Journal of Food Legumes
37(2): 205-210, 2024
DOI: 10.59797/jfl.v37.i2.196

Research Paper

Weather variable selection for whitefly population prediction modeling by using backward elimination regression

Hemant Kumar[1], Anup Chandra[1]*, Man Mohan Deo[1] and Kaushik Bhagawati[2]

[1]*ICAR-Indian Institute of Pulses Research,Kalyanpur, Kanpur 208024, Uttar Pradesh*
[2]*ICAR Research Complex for NEH Region, AP Centre, Basar-791101, Arunachal Pradesh*

***Corresponding authors e-mail**: anup.ento@gmail.com*

***Received:** 13 December 2023*
***Accepted:** 14 May 2024*

Handling Editor:
Dr. Jitendra Kumar, ICAR-Indian Institute of Pulses Research, Kanpur, India

ABSTRACT

The present investigation discusses the selection process of the most influencing weather variables for developing a prediction model for whitefly, *Bemisia tabaci* (Gennadius), based on the backward elimination method. This method aids in the selection of a model with fewer variables by eliminating those that are less pertinent, thereby enhancing precision and mitigating model complexity. In the pursuit of achieving a balance between simplicity and model fit, the conventional 5% level of significance (p-value ≤ 0.05) was utilized along with six weather variables viz., maximum temperature, minimum temperature, evaporation rate, sunshine hours, rainfall, and evening relative humidity. Through an iterative elimination process, it was determined that only three variables-minimum temperature, sunshine hours, and evening relative humidity-significantly contributed to the prediction model. Subsequently, these three variables were retained for predicting whitefly population counts, while the remaining less relevant variables were discarded. The model was found to be around 74 percent accurate in predicting the dynamics of whitefly.

Key words: Abiotic factors, *Bemisia tabaci*, Correlation, Minimum temperature, Rainfall, Relative humidity, Sunshine hours.

INTRODUCTION

Bemisia tabaci (Gennadius), commonly known as whitefly, is a polyphagous pest of global significance. In pulses, this pest is a major impediment in achieving potential productivity particularly due to its role as a major transmitter of Yellow Mosaic Disease (YMD) caused by *Begomoviruses* in *Vigna* crops (Varma 1952, Nene 1973). The impact of this pest on potential productivity is substantial, with severe cases leading to a complete cessation of production (Singh 1980). Abiotic factors play a key role in regulating the population dynamics of insect pests. Mapping population dynamics, as influenced by various weather variables, is crucial for developing an effective prediction model for targeted control of the pest. Out of the several weather variables influencing insect dynamics, the selection of relevant variables is often regarded as the most pivotal and challenging aspect of model development. A comprehensive understanding of insect pest dynamics is most crucial for effectively managing targeted insects and employing insecticides judiciously (Veeranna *et al.* 2023), particularly in conservation agriculture (Kumar *et al.* 2023). This understanding not only facilitates precise control of the intended insect pests but also plays a crucial role in minimizing the environmental impact arising from chemical residues. By gaining insights into the ecological interactions and life cycles of insect pests, practitioners can adopt more strategic and sustainable pest control measures, thereby striking a balance between pest management and environmental conservation. Statistical weather-based forecasting models play a pivotal role in monitoring and managing insect pests in crops. By analyzing the weather patterns, these models predict the occurrence and population dynamics of pest outbreaks. Apprehending the relationships between weather parameters and pest outbreaks helps to anticipate optimal conditions for pest development. A timely and accurate forecast empowers the implementation of targeted pest control measures, reducing reliance on broad-spectrum pesticides and minimizing environmental impact (Kawakita and Takahasi 2022). Forecasting models are indispensable tools in integrated pest management, fostering precision and efficiency in safeguarding crops from insect threats (Kleynhans *et al.* 2018). Developing accurate forecasting models for

insect pests poses several challenges, primarily due to the complexity of ecological interactions and the dynamic nature of weather patterns. Additionally, the lack of comprehensive long-term datasets and the inherent variability in insect populations further complicate model development. When the forecasting model depends on several interrelated weather parameters, the model becomes complicated to generalize and comprehend. Identifying the key weather parameters that influence insect behavior, is a nuanced task, as pests may respond differently to various climatic factors. Choosing relevant weather parameters demands a thorough understanding of the specific pest's biology and life cycle. To gain insights into its population behavior, it is imperative to comprehensively assess the influence of various weather parameters (Chandra *et al.* 2021). The development of a predictive model, contingent upon the judicious selection of pertinent weather elements or variables, holds promise for anticipating and managing the population dynamics of the whitefly with precision.

The process of identifying the variables to be included in the final model is referred to as variable selection, a crucial step in model development. Variable selection serves dual objectives. Firstly, it aids in identifying all variables relevant to the outcome, ensuring the model's completeness and accuracy. Secondly, it facilitates the selection of a model with a limited number of variables by eliminating irrelevant ones that contribute to decreased precision and increased complexity. Ultimately, variable selection strives to strike a balance between model simplicity and fit. The task of selecting appropriate variables for model inclusion is widely recognized as the most important and challenging aspect of model building. This article delves into the vital process of weather variable selection using the backward elimination method and its application in the development of a prediction model for whiteflies. Variable selection involves the careful consideration of multiple variables to determine which one should be included in a specific model, entailing the removal of those deemed unimportant in prediction modeling (Ranter 2010).

The objective of variable selection is to identify a set of variables that optimally contribute to model fitting, thereby enhancing the accuracy of predictions. Datasets frequently encompass numerous variables that are ultimately not incorporated in model development. The judicious choices of pertinent variables are essential to avoid the inclusion of extraneous variables, commonly referred to as noise, in the final model. Variable selection confers several advantages, including improved predictive model performance, expedited variable identification, cost-effectiveness through reduced training and utilization time, enhanced data visualization, and a more comprehensive understanding of the underlying data generation process (Guyon and Elisseeff 2003).

There are various practical reasons for variable selection, notably considerations of practicality. The inclusion of an extensive set of variables in a model is often impractical. Additionally, certain variables may exert negligible effects on outcomes, justifying their exclusion. A model with fewer variables translates to reduced computational time and complexity (Kuhn and Johnson 2013). In adherence to the principle of parsimony, models characterized by simplicity and fewer variables are preferred over complex models with an abundance of variables. The proliferation of variables in a model renders it more reliant on observed data (Hosmer *et al*. 2013). Simple models, being more straightforward to interpret, generalize, and apply in practical scenarios, align with the principle of parsimony (Steyerberg 2008). Nonetheless, it is imperative to ensure that pivotal variables are not overlooked in the pursuit of model simplicity. Ultimately, the selection of variables for inclusion in the final prediction model should be driven by a comprehensive analysis of the dataset.

MATERIALS AND METHODS

The experimental investigation was conducted at the Research Farm of the ICAR-Indian Institute of Pulses Research, situated in Kanpur, Uttar Pradesh, India, at coordinates 26°27′ N latitude and 80°14′ E longitude, with an elevation of 152.4 meters above mean sea level. The whitefly population was monitored at a weekly intervals from the 17th standard meteorological week (SMW) of 2020 (April 23, 2020) to the 16th SMW of 2021 (April 16, 2021) using yellow sticky traps. Yellow acrylic boards measuring 20 × 15 cm, coated with castor oil, were suspended on sticks approximately 70 cm above the ground, at a level of the crop canopy. These traps were randomly placed at a rate of 15 per hectare in greengram and blackgram crops at the main farm of ICAR-IIPR Kanpur. Whiteflies adhering to the sticky boards were counted 24 hours after trap installation. Daily meteorological data, including maximum and minimum temperature (°C), relative humidity, total rainfall (mm), evaporation, and sunshine hours, were collected from the meteorological section of

ICAR-IIPR Kanpur from April 2020 to April 2022.

Various methodologies exist for the execution of backward elimination, with one of the prevalent approaches involving the utilization of a p-value threshold. P-values serve as an indicator of the necessity of a variable in the model. A designated p-value threshold is established, beyond which a variable is deemed dispensable. Specifically, any variable presenting a p-value exceeding 0.05 is eliminated. The application of backward elimination necessitates the determination of p-values for each variable, followed by a comparison of these values against the predefined threshold of 0.05. Subsequently, variables with p-values surpassing the threshold are systematically removed. Additionally, data points were excluded when the whitefly count was zero or nonexistent as a dependent variable. Backward elimination is a variable selection technique in statistics used to refine predictive models. It involves starting with a model that includes all potential predictors and systematically removing the least significant variables. The iterative process of variable elimination in the study involves sequentially removing variables from the full model until only those deemed to have a significant contribution to the outcome remain (Ranter 2010). The variable with the smallest test statistic, indicating its impact on the model, and either a test statistic below the cut-off value or the highest p-value surpassing the cut-off, signifying the least significance, is initially excluded. Subsequently, the model is restructured without these variables, and test statistics or p-values are recalculated. This elimination-refitting cycle persists until each remaining variable attains significance at the predetermined cut-off value. This cut-off, often referred to as 'p-to-remove,' must be pre-established.

RESULTS AND DISCUSSION

Population dynamics of white fly and relationship with weather variable

During the investigation period, fluctuations in the abundance of trapped whiteflies were observed, ranging from 0 to 318.5 (Figure 1). The highest population density occurred during the 35th SMW of the study (318.5 individuals) in the last week of August to the first week of September. Concurrently, environmental parameters such as Maximum Temperature (TMAX), Minimum Temperature (TMIN), Evaporation (EV), Sunshine Hour (SSH), Rainfall, and Relative Humidity Evening (RHE) were recorded at 34°C, 27.5°C, 1, 5.5 hrs, 0 mm, and 76.3%, respectively. Subsequent notable peaks in whitefly abundance were identified during the 26th SMW (309.5) and the 25th SMW (303), which corresponded to June through July. Notably, the insect exhibited sustained activity until September, gradually diminishing from October to December, with the lowest recorded levels persisting from January to mid-March. The findings presented herein align with the observations made by Ali *et al.* (2005), indicating a progressive increase in the whitefly population from mid-July onwards, with the peak occurring on August 19. The whitefly population exhibited activity spanning from the 30th to the 39th SMW (week till month of September) followed by a rapid decline. Garg and Patel (2018) reported the maximum population of whitefly during 36th SMW in greengram, blackgram and soybean. Similarly, Gupta *et al.* (2009) reported a comparable activity pattern for the whitefly.

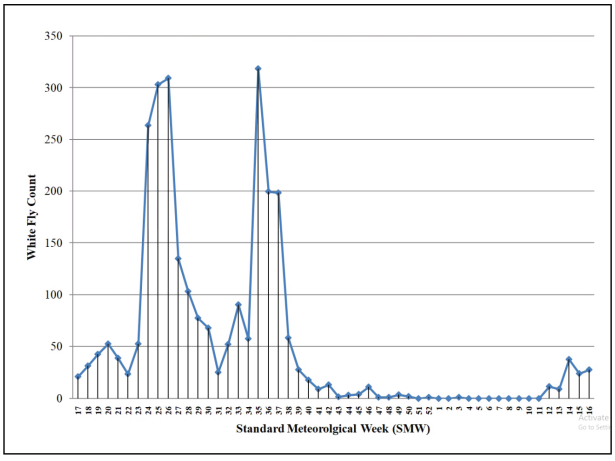

Fig. 1: White fly count with highest peak during 35th SMW

Observations revealed discernible variations in the cumulative whitefly count throughout the study duration, and these distinctions may be ascribed to fluctuations in weather conditions throughout the respective months. The correlation analysis between the assessed whitefly population and meteorological variables is detailed in Table 1.

In our study, the correlation analysis indicated a positive association between the whitefly population and minimum temperature (TMIN), evening relative humidity (RHE), and sunshine hours (SSH). These outcomes agree with the findings of Muthukumar and Kalyanasundaram (2003). The significant positive correlation found with minimum temperature is also supported by findings of Sousheel *et al.* (2022) but, contrastingly they found a negative correlation with evening

Table 1. Correlation (r) between weather parameters and whitefly population

	WF	TMIN	EV	RHE	RAIN	TMAX	SSH
WF	1						
TMIN	**0.5704**	1					
EV	0.0881	0.4984	1				
RHE	**0.2582**	-0.2964	-0.8062	1			
RAIN	-0.0228	-0.1724	-0.1940	0.2292	1		
TMAX	-0.0727	0.8566	0.7340	-0.5812	0.0154	1	
SSH	**0.3603**	0.4618	0.7719	-0.6866	-0.3359	0.6475	1

A significant value of r at 5% level of significance is 0.231 for 50 degree of freedom

relative humidity and sunshine hours. Zeshan *et al.* (2015) also reported temperature and relative humidity contributing towards whitefly population build-up, but with maximum temperature being more important than minimum. Sharma *et al.* (2017) also reported a positive correlation of the whitefly population with temperature and sunshine hours with a negative association with humidity and rainfall. Also, Dhole *et al.* (2023) reported temperature positively correlated but both rainfall and humidity had a negative correlation with the whitefly population. Rajnish *et al.* (2004) emphasized the favorable impact of sunshine hours on whitefly development, as evidenced by the observed positive correlation between them.

Variable selection by backward regression method

The initiation of the backward elimination process involves fitting a multiple linear regression model with the dependent variable being the whitefly population (WF), and the independent variables encompassing Maximum Temperature (TMAX), Minimum Temperature (TMIN), Evaporation (EV), Sunshine Hour (SSH), Rainfall, and Relative Humidity Evening (RHE). The model summary, as derived from the fitted regression, is presented in Tables 2 and 3.

Table 2 presents a highly significant F-value of 4.91 for the regression model, as evidenced by a corresponding p-value of 0.0012. This outcome suggests that at least one of the regression coefficients associated with the weather variables is non-zero, rendering the regression model meaningful. Thus, the presence of a pertinent linear relationship between the dependent variable and, at the very least, one of the six predictor weather variables integrated into the model is inferred. However, this does not imply the indispensability of all six predictors. In pursuit of variable selection, EV emerges as the least significant, boasting the largest p-value of 0.5215, exceeding our predetermined significance threshold of 0.05. Consequently, EV is excluded from the model, and the selection process recommences. A customary threshold for retaining predictors in a model is set at p = 0.05, denoting a minimum 95 percent likelihood that the predictor holds substantive significance.

Table 2. Analysis of variance for multiple linear regression model

	df	*SS*	*MS*	*F*	*Significance F(p-value)*
Regression	6	153459.26	25576.54	4.91	0.0012
Residual	32	166713.29	5209.79		
Total	38	320172.55			

Table 3. Coefficients of variables and p-value

Variable	Coefficients	Standard Error	t Stat	P-value
Intercept	90.82	201.13	0.45	0.6546
TMIN	13.65	3.74	3.65	0.0009
TMAX	-5.11	5.76	-0.89	0.3816
SSH	20.89	7.56	2.76	0.0094
RHE	0.63	1.50	0.42	0.0089
EV	9.25	14.28	0.65	0.5215
RAIN	-2.47	1.39	-1.78	0.0846

The regression model was fitted with the dependent variable Whitefly (WF) and all independent variables, excluding EV. The independent variables considered in the model encompass TMAX, TMIN, SSH, RHE, and RAIN. The summary of the fitted model is presented in Tables 4 and 5.

During the third step, the variable Maximum Temperature (TMAX) was excluded from the model due to its minimal significance, as indicated by the highest p-value of 0.4249. Given that this value surpasses our predetermined significance threshold of 0.05, the variable was systematically removed from the model, marking the initiation of subsequent steps. In the fourth, the variable RAIN with a p-value of 0.0854 was eliminated.

Consequently, the variable selection for the final model comprises TMIN, SSH, and RHE.

Table 4. Analysis of variance for regression model

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	152549.77	30509.95	6.01	0.0005
Residual	33	167622.79	5079.48		
Total	38	320172.55			

Table 5. Coefficients of variables and p-values

Variable	*Coefficients*	*Standard Error*	*t Stat*	*P-value*
Intercept	28.94	134.38	0.22	0.8308
TMIN	12.95	3.30	3.92	0.0004
TMAX	-4.38	5.42	-0.81	0.4249
RHE	0.23	7.12	0.70	0.0084
SSH	11.70	12.86	2.91	0.0086
RAIN	-2.42	1.37	-1.77	0.0854

Subsequently, these three variables were utilized in constructing the predictive model for the Whitefly (WF) population, as illustrated below.

WF (Population) = 28.94 + 12.95*TMIN + 0.23*RHE +11.70*SSH

The model was found to be on average 74 percent accurate in predicting the whitefly population. The comparison between the actual and predicted population is shown in Figure 2.

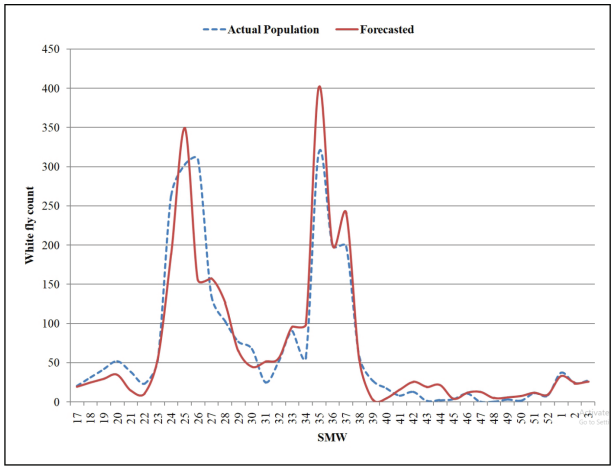

Fig 2. Comparison between the whitefly population forecasted by the model and the actual population recorded in the given period. The model was found to be 74 percent accurate in forecasting the whitefly population.

In forecasting the whitefly population, various analytical methods are available; however, the utilization of multiple regression analysis underscores the crucial process of selecting pertinent weather variables. The inclusion of suitable and logically chosen variables in predictive modeling holds paramount significance, as the model's performance is substantially contingent on the variables ultimately integrated into the model. The omission of appropriate variables can yield inaccurate results, leading to the model's failure to capture the true relationships inherent in the data between the outcome and the selected variables. Instances abound where the proper steps in adopting an appropriate method of variable selection were overlooked in predictive modeling. Researchers should familiarize themselves with and remain vigilant about these critical aspects of prediction modeling. In conclusion, weather parameters such as minimum temperature, sunshine hours, and evening relative humidity exert the most pronounced influence on the whitefly population compared to other factors.

REFERENCES

Ali S, Khan MA, Habib A, Rasheed S and Iftikhar Y. 2005. Correlation of environmental conditions with okra yellow vein mosaic virus and *Bemisia tabaci* population density. Int J Agric Bio **7**: 142-144.

Chandra A, Sujayanand GK and Kumar R. 2021. Influence of Sowing Dates and Host Crops on Population Incidence of Whitefly, *Bemisia tabaci* (Gennadius) in Greengram and Blackgram. National Academy Science Letters **44(5)**: 389–391.

Dhole RR, Singh RN, Dhanapal R, Singla S, Ramkumar G, Muthuusamy R, Salmen SH, Alharbi SA, Narayanan M and Karuppusamy I. 2023. Impact assessment of natural variations in different weather factors on the incidence of whitefly, *Bemisia tabaci* Genn. and yellow vein mosaic disease in *Abelmoschus esculentus* (L.) Moench. Environmental Research **231(2):** 116209.

Garg VK and Patel Y. 2018. Influence of weather parameters on population dynamics of whitefly in kharif legumes. Annals of Plant and Soil Research **20(4):** 371-374.

Gupta MP, Nayak MK and Srivastava AK. 2009. Studies on seasonal activity of whitefly (*Bemisia tabaci* Genn.) population and its association with weather parameters in Bundelkhand zone of Madhya Pradesh. Journal of Agrometeorology **11**: 175-179.

Guyon I and Elisseeff A. 2003. An introduction to variable and feature selection. Journal of Machine Learning Research **3**: 1157–1182.

Hosmer DW, Lemeshow S and Sturdivant RX. 2013. Applied logistic regression. John Wiley & Sons, New York.

Kawakita S and Takahasi H. 2022. Time-series analysis of population dynamics of the common cutworm, *Spodoptera litura* (Lepidoptera: Noctuidae), using an ARIMAX model. Pest Managment Science **78(6)**: 2423-2433.

Kleynhans E, Barton MG, Conlong DE and Terblanche JS. 2018. Population dynamics of *Eldana saccharina* Walker (Lepidoptera: Pyralidae): application of a biophysical model to understand phenological variation in an agricultural pest. Bulletin of Entomological Research **108(3):** 283-294.

Kuhn M and Johnson K. 2013. Applied predictive modeling, Springer, New York.

Kumar N, Hashim M, Nath CP, Hazra KK and Singh AK. 2023. Pulses in conservation agriculture: An approach for sustainable crop production and soil health. Journal of Food Legumes **36**(1): 1-9.

Muthukumar M and Kalyanasundaram. 2003. Influence of abiotic factors on the incidence of major insect pests in Brinjal (*Solanum melongena* L.). South Indian Horticulture **51**: 214-218.

Nene YL. 1973. Control of *Bemisia tabaci* Genn., a vector of several plant viruses. Indian Journal of Agricultural Sciences **43**: 433–436.

Rajnish KM, Rizvi SMA and Shamshad A. 2004. Seasonal and varietal variation in the population of whitefly (*Bemisia tabaci* Genn.) and incidence of yellow mosaic virus in urd and mungbean. Indian Journal of Entomology **66**: 155-158.

Sharma D, Maqbool A, Jamwal VV, Srivastava K and Sharma A. 2017. Seasonal dynamics and management of whitefly (*Bemesia tabaci* Genn.) in tomato (*Solanum esculentum* Mill.). Brazilian Archives of Biology and Technology **17**: 60.

Singh DP. 1980. Inheritance of resistance to yellow mosaic virus in blackgram (*Vigna mungo* L.). Theoretical and Applied Genetics **52**: 233–235.

Sousheel NY, Bhat BN, Devi GU, Yamini KN and Sridevi G. 2022. Population dynamics of whitefly population infesting chilli as influenced by various weather parameters. The Pharma Innovation Journal **11(9)**: 1381-1385.

Steyerberg EW. 2008. Clinical prediction models: A practical approach to development, validation, and updating. Springer, New York.

Varma PM. 1952. Studies on the relationship of the *Bhendi* yellow vein mosaic virus and its vector, the whitefly (*Bemisia tabaci* Gen.). Indian Journal of Agricultural Sciences **22:** 75–91.

Veeranna D, Fatima T, Kishore NS, Padmaja G, Rao PJ, Madhu M and Reddy RU. 2023. Bio-efficacy of certain new insecticides against pod borer complex in pigeonpea (*Cajanus cajan* L.). Journal of Food Legumes **36**(2&3): 178-182.

Zeshan MA, Khan MA, Safdar A and Arshad M. 2015. Correlation of conducive environmental conditions for the development of whitefly, *Bemisia tabaci* population in different tomato genotypes. Pakistan Journal of Zoology **47(6):** 1511-1515.