helpful to comprehend the seismic behaviour of the basin during an earthquake. Since the Dun Valley has been interpreted as a synclinal basin which is characterized by the Santaurgarh anticline to the NE and Mohand anticline to the SW[7], the present study was aimed at mapping the thickness of sediments/depth to bedrock in terms of V_s after validating the lithosections with V_s exposed along river-cuttings for the Dehradun fan, which is a part of the Dun Valley. In this study, we have considered the V_s with depth of 10 representative sites along a NE–SW section across the Dehradun fan to map the thickness of sediments/depth to bedrock. The sediments of the Dehradun fan have been described as Dun gravels due to the preponderance of gravels as a compositional material. Further, the Dun gravels have been divided into three lithological subunits, i.e. units A–C. These subunits have been assigned by V_s bands after validation with the exposed lithosections. Unit A shows a V_s band of 700–850 m/s, whereas units B and C represent 500–700 and < 500 m/s respectively. The V_s bands for the Upper and Middle Siwaliks have been derived after running the direct V_s profile at the top of the exposed section at Donga and Majhaun sites, where the Middle Siwalik sandstone and Upper Siwalik conglomerate were exposed along the river sections respectively. The Middle Siwalik sandstone represents the V_s band of > 750 m/s whereas the Upper Siwalik boulder conglomerate is characterized by V_s band of > 950 m/s. The corresponding V_s of the Dehradun fan will help earthquake engineers to characterize the seismic behaviour of the waves from the actual bedrock to the surface instead of doing the analysis on the basis of mean V_s of upper 30 m soil column.

1. Thakur, V. C. and Pandey, A. K., Late quaternary tectonics evolution of Dun in fault bend/propagated fold system, Garhwal sub-Himalaya. *Curr. Sci.*, 2004, **87**(11), 1567–1576.
2. Valdiya, K. S., The main boundary thrust zone of the Himalaya, India. *Ann. Tecton.*, 1992, **6**, 54–84.
3. Karunakaran, C. and Rao, A., Status of exploration of hydrocarbons in the Himalayan region – contributions to stratigraphy and structure. *Misc. Publ. Geol. Surv. India*, 1979, **41**, 1–66.
4. Nossin, J. J., Outline of the geomorphology of the Dun valley, Northern UP, India. *Z. Geomorphol.*, 1971, **12**, 18–50.
5. Nakata, T., Geomorphic history and crustal movements of the foot hills of the Himalayas. Report of Tohoku University, Japan, 7th Series (Geography), 1972, vol. 22, pp. 39–177.
6. Singh, A. K., Prakash, B., Mohindra, R., Thomas, J. V. and Singvi, A. K., Quaternary alluvial fan sedimentation in the Dehradun valley piggyback basin, NW Himalaya: tectonic and palaeoclimatic implications. *Basin Res.*, 2001, **13**, 449–471.
7. Thakur, V. C., Geology of Dun Valley, Garhwal Himalaya: Neotectonics and coeval deposition with fault-propagation folds. *J. Himalayan Geol.*, 1995, **6**(2), 1–8.
8. Raiverman, V., Srivastava, A. K. and Prasad, D. N., Structural style in the northwestern Himalayan foothills. *Himalayan Geol.*, 1994, **15**, 263–283.
9. Power, P. M., Lillie, R. J. and Yeats, R. S., Structural and shortening of the Knagra and Dehradun re-entrants, sub-Himalaya, India. *Geol. Soc. Am. Bull.*, 1998, **110**, 1010–1027.
10. Nazarian, S., Stokoe, K. H. and Hudson, W. R., Use of spectral analysis of surface wave method for determination of moduli and thickness of pavement. *Transp. Res. Rec.*, 1983, **930**, 38–45.
11. Park, C. B., Miller, R. D. and Xia, J., Multichannel analysis of surface waves. *Geophysics*, 1999, **64**, 800–808.
12. Xia, J., Miller, R. D. and Park, C. B., Estimation of near surface shear wave velocity by inversion of wave. *Geophysics*, 1999, **64**, 691–700.
13. Mahajan, A. K., Slob, S., Ranjan, R., Sporry, R., Champati Ray, P. K. and van Westen, C. J., Seismic microzonation of Dehradun city using geophysical and geotechnical characteristics in the upper 30-meters of soil column. *J. Seismol.*, 2007, **11**, 335–370.
14. Mahajan, A. K., NEHRP soil classification and estimation of 1-D site effect of Dehradun fan deposits using shear wave velocity. *Eng. Geol.*, 2008, **104**, 232–240.
15. Park, C. B., Miller, R. D. and Xia, J., Imaging dispersion curves of surface waves on multichannel records. Technical programme with biographies. In Society of Exploration Geophysics, 68th Annual Meeting, New Orleans, Louisiana, 1998.
16. Kumar, S. and Mahajan, A. K., The Uttarkashi earthquake of 20 October 1991: field observations. *Terra Nova*, 1993, **6**, 95–99.

ACKNOWLEDGEMENTS. We thank the Director, Wadia Institute of Himalayan Geology, Dehradun, for permission to carry out this work under the institute research project on shallow subsurface studies.

Received 3 July 2010; revised accepted 8 November 2010

A decision tree analysis for predicting the occurrence of the pest, *Helicoverpa armigera* and its natural enemies on cotton based on economic threshold level

M. Pratheepa[1,*], K. Meena[2], K. R. Subramaniam[3], R. Venugopalan[4] and H. Bheemanna[5]

[1]National Bureau of Agriculturally Important Insects, Bangalore 560 024, India
[2]Bharadhidasan University, Tiruchirappalli 620 024, India
[3]Department of MCA, Shrimathi Indira Gandhi College, Tiruchirappalli 620 002, India
[4]Indian Institute of Horticultural Research, Hessaraghatta, Bangalore 560 089, India
[5]University of Agricultural Sciences, Agricultural Research Station, Raichur 584 102, India

The cotton bollworm, *Helicoverpa armigera* (Hübner) is one of the most important pests affecting crop production globally. The data-mining technique, for predicting pest incidence using biotic and abiotic factors has not been developed so far. To identify the biotic

*For correspondence. (e-mail: mpratheepa@rediffmail.com)

and abiotic factors that play a role in the occurrence of the pest, the decision tree analysis in conjunction with Shannon information measure was explored. The developed classification model has the ability to successfull treat 'categorical' variables as well as 'continuous' variables in the database. The information-theoretic classification method used in the present study was aimed at finding a minimal set of database attributes involved in the induced model and was successfull in predicting pest incidence. It was found that there was 8.82% misclassified testing data. The confusion matrix for the testing set revealed that the classification was done more accurately using the training set. The developed prediction or classification model will be helpful in forewarning about pest incidence and also to identify the factors influencing the pest population density. Using this model, agricultural farmers can apply pest control strategies on time to reduce crop loss.

Keywords: *Chrysoperla carnea*, cotton, decision tree, economic threshold level, *Helicoverpa armigera*.

HELICOVERPA ARMIGERA (Hübner), commonly known as cotton bollworm, is a major pest damaging a wide range of crops. In 2001, cotton crop was heavily damaged in India by *H. armigera*, reducing the cotton seed yield to an unprecedented low level[1]. Globally, the losses due to *Helicoverpa* exceed US$ 2 billion and the cost of pesticides applied for controlling this pest exceed US$ 500 million annually[2]. Biological control is an important component in *H. armigera* management, especially in subsistence agriculture, given the known harmful effects of chemical pesticides[3]. The priority of research requires linking of pest incidence with weather parameters and to understand the contribution of natural enemies in pest incidence[1]. Data mining is a technique for extracting or mining of knowledge from a large database. Knowledge discovery (KD) is a nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in any data[4]. The decision tree analysis in the data-mining technique is a popular predictive model and is largely useful in classification applications, because it resembles human reasoning and can be easily understood[5]. The decision can be represented in terms of a set of rules[6].

The machine learning method relational decision tree model has been developed for the classification/prediction of pollen and seed dispersal of genetically modified (GM) and non-GM crops. The decision tree is a powerful tool for learning about the co-existence rules for GM and non-GM crops[7]. Knowledge-based systems were developed for solving problems concerned with pest identification, treatment prescription and strategic planning[8]. Research has been carried on the development of user-friendly, computerized, expert systems for the management of pests and diseases in the Jamaican coffee industry and for crop protection in India and abroad considering factors like climate, topography, soil type of the farm, agronomic practices, crop phenology, biology and damage due to potential pests, etc.[9,10]. However, there exists a gap in identifying the biotic and abiotic factors that play a role in the incidence of *H. armigera* on different crops. According to Singh[11], the main problem in addressing the issue of pest management is inadequate knowledge about the factors influencing the pest dynamics. The extent of damage may vary from crop to crop and season to season. Hence, the decision tree analysis for predicting *H. armigera* incidence on cotton crop, by analysing the abiotic and biotic factors has been proposed. The decision tree analysis in conjunction with Shannon information measure for the classification of pest incidence as high or low based upon economic threshold level (ETL) was explored for the occurrence of cotton bollworm, *H. armigera*.

The dataset was obtained from the Regional Agricultural Research Station, Raichur, Karnataka from the unsprayed experimental plots under All-India Coordinated Research Project (AICRP) on non-*Bt*-cotton. The sampling size was 25 plants/500 m^2 area. Weekly observations on mean number of *H. armigera* larvae present per 5 plants were recorded for the period 2005–2008. Natural enemies – spiders (NE1) and *Chrysoperla carnea* (NE2) per plant were recorded during this period. Weather parameters like maximum temperature (MaxT), minimum temperature (MinT), relative humidity (RH) and rainfall (RF) were taken based on weekly mean value. Pest incidence related with previous week biotic and abiotic factors was taken for decision tree analysis. The sample data are given in Table 1 and explanation of the attributes in the database, is given in Table 2.

Pest incidence (PI) was the dependent variable (also referred to as class or target variable) and it was predicted based on several independent variables (also referred to as features or attributes).

The overall dataset was divided into two parts, two-thirds of the records were chosen for the training phase (from July 2005 to January 2007) and one-third for the testing phase (from August 2007 to January 2008). A decision tree was built by initially selecting the instances from a training set. The intrinsic nature of the training set was then used by the algorithm to construct a decision tree for the testing set.

Pest incidence was grouped into two classes, namely high and low based upon ETL. According to Dhaliwal and Arora[12], pest incidence was considered as high if PI $\geq$ 1 larva/10 plants, and as low if PI < 1 larva/10 plants. Class values were assigned to the database based upon the user threshold input.

The discretization technique, equal binning method was used to convert all numerical attributes (continuous variables) like NE1, NE2, MaxT, MinT, RF and RH into categorical values/labels (A1–A5) as the option of totally five bins was given. The labels of A1–A5 for each numerical attribute have their own range values stored in

Table 1. Sample records/tuples from the database

Tuple	Season	Independent variable/feature/attribute NE1	NE2	MaxT	MinT	RF	RH	Class/target PI (dependent)
1	Monsoon	0.00	0.00	31.4	22.6	62.4	74	Low
2	Monsoon	0.00	0.00	30.7	22.7	47.6	82	Low
3	Post-monsoon	0.00	0.00	31.5	22.3	6.0	73	High
4	Post-monsoon	0.00	0.00	32.5	23.1	59.2	74	High
5	Monsoon	0.00	0.00	31.2	22.9	101.0	79	Low
6	Monsoon	0.22	0.00	31.3	22.3	0.0	72	Low

Table 2. Details of attributes

Attribute	Explanation
PI	Pest incidence per plant
Season	Observation period
NE1	Natural enemy 1 (no. of spiders per plant)
NE2	Natural enemy 2 (no. of *Chrysoperla carnea* per plant)
MaxT	Maximum temperature (°C)
MinT	Minimum temperature (°C)
RH	Relative humidity (%)
RF	Rainfall (mm)

the database. The mathematical representation of the binning method is as follows: (i) Get user input as the number of bins required (b); (ii) The number of bins $nb = n/b$, where n denotes the total number of records. (iii) Sort the numerical attribute values in ascending order. (iv) For each numerical attribute, assign the class label A1 into 1 to nb records, and A2 into $nb + 1$ to $nb + nb$, and A3 into $2nb$ to $2nb + nb$, and so on, until the counter number of bins becomes zero. (v) The range values of A1 to An was stored separately in the database for each numerical attribute. (vi) If extra records are available after creation of the bins, the last label, An, was assigned.

Given a database D consisting of t data samples, where $D = \{t_1, t_2, \ldots, t_n\}$ and a set of classes $C = \{c_1, c_2, \ldots, c_m\}$, the classification problem is to define a mapping $f : D \rightarrow C$, where each t_i is assigned to one class. D contains t_i tuples of class C_j for $j = 1$ to m, where m represents the number of classes and $m = 2$. The expected information needed to classify a tuple/record in D is given by:

$$\text{Info}(D) = \sum_{j=1}^{m} p_j \log_2(p_j),$$

where p_j is the probability that an arbitrary tuple in D belongs to class C_j and is estimated using $n(C_j, D)/n(D)$. Info(D) is also known as the entropy of D, i.e. total information value.

Entropy of attribute (feature) A with values $\{a_1, a_2, \ldots, a_v\}$ is used to split D into v subsets:

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{n(D_j)}{n(D)} \times \text{Info}(D_j).$$

$\text{Info}_A(D)$ is the expected information required to classify a tuple from D based on the partitioning by A.

Information is gained by branching an attribute A, and

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D).$$

Information gain is a measure of how good an attribute is for predicting the class of each of the training data. The attribute with the highest information gain was selected as the next split attribute according to the standard procedure[13].

The attribute selection measure based on Shannon information theory was used in the decision tree analysis. According to this, the maximum gain value attribute was chosen as the splitting attribute and based on the attribute the subsets were evaluated. Information measure was used recursively for each subset until the gain value or the entropy reached zero for the attribute and this was used as a stopping criterion. The process continued until the search was completed and the attributes and labels stored in the table for tree generation. The model used the filter method as it selecting the features/attributes before applying an induction algorithm. The information theoretic method used in this model implemented automatic feature selection 'on the fly' as a part of the learning process. Thus, a minimal subset of features was found in a single run of the induction algorithm.

The root of the decision tree was fixed as PI and it was at level 0. The maximum gain value of the attribute was 'season' (Table 3). Hence, tree generation started from the season at level-1 and subsets had been generated automatically based on the information theory. The decision tree is a non-backtracking algorithm and hence it was constructed in a top-down manner[13].

The data-mining process for construction of the decision tree is given step by step in Figure 1. The training and testing process is given in Figure 2.

The data obtained were subjected to regression analysis[14] for developing the prediction equations for pest occurrence in cotton crop based on biotic and abiotic factors using the Statistical Package for Social Sciences[15], ver. 17. The variable 'season' has been assigned numerical values 1–3 for monsoon, post-monsoon and winter seasons respectively.

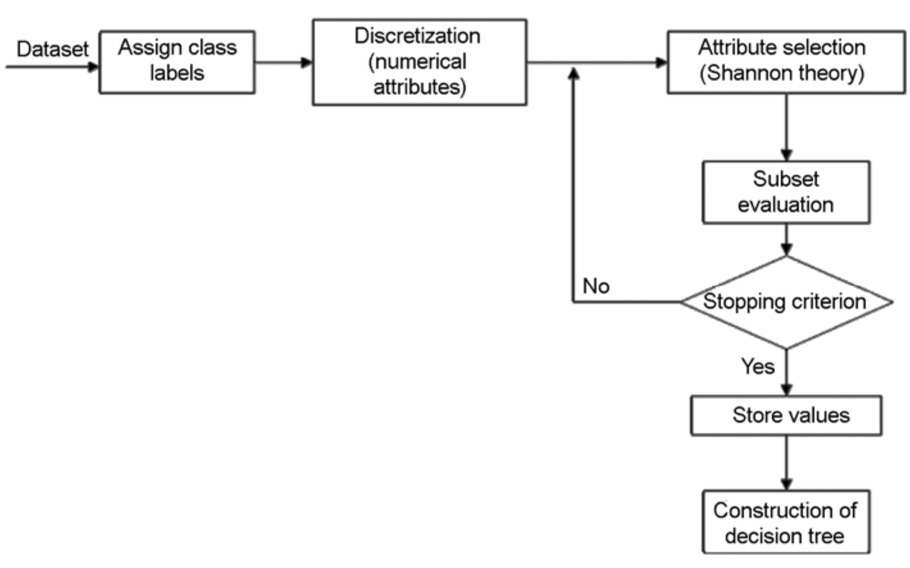

Figure 1. Data-mining process of construction of decision tree.

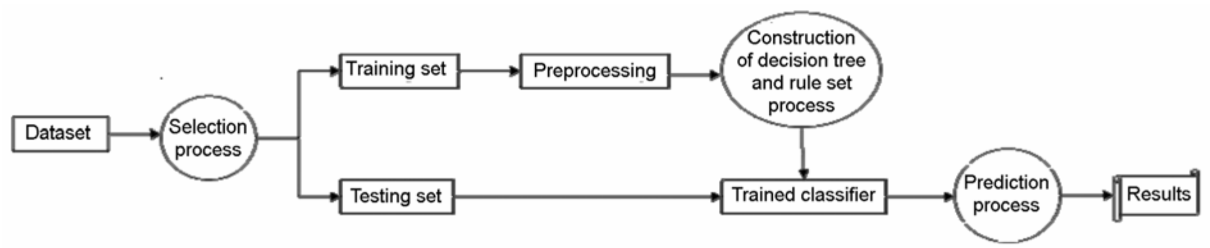

Figure 2. Training and testing process.

Table 3. Information gain values at level-1

Variable/feature/attribute	Gain value
Season	0.226
NE2	0.188
NE1	0.140
MaxT	0.120
MinT	0.104
RF	0.072
RH	0.055

The entropy or information gain value at level-1 is given in the Table 3.

The constructed binary tree had decision node as the condition and output/result of that condition derived as the 'yes' or 'no' options. The 'yes' option always grows as left child and 'no' option as right child in the tree. The node ended when the condition was not able to proceed further in the 'no' option as right child. The end of the leaf node denoted the class label of pest incidence. The classes were always defined in the left child of the tree i.e. 'yes' option of the decision node. Root had been fixed as PI and the categorical variable season had been taken as the first attribute/variable starting from level-1 based on the highest gain value. The next attribute was selected based on the highest information gain value while in the process of feature subset selection as level-2, and so on until the information gain value became zero. The decision tree diagrams for *H. armigera* incidence during different seasons related with weather factors and natural enemies are given in Figures 3–5.

IF–THEN rules derived from Figures 3 to 5 are given in Tables 4–6 for monsoon, post-monsoon and winter seasons respectively.

During monsoon season, when the maximum temperature ranged from 28.30°C to 30.80°C and minimum temperature ranged from 21.24°C to 22.59°C, pest incidence was high (Table 4). But, when maximum temperature ranged from 28.30°C to 30.80°C and NE1(spiders) was in the range 2.12–4.10, then pest incidence was low. Similarly, when the maximum temperature ranged from 30.81°C to 31.39°C, pest incidence was high when there was no rainfall (Table 4). But, when the maximum temperature ranged from 30.81°C to 31.39°C and rainfall ranged from 3.10 to 156.20 mm, the pest incidence was low (Table 4). Similarly, when the maximum temperature > 33°C with rainfall 0 mm, pest incidence was low, but when the maximum temperature > 33°C with rainfall in the range 18.40–156.20 mm, pest incidence was high (Table 4).

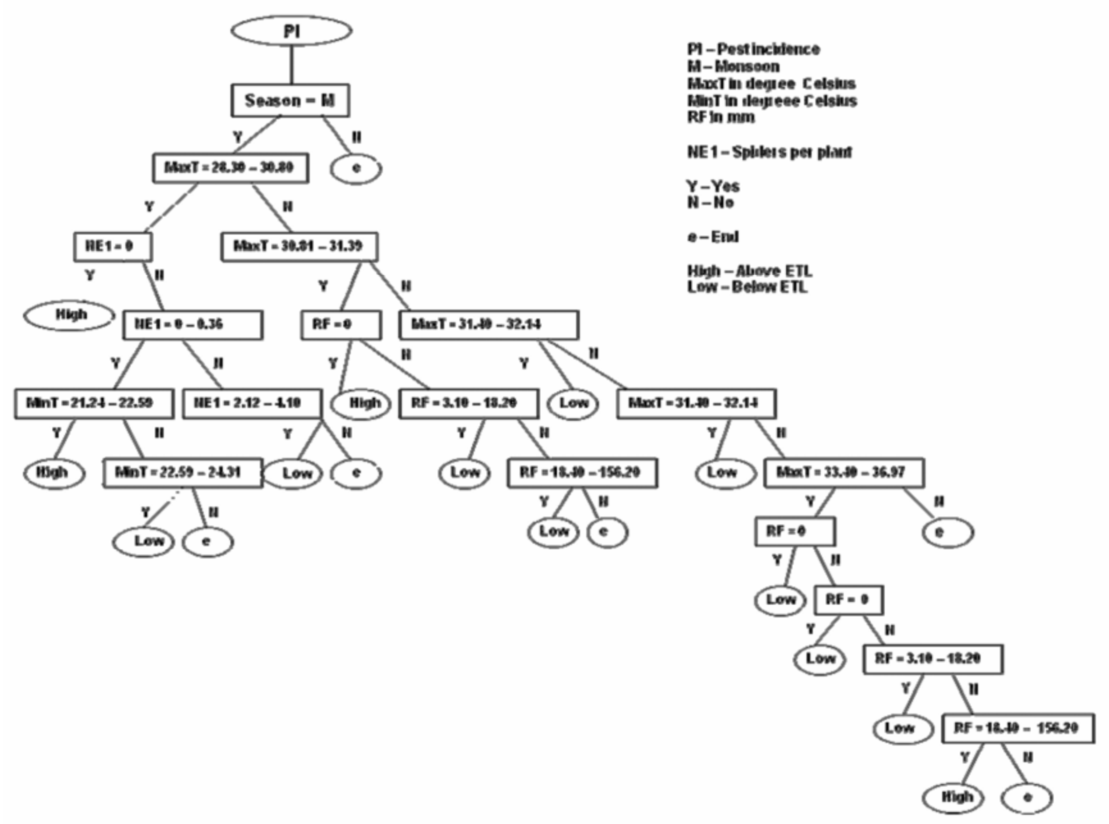

Figure 3. Decision tree with numerical attributes categorized into five bins for monsoon season.

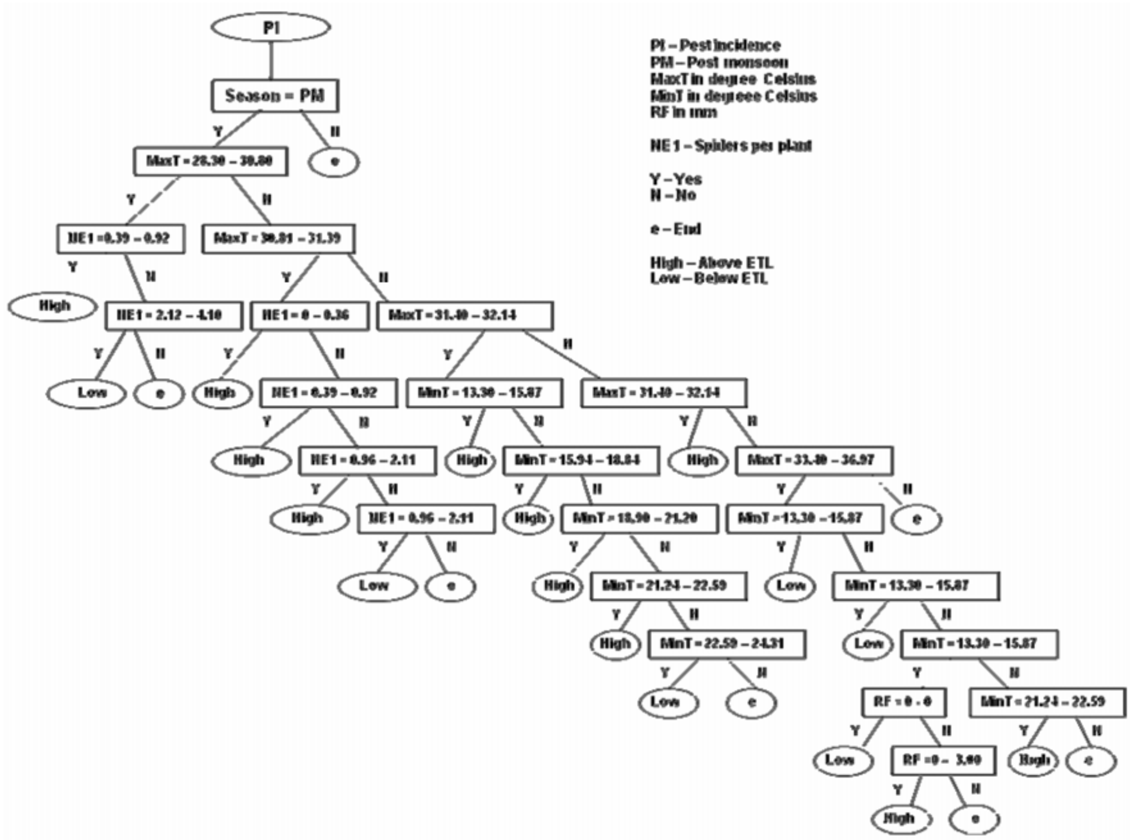

Figure 4. Decision tree with numerical attributes categorized into five bins for monsoon season.

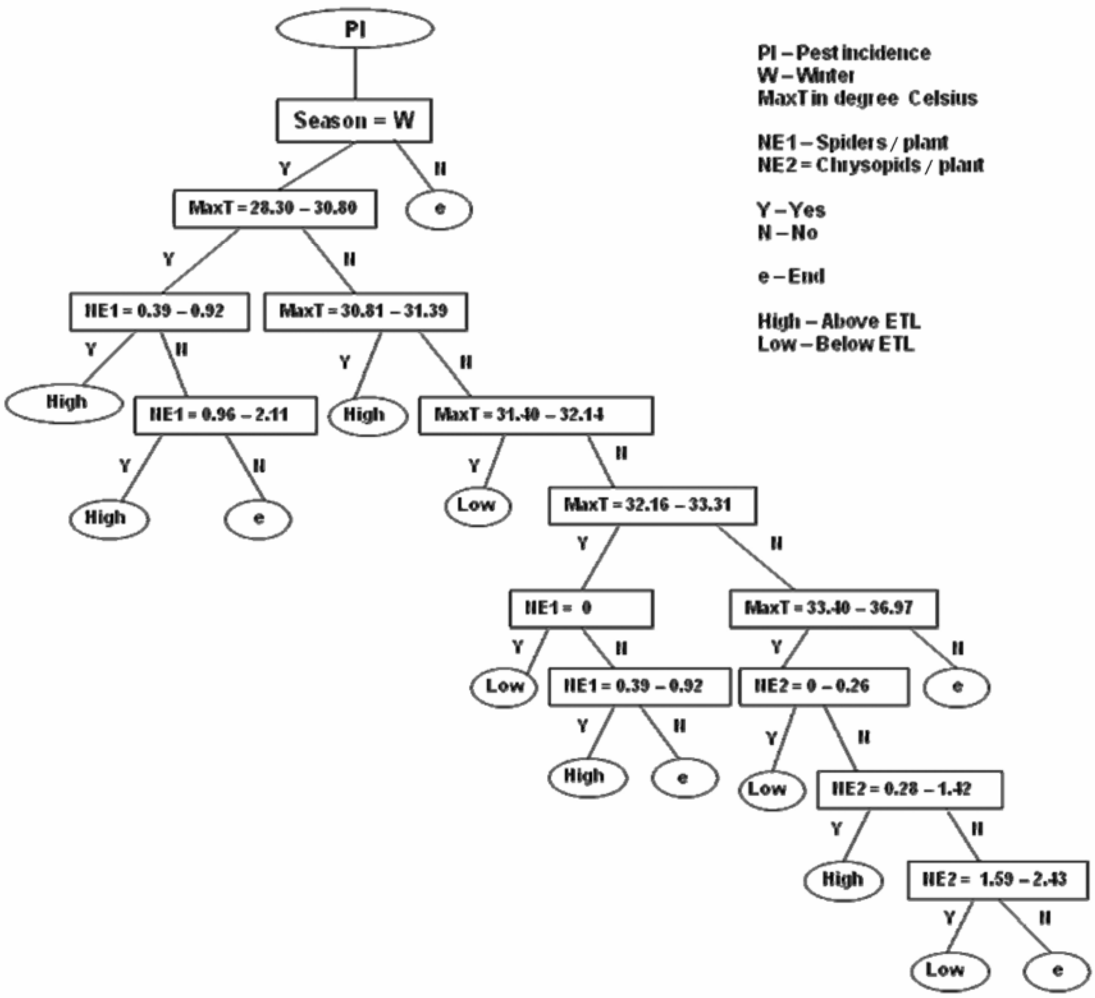

Figure 5. Decision tree with numerical attributes categorized into five bins for winter season.

Table 4. IF–THEN rules for monsoon season

	Condition		Class/target
IF	Season = monsoon, MaxT = 28.30–30.80°C and NE1 = 0	THEN	High
IF	Season = monsoon, MaxT = 28.30–30.80°C, NE1 = 0–0.36 and MinT = 21.24–22.59°C	THEN	High
IF	Season = monsoon, MaxT = 28.30–30.80°C, NE1 = 0–0.36 and MinT = 22.59–24.31°C	THEN	Low
IF	Season = monsoon, MaxT = 28.30–30.80°C and NE1 = 2.12–4.10	THEN	Low
IF	Season = monsoon, MaxT = 30.81–31.39°C and RF = 0 mm	THEN	High
IF	Season = monsoon, MaxT = 30.81–31.39°C and RF = 3.10–18.20 mm	THEN	Low
IF	Season = monsoon, MaxT = 30.81–31.39°C and RF = 18.40–156.20 mm	THEN	Low
IF	Season = monsoon, MaxT = 33.40–36.97°C and RF = 0 mm	THEN	Low
IF	Season = monsoon, MaxT = 33.40–36.97°C and RF = 18.40–156.20 mm	THEN	High

Table 5. IF–THEN rules for post-monsoon season

	Condition		Class/target
IF	Season = post-monsoon, MaxT = 28.30–30.80°C and NE1 = 0.39–0.92	THEN	High
IF	Season = post-monsoon, MaxT = 28.30–30.80°C and NE1 = 2.12–4.10	THEN	Low
IF	Season = post-monsoon, MaxT = 31.40–32.14°C and MinT = 13.30–15.87°C	THEN	High
IF	Season = post-monsoon, MaxT = 31.40–32.14°C and MinT = 15.94–18.84°C	THEN	High
IF	Season = post-monsoon, MaxT = 31.40–32.14°C and MinT = 22.59–24.31°C	THEN	Low
IF	Season = post-monsoon, MaxT = 33.40–36.97°C and MinT = 13.30–15.87°C	THEN	Low
IF	Season = post-monsoon, MaxT = 33.40–36.97°C and MinT = 21.24–22.59°C	THEN	Low

Table 6. IF–THEN rules for winter season

	Condition		Class/target
IF	Season = winter, MaxT = 28.30–30.80°C and NE1 = 0.39–0.92	THEN	High
IF	Season = winter, MaxT = 28.30–30.80°C and NE1 = 0.96–2.11	THEN	High
IF	Season = winter, MaxT = 28.30–30.80°C	THEN	High
IF	Season = winter, MaxT = 31.40–32.14°C	THEN	Low
IF	Season = winter, MaxT = 32.16–33.31°C and NE1 = 0	THEN	Low
IF	Season = winter, MaxT = 32.16–33.31°C and NE1 = 0.39–0.92	THEN	High
IF	Season = winter, MaxT = 33.40–36.97°C and NE2 = 0–0.26	THEN	Low
IF	Season = winter, MaxT = 33.40–36.97°C and NE2 = 0.28–1.42	THEN	High
IF	Season = winter, MaxT = 33.40–36.97°C and NE2 = 1.59–2.43	THEN	Low

During post-monsoon season, when the maximum temperature ranged from 31.40°C to 32.14°C and from 28.30°C to 30.80°C, there was a greater chance of pest incidence (Table 5). Similarly, when minimum temperature ranged from 13.30°C to 15.87°C and from 15.94°C to 18.84°C, there was a greater chance of pest incidence. When minimum temperature ranged from 21.24°C to 22.59°C and from 22.59°C to 24.31°C, there was less chance of pest incidence during post-monsoon season (Table 5). Similarly, when NE1 was in the range 2.12–4.10 per plant, pest incidence was low, which implied that spiders play a role in minimizing the pest occurrence (Table 5).

During winter season, when the maximum temperature ranged from 28.30°C to 30.80°C and from 32.16°C to 33.31°C, the pest population was high. When the maximum temperature ranged from 33.40°C to 36.97°C, the pest population was low. But, when the maximum temperature ranged from 33.40°C to 36.97°C and NE2 (*C. carnea*) ranged from 1.59 to 2.43 per plant, the pest population was low (Table 6). These findings of pest incidence during different seasons considering abiotic factors like maximum and minimum temperature were similar to those of earlier reports[16–21].

The above results show that when maximum temperature ranged from 33.40°C to 36.97°C, the pest population was low during monsoon and post-monsoon seasons. But, during monsoon with same range of maximum temperature and with rainfall, the pest population was high. During winter, with the maximum temperature in the same range, pest incidence was low as well as high. During monsoon and post-monsoon seasons, spiders played a role in reducing pest incidence, but during winter *C. carnea* played a role in reducing pest incidence.

The prediction equation for *H. armigera* occurrence along with the coefficient of determination (R^2) values and mean square error (MSE)[22] are given in Table 7.

The results of regression analysis revealed that about 66%, 21% and 40% of *H. armigera* incidence could be attributed to two natural enemies and the weather factors respectively (Table 7).

A comparison of Shannon information gain value, correlation analysis and regression analysis is given in Table 8.

The Shannon information theory has shown that the attribute/variable or the factor 'Season' played an important role in pest incidence. The same had been proved with correlation analysis. Among the weather parameters, maximum temperature plays a major role in pest incidence because its *r* value is significant and also that its information gain value is more than other weather parameters values. Regression analysis also revealed that 'Season' played a major role in pest incidence among all other parameters (Table 8).

The testing set was used for finding the accuracy of the classification. It was found that the misclassified testing data were 8.82%. The confusion matrix derived for training set of data and testing set of data is given in Tables 9 and 10 respectively.

The dataset for the period from July 2005 to January 2006 (training data) used for training the model such that assignment of classes based on PI threshold values, categorization of numerical attributes based on the binning method, storage of range values for the numerical attributes, finding information gain table and construction of decision tree. The dataset for the period from August 2007 to January 2008 (testing data) was given to the model for testing. In the testing phase, the class values assigned for PI, categorization of numerical attributes carried out based on training data information. Then, the information gain table was calculated for testing records and the decision tree was constructed. Confusion matrix was derived for the testing the records after the classification process and the percentage of misclassification was 8.82. Hence, the testing set revealed that prediction or classification of pest incidence was done 91.18% accurately. Moreover, the decision tree visualized the biotic and abiotic factors with range values playing a role in pest incidence (either 'LOW' or 'HIGH').

The occurrence of cotton bollworm *H. armigera* (Hübner) was greatly influenced by its natural enemies, viz. spiders and *C. carnea* and by abiotic factors. In the present work, the population dynamics of the pest and its natural enemies was studied using Shannon information measure with decision tree induction approach. The developed classification model has the ability to successfully treat 'categorical' variables as well as 'continuous' variables in the database. Pest incidence had been classified

Table 7. Prediction equation for *H. armigera* occurrence

Class/model	Regression equation	R^2	MSE
Low	Y = 0.559 + 0.009 Season + 0.005 NE1 + 0.013 NE2 – 0.010 MaxT – 0.004 MinT – 0.002 RH	0.66	0.030
High	Y = 2.518 + 0.085 Season + 0.057 NE1 + 0.033 NE2 – 0.081 MaxT + 0.042 MinT – 0.007 RH	0.21	0.262
Overall	Y = 3.679 + 0.053 Season – 0.042 NE1 + 0.118 NE2 – 0.091 MaxT + 0.028 MinT + 0.001 RF – 0.018 RH	0.40	0.271

Y, *H. armigera*; NE, Spiders; NE2, Chrysopids; MaxT, Maximum temperature; MinT, Minimum temperature; RF, Rainfall; RH, Relative humidity, Season = 1 (monsoon), Season = 2 (post-monsoon), Season = 3 (winter).

Table 8. Comparison of information gain values, correlation analysis and regression analysis

Attribute	Information gain value	Correlation analysis (r)	Regression analysis
Season	0.226	0.280*	Y = 3.679 + 0.053 Season – 0.042 NE1 + 0.118 NE2 – 0.091 MaxT + 0.028 MinT + 0.001 RF – 0.018 RH
NE2	0.188	0.376**	
NE1	0.140	0.154	
MaxT	0.120	–0.280*	
MinT	0.104	–0.208	
RF	0.072	–0.200	
RH	0.055	–0.264*	

r values with * are significant at $P < 0.05$ level and with ** are significant at $P < 0.01$ level.

Table 9. Confusion matrix for training data

	Predicted class		
	Low	High	Total
True class			
Low	31	1	32
High	2	35	37
Total	36	33	69

Table 10. Confusion matrix for testing data

	Predicted class		
	Low	High	Total
True class			
Low	18	3	21
High	1	12	13
Total	19	15	34

into two classes, viz. low and high based upon ETL. R^2 for low and high class values was 0.66 and 0.21 respectively. Our studies showed that season influenced the population dynamics of *H. armigera* among all the factors. It was found that the misclassified testing data were 8.82%. The confusion matrix for the testing set revealed that the classification was done more accurately using the training set. Hence, this approach could be successfully utilized to understand the role of natural enemies and weather factors on the occurrence of *H. armigera* (Hübner) as well as prediction of this pest.

1. Trivedi, T. P., Yadav, C. P., Vishwadhar, Srivastava, C. P., Dhandapani, A., Das, D. K. and Singh, Joginder, Monitoring and forecasting of *Heliothis*/*Helicoverpa* population. In *Heliothis/Helicoverpa Management – Emerging Trends and Strategies for Future Research*, 2005, pp. 119–140.
2. Sharma, H. C., Cotton bollworm/legume pod borer, *Helicoverpa armigera* (Hübner) (Noctuidae; Lepidoptera): Biology and management. In *Crop Protection Compendium*, CAB International, Wallingford, UK, 2001, p. 72.
3. Youm, O., Gilstrap, F. E. and Teetes, G. L., Pesticides in traditional farming systems in West Africa. *J. Agric. Entomol*., 1990, **7**(3), 171–181.
4. Fayyad, U., Piatetsky Shapiro, G. and Smyth, P., Knowledge discovery and data mining: towards unifying framework. In Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining, KDD-96, Oregon, 1996.
5. Zhao, H. and Ram, S., Constrained cascade generalization of decision trees. *IEEE Trans. Knowledge Data Eng*., 2004, **16**(6), 727–739.
6. Basak, J. and Krishnapuram, R., Interpretable hierarchical clustering by constructing an unsupervised decision tree. *IEEE Trans. Knowledge Data Eng*., 2005, **17**(1), 121–132.
7. Aneta, I., Celine, V., Nathalie, C., Marko, D. and Saso, D., The feasibility of co-existence between conventional and genetically modified crops: using machine learning to analyse the output of simulation models. *Ecol. Model*., 2008, **215**, 262–271.
8. Edwards Jones, G., Knowledge-based systems for crop protection: theory and practice. *Crop Protect*., 1993, **12**(8), 565–578.
9. Chakraborty, P. and Chakrabarti, D. K., A brief survey of computerized expert systems for crop protection being used in India. *Prog. Nat. Sci*., 2008, **18**(4), 469–473.
10. Mansingh, G., Reichgelt, H. and Osei Bryson, K. M., CPEST: An expert system for the management of pests and diseases in the Jamaican coffee industry. *Expert Syst. Appl*., 2007, **32**(1), 184–192.
11. Singh, N. B., *Helicoverpa* Menace in the Indian Subcontinent. In *Heliothis/Helicoverpa Management – Emerging Trends and Strategies for Future Research*, 2005, pp 39–43.
12. Dhaliwal, G. S. and Arora, R. (eds), Integrated pest management: Achievements and challenges. In *Principles of Insect Pest Management*, NATIC, India, 1996, pp. 308–355.
13. Gupta, G. K., Classification. In *Introduction to Data Mining with Case Studies*, Prentice-Hall of India, 2006, pp. 106–136.
14. Rayan, P. T., *Modern Regression Methods*, John Wiley, New York, 1997.
15. SPSS V 17.0, Statistical Package for Social Sciences, SPSS Inc., Illinois, Chicago, USA, 2008.

16. Ballal, C. R., Studies on population dynamics of *Helicoverpa armigera* and its natural enemies. Studies on the feasibility of using biocontrol measures for developing a bio-intensive IPM programme for the management of *Helicoverpa armigera* (Hübner) (Lepidoptera: Noctuidae). Ph D thesis, Mysore University, 1998, pp. 43–108.
17. Dahiya, K. K., Chauhan, R., Ombir and Malik, V. S., Effect of abiotic factors on the incidence of pod borer, *Helicoverpa armigera* (Hübner) on early maturing varieties of red gram. *Crop Res. Hisar*., 1997, **14**(1), 151–154.
18. Deshmukh, A. Y., Khan, M. I. and Khande, D., Seasonal incidence of pigeonpea pod borers under Akola conditions (Maharashtra). *Insect Environ*., 2003, **9**(3), 127–128.
19. Dhaliwal, L. K., Kooner, B. S., Singh, J. and Sohi, A. S., Incidence of *Helicoverpa armigera* (Hübner) in relation to meteorological parameters under Punjab conditions. *J. Agrometeorol*. (*Spec. Issue*), 2004, **6**, 115–119.
20. Dubey, A., Kanaujia, K. R. and Sudha, K., Efficiency of different pheromonal blends to monitor *Helicoverpa armigera* (Hübner) moth catches. *Indian J. Plant Protect*., 2004, **32**(1), 147–148.
21. Kumar, P. and Kumar, S., Mating and oviposition cage for *Helicoverpa* during winter season. *Entomon*., 2001, **26**(3/4), 331–335.
22. Kvalseth, T. O., Cautionary note about R^2. *Am. Stat*., 1985, **39**, 279–285.

ACKNOWLEDGEMENT. This work forms part of the Ph D work of M.P. at Shrimathi Indira Gandhi College, Bharathidasan University, Tiruchirappalli. We thank the anonymous referees for their critical comments that helped improve the quality of the manuscript.

Received 23 December 2009; revised accepted 28 October 2010

Evidence of a random ovipositional strategy by female fruit fly *Bactrocera dorsalis* (Tephritidae: Diptera) with reference to host quantum

Abraham Verghese*, M. S. Uma, P. D. Kamala Jayanthi, Ranju Mouly and Mahiba Helen

Indian Institute of Horticultural Research, Hessarghatta Lake PO, Bangalore 560 089, India

Mango (*Mangifera indica* L.) is a favoured host of the fruit fly, *Bactrocera dorsalis* (Hendel), on which the females oviposit. Detailed laboratory studies were carried out to study the role of host biomass (= weight) on the oviposition decision of the tephritid fruit fly, *B. dorsalis.* The hypothesis that the biomass of mango fruit influences the decision of the mother (= gravid) fruit fly was tested on three varieties, *Alphonso*, *Totapuri* and *Banganpalli*, in three different but identical experiments. The results clearly showed that the number of oviposition punctures by a female fruit fly did not show significant differences among the different weight groups within a variety. The correlation coefficients between fruit weight and number of oviposition punctures were also not significant, i.e. the weight of the fruit did not influence the number of ovipositional punctures by the female fruit flies. This was true in all the three varieties. Further, the fruit weight did not influence the number of pupae that metamorphosed, number of adults that emerged and sex ratio in all the three experiments. The implications of these results on oviposition selection strategy of the female fruit fly are discussed. The study clearly indicated that after selecting a host (mango) the female fruit fly does not show non-random selection for a mango fruit with more quantum of pulp, as is expected if it were to maximize its survival and fitness.

Keywords: *Bactrocera dorsalis*, host biomass, mango, oviposition strategy.

AN organism exploits its resources to maximize its fitness. This is also true in the case of frugivorous insects. An insect, like a true fruit fly (Tephritidae: Diptera) ensures its fitness by favouring development of maximum number of progenies, through appropriating optimally the available food resource. The present study is on the tephritid, *Bactrocera dorsalis* (Hendel), a fruit fly that develops in mango (*Mangifera indica*) fruits. Earlier studies have found that food resource affects the development of the flies, as evident from differential pupal weights; poor food hosts result in lower pupal weight[1]. A fly would in all probability exploit its available food resources in such a manner that all its progenies have equal chance of survival. So, one would expect a fly to appropriate the pattern of egg-laying (oviposition) on a fruit depending on the quantum of a given food that would be available to its offspring, which would offset intraspecific competition. In the present context of the fruit fly, *B. dorsalis*, the adult female is expected to choose a mango fruit that can provide optimum quantum of food (fruit pulp) for maximum development of the maggots (the apodus pulp-feeding stage of a fruit fly) and ensure successful adult eclosion to the next generation. Thus all the maggots which hatch from the eggs are expected to have sufficient food for development into puparia, and metamorphose to fertile adults, ensuring a fit population in the next generation. If the food is inadequate, the maggots which manage to get sufficient food become fitter than the other sibling maggots and eventually pupate successfully, more by chance, as each individual is as fit as the other on hatching and hence has equal probability of feeding and developing.

The tephritid, *B. dorsalis* is a polyphagous fruit-feeding maggot, but dominantly feeds on mango[2,3]. The adult females lay eggs below the rind of the mango using their piercing ovipositor. On hatching the maggots feed

*For correspondence. (e-mail: avergis@iihr.ernet.in)