J. Hort. Sci.
Vol. 1 (1): 68-70, 2006

Statistical modelling for pre-harvest forecast: an illustration with rose

K. S. Shamasundaran and R. Venugopalan
Section of Economics and Statistics
Indian Institute of Horticultural Research
Hessaraghatta Lake Post, Bangalore-560 089, India
E-mail: sham@iihr.ernet.in

ABSTRACT

Crop yield forecast plays a vital role in arriving at pre-harvest yield estimate of a standing crop and to identify the stage at which reliable forecasting could be made before final harvest. In this paper, an attempt has been made to apply the regression technique for prediction of yield in rose. Rose, is an important flower crop not only for internal market but is also intended for export, and since it shrivels, estimation of yield of a standing crop before its actual harvest is essential. Based on results a model was developed, which showed that information from the first two pickings of a standing crop could be used to forecast rose yield to an extent of 77% two months before final harvest. It is also suggested to have a minimum sample size of 20% to develop such a forecast model.

Key words: Goodness of fit statistics, statistical modelling, yield forecast

INTRODUCTION

Commercial cultivation of roses has gained importance in recent years in India due to a growing demand for these flowers in both domestic and export markets. India, blessed with diverse agro-climatic conditions, has an immense potential to increase the productivity and, in turn, yields maximum return, in overseas market for this crop. This can be achieved by developing a suitable model to predict the actual yield of a standing crop and subsequently identify the stage within which forecasting could be made to the desired extent. To this end, it is imperative to develop a model through which growers and policy makers could frame suitable management strategies for maximizing crop productivity and net return. In this regard, statistical modelling plays a vital role in developing appropriate forecast models, on a strong scientific footing, for crop yield prediction. Shamasundaran and Singh (2003) made a beginning in this direction.

In the present study, an attempt has been made to develop multiple regression models for obtaining a pre-harvest estimate of yield of rose based on information pertaining to several pickings. Goodness of fit of the models developed was carried out by statistically testing the computed regression coefficients and working out measures model adequacy.

MATERIAL AND METHODS

An investigation was carried out at the Indian Institute of Horticultural Research, Bangalore during 1994-95 for yield prediction in rose cv. Happiness. Two hundred and fifty six samples were used in this study. All the recommended cultural practices with a spacing of 75 cm x 75 cm were followed uniformly for the entire plot. Data on yield in terms of number of flowers/plant from several pickings were recorded and consolidated. The first picking was made eleven months after planting. Subsequent pickings were made at an interval of 45 days.

Linear correlation coefficient among harvests done during several pickings and total yield were computed and statistically tested. Further, multiple regression models were developed by regressing harvest pertaining to different pickings with the cumulative yield by utilizing the principle of least squares (Lewis-Beck, 1993). The following measures of goodness of fit statistics were used to judge the adequacy of the model developed (Agostid'no and Stephens, 1986):

Mean squared error (MSE)

$$\text{MSE} = [\,\Sigma\,(Y_t - \hat{Y}_t)^2 / n\,]$$

Coefficient of Determination (R^2)

$$R^2 = 1 - [\Sigma(Y_t - \hat{Y})^2 \;/\; [\Sigma(Y_t - \bar{Y}_t)^2]$$

where Y_t represents the harvest/yield at time t. However, while fitting regression models to the data considered, it may be noted that even an addition of one more independent variable to the model would result in increase in R^2 value (Kvelsth, 1985). Hence, to test the significance of the added variable, regression coefficients were subjected to t-test statistic analysis (Lewis-Beck, 1993).

RESULTS AND DISCUSSION

Linear [simple(r) and multiple(R)] correlation among yield (total) and individual pickings yield were computed are presented in Tables 1 and 2. Results revealed that there existed a highly significant relationship in almost all the pickings at 1% level, either individually or in combination with total yield. Further, it was noticed that the first picking gave rise to R^2 of 70% followed by others and the least was noticed with the fifth picking. When multiple correlation and regression was carried out, it revealed that all the pickings, individually or in combination, had significantly higher association with total yield ranging from 0.3889 to 0.9060. It was found that more than 80% of R^2 noticed with all the pickings together followed by first three and first four pickings. The first two pickings and the same along with four pickings; the first five pickings except second gave rise to an R^2 of more than 77% yield prediction. Further, as discussed earlier, inclusion of additional information about the harvest obtained in every pickings, R^2 value tends to increase further. To this end, regression coefficients derived by including an additional variable were tested for its statistical significance. Results presented in table 3 indicate that inclusion of X_3 variable into the model yielded non-significant regression coefficient, as indicated by t-statistic value of 1.04, which falls outside the acceptance region. Similarly, it may be further observed

Table 1. Results of correlation (r) among individual pickings and total yield

DV	IV	r	a	b_1	b_2	b_3	b_4	b_5	R^2(%)
6	1	0.84**	2.5487	0.3682	-	-	-	-	69.83
	2	0.51**	-0.0172	-	0.0759	-	-	-	25.96
	3	0.53**	-0.4903	-	-	0.2385	-	-	27.80
	4	0.39**	0.0992	-	-	-	0.1139	-	15.12
	5	0.15	0.4205	-	-	-	-	0.0387	2.20

** Significant at 1%

DV- Dependent Variable IV-Independent Variable

1-First picking (X_1) 2-Second picking (X_2) 3-Third picking (X_3)

4-Fourth picking (X_4) 5- Fifth picking (X_5) 6.Total yield (X_6)

Table 2. Results of multiple correlation (R) among pickings and total yield

DV	IV	R	a	b_1	b_2	b_3	b_4	b_5	R^2 (%)
6	1,2	0.88**	-1.6575	1.7039	1.9354	-	-	-	77.43
	1,3	0.72**	3.4425	1.2992	-	0.8668	-	-	60.09
	1,4	0.81**	3.2340	1.2362	-		1.0299	-	65.85
	1,5	0.79**	1.6152	1.4667	-	-	-	1.2010	62.56
	2,3	0.72**	6.2107	-	3.3277	1.1362	-	-	52.37
	2,4	0.75**	7.0901	-	1.3168		1.5209	-	58.90
	2,5	0.57**	7.9559		1.2598	-	-	1.0325	32.11
	3,4	0.74**	5.8523	-	-	1.4238	1.8123	-	54.60
	3,5	0.51**	8.5244	-	-	1.3072		1.0150	25.68
	4,5	0.39**	9.7716	-	-	-	1.2908	0.1404	15.25
	1,2,3	0.89**	-1.1389	1.4853	2.0963	0.3648	-	-	79.30
	1,2,4	0.88**	2.7167	1.0018	0.9228	-	1.1600		77.92
	1,2,5	0.84**	1.2488	1.2886	0.7775	-	-	1.2007	71.28
	2,3,4	0.81**	4.5190	-	2.4385	1.3419	1.3873	-	66.33
	2,3,5	0.73**	5.9351	-	1.3418	1.4194		1.0604	53.97
	3,4,5	0.74**	5.8802	-	-	-	-	-	54.64
	1,2,3,4	0.91**	-0.8506	1.2582	1.8455	0.5842	0.6845	-	82.08
	1,2,3,5	0.90**	0.6815	1.1200	0.8992	1.0148	-	1.1987	81.78
	1,3,4,5	0.89**	1.1059	1.1847		0.8509	0.8678	0.9536	79.58
	2,3,4,5,	0.85**	4.3604	-	3.4531	1.3611	1.5174	1.0970	71.72
	1,2,3,4,5	0.91**	-0.8571	1.2599	1.8408	0.5831	0.6835	0.0043	82.08

** Significant at 1%

DV- Dependent Variable IV-Independent Variable

1-First picking (X_1) 2-Second picking (X_2) 3-Third picking (X_3)

4-Fourth picking (X_4) 5- Fifth picking (X_5) 6.Total yield (X_6)

Table 3. Results of goodness of fit statistics along with the selected models

IV	R^2	MSE	Model and (t-statistic)	Significant IV
1,2	0.88	6.05	$Y = -1.66+1.7X_1+1.93X_2$ (5.44) (2.09)	X_1, X_2
1,2,3	0.89	6.01	$Y = -1.14+1.48X_1+2.09X_2+0.36X_3$ (3.95) (2.24)	X_1, X_2
1,2,3,4	0.82	5.69	$Y = -0.85+1.26X_1+1.84X_2+0.58X_3+0.68X_4$ (3.1) (1.98) (1.54) (1.3)	X_1, X_2
1,2,3,4,5	0.82	6.52	$Y = -0.85+1.26X_1+1.8X_2+0.58X_3+0.68X_4+0.004X_5$ (1.1) (0.005)	X_1

Figures in parentheses indicate t-statistic values
DV- Dependent Variable IV-Independent Variable
1-First picking (X_1) 2-Second picking (X_2) 3-Third picking (X_3)
4-Fourth picking (X_4) 5- Fifth picking (X_5)\ 6.Total yield (X_6)

that inclusion of an additional variable into the model results in non-significant regression estimates. Thus, results indicate that information from two pickings could predict the yield to an extent of 77 %. Further, corresponding regression coefficients were significant as indicated by the t-statistic values, which fall inside the acceptance region of 1.96. Moreover, the mean square error in reduction also strengthens our conclusion for identifying a model based on the first two pickings. Hence, the model developed showed that information from the first two pickings of a standing crop could be used to forecast rose yield considerably two months before final harvest. It may also be stressed here that as reported by Shamasundaran *et al* (2003), a minimum sample size of 20% of the population is required to get a good estimate to develop such a forecast model.

ACKNOWLEDGEMENT

The authors are grateful to Director, Indian Institute of Horticultural Research, Bangalore for providing all facilities to conduct this investigation.

REFERENCES

Agostid'no, R.B. and Stephens M.A.1986. *Goodness of Fit Techniques.* Marcel Dekker, NewYork 576p

Lewis-Beck, S. M.1993. Regression Analysis. Sage Publ., New York. 433p

Kvelseth,T.O 1985. Cautionary note about R^2. *The Amer. Stat.,* **39**:279-85.

Shamasundaran, K. S. and.Singh, K. P. 2003. Yield forecasting in tuberose (*Polyanthes tuberosa* Linn.) as effected by association of various characters. *J. Orn. Hort.,* **6**:372-75.

Shamasundaran, K. S. Venugopalan, R and Singh, K. P. 2003. Optimum sample size for yield estimation in certain commercial crops. *J. Orn. Hort.,* **6**:244-47

(MS Received 16 February, 2006 Revised 6 June, 2006)