Jour. Ind. Soc. Ag. Statistics
51(1), 1998 : 61–72

Forecasting of Crop Yields using Second Order Markov Chains

R.C. Jain and V. Ramasubramanian[1]
IASRI, New Delhi-12
(Received : January, 1997)

SUMMARY

A second order Markov chain model has been developed for forecasting of sugarcane yield through which, it was possible to use data from two stages simultaneously. This model has been found better than the models in use i.e. first order Markov chain model and the regression model.

Key words: Second order Markov chain (SOMC), Transition probability matrix (t.p.m.), Composite stage, Yield forecast.

1. Introduction

The need for pre-harvest forecasting of crop yields need hardly be emphasized. Most of the earlier studies carried out in this respect utilise regression models (Agrawal *et al* [1], Jain *et al* [4], [5]). Hocking and Pendleton [3] and several other authors have discussed problems faced by using regression model, and again the remedies given by them are difficult to implement and also not very satisfactory. Matis *et al* [8] proposed an alternative approach based on Markov chain theory to forecast crop yields. This method overcomes some of the drawbacks of the regression model. This method is completely non-parametric and is robust against outliers and extreme values.

A Markov chain is constructed to provide a forecast distribution of crop yield for the various crop condition states at selected stages of plant life. Matis *et al* ([8], [9]) and Jain and Agrawal [6] developed Markov chain model using one stage data at a time for forecasting of crop yields. This paper attempts to develop second order Markov chain model through which it is possible to use data from two stages simultaneously.

As opposed to first order Markov chains whose simple dependence structure is very restrictive, the second order Markov chains (SOMC) can provide a more realistic model which assumes that the future depends on the

[1] Ph.D. student, IASRI, New Delhi-12.

present as well as on the most recent past. We were motivated to use second order Markov chains (SOMC) based on our earlier studies (Jain *et al* [4], [5], [7]) in which inclusion of data from two or more stages has improved the forecast model.

2. *Second Order Markov Chains (SOMC)*

In second order Markov chain $\{X_n, n = 0, 1, 2, \ldots\}$ with states (1, 2), transition probabilities can be written as

$$p_{ijk} = P(X_n = k/X_{n-1} = j, X_{n-2} = i) \ (i, j, k = 1, 2)$$

It is assumed that the future depends on present as well as on the most recent past. The one-step probabilities can be arranged as

(k) (ij)	1	2
11	p_{111}	p_{112}
12	p_{121}	p_{122}
21	p_{211}	p_{212}
22	p_{221}	p_{222}

Following Bhat [2], the above transition probability matrix (t.p.m.) can be written in a convenient form for matrix operations as

(k) (ij)	11	12	21	22
11	p_{111}	p_{112}	0	0
12	0	0	p_{121}	p_{122}
21	p_{211}	p_{212}	0	0
22	0	0	p_{221}	p_{222}

This t.p.m. is that of a first order Markov chain whose states are the composite states {11, 12, 21, 22} and its analysis can be done in the usual manner. When the number of states is m (in our case, two), then the number of probability elements which could be non-zero in the t.p.m. of a q-th (in our case 2nd) order Markov chain will be m^{q+1} (here $2^{2+1} = 8$) arranged in a matrix of size $m^q \times m^q$ (here $2^2 \times 2^2 = 4 \times 4$). We note here that the technique above can always be extended to higher order Markov chains in a straight forward manner.

3. *Some Terminology*

Various terms used in SOMC are defined below.

3.1 Composite Stages

Data on the various stages, when combined two by two, give rise to composite stages i.e. the composite stages of SOMC are denoted as S_r for $r = 1, 2, 3, 4$ with S_r obtained through combination of original stages s_r and s_{r+1} with r ranging from 1 to 4. S_5 denotes the harvest stage i.e. the original stage s_6.

3.2 Composite States

States in a composite stage are the combination of states of individual stages involved in the composite stage. For example, in S_1 two stages s_1 and s_2 were combined. Suppose stage s_1 has m states and stage s_2 has n states then S_1 will contain mxn states, called as composite states. S_5 will have ten states if deciles are used for yield intervals (pertaining to stage s_6).

3.3 Forecast Distribution

Let n_r, for $r = 1, 2, ..., r_1$ denote the number of composite states at the commencement of composite stage r. Let $A_{r,\ r+1}$ $(r = 1, 2, ..., r_1 - 1)$ denote the $n_r \times n_{r+1}$ transition matrices which gives the transition probabilities of a group of plants moving from any possible composite state of composite stage r to any possible composite state of composite stage r+1, each row summing to unity. These t.p.m. 's will then be used to construct a final forecast matrix. For details see Jain *et al* [6]. This forecast matrix can be used to forecast crop yields. Each row of it represents a crop composite state at a given crop composite stage. Each column of it represents a different yield interval. The values in each row of it are the estimated probabilities of the crop producing a final yield within each of the yield intervals. Thus, each row of it is a predicted yield distribution for a given composite state and composite stage, which may be analysed to get mean and standard error of the forecast.

3.4 Standard Error of Forecast

Standard error of mean yield forecast at composite stage r was worked out as

$$\text{S.E.} = \frac{1}{\left[\sum_{k=1}^{n_r} f_{rk}\right]^{\frac{1}{2}}} \frac{\left[\sum_{k=1}^{n_r} f_{rk}\ (Y_{Frk} - Y_{Ork})^2\right]^{\frac{1}{2}}}{\left[\sum_{k=1}^{n_r} f_{rk}\right]^{\frac{1}{2}}}$$

where f_{rk}, Y_{Frk} and Y_{0rk}denote number of observations, yield forecast and observed yield respectively, corresponding to the k-th composite state of the r–th composite stage.

4. *Illustration*

The two year data on biometrical characters and yield collected by Indian Agricultural Statistics Research Institute, New Delhi, in 1977-78 and 1978-79, under the pilot study on pre-harvest forecasting of sugarcane yield in Meerut district in Uttar Pradesh (U.P.) were utilised for the study. A stratified multistage sampling design was adopted for collection of field data. The biometrical characters included in the study were: number of shoots/millable canes per plot, plant height, girth of cane and width of third leaf from the top. The first character was measured on a whole plot basis while for the other ones, two clumps located at the diagonally opposite corners of the plot were used. Plot size of 3 crop rows × 4m (approx. 7.8 m^2) was used in the study. The first observation was recorded at about 3 months after planting and thereafter observations were recorded at an interval of one month upto 8 to 9 months of crop growth. The recording of last observation coincided with the harvest, except for width of the third leaf which was recorded up to 6 months after planting. Girth denotes circumference of the cane at the middle point and was measured 5 months onwards. At harvest, weight of canes was also recorded. Sampling units and plants were kept fixed for all the successive occasions. The biometrical data were collected by village-level workers.

In all 144 fields data were available in 1977-78 whereas 156 fields data were available in 1978-79. The various periods of observations i.e. 3-4, 4-5, 5-6, 6-7, 7-8 months after planting and at harvest have been denoted as stages s_1, s_2, s_3, s_4, s_5 and s_6 respectively.

5. *Model Formulation*

In the present study, variables have been selected on the basis of scatter plots. In the first stage the variable selected was number of shoots/millable canes per plot (referred here after as plant population) whereas plant height and plant population were the variables selected at other stages of crop growth.

Let X_{ij} (i = 1, 2; j =1, 2, 3, 4, 5) denote the selected biometrical character where the first subscript denotes the biometrical character and the second, the stage in which observations on that character were taken. The states were formed only on the basis of medians. Earlier workers in their approaches have also formed the states on the basis of quartiles and again based on combination of medians and quartiles. However, in the present study such possibilities were not attempted as it would have increased the number of states within a composite stage very rapidly and might have led to large number of zeroes appearing in the t.p.m.'s.

The model was developed upon 1977-78 data and forecasting of yield based on this was made for 1978-79.

6. *Results and Discussion*

6.1 Yield Distributions

Yield distribution was made on the basis of deciles of observations on yield of the 144 fields data available on 1977-78 upon which the SOMC model was to be built up. These quantitative intervals of yield will represent the columns of the forecast matrix. The midpoints of these class intervals will be used to find the predicted yield distribution at each composite stage.

Table 1 : Yield distribution, 1977-78 data (kg/plot)

Class interval	Frequency
11.00 – 41.90	14
41.90 – 50.35	15
50.35 – 55.93	14
55.93 – 60.81	15
60.81 – 66.51	14
66.51 – 71.03	15
71.03 – 76.45	14
76.45 – 83.20	15
83.20 – 94.57	14
94.57 – 113.40	14

Observed mean yield = 66.28 kg/plot

6.2 Plant Condition States

The plant condition states were defined on the basis of medians (i.e.)-Medians of plant population X_{1j} × Medians of plant height X_{2j} (j = 1, 2, 3, 4, 5). Note that X_{ij} denotes the i-th biometrical character in the original j-th stage for i = 1, 2; j =1, 2, 3, 4, 5.

The chain of stages (read composite stages) which form the second order Markov chain are

Stage S_1 (stages s_1 & s_2 of the original data combined)

Stage S_2 (stages s_2 & s_3 of the original data combined)

Stage S_3 (stages s_3 & s_4 of the original data combined)

Stage S_4 (stages s_4 & s_5 of the original data combined)

Stage S_5 (stage s_6, the harvest stage)

The following table gives details about the plant condition states defined for the 1977-78 data.

Table 2 : Plant condition states using medians of characters in the Markov chain model, 1977-78 data

Stages combined	No. of states using medians of characters	Characters used	Defining quantiles for states Q_2 (Median)
S_1 (1 & 2)	8	X_{11}	146.000
		X_{12}	156.500
		X_{22}	0.750
S_2 (2 & 3)	16	X_{12}	154.500
		X_{22}	0.750
		X_{13}	103.500
		X_{23}	1.255
S_3 (3 & 4)	16	X_{13}	103.500
		X_{23}	1.255
		X_{14}	109.000
		X_{24}	1.570
S_4 (4 & 5)	16	X_{14}	109.000
		X_{24}	1.570
		X_{15}	109.000
		X_{25}	1.800
S_5 (6)	10	Yield classes of Y	As given in Table 1

6.3 Transition Probability Matrices

For illustration purpose, we proceed to compute A_{12} (i.e.) the t.p.m. from stage S_1 to stage S_2. At stage S_1 since the biometrical characters X_{11}, X_{12}, X_{22} (refer Table 2) are divided on the basis of medians we get $(2 \times 2 \times 2)$ i.e. 8 states. Again, at stage S_2 since the biometrical characters X_{12}, X_{22}, X_{13}, X_{23} (refer Table 2) are divided on the basis of medians we get $(2 \times 2 \times 2 \times 2)$ i.e. 16 states. Hence A_{12} will evidently be a 8 × 16 matrix. A_{12} has been obtained by classifying the observations of each of the 8 condition (composite) states of composite stage S_1 into 16 condition (composite) states of composite stage S_2.

If we denote the medians of X_{11}, X_{12}, X_{22} as a, b, c respectively then the eight composite states of the composite stage S_1 are (here we have a =146; b = 154.5; c =0.75)

$$X_{11} \leq a,\ X_{12} \leq b,\ X_{22} \leq c$$
$$X_{11} \leq a,\ X_{12} \leq b,\ X_{22} > c$$
$$X_{11} \leq a,\ X_{12} > b,\ X_{22} \leq c$$
$$X_{11} \leq a,\ X_{12} > b,\ X_{22} > c$$
$$X_{11} > a,\ X_{12} \leq b,\ X_{22} \leq c$$
$$X_{11} > a,\ X_{12} \leq b,\ X_{22} > c$$
$$X_{11} > a,\ X_{12} > b,\ X_{22} \leq c$$
$$X_{11} > a,\ X_{12} > b,\ X_{22} > c$$

Likewise we can form the 16 composite states of composite stage S_2. This results in classification of observations in a 8 × 16 matrix. Then A_{12} (as given below) can be obtained by dividing each of the frequencies in a row by the corresponding row total.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
(1)	.857	.029	.114	0	0	0	0	0	0	0	0	0	0	0	0	0
(2)	0	0	0	0	.105	.474	.105	.316	0	0	0	0	0	0	0	0
(3)	0	0	0	0	0	0	0	0	.5	0	.4	.1	0	0	0	0
(4)	0	0	0	0	0	0	0	0	0	0	0	0	0	.333	0	.667
(5)	.5	0	.5	0	0	0	0	0	0	0	0	0	0	0	0	0
(6)	0	0	0	0	0	0.25	0	.75	0	0	0	0	0	0	0	0
(7)	0	0	0	0	0	0	0	0	.423	.077	.385	.115	0	0	0	0
(8)	0	0	0	0	0	0	0	0	0	0	0	0	.074	0	.074	.852

Thus the t.p.m. A_{12} is an 8 × 16 matrix. Likewise the t.p.m.'s A_{23}, A_{34}, A_{45} are of order 16 × 16, 16 × 16, 16 × 10 respectively.

6.4 Predicted Yield Distributions

The final transition matrix F can be obtained by multiplying the transition probability metrices as mentioned in the F matrix in Section 2. The matrices (A_{12} A_{23} A_{34} A_{45}), (A_{23} A_{34} A_{45}), (A_{34} A_{45}), (A_{45}) give the yield distributions for the 1977-78 data at different composite stages. Means of predicted yield distributions for each of the composite states of a composite stage were worked out. In order to get an idea, the results of composite stages using Medians of X_{11}, X_{12} and X_{22} at composite stage S_1 are presented below:

Table 3 : Means of predicted yield distributions at composite stage S_1 using Medians of X_{11}, X_{12} and X_{22}

Sl. No.	Plant condition stages (a = 146; b = 154.5; c = 0.75)	Predicted Means from stages S_1 (kg/plot)
1.	$X_{11} \leq$ a, $X_{12} \leq$ b, $X_{22} \leq$ c	48.49
2.	$X_{11} \leq$ a, $X_{12} \leq$ b, $X_{22} >$ c	65.51
3.	$X_{11} \leq$ a, $X_{12} >$ b, $X_{22} \leq$ c	38.96
4.	$X_{11} \leq$ a, $X_{12} >$ b, $X_{22} >$ c	67.06
5.	$X_{11} >$ a, $X_{12} \leq$ b, $X_{22} \leq$ c	55.31
6.	$X_{11} >$ a, $X_{12} \leq$ b, $X_{22} >$ c	71.29
7.	$X_{11} >$ a, $X_{12} >$ b, $X_{22} \leq$ c	41.85
8.	$X_{11} >$ a, $X_{12} >$ b, $X_{22} >$ c	72.27

6.5 Yield Forecasts and Their Standard Errors

To forecast the yield of 1978-79, the 1978-79 data were classified using observed values of biometrical characters as per the states of a composite stage of 1977-78. This resulted in number of observations of 1978-79 data falling in different states of a particular composite stage in 1977-78 data. Weighted mean of means of predicted yield distributions for each of the states of a composite stage were worked out, weights being the number of observations of 1978-79 data in different states of 1977-78 data. These forecasts were worked out at different composite stages of crop growth. Per cent standard errors of forecasts were also calculated. These are presented in Table 4.

Table 4 : Yield forecasts of 1978-79 data using 1977-78 data model

Composite stage	Original stages used for combination	Yield forecast (kg/plot)
S_1	1 & 2	52.18 (0.69)
S_2	2 & 3	48.49 (1.11)
S_3	3 & 4	54.27 (0.75)
S_4	4 & 5	53.60 (0.33)

Observed mean yield - 51.82 kg/plot

Note : Figures in parentheses indicate % standard errors.

Perusal of Table 4 indicates that this method can be successfully used in crop yield forecasting. As the crop growth advances, it is expected that the forecasts will stabilise and here the same phenomenon has been observed. This shows that as we go towards maturity, the forecast becomes more reliable.

6.6 Comparison of the Results Obtained Through the Study with Those from the Existing Methods

6.6.1 Comparison with regression model

Since multiple regression models have been employed in the past, it would be worthwhile to compare their forecasting ability with that of models based on SOMC. For that, yield was used as regressand and the biometrical characters as regressors at various stages of crop growth. Thus the model was built at each individual stage. The results thus obtained are presented in Table 5.

Table 5 : Yield forecasts of 1978-79 using 1977-78 regression model (kg/plot)

Stage	Forecasts kg/plot
1	59.86 (2.01)
2	**57.17 (1.64)**
3	56.34 (1.54)
4	54.84 (1.30)
5	53.15 (1.23)

Observed mean yield - 51.82 kg/plot

Note : Figures in parentheses indicate % standard errors.

A comparison between the forecasts through the present study and that of the regression model suggests that our model is more reliable for forecasting. Here again, we should note that the results of the regression model were satisfactory because outliers and extreme values were removed from the data in earlier study from where the data have been taken for the present study. But in case some outliers would have been present in the data then they would have distorted the parameter estimates and hence the forecasts, but the results obtained through Markov chain approach would remain stable as this is robust against outliers and other small disturbances. We note here that forecasts at the composite stage S_1 of our study and the forecasts of regression model at stage 2 are appropriate forecast to be compared as the former consists of stages 1 & 2 of the original data. Our forecast at stage S_1 was 52.82 kg/plot when compared with that at stage 2 of regression model which was 57.17 kg/plot. Again the per cent standard errors being small in the present case revealed that SOMC model is more reliable. A glance at Tables 4 & 5 will suffice to bring about the improvement of forecasts in the present case over regression model.

6.6.2 Comparison with first order Markov chain model

For comparing the results of our study with those obtained from the Markov chain model built using one stage data, the results obtained by Jain *et al* [6] (who used the same data which is used in the present study) are taken in Table 6.

Table 6 : Yield forecasts of 1978-79 using 1977-78 Markov chain model using one stage data at a time

Stage	States are defined as			
	M×M	Q×M	M×Q	Q×Q
1	63.56 (1.98)	64.78 (2.12)	65.37 (2.26)	62.81 (1.82)
2	**61.11** **(1.60)**	61.66 (1.67)	62.33 (1.73)	60.32 (1.52)
3	58.02 (1.23)	58.09 (1.26)	58.70 (1.14)	57.87 (1.34)
4	56.04 (1.04)	56.42 (1.07)	57.64 (1.04)	55.49 (1.06)
5	54.24 (0.88)	54.04 (0.78)	54.82 (0.57)	53.75 (0.97)

Observed mean yield = 51.82 kg/plot

Note : 1. Figures in brackets indicate % standard errors

2. M and Q stands for median and quartile respectively

The forecasts of 1978-79 using SOMC model developed on 1977-78 data is 52.18 kg/plot at composite stage S_1 (stages s_1 & s_2 of the original data combined) as against 61.11 kg/plot, the forecasts at stage s_2 which were obtained through Markov chain model which uses one stage data at a time using medians of characters for classifications of states. The results of other classifications of states in one stage Markov chain model are also inferior. The observed mean yield is 51.82 kg/plot. Also per cent standard error was small. Further, comparison between the forecasts of SOMC model at stages S_2, S_3, S_4 and one stage Markov chain model at corresponding stages s_3, s_4, s_5 reveals superiority of SOMC model both in terms of forecasts and per cent standard errors. Thus it becomes evident that SOMC model which uses combination of stages is better than Markov chain model which uses one stage data at a time.

7. Conclusion

Thus this study reveals that second order Markov chain model can be used for crop yield forecasting in preference over regression model and first order Markov chain model.

ACKNOWLEDGEMENTS

The authors thank the referees for their constructive suggestions and valuable comments.

REFERENCES

[1] Agrawal, Ranjana, Jain, R.C. and Jha, M.P., 1986. Models for studying rice crop-weather relationship, *Mausam*, **37(1)**, 67-70.

[2] Bhat, U.N., 1984. *Elements of applied stochastic processes*, John Wiley and Sons, New York, 125-137.

[3] Hocking, R.R. and Pendleton, O.J., 1983. The regression dilemma. *Comm. Stat. Theory and Methods*, **2(5)**, 497-527.

[4] Jain, R.C., Sridharan, H. and Agrawal, R., 1984. Principal component technique for forecasting Sorghum yield, *Ind. J. Agric. Sci.*, **54(6)**, 467-470.

[5] Jain, R.C., Agrawal, R. and Jha, M.P., 1985. Use of growth indices in yield forecast, *Biom. J.*, **27(4)**, 435-439.

[6] Jain, R.C. and Agrawal, R., 1992. Probability model for crop yield forecasting, *Biom. J.*, **34(4)**, 501-511.

[7] Jain, R.C., Garg, R.N., Gurcharan Singh and Ranjana Agrawal, 1994. Model to forecast yield of rice *(Oryza sativa)* using agro-spectral data, *Ind. J. Agri. Sci.*, **64(5)**, 320-323.

[8] Matis, J.H., Saito, T., Grant, W.E., Twig, W.C. and Ritchie, J.T., 1985. A Markov chain approach to crop yield forecasting, *Agricultural Systems,* **18,** 171-187.

[9] Matis, J.H., Birkett, T. and Bourdeaux, D., 1989. An application of Markov chain approach to forecasting cotton yield from surveys. *Agricultural Systems,* **29**, 357-370.