Indian Journal of Agricultural Sciences **81** (10): 945–7, October 2011

Effect of outliers in statistical modelling for predicting the outbreak of anthracnose in grapes (*Vitis vinifera*)

R VENUGOPALAN[1] and R D RAWAL[2]

Indian Institute of Horticultural Research, Hesseraghatta Lake, PO, Bangalore, Karnataka 560 089

Received: 6 September 2010; Revised accepted: 3 August 2011

ABSTRACT

Role of outliers while developing statistical models for disease prediction is studied. Specifically, statistical models were developed to optimize the role of weather factors and simultaneously to predict the incidence of anthracnose in grapes (*Vitis vinifera*) by eliminating the effect of aberrant /outlier data. It was observed that about 93.3% of the disease incidence was collectively explained by weather factors with a time lag of one week prior to incidence as expressed by the equation (Y = –86.87 +8.62 max.temp –0.98 min.temp –1.4 RH1 +0.64 RH2+6.9 ws–8.6 Evap+0.07 Rf–0.08 no.rd–0.6 Br.SS–9.2 VPD) of the variability in PDI. Out of this, two factors namely, wind speed and vapor pressure deficit could explain about 80.1 % of the anthracnose incidence, as explained by the equation Y=5.2+5.95ws–8.1VPD. It was noted that effect of 13 outliers/ aberrant observations when removed increased the prediction power of the model by 14 %. Moreover, as a measure of goodness-of-fit, the coefficient of determination (R^2) was used to evaluate the empirical models developed. Before taking final conclusion about the model, the model-generated residuals were tested for their robustness using statistical techniques.

Key words: Anthracnose, Coefficient of determination, Model, Weather factors

Horticultural crops in general and fruit crops in particular play a major role in the nutritional and economic security of the country. Among them, mango and grapes (*Vitis vinifera* L.) are the major crops, which help the country in foreign exchange earnings. Among the fruit crops, grapes occupy a major place due to its domestic demand as well as export potential. During the past few years grapes produced in India has entered the International market and commanding high prices. Now, grapes are being exported to more than 35 countries. Increased trade and the rapid dissemination of information have created a new competitive environment, which demands changed approach to the industry itself. In spite of these, productivity rate suffer badly because of different diseases, which also hamper the quality standards. Among the commercial varieties, those derived from *V. vinifera* are the most susceptible to the diseases infecting all the aerial parts of the plants depending upon the weather conditions .

These diseases lower vine production and increases cost of cultivation to severe losses to growers. The control of these diseases is necessary to maintain the required quality of grapes and grape products. No commercial fruit crop receives more chemical fungicides and pesticides than grape. It is also considered as the most liming factor in grapes production globally. Research to explore the severity of the disease in relation to the prevailing weather conditions is useful not only help in taking management decisions but also in execution of need-based management practices. In this direction, statistical models are considered as efficient tools in quantitative epidemiology and, therefore, may be developed to both describe and predict disease progress (Jeger, 1984, Teng 1985). Research to quantify the risk of an epidemic in relation to weather factors (Johonson *et al*. 1996; Papastamati *et al*. 2002, Chandel and Chandel, 2010, Biswas *et al*. 2011, Prasad *et al*.,2011) in several agricultural crops was studied in detail. Venugopalan *et al*. (2006) developed a statistical model for ascertaining the influence and reliability of weather parameters on incidence of blossom blight in mango (*Mangifera indica* L.). Recently, similar attempts were also made to develop statistical prediction models for downy and powdery mildew incidences in Grapes (Rawal *et al*. 2008 (a, b)). Growth models also play a vital role in understanding the dynamics of growth pattern over time. It is generally observed that a plant disease progress at a slow rate in the initial stage and reaches a peak, called as a point of inflection and then it starts again decreasing. Isolated attempts were also made in the past to develop growth models to delineate the epidemiology of crop diseases (Sinha *et al*. 2002). In view of the foregoing importance of this problem, it has become imperative to develop a holistic model, which

[1]Senior Scientist (e mail: venur@iihr.ernet.in) (Agril. Statistics), Section of Economics and Statistics; [2]Ex-Principal Scientist, (e mail: rawal.rd@gmail.com), Division of Plant Pathology

captures the change in climate factors in relation to disease incidence. To this end, in the present communication, statistical model was developed to optimize the role of weather variables and simultaneously to predict anthracnose outbreak in grapes (cv. Anab-E-Shahi) at Indian Institute of Horticultural Research, Bangalore.

MATERIALS AND METHODS

Grapes (cv. Anab-E-Shahi) were surveyed regularly during 2005–07 to record the anthracnose disease initiation and further progression. Disease ratings were recorded at weekly interval by following 0 – 5 scale, where 0 = nil PDI; 1= 0> PDI =10; 2= 11= PDI = 25; 3= 26= PDI =50; 4= 51 = PDI = 75 and 5 = = 76% diseases intensity (PDI). Data thus recorded were converted to per cent disease index by following Mckenny (1923). The weekly weather data such data as maximum temperature (°C) (X_1), minimum temperature (°C) (X_2), relative humidity (%) (7.30 h) (X_3); X4: relative humidity (%) (14.30 h), (X_4), evaporation (mm) (X_5), wind speed (Kph) (X_6), rainfall (mm) (X_7) , number of rainy days (X_8), bright sunshine (hr) (X_9) and vapour pressure deficit (VPD (Kpa) (X_{10}) were collected from IIHR meteorological observatory for the same period.

To select the weather parameters influencing observed variability in anthracnose, step-wise regression procedure (Draper and Smith 1981) was employed. This technique essentially consists in identifying, stage by stage, the independent variables, which are significantly correlated with anthracnose incidence (Y). During each stage, making use of F test, an independent variable would enter into the equation if the significance level of its F value is less than or equal to 0.05, and would be removed if the significance level is greater than or equal to 0.1(Draper and Smith 1981). Further, while dealing with a biological data, it has been observed that the regression model developed based on different subsets of data produce very different results, raising questions of model stability. One of the reasons that could be attributed to this fact is that the real data may contain one or more outliers or aberrant observations. Hence, data analyst while developing regression models must check initially for the presence or absence of an aberrant observations, which will also ensure meaningful interpretation of regression estimates in addition to avoid drawing any erroneous conclusions, thus increasing the power of prediction.

As a measure of goodness-of-fit, the value of co-efficient of determination (R^2) (Kvalseth 1985) was calculated as below:

Coefficient of determination (R^2)

$$R^2 = 1 - [\sum(Y_t - \hat{Y})^2/[\sum(Y_t - \dot{Y})^2],$$

where Y_t represents the percent disease incidence during the period *t*.

However, inclusion of an additional independent variable into the selected candidate model will always boost the computed R^2 value (Kvalseth 1985). Hence, to ensure the statistical significance of the computed regression coefficients, they were subjected to t-test statistic analysis (Draper and Smith 1981). Before taking any final decision about the statistical adequacy of the selected model, residual analysis was also carried out using the one sample run-test (Siegel and Castellan 1988).

RESULTS AND DISCUSSION

Results of correlation analysis among per cent disease incidence observed over time and different weather factors showed that per cent disease incidence (PDI) is positively related to wind speed (r=0.76), minimum temperature (r=0.49) and with both the relative humidity's (r=0.32 and r=0.44 respectively) and it is inversely related to vapour pressure deficit (r=-0.47). Correlation results also revealed about the presence of inter-relationship among the weather factors which may indirectly influence disease incidence in combination. Further, to account for the collective influence of various weather factors on percent disease incidence a statistical model relative per cent disease incidence over time period with weather factors was developed. About 93.3% variability in disease incidence was collectively due to all the weather factors (Table 1). However, based on the t-statistic value of the regression coefficients when tested for their significance showed that some of the factors (not all) were significantly influencing per cent disease incidence (as the corresponding t-statistic values are well inside the rejection region value of 1.96). Further, step-wise regression approach

Table 1 Results of statistical models along with goodness of fit statistics

Model type		Statistical model (with standard error of b_i))										R^2 (%)
Full regression		Y = -86.87 + 8.62 X_1	-0.98 X_2	-1.4 X_3	+0.64 X_4	+6.9 X_5	-8.6 X_6	+0.07 X_7	-0.08 X_8	-0.6 X_9	-9.2 X_{10}	93.3
model (all weather		(0.77)	(0.46)	(0.19)	(0.16)	(0.5)	(1.8)	(0.08)	(0.77)	(0.56)	(0.86)	
parameters)	t-stat	11.3	2.1	7.5	4.1	14.5	10.1	0.94	0.11	1.1	10.7	
Optimized		Y=5.2 +5.95 wind speed -8.1 VPD										80.1
model using		(0.69)	(1.34)									
significant factors	t-stat	8.59	5.89									

Values in parentheses are standard error of regression estimates b_i and bolded t-stat values indicate significance of b_i (P<0.05)

was used to construct a model having only statistically significant weather factors. An optimized model (Y=5.2 +5.95 wind speed -8.1 VPD) showed wind speed and vapour pressure deficit (kpa) could themselves predict the incidence to the extent of 80.1% ((R^2). It was also noted that effect of 13 outliers/ aberrant observations when removed increased the prediction power of the model by 14%. Further, it may be noticed that the regression coefficient corresponding to the optimized parameters were statistically significant as they exceeded the critical value of 1.96 (P<0.05). Also, for the optimized model, the run-test statistic value (0.402) was well within the acceptance region (P<0.05), further strengthened the statistical validity of the model. The pictorial representation of the optimized model is depicted in Fig 1.

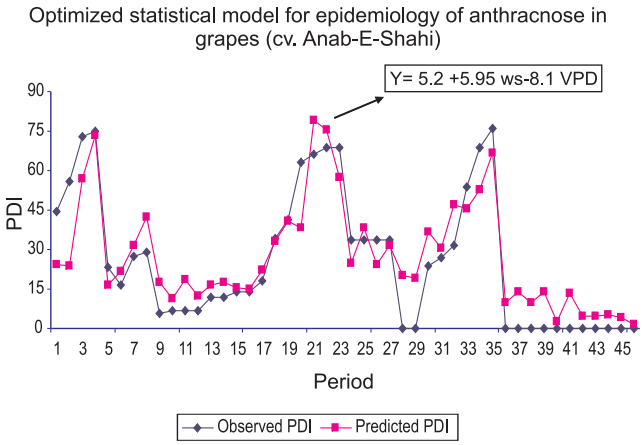

Fig 1 Graphical representation of statistical model

REFERENCES

Biswas B, Dhaliwal L K, Chahal S K and Pannu PPS. 2011. Effect of meteorological factors on rice sheath blight and exploratory development of a predictive model. *The Indian Journal of Agricultural Sciences* **81**(3): 256–60.

Chandel S and Chandel V. 2010. Correlation of disease with meteorological factors and management of Septoria leaf spot of chrysanthemum (*Chrysanthemum grandiflorum*). *The Indian Journal of Agricultural Sciences* **80**(1): 54–8.

Draper N R, and Smith H. 1981. *Applied Regression Analysis*. 2nd edn. 709 pp. John Wiley & Sons, New York.

Jeger M J. 1984. The use of mathematical models in plant disease epidemiology. *Scientific Horticulture*, **35**: 11–27.

Johnson D A, Alldredge J R, and Vakoch D L. 1996. Potato late blight forecasting models for the semiarid environment of south-central Washington. *Phyto Pathology*, **86**: 480–4.

Kvalseth T O. 1985. Cautionary note about R^2. *The American Statistician*. **39**: 279–85.

Mckenny H H. 1923. Influence of soil temperature and moisture on infection of wheat seedlings by *Helminthosporium sativum. Journal of Agricultural Research.* **26**: 195–217.

Papastamati K, van den Bosch F, Wewlham S J, Fitt B D L, Evans N, Steed J M. 2002. Modelling the daily progress of light spot epidemics on winter oilseed rape (*Brassica napus*), in relation to *Pyrenopeziza brassicae* inoculum concentrations and weather factors. *Ecological Modeling*. **148**: 169–89.

Prasad Y G, T Nagalakshmi M, Prabhakar D Yella, Reddy N H P, Rao M P, Prasad Rao M, Sinivasa Rao, S Desai, G G S N Rao and Y S Ramajwswna. 2007. Development of prediction models for bacterial leaf blight (*Xanthomonas aronopodis* pv *malvacearum*) of cotton (*Gossypium hirsutum*) in Andhra Pradesh and Maharashtra. *The Indian Journal of Agricultural Sciences*. **77**(11): 752–5.

Rawal R D, Venugopalan R, and Saxena A K. 2008. A statistical model for predicting the outbreak of powdery mildew in grapes. *Acta Horticulture* (ISHS) **785**: 293–6.

Rawal R D, Venugopalan R, and Saxena A K. 2008. A statistical model for describing the epidemiology of incidence of downy mildew in grapes. *Acta Horticulture* (ISHS) **785**: 279–84.

Siegel S, and Castellan N D. 1988. *Nonparametric Statistics for Behavioral Sciences*. 389 pp. McGraw-Hall.

Sinha P, Prajneshu and Verma A. 2002. Growth models for powdery mildew development of mango. *Annals of Plant. Protection Science.* **10**: 84–7.

Teng P S. 1985. A comparison of simulation approaches to epidemic modeling.. *Annual Review of Phytopathology* **23**: 351–79.

Venugopalan R, Rawal R D and Saxena A K. 2006. A statistical model for ascertaining the influence and reliability of weather parameters on incidence of blossom blight in mango. *Journal of Horticultural Sciences*, **1**(1): 64–7.