

Strategies for Composite Forecast

S.C. Mehta, Ranjana Agrawal and V.P.N. Singh
Indian Agricultural Statistics Research Institute, New Delhi
(Received : January, 1998)

SUMMARY

An attempt has been made to develop methodology for obtaining a composite forecast by combining the forecasts obtained from different models. Various strategies of assigning appropriate weights (to the component forecasts) viz. equal weights, weights proportional to inverse of variances and weights based on variances and co-variances are suggested. The selection of a strategy for a particular situation depends upon the values of the parameters viz. the correlation coefficient and the variance ratio of the errors of the individual forecasts. Suitable strategies for various ranges of the parameters are given. The procedure has been illustrated through a case study on sugarcane crop.

Key words : Composite forecast, Growth-index.

1. Introduction

The paramount necessity of crop-yield forecasting especially for planning of agricultural produce needs no emphasis. Since the crop yield depends on many types of variables viz. weather factors, performance of plant during crop growth stages, agricultural inputs, it will be better if all these aspects are considered while forecasting crop yield. But sometimes it may not be feasible to develop a single model based on different types of data. Thus there is a need to develop statistical procedure for combining the forecasts of crop yield obtained from different procedures to get a composite forecast using all types of available independent information.

Dickson [2] studied composite forecast, obtained by combining different forecasts using a minimum variance criterion, had shown that composite forecast can be computed with smaller error variance than any of the components and also considered the sampling distribution of the weights to be attached to the components and of the error variance of the combined forecast. He derived exact expression for the minimum variance weight vector, provided a proof that error variance of the composite forecast is no greater than that of any

of the component forecasts and examined probability distribution of the weight estimators Dickson [2].

In the present paper, various strategies to combine the forecasts of crop yield obtained from different models have been suggested. The behaviour of the strategies and their theoretical as well as empirical comparisons has been studied.

2. Composite Forecast

The minimum variance composite forecast- y_c (Dickson [2], [3]) obtained by a linear combination of 'n' forecasts - y_i ; $i = 1, 2, \dots, n$; is as follows :

$$y_c = \sum_{i=1}^n w_i y_i = \mathbf{w}' \mathbf{y}; \mathbf{y} = (y_1, y_2, \dots, y_n)', \mathbf{w} = (w_1, w_2, \dots, w_n)'$$

where y_i = The forecast computed from i^{th} approach with error e_i distributed as $N(0, \sigma_{ii})$

w_i = The weight assigned to the i^{th} forecast

$$\mathbf{w} = (\Sigma^{-1} \mathbf{1}_n) (\mathbf{1}'_n \Sigma^{-1} \mathbf{1}_n)^{-1}; \mathbf{1}_n = (1, 1, \dots, 1)'$$

Σ = covariance matrix of the errors of the component forecasts

The variance of the errors of the composite forecast i.e. $(\mathbf{1}'_n \Sigma^{-1} \mathbf{1}_n)^{-1}$ is less than the variance of the errors of the forecast obtained from the individual model based on any of the n approaches (Dickson [2]).

3. Strategies to Determine the Weights

Though the value of w is optimum theoretically but in practice simple strategies of assigning the weights to component forecasts can be used which are discussed below :

3.1 Equal Weights

The composite forecast and its variance of the errors are given by

$$y_{cs1} = \frac{1}{n} \mathbf{1}'_n \mathbf{y} \text{ and } V_1 = \frac{1}{n^2} (\mathbf{1}'_n \Sigma \mathbf{1}_n)$$

This strategy is simple and its variance can be correctly estimated.

3.2 Weights Proportional to Inverse of the Variances

The composite forecast is developed as

$$y_{cs2} = \mathbf{b}'\mathbf{y}, \mathbf{b} = \sum_{i=1}^n \frac{1}{\sigma_{ii}} \left[\frac{1}{\sigma_{11}}, \frac{1}{\sigma_{22}}, \dots, \frac{1}{\sigma_{nn}} \right]$$

It gives larger weight to the forecast with smaller variance of errors in the forecast. The variance of the errors in the composite forecast is as follows

$$V_2 = (\mathbf{1}'_n \Sigma \mathbf{1}_n) \prod_{i=1}^n \sigma_{ii} \left[\sum_{i=1}^n \left\{ \prod_{\substack{j=1 \\ j \neq i}}^n \sigma_{jj} \right\} \right]^{-2}$$

3.3 Weights Depend on Variances and Co-variances

The weights are proportional to the row sums of the inverse of the co-variance matrix.

i.e. $\mathbf{w} = k \Sigma^{-1} \mathbf{1}_n, k = (\mathbf{1}'_n \Sigma^{-1} \mathbf{1}_n)^{-1}$

Under this set up, the composite forecast and its variance of errors are as follows :

$$y_{cs3} = k \mathbf{1}'_n \Sigma^{-1} \mathbf{y} \text{ and } V_3 = (\mathbf{1}'_n \Sigma^{-1} \mathbf{1}_n)^{-1}$$

Theoretically, this is the minimal variance composite forecast.

4. Composite Forecast in Particular Case when n = 2

Using the three strategies discussed in the preceding section, the composite forecasts and their variances of errors, for n = 2 are given below :

$$y_{cs1} = 0.5 y_1 + 0.5 y_2, V_1 = 0.25 (\sigma_1^2 + \sigma_2^2 + 2 r \sigma_1 \sigma_2)$$

$$y_{cs2} = \left[\frac{1}{\sigma_1^2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \right] y_1 + \left[\frac{1}{\sigma_2^2} \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) \right] y_2$$

$$V_2 = \sigma_1^2 \sigma_2^2 [\sigma_1^2 + \sigma_2^2 + 2 r \sigma_1 \sigma_2] (\sigma_1^2 + \sigma_2^2)^{-2}$$

$$y_{cs3} = [(\sigma_2^2 - r \sigma_1 \sigma_2) / (\sigma_1^2 + \sigma_2^2 - 2 r \sigma_1 \sigma_2)] y_1 + [(\sigma_1^2 - r \sigma_1 \sigma_2) / (\sigma_1^2 + \sigma_2^2 - 2 r \sigma_1 \sigma_2)] y_2$$

$$V_3 = \sigma_1^2 \sigma_2^2 (1 - r^2) / (\sigma_1^2 + \sigma_2^2 - 2 r \sigma_1 \sigma_2)$$

where r = corr. coeff. between errors in the first set of forecasts and those in the second.

5. Behaviour of Different Strategies-Theoretical Comparison

To see the performance of the strategies in different situations, the relative efficiencies E_{21} , E_{31} and E_{32} were computed as follows

$$E_{21} = \frac{V_1}{V_2} = \frac{(1 + VR)^2}{4VR}$$

$$E_{31} = \frac{V_1}{V_3} = \frac{(1 + VR)^2 - 4r^2 VR}{4VR(1 - r^2)}$$

$$E_{32} = \frac{V_2}{V_3} = \frac{(1 + VR)^2 - 4r^2 VR}{(1 + VR)^2(1 - r^2)}$$

where E_{ji} is the relative efficiency of j^{th} strategy over the i^{th} one and $VR = \sigma_1^2 / \sigma_2^2$

For different hypothetical values of the correlation coefficient 'r' and variance ratio VR of the errors of the forecasts from individual approaches, the values of the relative efficiencies E_{21} , E_{31} and E_{32} are presented in Table 1. These are computed when $VR \geq 1$, however, for $VR < 1$, the same Table can be used by taking $1/VR$ instead of VR. Although, the variance from the third strategy is always least but we may go for first or second strategy merely by foregoing a minor gain in efficiency over the third one. The choice of a strategy under various situations are discussed below :

- (i) The **first strategy** may be adopted, merely by sacrificing a maximum gain of 8% over third strategy in the situations

$$r \leq 0.70 \text{ \& \ } VR \leq 1.5 \text{ \ or \ } r \leq 0.95 \text{ \ \& \ } VR \leq 1.2$$

- (ii) When $r \leq 0.5$ and $2 \leq VR \leq 3$ or $r = 0.6$, $1.8 \leq VR \leq 2.5$ the **second strategy** appears to be a proper choice because the gain in efficiency over the first strategy will be in the range of 10-33% whereas the meager sacrifice of gain over the third strategy will be upto 8% only. In case $r = 0$ & $VR \geq 2$, still the second strategy may be taken because the increase in gain over the first strategy will increase at the rate of more than 20% for a unit increase in VR.

- (iii) The **third strategy** will be appropriate in the remaining cases.

Table 1. Relative efficiencies of the strategies for hypothetical values of 'r' & 'VR'

'r'	Relative efficiencies E_{ji}	'VR'									
		1.0 1/1	1.17 7/6	1.20 6/5	1.25 5/4	1.33 4/3	1.50 3/2	2.00 2/1	3.00 3/1	4.00 4/1	5.00 5/1
0.0	E_{21}	1.00	1.01	1.01	1.01	1.02	1.04	1.13	1.33	1.56	1.80
	E_{31}	1.00	1.01	1.01	1.01	1.02	1.04	1.13	1.33	1.56	1.80
	E_{32}	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.50	E_{21}	1.00	1.01	1.01	1.01	1.02	1.04	1.13	1.33	1.56	1.80
	E_{31}	1.00	1.01	1.01	1.02	1.03	1.05	1.17	1.44	1.75	2.07
	E_{32}	1.00	1.00	1.00	1.00	1.01	1.01	1.04	1.08	1.12	1.15
0.60	E_{21}	1.00	1.01	1.01	1.01	1.02	1.04	1.13	1.33	1.56	1.80
	E_{31}	1.00	1.01	1.01	1.02	1.03	1.06	1.19	1.52	1.88	2.25
	E_{32}	1.00	1.00	1.01	1.01	1.01	1.02	1.06	1.14	1.20	1.25
0.70	E_{21}	1.00	1.01	1.01	1.01	1.02	1.04	1.13	1.33	1.56	1.80
	E_{31}	1.00	1.01	1.02	1.02	1.04	1.08	1.24	1.65	2.10	2.57
	E_{32}	1.00	1.01	1.01	1.01	1.02	1.04	1.11	1.24	1.35	1.43
0.80	E_{21}	1.00	1.01	1.01	1.01	1.02	1.04	1.13	1.33	1.56	1.80
	E_{31}	1.00	1.02	1.02	1.03	1.06	1.11	1.35	1.93	2.56	3.22
	E_{32}	1.00	1.01	1.01	1.02	1.04	1.07	1.20	1.45	1.64	1.79
0.90	E_{21}	1.00	1.01	1.01	1.01	1.02	1.04	1.13	1.33	1.56	1.80
	E_{31}	1.00	1.03	1.04	1.06	1.11	1.22	1.66	2.76	3.96	5.21
	E_{32}	1.00	1.03	1.04	1.05	1.09	1.17	1.47	2.07	2.53	2.89
0.95	E_{21}	1.00	1.01	1.01	1.01	1.02	1.04	1.13	1.33	1.56	1.80
	E_{31}	1.00	1.06	1.08	1.13	1.21	1.43	2.38	4.42	6.77	9.20
	E_{32}	1.00	1.05	1.08	1.11	1.19	1.37	2.03	3.32	4.33	5.11

N.B. E_{ji} : Relative Efficiency of the j-th strategy over i-th one.

'r' : The correlation coefficient between errors of forecasts based on two different models.

'VR' : The ratio of the variances of the errors of the forecasts based on two different models.

6. Illustration—A Case Study

The methodology discussed in the preceding sections has been applied for computing composite forecast of sugarcane yield by combining the two forecasts – one based on biometrical characters of the plants observed during different growth stages while the other on time series data of various weather factors.

6.1 Material

To develop the model based on biometrical characters, the secondary data on the variables – number of canes per plot, height of the cane, girth of the cane, length and width of the third leaf (from top) recorded at various stages of crop growth during the periods -II (7-8 months after plantation) to V (10-11 months after plantation) pertaining to the project– a pilot survey on pre-harvest forecasting of yield of sugarcane in Kohlapur district of Maharashtra, from 1977-78 to 1979-80, conducted by Indian Agricultural Statistics Research Institute, New Delhi, was utilised.

For the development of the model using weather factors, the daily data (for the district) on weather factors – maximum temperature, minimum temperature, relative humidity, rainfall, number of rainy days and wind speed were procured, for the period 1951-80, from Regional Meteorology Centre, Bombay & Indian Meteorological Department, Pune. The corresponding yield data was obtained from the various reports of Agricultural Situation in India, Directorate of Economics and Statistics, New Delhi.

6.2 Model Based on Growth Indices of Biometrical Characters

The models were developed through regression approach taking the growth indices as regressors. The growth indices were obtained as weighted accumulations of two derived biometrical characters – volume of the cane and a variable proportional to area of the third leaf, the weights used being the correlation coefficients between the yield and the relevant biometrical characters in various periods of the crop growth phases (Jain *et al.* [4]). Thus the model was of the form :

$$\text{Model 6.2} \quad y = \beta_0 + \sum_{i=1}^p \beta_i G_i + e, \quad G_i = \sum_{s=n_1}^{n_2} r_{is} x_{is}$$

where y = Sugarcane yield (q/ha.)

G_i = Growth index of the i -th derived biometrical character

s = period-identification

n_1 = initial period

n_2 = final period

p = number of derived biometrical character

r_{is} = correlation coefficient between yield and the i -th derived character in s -th period

e = random error term distributed independently and normally with mean zero and constant variance

x_{is} = value of i -th character in the s -th period

These models were developed for appropriate combination of periods viz. II & III; II to IV, III & IV; II to V, III to V and IV & V.

6.3 Model Based on Weather Factors

Before taking up the development of forecast model based on weather variables, the trend in the yield data was examined by fitting the regression equation with yield as dependent variable and year as independent variable. With the help of stepwise regression technique, the following type of model was developed taking regressors as appropriate simple and weighted accumulations of different weather variables/various products of weather variables, the weights used being correlation coefficients between yield (adjusted for the trend, if any) and the relevant weather variable/product of two weather variables (Agrawal *et al.* [1]).

Model 6.3

$$y = \beta_0 + \sum_{i=1}^p \sum_{j=0}^l \beta_{ij} \sum_{t=1}^f (r_{it}^j x_{it}) + \sum_{\substack{i,i'=1 \\ i \neq i'}}^p \sum_{j=0}^l \beta_{ii'j} \sum_{t=1}^f (r_{ii't}^j x_{it} x_{i't}) + \gamma T + e$$

where y = sugarcane yield (q/ha.)

t = fortnight-identification

p = number of weather variables

f = fortnight of forecast

T = year

e = random error term distributed independently and normally with mean zero and constant variance

x_{it} = value of i -th weather variable in t -th fortnight

r_{it} = correlation-coefficient between x_i and y in t -th fortnight after adjusting yield for trend effect, if any

$r_{ii't}$ = correlation-coefficient of yield and the product of the two weather variables x_i and $x_{i'}$ in t -th fortnight after adjusting yield for trend effect, if any

Here $i = 1, 2, \dots, 7$ corresponds to maximum temperature, minimum temperature, relative humidity, rainfall, number of rainy days, wind speed and temperature difference.

6.4 Results and Discussion

The models 6.2 and 6.3 were developed using stepwise regression technique. The talukwise forecasts were worked out by substituting the corresponding regressors in both the models. The deviations of these talukwise forecasts from the observed ones were computed. Using these taluk-wise deviations, the correlation coefficient 'r' between the deviations (errors) as well as the variances σ_1^2 and σ_2^2 of the errors of the forecasts from two models were worked out.

For studying the behaviour of different strategies (Section 5), the ratio of the variances i.e. $VR = \sigma_1^2/\sigma_2^2$ were also computed.

Using the three strategies (Section 4) the composite forecast yield had been worked out and compared with the individual forecasts computed from the models based on growth indices of derived bio-characters (Section 6.2) and the weather factors (Section 6.3). These comparative results can be seen in Table 2. A closer look on the standard errors of the forecasts confirmed the theoretical aspect – the standard error of the composite forecast obtained by using the third strategy is least when compared with those associated with the cases relating to the individual forecasts as well as the standard errors of the composite forecasts computed by using first and second strategies. The criterion (Section 5), for selection of a strategy has been applied for computing composite forecast for periods III, IV & V in the following cases :

- Case A :** Year of forecast 1978-79 based on the models for the years-1977-78 (biometrical characters'-data) & 1951-77 (weather-data)
- Case B :** Year of forecast 1979-80 based on the models for the years-1977-78 (biometrical characters'-data) & 1951-77 (weather-data)
- Case C :** Year of forecasts 1979-80 based on the models for the years-1978-79 (biometrical characters'-data) & 1951-78 (weather-data)

The results are presented in Table 3. The selection of a strategy for a particular situation [viz. periods III to V in cases A, B and C] depends upon the values of the parameters $\{r, VR\}$, where 'r' is the correlation coefficient

Table 2 : Forecasts from the models based on biometrical characters' data & weather data alongwith composite forecasts from three strategies.

Year of forecast	Period of forecast	Forecast Yield (q/ha.)					Ob- served yield
		Y_{cs1}	Y_{cs2}	Y_{cs3}	Y_b	Y_w	Y_0
1978-79*	III	806.98 (37.48)	799.65 (36.90)	757.40 (35.29)	848.54 (42.25)	765.43 (35.35)	806.37
	IV	791.33 (36.86)	785.94 (36.70)	770.26 (36.50)	848.78 (41.29)	733.89 (37.58)	
	V	790.50 (34.94)	793.19 (34.90)	799.04 (34.87)	850.09 (37.19)	730.91 (38.91)	
1979-80*	III	850.54 (48.16)	854.87 (48.02)	876.17 (47.71)	908.12 (48.40)	792.96 (52.18)	878.39
	IV	842.88 (53.66)	863.77 (51.52)	906.51 (49.69)	917.54 (49.80)	768.23 (66.40)	
	V	836.00 (56.66)	865.39 (53.12)	914.55 (50.71)	920.50 (50.75)	751.50 (72.95)	
1979-80**	III	884.46 (53.23)	863.86 (51.80)	827.15 (50.79)	974.08 (65.18)	794.84 (51.57)	878.39
	IV	877.94 (55.59)	871.89 (55.50)	867.45 (55.48)	983.09 (67.77)	772.80 (63.98)	
	V	866.62 (64.35)	859.46 (64.22)	852.23 (64.18)	978.94 (76.58)	754.31 (71.84)	

N.B. : Figures in paranthesis are the respective standard errors.

(*) : Forecasts based on the models 1977-78 (biometrical characters' data) & 1951-77 (weather data)

(**) : Forecasts based on the models 1978-79 (biometrical characters' data) & 1951-78 (weather data)

and 'VR' is the variance ratio of the errors of the individual forecasts. For the sake of brevity, [IV(B)-{0.5, 2.00}] will represent the period IV in case B with values $r = 0.5$ & $VR = 2.00$. A persual of the findings reveal that first strategy will be a proper choice for computing the composite forecast in periods IV & V of cases A & C and III of case B because [V(A)-{0.69, 1.10}], [IV(C)-{0.42, 1.12}] & [V(C)-{0.50, 1.14}] and [IV(A)-{0.75, 1.20}] & [III(B)-{0.83, 1.16}] satisfy the conditions [$r \leq 0.70$ & $VR \leq 1.5$] and [$r \leq 0.95$ & $VR \leq 1.2$] respectively. In [III(A)-{0.87, 1.43}], where the correlation coefficient is high, third strategy appears to be appropriate because

Table 3 : Values of 'r', 'VR' and relative efficiencies of three strategies in different years/periods of forecast.

Year of fore-cast	Period of forecast - III				
	r	VR	E ₂₁	E ₃₁	E ₃₂
1978-79(*)	0.87	1.43	1.03	1.13	1.10
1979-80(*)	0.83	0.86	1.01	1.02	1.01
1979-80(**)	0.66	1.60	1.06	1.10	1.04
	Period of forecast - IV				
1978-79(*)	0.75	1.20	1.01	1.02	1.01
1979-80(*)	0.70	0.56	1.09	1.17	1.08
1979-80(**)	0.42	1.12	1.00	1.00	1.00
	Period of forecast - V				
1978-79(*)	0.69	0.91	1.00	1.00	1.00
1979-80(*)	0.67	0.48	1.14	1.26	1.10
1979-80(**)	0.50	1.14	1.00	1.01	1.00

N.B. (*) : Based on the models 1977-78 (biometrical characters' data) and 1951-77 (weather data)

(**) : Based on the models 1978-79 (biometrical characters' data) and 1951-78 (weather data)

the gain in efficiency will be 13% over first and 10% over second strategy. For [IV (B)—{0.70, 1.79}], selection of second strategy gives 9% gain over first, however, if we go for third strategy the gain over first and second will be 17% and 8% respectively. The similar argument is applicable in the identification of a strategy in [V (B)—{0.67, 2.08}] where if we choose second strategy, the gain (over first) will be 14% whereas if we apply third strategy the efficiency will be 26% and 10% over first and second strategy respectively. The selection of strategy in [III (C)—{0.66, 1.60}] gives rise to a marginal situation — the second strategy over the first one merely gives 6% gain whereas if we go for the third strategy the gain will be only 10% and 4% over first and second strategy. In such a situation strategy third may be best though the gain is marginal but for the sake of simplicity even the first strategy can also be adopted. Thus the selection of a strategy for a particular situation depends upon the values of the statistical parameters viz. the correlation coefficient between the errors of the individual forecasts obtained from different models as well as the ratio of their variances.

REFERENCES

- [1] Agrawal, Ranjana, Jain, R.C., Jha, M.P. (1986). Models for studying rice crop-weather relationship. *Mausam*, **37**(1), 67-70.
- [2] Dickson, J.P. (1973). Some statistical results in the combination of forecasts. *Operational Research Quarterly*, **24**(2).
- [3] Dickson, J.P. (1975). Some comments on the combination of forecasts. *Operational Research Quarterly*, **26**(1).
- [4] Jain, R.C., Agrawal, Ranjana, Jha, M.P. (1985). Use of growth indices in yield forecast. *Biometrical Journal*, **27**(4), 435-439.