0.0244	0.9805
eta_6	3.8965e	1.4598e03	0.0000	1.0000
Intercept	5.3754e-01	2.3028e00	0.2334	0.8154
Soybean tobacco caterpillar				
Intercept	8.7201e-07	2.2238e00	0.0000	1.0000
beta_1	7.0849e-01	1.0786e01	6.5688	5.071e-11
eta_1	8.2318e-12	5.2669e02	0.0000	1.0000
eta_2	5.7351e-09	6.1973e02	0.0000	1.0000
eta_3	5.3904e-07	7.3661e03	0.0001	0.9999
eta_4	3.0883e-03	7.6916e03	0.4015	0.6880
eta_5	1.9162e-03	1.2062e02	0.1589	0.8738
eta_6	2.0290e-04	1.6725e03	0.1213	0.9034
Intercept	8.7201e-07	2.2238e00	0.0000	1.0000

Table 3: Parameter specifications of SVRX and ANNX models considered for model development in Incidence of Lepidoptera species on leguminous crops

ANNX			
	Chickpea pod borer	**Red gram pod borer**	**Soybean tobacco caterpillar**
Input lag	15	10	9
Dependent variable	1	1	1
Hidden layer	1	1	1
Hidden nodes	10	2	8
Exogenous variables	6	6	6
	21-10-1	16-2-1	15-8-1
Number of parameters	231	181	137
Network type	Feed Forward	Feed Forward	Feed Forward
Activation function I: H	Sigmoidal	Sigmoidal	Sigmoidal
Activation function H: O	Identity	Identity	Identity
SVMX			
Kernel function	Radial	Radial	Radial
No. of Support Vectors	170	140	111
Cost	1	1	1
Gamma	0.17	0.17	1.16
Epsilon	0.1	0.1	0.1
Random Forest			
Number of trees	500		
No. of variables tried at each split	2		
Mtree	1		
Ntree	1		

Table 4: Comparison of prediction models with consideration of training and testing data sets

Model	Accuracy measures	Training Set	Testing Set	Box-Pierce test
Chickpea pod borer				
Poison	RMSE	0.206	1.06	χ^2 =48.132 p-value = 3.98e-12
	MAE	0.175	1.15	
ANNX	RMSE	0.002	0.01	χ^2 =0.581 p-value=0.445
	MAE	0.001	0.01	
SVMX	RMSE	0.29	0.37	χ^2 =52.25 p-value=4.89e-10
	MAE	0.085	0.31	
Random Forest	RMSE	0.30	0.29	χ^2 =16.41 p-value=45.1e-10
	MAE	0.28	0.33	
Red gram pod borer				
Poison	RMSE	0.210	1.209	χ^2 =14.57 p-value=0.0001
	MAE	0.170	1.019	
ANNX	RMSE	0.050	0.043	χ^2 =0.23 p-value=0.654
	MAE	0.038	0.036	
SVMX	RMSE	0.227	0.334	χ^2 =34.396, p-value=4.497e-09
	MAE	0.177	0.261	
Random Forest	RMSE	0.290	0.290	χ^2 =11.415, p-value=0.0007
	MAE	0.245	0.259	
Soybean tobacco caterpillar				
Poison	RMSE	0.223	1.081	χ^2 =5.26 p-value=0.02
	MAE	0.191	1.040	
ANNX	RMSE	0.0034	0.053	χ^2 =0.56 p-value=0.45
	MAE	0.002	0.034	
SVMX	RMSE	0.258	0.352	χ^2 =27.301, p-value=1.741e4
	MAE	0.192	0.295	
Random Forest	RMSE	0.318	2.42	χ^2 =49.33, p-value=0.41
	MAE	0.278	1.23	

Table 5: Diebold–Mariano test for comparison of performance of different models

	Test statistic	p-value	Test statistic	p-value	Test statistic	p-value
	Soyabean		Red gram		Chickpea	
ANNXvs Random Forest	-5.14	5.85e-05	1.22	0.231	0.21	0.84
ANNX vsSVMX	-3.40	0.001	1.567	0.119	-0.185	0.854
ANNXvsINGARCH	-7.781	2.52e-07	3.10	0.002	8.14	9.521e-09
SVMXvs Random Forest	1.32	0.204	-0.21	0.84	1.78	0.08
SVMXvsINGARCH	-6.756	1.87e-06	1.45	0.149	21.76	<2.2e-16
Random forest vsINGARCH	-7.30	6.366e-07	1.82	0.070	26.11	<2.2e-16

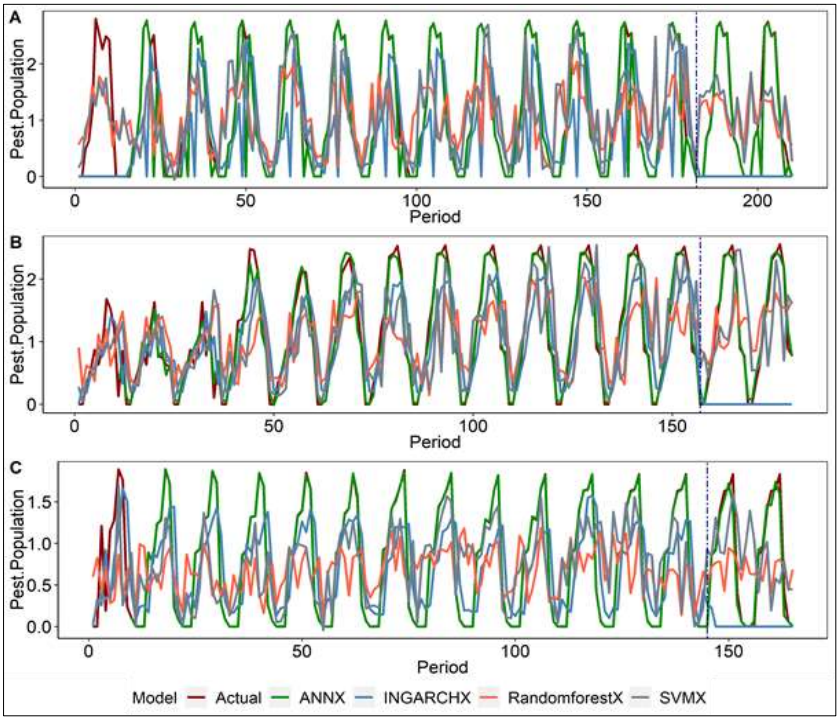

Fig 2: Actual vs. fitted plot for Incidence of (a) Chickpea Pod borer; (b) Red gram Pod Borer; Fig 2.c Soybean Tobacco Caterpillar)

Conclusion

- The study focused on developing efficient forecasting models for lepidopteran pests in soybean, pigeon pea, and chickpea crops.
- Both machine learning and count time series techniques were used based on climatological input variables.
- Machine learning models outperformed count time series models in predicting lepidopteran pest occurrences due to the highly nonlinear and heterogeneous nature of the data.
- Among the various machine learning models, the ANNX model was found to be the most effective in modeling and predicting the incidence of all lepidopteran pests based on time series data.
- The study highlighted that the use of machine learning techniques, such as ANN with exogenous variables, can increase the prediction accuracy of count time series.
- The Diebold Mariano test statistics revealed the superiority of ANNX models over INGARCH, SVMX, and Random Forest models.
- It is anticipated that machine learning techniques will be increasingly employed in modeling count time series of pests in other crops in the future.

References

1. Alam W, Ray M, Kumar RR, Sinha K, Rathod S, Singh KN. Improved ARIMAX modal based on ANN and SVM approaches for forecasting rice yield using weather variables. Indian J Agric. Sci. 2018;88(12):1909-1913.
2. Alam W, Sinha K, Kumar RR, Ray M, Rathod S, Singh KN, *et al*. Hybrid linear time series approach for long term forecasting of crop yield. Indian J Agric. Sci. 2019;88:1275-1279.
3. Anonymous. E-pulses Data Book, Area, Production & Yield of different pulses in India. 2022. https//IIPR.ICAR.gov.in.
4. Anonymous; c2020. https://www.agrifarming.in/district-wise-crop-production-in-karnataka-list-of-crops-grown-in-karnataka.
5. Anonymous. Directorate of Economics and Statistics, Department of Agri. Coop & Far. Welfare, Ministry of Agriculture, Govt of India; c2019.
6. Anonymous. Fully revised estimates of area, production & yield of principal crops in Karnataka. Directorate of Economics, Bangalore; c2018.
7. Anonymous. NABARD.PLP Data/Information from State Agricultural Department (Area, Production and Productivity of Bidar); c2022.
8. Arya Paul RK, Kumar A, Singh KN, Sivaramne N, Chaudhary P. Predicting pest population using weather variables an ARIMAX time series framework. bug, *Clavigralla gibbosa* Spinola (Hemiptera: Coreidae) on long duration pigeonpea. J Entomol. Zool. Stud. 2015;5(4):433-437.
9. Chitikela G, Admala M, Ramalingareddy VK, Bandumula N, Ondrasek G, Sundaram RM, *et al*. Artificial-intelligence- based time-series intervention models to assess the impact of the COVID-19 pandemic on tomato supply and prices in Hyderabad, India. Agronomy. 2021;11:1878.
10. Dhingra S, Kodandaram RS, Hegde S, Srivastava C. Evaluation of different insecticide mixture against third instar larvae of Helicoverpa armigera. Ann. Plant Protection Sci. 2003;11:274-276.
11. Diebold FX, Mariano RS. Comparing predictive accuracy. J. Bus. Econ. Stat. 1995;13:253-263.
12. Gorlapalli A, Kallakuri S, Sreekanth PD, Patil R, Bandumula N, Ondrasek G, *et al*. Characterization and prediction of water stress using time series and artificial intelligence models. Sustainability. 2022;14(11):6690.
13. Huang T, Yamg R, Huang W, Huang Y, Qiao X. Detecting sugarcane borer diseases using support vector machine. Inf. Process. Agric. 2018;5:74-82.
14. Kim YH, Yoo SJ, Gu YH, Lim JH, Han D, Baik SW. Crop pests prediction method using regression and machine learning technology: survey. IERI Procedia. 2014;6:52-56.
15. Liu Y, Wang Y, Zhang J. New machine learning algorithm: Random forest. In International Conference on Information Computing and Applications. 2012, 246-252.
16. Paul RK, Vennila S, Bhat MN, Yadav SK, Sharma VK, Nisar S, *et al*. Prediction of early blight severity in tomato (*Solanum lycopersicum*) by machine learning technique. Indian J Agric. Sci. 2019;89:1921-1927.
17. Prasad YG, Gayathri M, Prabhakar M, Jeyakumar P, Vennila S, Subba Rao AVM, *et al*. Population dynamics of Spodopteralitura outbreak on soybean vis-à-vis rainfall events. J Agromet. 2013;15(1):37-40.
18. Rathod S, Paramesha V. Time Series Analysis using Machine Learning Techniques. In Vadivel, A., Paramesh, V., Uthappa, A. and Kumar, R.P. (Ed.). Ecosystem service analysis: concepts and applications in diversifed coconut and arecanut gardens. ICAR- Central Coastal Agricultural Research Institute, India. 2022. p. 262-292.
19. Rathod S, Yerram S, Arya P, Katti G, Rani J, Padmakumari AP, *et al*. Climate-Based Modeling and Prediction of Rice Gall Midge Populations Using Count Time Series and Machine Learning Approaches. Agronomy. 2021;12(1):22.
20. Rathod S, Singh KN, Patil SG, Naik RH, Ray M, Meena VS. Modeling and forecasting of oilseed production of India through artificial intelligence techniques. Indian J. Agric. Sci. 2018;88:22-27.
21. Sasvihalli BP, Naik CM, Nataraj K. Efficacy of Bioagents, Botanicals and Insecticides in Suppression of Spodopteralitura on Vegetable Soybean. Soybean Res. 2017;15(1):40-45.
22. Sujithra M, Chander S. Seasonal incidence and damage of major insect pests of pigeonpea, *Cajanus cajan* (L.). Indian J Entomol. 2014;76:202-206.
23. Zhang GP. Time-Series Forecasting using a Hybrid ARIMA and Neural Network Model. Neurocomputing. 2003;50:159-175.