Statistics and Applications
Volume 12, Nos. 1&2, 2014 (New Series), pp. 1-13

Crop Yield Forecasting by Multiple Markov Chain Models and Simulation

Ramasubramanian V.[1] and Lalmohan Bhar[2]
[1]*Central Institute of Fisheries Education, Mumbai*
[2]*Indian Agricultural Statistics Research Institute, New Delhi*

Abstract

Markov chain models provide objective pre-harvest forecasts of crop yields with reasonable precisions well in advance aiding timely decisions. However, these models require sizable dataset for them to be stable and reliable. If the dataset is small, the estimated probabilities may not be precise with many zeroes occurring in the transition probability matrices. This will be more so with increase in the order of the Markov chain, because in such cases the number of states increases very rapidly. The present study deals with development of yield forecast models for sugarcane crop based on higher order (multiple) Markov chains built on a massive database. The results revealed that use of such models advanced the time of forecast for the same precision and the forecasts were found to be better when compared to that of first order Markov chain and regression based models. Moreover, when the order of Markov chain increases and/or the definition of states became finer, the mean yield forecasts approach the actual yield justifying the development of models with finer definitions of states of plant conditions. For the data under study, the principal component based third order Markov chain models are the models that give better forecasts.

Keywords: Crop yield; forecast; higher order (multiple); Markov chain; simulation; growth index; principal component; massive datasets

1 Introduction

Crop yield forecasts are quite useful in formulation of policies regarding stock, distribution and supply of agricultural produce to different areas of any country. Prominent among the methods of forecasting are based on models which employ regression, time series or stochastic approaches. These methods exploit data on crop biometrical characters, weather parameters, farmers' eye estimates, agro-meteorological conditions and/ or remotely sensed crop reflectance observations etc., utilized either separately or in an integrated fashion for forecasting crop yields at successive stages in the crop growing season. One among the various statistical approaches in vogue is the probability model based on Markov chain theory, which overcomes some of the well known drawbacks of the widely used regression model.

Matis *et al.* (1985) proposed a statistical methodology for forecasting crop yields at successive stages of the growing season of any crop using Markov chain theory. Later, Matis *et al.* (1989) applied the Markov chain approach for forecasting cotton yields. Jain and Agrawal (1992) developed First Order Markov Chain (FOMC) model for forecasting sugarcane yields. Agrawal and Jain (1996) demonstrated the performance of Markov chain model in forecasting sugarcane yields by using farmers' eye estimates in addition to the biometrical characters. Singh and Ibrahim (1996) attempted use of remotely sensed spectral data in Markov chain model for

obtaining pre–harvest wheat yield forecasts. Jain and Ramasubramanian (1998) developed a Second Order Markov Chain (SOMC) model using coarser states within the stages of the chain. Ramasubramanian and Jain (1999) have developed Markov chain models by using growth indices of the biometrical characters, through which, it was possible to use data from two stages simultaneously for forecasting sugarcane yields. Ramasubramanian *et al*. (2010) have developed yield forecast models for sugarcane crop using higher order (multiple) Markov chains by considering different possible combinations of various aspects viz., orders of Markov chain used, number of biometrical characters used and percentile definitions of plant condition states of biometrical characters within stages albeit using a smaller dataset. Patel *et al*. (2013) have developed user-friendly software for fitting Markov chain models.

The only assumption under an FOMC (or simply a Markov chain) set up is that the past (crop) conditions are statistically uninformative for predicting the future (crop yield forecasts), after the present (crop) conditions are known. As regards to higher order (say r-th order) Markov chains, the only assumption is that, after the present r stages conditions are known, the conditions previous to them are uninformative for predicting the future. However, the Markov chain methodology requires sizable dataset to estimate the transition probability matrices (TPMs). When the dataset is small, the estimated transition probabilities will not be precise. Hence, the stability and reliability of the model have been examined in this study, by using a massive dataset generated from the available dataset by simulation in order to find whether any improvement can be made in the forecast and its efficiency.

The higher order (multiple) Markov chains provide a more realistic and elaborate model as against the simple dependence structure of the FOMC model utilized by earlier workers. However, the use of multiple Markov chains will result in increase in the number of states very rapidly with increase in the order of Markov chain. So an attempt has also been made to study the effect of coarser or finer definition of plant condition states (and hence smaller or larger number of states) on the ultimate precision of the forecasts.

2 Notations and Preliminaries

Two years data on biometrical characters and yield collected by Indian Agricultural Statistics Research Institute, New Delhi under the pilot study on pre–harvest forecasting of sugarcane in Meerut district of U.P. state (Jha *et al.* 1979) were utilized for the study which is a standard data set, collected through scientific sampling, to represent actual growth characteristics and used by many researchers for developing/testing various forecast models (Jain and Agrawal, 1992; Agrawal and Jain, 1996; Jain and Ramasubramanian, 1998; Ramasubramanian and Jain, 1999; Ramasubramanian *et al*., 2010; Patel *et al*., 2013) . As the aim was to propose a simulation based Markov chain based forecasting methodology upon a massive dataset generated from this basic dataset and the proposed models can be conveniently compared with earlier models for judging their forecasting performance.

In all, 144 plots data were available in 1977–78 (hereinafter called first year) whereas 156 plots data were available in 1978–79 (second year). The selected biometrical characters are:– number of plants per plot (X_1) and average plant height per plant (X_2). The various stages of observations on X_1 and X_2 are 3–4, 4–5, 5–6, 6–7 and 7–8 months after planting. At harvest (12 months after planting), the actual yield i.e. weight of canes per plot (Y) were also available. These original stages are denoted by s_1, s_2, s_3, s_4, s_5 and the harvest stage, by s_6. Let X_{bi} denote the b^{th} biometrical character (b = 1, 2) in the i^{th} original stage (i = 1, 2, 3, 4, 5). For a multiple Markov chain model, say, an SOMC model developed, first year data on these stages, when combined two by two, gave rise to four composite stages. These composite stages of SOMC are denoted by S_r for r = 1, 2, 3, 4 with S_r obtained through combination of original stages as respectively (s_1, s_2),

(s_2, s_3), (s_3, s_4) and (s_4, s_5). The harvest stage s_6 is denoted as the stage S_5 in the SOMC chain. Note also that states in a composite stage are the combination of states of individual stages involved in the composite stage. For example, consider the composite stage S_1 of SOMC model that has been obtained by combining the two original stages s_1 and s_2 (of FOMC). Suppose original stage s_1 has k states and stage s_2 has l states then composite stage S_1 will contain (kxl) states as composite states with the final stage S_5 having ten states (as deciles of yield were used for defining states). Similarly, let T_1, T_2, T_3 denote the composite stages of a Third Order Markov Chain (TOMC) model formed by combining the original stages as respectively (s_1, s_2, s_3), (s_2, s_3, s_4) and (s_3, s_4, s_5). Let T_4 denote the harvest stage of TOMC model (original stage s_6)

Several Markov chain models have been developed in the present study based on different combinations of the following three criteria: (restricting the number of states in any TPM considered being at the most sixteen only for the present study)

(i) Order of Markov chain restricted to one, two or three (i.e.) FOMC, SOMC or TOMC.

(ii) Definition of (plant condition) states on the basis of median (M) of the biometrical character plant population (X_1) and median (M) of the biometrical character average plant height (X_2), refer this as definition MxM or quartiles (Q) of the biometrical character X_1 and quartiles (Q) of the biometrical character X_2, refer this as definition QxQ. (In addition, the options of considering median (M) of the biometrical character X_1 and quartiles(Q) of the biometrical character X_2, refer this as definition MxQ or quartiles(Q) of the biometrical character X_1 and median (M) of the biometrical character X_2, refer this as definition QxM have also been considered but only in developing FOMC models because they are not valid in case of higher order Markov chain models as such definitions of states will not make the latter's stages to follow conditional dependence of the Markov chain property).

(iii) Untransformed data (simulated large dataset considered as such) or transformed data viz. growth indices and principal components of the simulated large dataset (discussed subsequently).

Use of Growth Indices (GIs) and Principal Components (PCs) of the biometrical characters in multiple Markov chain models facilitates reduction in the number of variables within a composite stage so that the number of states in the multiple Markov chains is kept at manageable level. Thus by using GI or PC, it is possible to use both X_1 and X_2 together in the model even when defining finer plant condition states on the basis of quartiles(QxQ).

The growth indices are obtained as weighted accumulations of observations on biometrical characters in different stages, weights being the partial correlation coefficients between yield (Y) and biometrical characters (X_1 or X_2) at different stages of crop growth. The growth index of the b^{th} biometrical character is given by

$$G_{bi} = \sum_{c} r_{bc}\, x_{bc} \quad (i = S_1, S_2, S_3, S_4)$$

where i is used for composite stage identification, the summation extends over the initial and final stages considered in developing the index of the character, r_{bc} is the partial correlation coefficient between yield and b^{th} biometrical character at c^{th} stage, X_{bc} is the value of the b^{th} biometrical character at the c^{th} stage. At composite stage S_1, there will be GIs G_{11} and G_{21} instead of (X_{11}, X_{12}) and (X_{21}, X_{22}) and similarly for other stages S_2, S_3 and S_4.

Instead of using growth indices for developing multiple Markov chain models, alternatively PCs can also be used. Principal components are generally not warranted in situations where there are only two variables. However, considering variable-set for each composite stage and

transforming the original two variables variable-wise separately into PCs and taking only the first (significant) PC from each of them will rapidly reduce the number of states within (composite) stages of the multiple Markov chain models considered. That is, if say a PC based SOMC model is to be developed and if X_1 and X_2 are the two variables at each of the stages s_1 and s_2 then the PCs of X_{11} and X_{12} are obtained separately and that of X_{21} and X_{22} obtained separately. Let PC_{bi} denote the first principal component of the b^{th} biometrical character (b = 1, 2) for the i^{th} composite stage (i = S_1, S_2, S_3, S_4) for an SOMC model. At stage S_1 (stages s_1 ans s_2 combined), the PCs $PC_{11(1)}$ & $PC_{11(2)}$ from variables (X_{11}, X_{12}) and PCs $PC_{21(1)}$ & $PC_{21(2)}$ from variables (X_{21}, X_{22}) were obtained as

$$\begin{pmatrix} PC_{11(1)} \\ PC_{11(2)} \end{pmatrix} = \mathbf{K_1} \begin{pmatrix} X_{11} \\ X_{12} \end{pmatrix} \text{ and } \begin{pmatrix} PC_{21(1)} \\ PC_{21(1)} \end{pmatrix} = \mathbf{K_2} \begin{pmatrix} X_{21} \\ X_{22} \end{pmatrix}$$

where $\mathbf{K}_1$ is the matrix with rows as the characteristic vectors of the corresponding covariance matrix $\mathbf{\Sigma_1}$ (uncorrected for mean) of (X_{11}, X_{12}), the variance of $PC_{11(1)}$ is the 1^{st} characteristic root λ_1 of matrix $\mathbf{\Sigma_1}$, accordingly for $PC_{11(2)}$ etc. Hereinafter $PC_{11(1)}$ will be referred to as PC_{11} and so on. Accordingly, at composite stage S_1 of SOMC model the variables are PC_{11} and PC_{21} instead of (X_{11}, X_{12}) and (X_{21}, X_{22}) and similarly for other stages S_2, S_3 and S_4. Similar treatment is followed for developing PC based TOMC models.

3 Model Development

Multiple Markov chains have been developed by constructing Markov chains with orders greater that one which are collectively referred to as higher orders. Let any multiple Markov chain model have s (composite) stages. Let m_i, i = 1, 2, …, s denote the number of states defined on the basis of percentiles of observations of the selected biometrical characters within stage i. Let $A_{i,i+1}$ = (p_{kg}) (i = 1, 2, …, s – 1; k=1, 2,…, m_i; g=1, 2, …, m_{i+1}) denote the (m_ixm_{i+1}) TPMs which give the transition probabilities p_{kg} of a group of plants moving from any possible state k of stage i to any possible state g of stage (i+1), each row summing to one. For a given stage i, the predicted yield distributions (PYDs) are obtained as the product $\prod_{a=i}^{s-1} A_{a,a+1}$ which is of order (m_ixm_s). This matrix thus will give as many PYDs as the number of states in stage i. In the final (harvest) stage s, the yield information is available in as many class intervals as is the number of states in it i.e. m_s. The midpoints of these yield classes can be formed as a summary mean vector $\mathbf{y_m}$ of order (m_s x 1).

Means of PYDs for each of the states of a stage for the first year can be calculated by simply multiplying the PYDs with $\mathbf{y_m}$. Thus at each stage i, means of PYDs can be obtained as

$$\mathbf{y}^{(i)} = \left(\prod_{a=i}^{s-1} A_{a,\,a+1} \right) \mathbf{y}_m$$

of order (m_i x 1). The vector $\mathbf{y^{(i)}}$ will contain elements as $y_{ij's}$ with j ranging from 1 to m_i.

To forecast yield of second year, the second year data can be classified as per the states of a stage in first year. This will result in number of observations of second year say f_{ij}, falling in various states 1, 2, …, m_i of a particular stage i in first year. Weighted mean of means of predicted yield distributions for each of the states of a stage can be worked out, weights being the number of observations of second year falling in different states/stages of the first year data, which would give mean yield forecast Y_{Fi} at each stage i for the second year. This (weighted) mean yield forecast at stage i can be mathematically written as

$$y_{F_i} = \frac{\sum_{j=1}^{m_i} f_{ij}\, y_{ij}}{\sum_{j=1}^{m_i} f_{ij}} = \frac{1}{n}\sum f_{ij}\, y_{ij} \text{ (say)},$$

where $n = \sum_{j=1}^{m_i} f_{ij}$ is the total number of observations available in the second year at each stage i, however, since irrespective of any stage same number of observations are used hence 'n' not suffixed by i.

As the actual yield data upon second year were already available the forecast errors (FE) can be calculated at different stages as

$$\text{F.E.}(\bar{Y}_{Fi}) = \frac{1}{\left(\sum_{j=1}^{m_i} f_{ij}\right)^{1/2}} \frac{\left[\sum_{j=1}^{m_i} f_{ij}\,(Y_{Fij} - Y_{Oij})^2\right]^{1/2}}{\left(\sum_{j=1}^{m_i} f_{ij}\right)^{1/2}} \qquad \text{--- (1)}$$

where f_{ij}, Y_{Fij} and Y_{Oij} denote the number of observations, yield forecast and observed yield respectively, corresponding to the *j*th state of the *i*th stage.

4 Simulation of Massive Dataset

Multiple Markov chain modeling requires sizable dataset for the developed models to be stable and reliable. If the dataset is small, the estimated probabilities may not be precise with many zero probabilities occurring in the TPMs. This will be more so with increase in the order of the Markov chain, because in such cases the number of states increases very rapidly resulting in very large and unmanageable TPMs. Inflation of data through simulation enables all types of variations in the crop growth system to reflect in the dataset in order that it can mimic natural happenings connected with growth of sugarcane crop in question. Thus any data point that might have occurred but did not occur in the collected sample but for the limited sample size can also be taken into account. The simulated dataset is generated to appear similar to the available sample dataset.

To start with, at every stage in the actual crop growth process, the central limit theorem considerations render procedures that treat the data to have come from distributions as if it has come from distributions somewhat close to normal distribution as the standards due to large number of data points upon the available variables. Having assumed that the sample have come from a Gaussian population, the simulations can be based exclusively on the mean vector and the covariance matrix over all stages as the latter takes care of the interrelationships between the variables at various stages under consideration. Attempts were made to include transition probability matrices also as additional parameters but by doing so we could only get more of the same data with replications rather than similar data hence could not be considered. Thus even though the basic setting of the multivariate simulation lends itself very naturally to the utilization of normal distribution, the Markov conditional dependencies between the stages of the chain compel one to put restrictions between each of the same variables observed at different stages. In effect, parameter estimates of the simulated data set compared well with those of the original data set, notwithstanding the specificity of the underlying distribution which initially was used in the simulation. Thus it is relevant that the simulation algorithm used with the given restrictions represent actual growth characteristics up to the extent where complexities in modeling such

systems mathematically can be coped with. Thus a massive population similar to the one of the original dataset has been arrived.

The original dataset has 144 data points for first year upon eleven variables viz. Y, X_{ij} (b=1,2 ; i = 1,2,3,4,5) where b represents biometrical character and i the stage at which the character has been observed. Shapiro-Wilk test for normality revealed that all the variables follow normal distribution except for the variables X_{11} and X_{21} at stage-1. The first year dataset has been distended into a massive dataset of 5000 data points upon the same eleven variables preserving Markov chain properties. For achieving this, a computer program for simulation was written in Fortran language to generate eleven-variate normally distributed population by using the estimates of parameters mean $\boldsymbol{\mu}$ vector and variance-covariance $\boldsymbol{\Sigma}$ matrix calculated from the available dataset. In order to preserve Markov chain property in the simulated dataset as well the following restrictions were also imposed:

(i) The biometrical character plant population i.e.X_{1i} (i =1,2,3,4,5) is restricted to lie within the corresponding range (i.e. between maximum and minimum values) of the baseline dataset.
(ii) The biometrical character average plant height is restricted to fulfil the inequality condition $X_{21} \le X_{22} \le X_{23} \le X_{24} \le X_{25}$

Though the simulation has been subjected to the ascending inequality restriction for the variable plant height, the same could not be made for the other variable plant numbers as such a condition does not suit the latter. This is because, in the growth process of the crop, the number of plants per plot may increase or decrease according as new plants emerge out or some plants wilt/die. This causes the plant population to change with the growth of the plant, more so in the initial stages but not in an increasing manner. However, to simulate data set with similar settings, this variable is restricted to lie between its corresponding range in the available dataset.

The massive population has been obtained by using the algorithm given by Scheuer and Stoller(1962) which is briefly discussed here. Let the 11-variate vectors to be generated 5000 times be denoted by ($\mathbf{v} + \boldsymbol{\mu}$). To start with, assuming normality, generate 5000 standard normal multivariate vectors $u \sim N(\mathbf{0}, I_{11})$ whose elements $u_1, u_2, \ldots, u_{11}$ are independent standard normal variates. Then by setting

$$v = Cu \sim N(\mathbf{0}, CC') \equiv N(\mathbf{0}, \Sigma) \text{ (say), where } \mathbf{C} = (c_{ij}) \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} \text{ with n=11 and}$$

assuming $\Sigma = CC'$ calculate the elements of **C** matrix as follows:

(i) $c_{i1} = \sigma_{i1} / \sqrt{\sigma_{11}}$, $1 \le i \le 11$ (ii) $c_{ii} = \sqrt{\sigma_{ii} - \sum_{k=1}^{i-1} \mathbf{c}_{ik}^2}$, $1 < \mathbf{i} \le 11$

(iii) $c_{ij} = \left(\sigma_{ij} - \sum_{k=1}^{i-1} c_{ik}\, c_{jk} \right) \Big/ c_{jj}$, $1 < j < i \le 11$ (iv) $c_{ij} = 0$, $i < j \le 11$

Once c_{ij}'s are obtained, then v_i's can be obtained as $v_i = \sum_{j=1}^{i} c_{ij}\, u_j$, $i = 1,2,...11$ to obtain $\mathbf{v} = \mathbf{Cu}$.

Then $\mathbf{v} \sim N(\mathbf{0}, \Sigma)$ and hence $\mathbf{v} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \Sigma)$. Thus 5000 such **v** vectors each of size 11x1 were generated. The known estimates of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ from the available first year data were utilised to get the massive Markov chain population of 5000 data points upon the same eleven

variables. The estimates of μ and Σ of simulated population (Table 1) compared well with that of those obtained from the available first year dataset.

5 Results and Discussion

Summary statistics of the simulated first year data and available second year data are given in Table 1. Table 2 gives the percentiles of biometrical characters of untransformed and of their transformations viz. principal components and growth indices used in the definition of plant condition states for various Markov chain models developed upon simulated first year data. The results (mean yield forecasts) for second year based on various first year Markov chain models using simulated first year data are presented in Table 3. It also provides description about the definition and number of states, name and type of model developed. As an illustration, the model MMC6 (Table 3) is taken to discuss the steps involved in model development. MMC6 is an SOMC model wherein growth indices of the two biometrical characters X_1 (plant population) and X_2 (average plant height) are used with definition of plant condition states as MxM i.e. medians(M) of growth indices (Table 2). Growth indices are formed by using partial correlation coefficients between yield(Y) and biometrical characters of simulated data of first year at various stages i=1,2,3,4,5 (0.46387, 0.54700, 0.54192, 0.54660, 0.55892 for Y with X_{1i} and 0.20579, 0.42559, 0.35262, 0.31053, 0.41122 for Y with X_{2i}).

Thus, at composite stage S_1 (original stages s_1 and s_2 combined), the growth indices are given by $G_{11} = 0.46387\ X_{11} + 0.54700\ X_{12}$ and $G_{21} = 0.20579\ X_{21} + 0.42559\ X_{22}$. The different composite states within the composite stage, for say, S_1 , are accordingly formed by combination of the following conditions:

(i) G_{11} is classified on the basis of median (159.02) and thus we get two classes viz.

$G_{11} \leq 159.02$ and $G_{11} > 159.02$

(ii) G_{21} is classified on the basis of median (0.42) and thus we get two classes viz.
$G_{21} \leq 0.42$ and $G_{21} > 0.42$

In all, we get four 'composite states' within the composite stage S_1

(i) $G_{11} \leq 159.02$, $G_{21} \leq 0.42$
(ii) $G_{11} \leq 159.02$, $G_{21} > 0.42$
(iii) $G_{11} > 159.02$, $G_{21} \leq 0.42$
(iv) $G_{11} > 159.02$, $G_{21} > 0.42$

Similarly, for other composite stages in the SOMC model such composite states were defined. The observed frequencies of the plants moving from one plant condition class(composite state) of a composite stage, say, S_i (i=1, 2, 3, 4) to different condition classes(composite state) of the next composite stage S_{i+1} (i=1, 2, 3) were calculated. For instance, the frequency matrix of transition from S_1 to S_2 is given by,

		States of stage S_2			
		(i)	(ii)	(iii)	(iv)
	(i)	1189	80	125	11
States of stage S_1	(ii)	150	815	20	111
	(iii)	165	4	805	125
	(iv)	6	90	95	1210

And other transition frequency matrices i.e. from S_2 to S_3 and from S_3 to S_4 were obtained as matrices of order 4. The final frequency matrix i.e. from S_4 to S_5 has been obtained as a 4x10 matrix which is given by

States of stage S_4	States of stage S_5 (i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)
(i)	394	340	295	230	188	125	84	39	20	0
(ii)	45	65	100	140	135	75	105	75	35	10
(iii)	35	60	85	85	125	85	85	95	60	89
(iv)	10	24	41	64	100	195	254	254	329	424

These frequencies were then utilised to compute various transition probabilities by noting that row sum of any TPM is unity. Thus each row element was divided by its corresponding row sum. This gave rise to TPM's $A_{i\,,\,i+1}$ (i= S_1, S_2, S_3, S_4). Each matrix will be of order 4x4 except for the last TPM whose order is 4x10 because in the final stage, only information upon Y is there and it has been classified on the basis of deciles. Thus from the above frequency matrix, the TPM of transition from S_1 to S_2 has been obtained as

0.846	0.057	0.089	0.008
0.137	0.744	0.018	0.101
0.150	0.004	0.732	0.114
0.004	0.064	0.068	0.864

Similarly, other TPMs i.e. $A_{i,\,i+1}$ for i = S_2 and S_3 as (4x4) matrices and TPM $A_{i,\,i+1}$ for i=S_4 as a (4x10) matrix have been obtained which are not presented here for brevity.

Predicted yield distributions (PYDs) from first year data multiple Markov chain models were used to forecast yield of second year at various stages of crop growth. The product $\prod A_{i,i+1}$ for i= S_1, S_2, S_3, S_4 gives a 4x10 matrix which gives four PYDs, one for each of the four composite states in composite stage S_1. The product $\prod A_{i,i+1}$ for i= S_2, S_3, S_4 again gives a 4x10 matrix which gives four predicted yield distributions, one for each of the four composite states in composite stage S_2. Likewise PYDs for S_3 and S_4 have also been obtained. Means of PYDs for each of the composite states of a composite stage for the first year were worked out by simply multiplying these PYDs separately with the midpoints of the yield class intervals formed on the basis of deciles (which is a 10x1 vector; refer Table 2). Thus at each composite stage, means of PYDs are obtained as

composite states →	(i)	(ii)	(iii)	(iv)
Stage S_1 →	59.64	58.02	55.51	53.68
Stage S_2 →	67.68	67.88	65.77	65.34
Stage S_3 →	67.96	68.24	71.05	70.95
Stage S_4 →	77.71	79.27	81.39	83.18

To forecast yield of second year, the second year data were classified as per the composite states of a composite stage in first year. This resulted in number of observations falling in various composite states of a particular composite stage in first year. It may be noted that the row sum is equal to the number of data points (i.e. 156) in second year at each stage.

composite states →	(i)	(ii)	(iii)	(iv)
Stage S_1 →	105	24	21	6
Stage S_2 →	112	21	17	6

Stage $S_3 \rightarrow$	133	14	5	4
Stage $S_4 \rightarrow$	128	9	12	7

Weighted mean of means of predicted yield distributions for each of the states of a stage was worked out, weights being number of observations of second year in different states/stages of first year. This gave mean yield forecasts at each stage viz. at stage S_1 as 62.69 kg/plot, at stage S_2 as 61.28 kg/plot, at stage S_3 as 57.59 kg/plot and at stage S_4 as 57.01 kg/plot, for second year (in which actual yield was 51.82 kg/plot) based on the particular first year data multiple Markov chain model MMC6. The forecast errors (S.E.) were also calculated at different stages by using equation (1) of section 3 with the values of f_{ij}'s as the number of observations of second year falling in various states of a particular stage in first year (given above). The values of y_{ij}'s are nothing but the four mean p.y.d.'s at the j^{th} state of the i^{th} stage. And the forecast errors were 0.35, 0.29, 0.26 and 0.15 for the mean yield forecasts at stages S_1, S_2, S_3 and S_4 respectively. In the same fashion, the mean yield forecasts for second year along with their forecast errors can be obtained by developing other multiple Markov chain models upon first year data (Table 3). The results obtained using multiple regression models for second year by Ramasubramanian and Jain (1999) wherein the model was built on available first year data at each individual stage with yield as regressand and the biometrical characters as regressors are presented in Table 3 as model REG for comparison purposes.

The forecast at composite stage T_1 (original stages s_1, s_2 and s_3 combined), composite stage S_2 (original stages s_2 and s_3 combined) and original stage s_3 are appropriate stages to be compared as all these stages consist of the common ultimate stage s_3 in them. Likewise the stages (s_2 and S_1), (s_4, S_3 and T_2) and (s_5, S_4 and T_3) are the appropriate stages for comparisons with common ultimate stages s_2, s_4 and s_5 respectively in each of them. Perusal of the Table 3 reveals that considerable improvement in forecasts can be obtained by using higher orders viz. two (SOMC) and three (TOMC) in preference to first order Markov chain models. Thus it can be inferred that in most of the cases, finer definitions of states can give better forecasts. The table also reveals mostly better forecasts when SOMC models (GI or PC based) are used instead of FOMC models when comparing at corresponding same definitions of states. Both GI based and PC based SOMC models perform at par when compared among them as far as forecasts are concerned. The TOMC models MMC12 and MMC13 turn out to be the best models as the differences between observed mean yield i.e. 51.82 kg/plot and the forecast values 52.54, 52.49 and 52.48 kg/plot for MMC12 i.e. PC based TOMC model using MxM definition and 52.72, 52.53 and 52.30 kg/plot for MMC13 i.e. PC based TOMC using QxQ definition seem to be very small with lower forecast error.

6 Concluding Remarks

When the order of Markov chain increases and/or the definition of states became finer, the mean yield forecasts approach the actual yield justifying the development of multiple Markov chain models with finer definitions of states of plant conditions. Hence there is advancement in the time of forecast when multiple Markov chain models are used in preference to the existing models. For the data under study, the principal components based TOMC models are the models that give better forecasts.

There are many advantages of using Markov chain models over other conventional approaches such regression, time series modeling etc. Firstly, they require less stringent assumptions over other models. Rather, the only assumption it requires is that the future scenarios depend only upon the present conditions which is supposed to contain all information about the past due to the Markovian property and renders the past uninformative and how the present has been arrived at from the past is of no consequence. Moreover, quantiles (medians, quartiles etc.) are used in this approach as against the usually used mean elsewhere, hence in certain situations

such as presence of outliers, extreme values etc., this method is unaffected. In this era of remote sensing, multi-spectral data of reflectance from crops over the various stages of the growing period can be conveniently used for forming transition matrices to build Markov chain models (Singh and Ibrahim, 1996) rather than visiting the fields for taking measurements. Moreover, at every stage of the crop growth period, forecasts can be obtained and hence more informative, even though intuitively, the forecasts at later periods should be more reliable. In addition, this method can be said to be 'model-free' with just the states, stages and transition probability matrices and final conditions yielding the crop forecasts, instead of imposing model equations for the data collected. However, there are certain limitations of Markov chain models as well. While much of theory is well established for usual methods like regression, the properties of Markov chain forecasts are yet to be studied in depth. For instance, for calculating prediction interval of future observations during model fitting itself (as can be obtained for regression forecasts, see Montgomery *et al*., 2012, pages 33-34), such formula are not readily available in case of Markov chain approach and one has to get contented with standard error of forecasts (as used in this study) for which the actual values of the future observations are also required. Thus the precision in the forecast using this approach requires separate evaluation and further study. Also the cross sectional cum time series data structure of this approach sometimes makes it difficult for the practitioners for fitting such models. However, customized software (Patel *et al*., 2013) on Markov chain modeling are available nowadays for fitting these models very easily once the data set is ready.

References

Agrawal, R. and Jain, R..C. (1996). Forecast of sugarcane yield using eye estimate along with plant characters, *Biom. J.*, **38**, 731–39.

Jain, R.C. and Agrawal, R. (1992). Probability model for crop yield forecasting, *Biom. J.*, **34**, 501–11.

Jain, R,.C. and Ramasubramanian, V. (1998). Forecasting of crop yields using second order Markov chains, *Jour. Ind. Soc. Agric. Stat.*, **51**, 61-72.

Jha, M.P., Iyer, V.N., Agrawal, R. and Chandrahas (1975–79). *Pilot studies on pre–harvest forecasting of yield of sugarcane in Meerut district of U.P.*, Project report, I.A.S.R.I., New Delhi.

Matis, J.H., Saito, T., Grant, W.F. Iwig, W.C. and Ritchie, J.T. (1985). A Markov chain approach to crop yield forecasting, *Agricultural Systems*, **18**, 171–87.

Matis, J.H., Birkett, T. and Boudreaux, D. (1989). An application of the Markov chain Approach to forecasting cotton yield from surveys, *Agricultural Systems*, **29**, 357–70.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2012*). Introduction to linear regression analysis*, fifth edition, John Wiley & Sons, New Jersey.

Patel, R.M., Goyal, R. C., Ramasubramanian, V. and Marwaha, S. (2013). Markov chain based crop forecast modeling software, *Jour. Ind. Soc. Agric. Stat.*, **67(3)**, 371-79.

Ramasubramanian V., Agrawal, R. and Bhar, L. (2010). Crop forecasting using multiple Markov chains, *Assam Stat. Rev.*, **24(1)**, 37-56.

Ramasubramanian, V. and Jain, R.C. (1999). Use of growth indices in Markov chain models for crop yield forecasting, *Biom. J.*, **41**, 99–109.

Scheuer, E M. and Stoller, D. S. (1962), On the generation of normal random vectors, *Technometrics*, **4**, 278 – 81.

Singh, R. and Ibrahim, A.E.I. (1996). Use of spectral data in Markov chain model for crop yield forecasting. *Jour. Ind. Soc. Remote Sensing*, **24**, 145–52.

Table 1: Summary information about available two years data upon yield and biometrical characters of sugarcane of Meerut district, U.P.

(i) Minimum, maximum and arithmetic mean values

	Simulated first year data			Available second year data		
Variable	Minimum	Maximum	Mean	Minimum	Maximum	Mean
Y	15.79	113.39	68.68	15.69	90.63	51.82
X_{11}	36.00	327.00	157.28	30.00	273.00	120.47
X_{21}	0.10	0.88	0.37	0.09	0.99	0.30
X_{12}	41.00	319.00	159.01	41.00	221.00	128.63
X_{22}	0.26	1.68	0.83	0.19	1.49	0.67
X_{13}	13.00	208.00	108.40	30.00	142.00	83.28
X_{23}	0.51	2.04	1.28	0.53	1.83	1.07
X_{14}	30.00	195.00	112.08	34.00	137.00	89.46
X_{24}	0.71	2.46	1.62	0.15	2.05	1.33
X_{15}	34.00	194.00	112.72	34.00	140.00	90.53
X_{25}	0.79	2.76	1.85	0.75	2.13	1.47

(ii) Variance–covariance matrix of simulated first year data

	Y	X_{11}	X_{21}	X_{12}	X_{22}	X_{13}	X_{23}	X_{14}	X_{24}	X_{15}	X_{25}
Y	317.90	433.68	0.71	409.32	1.55	327.93	2.46	300.54	2.84	302.08	3.56
X_{11}	433.68	2412.83	1.51	1537.42	3.63	990.85	5.04	832.21	5.14	810.67	5.99
X_{21}	0.71	1.51	0.02	-0.89	0.03	1.45	0.02	1.19	0.02	1.08	0.02
X_{12}	409.32	1537.42	-0.8	2097.37	-0.3	741.27	2.09	660.43	3.22	657.95	4.43
X_{22}	1.55	3.63	0.03	-0.32	0.06	2.56	0.06	2.36	0.05	2.24	0.05
X_{13}	327.93	990.85	1.45	741.27	2.56	871.58	3.14	676.17	3.53	651.10	4.06
X_{23}	2.46	5.04	0.02	2.09	0.06	3.14	0.08	3.31	0.08	3.22	0.08
X_{14}	300.54	832.21	1.19	660.43	2.36	676.17	3.31	654.31	3.74	619.54	4.10
X_{24}	2.84	5.14	0.02	3.22	0.05	3.53	0.08	3.74	0.09	3.54	0.10
X_{15}	302.08	810.67	1.08	657.95	2.24	651.10	3.22	619.54	3.54	613.85	4.07
X_{25}	3.56	5.99	0.02	4.43	0.05	4.06	0.08	4.10	0.10	4.07	0.11

(iii) Variance–covariance matrix of available second year data

	Y	X_{11}	X_{21}	X_{12}	X_{22}	X_{13}	X_{23}	X_{14}	X_{24}	X_{15}	X_{25}
Y	246.04	397.67	1.07	335.74	2.09	256.21	2.63	262.75	2.86	262.14	3.02
X_{11}	397.67	1399.30	3.03	915.19	4.53	612.37	4.93	598.84	4.42	584.68	4.28
X_{21}	1.07	3.03	0.03	-0.96	0.04	1.77	0.03	1.40	0.02	1.35	0.02
X_{12}	335.74	915.19	-0.9	1378.00	0.51	439.13	2.04	485.10	2.90	480.52	3.34
X_{22}	2.09	4.53	0.04	0.51	0.06	2.96	0.05	2.53	0.04	2.47	0.04
X_{13}	256.21	612.37	1.77	439.13	2.96	423.17	3.21	395.83	3.04	392.60	3.08
X_{23}	2.63	4.93	0.03	2.04	0.05	3.21	0.06	3.00	0.06	2.93	0.05
X_{14}	262.75	598.84	1.40	485.10	2.53	395.83	3.00	401.01	2.98	400.17	3.02
X_{24}	2.86	4.42	0.02	2.90	0.04	3.04	0.06	2.98	0.06	2.95	0.06
X_{15}	262.14	584.68	1.35	480.52	2.47	392.60	2.93	400.17	2.95	406.81	3.00
X_{25}	3.02	4.28	0.02	3.34	0.04	3.08	0.05	3.02	0.06	3.00	0.07

Table 2: Quantiles of biometrical characters of untransformed and of their transformations viz. principal components and growth indices used in the definition of plant condition states for various Markov chain models developed upon simulated first year data

Stage	Biometrical character	Quantile	Values		
			Q_1	Q_2	Q_3
(i) Untransformed					
s_1	X_{11}	Quartile	123	156	190
	X_{21}	Quartile	0.25	0.35	0.47
s_2	X_{12}	Quartile	129	158	192
	X_{22}	Quartile	0.65	0.81	1.00
s_3	X_{13}	Quartile	88	108	128
	X_{23}	Quartile	1.07	1.27	1.49
s_4	X_{14}	Quartile	94	112	130
	X_{24}	Quartile	1.39	1.61	1.83
s_5	X_{15}	Quartile	96	112	130
	X_{25}	Quartile	1.61	1.84	2.09
(ii) Transformed (Principal components)					
S_1	PC_{11}	Quartile	180.39	222.04	264.53
	PC_{21}	Quartile	0.7	0.9	1.1
S_2	PC_{12}	Quartile	157.79	190.17	226.39
	PC_{22}	Quartile	1.27	1.53	1.78
S_3	PC_{13}	Quartile	129.5	154.97	181.13
	PC_{23}	Quartile	1.78	2.06	2.35
S_4	PC_{14}	Quartile	133.89	158.39	183.83
	PC_{24}	Quartile	2.14	2.45	2.77
T_1	PC_{11}	Quartile	202.15	246.32	292.55
	PC_{21}	Quartile	1.30	1.57	1.83
T_2	PC_{12}	Quartile	186.14	220.65	259.82
	PC_{22}	Quartile	1.90	2.22	2.54
T_3	PC_{13}	Quartile	160.61	191.78	223.15
	PC_{23}	Quartile	2.40	2.77	3.13
(iii) Transformed (Growth indices)					
S_1	GI_{11}	Quartile	130.27	159.02	188.89
	GI_{21}	Quartile	0.33	0.42	0.52
S_2	GI_{12}	Quartile	121.09	144.33	171.93
	GI_{22}	Quartile	0.67	0.80	0.95
S_3	GI_{13}	Quartile	99.08	119.13	140.95
	GI_{23}	Quartile	0.81	0.95	1.09
S_4	GI_{14}	Quartile	103.96	123.27	143.17
	GI_{24}	Quartile	1.10	1.25	1.42
T_1	GI_{11}	Quartile	180.20	217.44	258.88
	GI_{21}	Quartile	0.72	0.87	1.04
T_2	GI_{12}	Quartile	174.43	205.09	242.28
	GI_{22}	Quartile	1.10	1.30	1.51
T_3	GI_{13}	Quartile	152.08	182.10	213.68
	GI_{23}	Quartile	1.48	1.70	1.94
s_6 / S_5 / T_4	Y	Decile	45.04, 53.29, 58.67, 63.92, 69.06, 73.52, 78.70, 84.25, 92.29		

Table 3: Mean yield forecasts for second year based on various first year Markov chain models using simulated first year data

(i) Models upon untransformed data

Name and type of model	Data type	Definition/ no. of States	s_1	s_2/S_1	$s_3/S_2/T_1$	$s_4/S_3/T_2$	$s_5/S_4/T_3$
REG	Available	-	59.86	57.17	56.34	54.84	53.15
Regression	dataset		(1.68)	(1.72)	(1.60)	(1.64)	(1.62)
MMC1	Untransformed	MxM	66.10	64.13	62.03	58.27	56.26
FOMC		4	(0.30)	(0.27)	(0.24)	(0.18)	(0.15)
MMC2	Untransformed	QxM	64.32	61.70	58.15	56.61	55.02
FOMC		8	(0.46)	(0.44)	(0.31)	(0.20)	(0.16)
MMC3	Untransformed	MxQ	64.26	62.05	59.06	56.51	54.51
FOMC		8	(0.51)	(0.47)	(0.32)	(0.21)	(0.16)
MMC4	Untransformed	QxQ	64.29	61.83	59.25	57.13	56.31
FOMC		16	(0.28)	(0.24)	(0.22)	(0.19)	(0.14)
MMC5	Untransformed	MxMxMxM	–	61.68	58.91	57.54	56.99
SOMC		16		(0.14)	(0.10)	(0.07)	(0.05)
(ii) Models upon transformed data							
MMC6	G.I.	MxM	–	62.69	61.28	57.59	57.01
SOMC		4		(0.35)	(0.29)	(0.26)	(0.15)
MMC7	G.I	QxQ	–	62.23	61.31	59.05	58.72
SOMC		16		(0.45)	(0.39)	(0.23)	(0.17)
MMC8	P.C	MxM	–	62.82	61.65	58.90	57.12
SOMC		4		(0.35)	(0.38)	(0.20)	(0.14)
MMC9	P.C	QxQ	–	59.61	58.58	56.23	54.83
SOMC		16		(0.26)	(0.22)	(0.14)	(0.11)
MMC10	G.I	MxM	–	–	60.22	60.07	58.43
TOMC		4			(0.25)	(0.21)	(0.12)
MMC11	G.I	QxQ	–	–	59.97	59.43	57.20
TOMC		16			(0.19)	(0.15)	(0.09)
MMC12	P.C	MxM	–	–	52.54	52.49	52.48
TOMC		4			(0.29)	(0.23)	(0.15)
MMC13	P.C	QxQ	–	–	52.72	52.53	52.30
TOMC		16			(0.20)	(0.15)	(0.10)

Observed mean yield = 51.82 kg/plot

Author for correspondence

Lalmohan Bhar
Indian Agricultural Statistics Research Institute,
Library Avenue, PUSA,
New Delhi
email: lmbhar@iasri.res.in

Received: 5 December, 2013 *Revised: 28 January, 2014* *Accepted: 10 November, 2014*