1. Introduction
Forecasting techniques in agriculture include, inter alia, forecasting of production/ yield/ area of crops and forewarning of incidence of crop pests and diseases. Crop yield forecasts are extremely useful in formulation of policies regarding stock, distribution and supply of agricultural produce to different areas in the country. Moreover, considerable losses in crop produce occur due to the infestation of various crop pests and diseases. Such losses can be reduced to a great extent if their occurrence is known in advance so that timely remedial measures can be taken. Thus there is also a need to develop forewarning systems which can provide advance information for outbreak of pests/ diseases attack so that protection measures can be implemented before the actual onset of the damage. However, statistical techniques employed should be able to provide objective crop forecasts with reasonable precisions well in advance before harvests for taking timely decisions. Various approaches have been used for forecasting such agricultural systems. Prominent among the methods of forecasting are based on models that utilize data on crop biometrical characters, weather parameters, farmers’ eye estimates, agrometeorological conditions and remotely sensed crop reflectance observations etc., utilized either separately or in an integrated approach.

2. Statistical Forecast Models in Agriculture
The following statistical forecast models are taken up for discussion for their applications in forecasting agricultural systems. However, the same cannot be claimed to be complete and exhaustive as far as the forecast models and/ or the application areas are concerned. Nevertheless, the essentials are hopefully covered.

Regression Models
- Multiple Linear Regression (MLR) models using plant characters (for forecasting crop yields)
- Weather indices based MLR models (for forecasting crop yields/ crop pest counts)
- Logistic regression models (Models for forecasting/ forewarning qualitative response variable like low or high crop yields, presence or absence of crop pests/ diseases etc.)

Time Series Models
- Exponential smoothing models (for forecasting area/ production of crops)
- Auto-Regressive Integrated Moving Average (ARIMA) models (for forecasting area/ production of crops)

Probabilistic Models
- Markov chain models (for forecasting crop yields)
2.1 Regression Models

Let the response variable (variable to be forecasted) be denoted by $Y$ and the set of predictor variables $X_1, X_2, \ldots, X_p$, where $p$ denotes the number of predictor variables. The true relationship between $Y$ and $(X_1, X_2, \ldots, X_p)$ can be approximated by a multiple linear regression model given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon.$$ 

$\beta_0$ and $\beta_i$ ($i=1,2,\ldots,p$) are parameters to be estimated and $\varepsilon$ is random error. Some assumptions are made about this model like the relationship of the response $Y$ to the predictors $X_1, X_2, \ldots, X_p$ is linear in the regression parameters $\beta_0, \beta_1, \ldots, \beta_p$, the errors are assumed to be independently and identically distributed (iid) normal random variables with mean zero and a common variance $\sigma^2$, the errors are independent of each other (their pair-wise covariances are zero) and that the predictor variables $X_1, X_2, \ldots, X_p$ are non-random and measured without error.

2.1.1 Multiple Linear Regression (MLR) Models using Plant Characters

In the crop yield forecasting context, $Y$ and $X_i$ ($i=1,2,\ldots,p$) are crop yield and plant biometrical characters respectively. These plant characters may be used in original scale or some suitably transformed variables of these can be used. The plant characters could be average plant height, number of ear heads / panicles etc. depending upon the crop in question. Even though, models using what are called ‘growth indices’ of plant characters require data at intermediate periods of crop growth, usually these models are developed in a simple manner by utilizing data at one point of time only during the crop growth period (Jain et al. 1985; Agrawal and Jain, 1996).

Exercise 2.1: Consider the Dataset -1 given in the Appendix with $X_1$ (number of plants/plot) and $X_2$ (average sugarcane height in metres) at seven months after planting & $Y$ (yield in kg/ha) at harvest. Fit a MLR model of $Y$ upon $X_1$ and $X_2$ using the first 15 data points.

In SPSS, go to Analyse -> Regression -> Linear and choose method as ‘Enter’. The fitted model will be like $Y=b_0+b_1X_1+b_2X_2$. Now, for the $16^{th}$ observation plug the values of $X_1$ and $X_2$ in this fitted model to get forecast for the $16^{th}$ observation and compare it with the actual value of 64.87 kg/ha.

2.1.2 Weather Indices based MLR Models

This is a model similar to the one used in previous subsection. However, instead of using the predictor weather variables as such suitable transformations are made to construct weather indices which in turn are used as predictor variables. The model is given by

$$Y=a_0 + \sum_{i=1}^{p} a_i Z_i + \sum_{i \neq j}^{p} a_{ij} Z_{ij} + e$$

where $Z_i = \sum_{w=1}^{m} r_{iw} X_{iw}$, $Z_{ij} = \sum_{w=1}^{m} r_{ijw} X_{iw} X_{jw}$

with $Y$ and $e$ denote the crop yield/ pest count (dependent variable) and random error, $r_{iw}$ denotes correlation coefficient of $Y$ with $i^{th}$ weather variable, product of $i^{th}$ and $j^{th}$ weather variables in $w^{th}$ week and $m$ and $p$ denote week of forecast and number of weather variables used respectively. Here $Z_i's$ and $Z_{ij}'s$ are the independent variables which are functions of the basic weather variables like, say, maximum temperature.
Minimum temperature, rainfall, relative humidity etc. These models utilize data on weather parameters over weeks within years during the crop growth, response variable, if crop yield is available, over years and if pest count is available, over weeks for various years. (Agrawal et al. 1983; Ramasubramanian et al. 2006)

Exercise 2.2: Consider Dataset-2 which gives weekly data on rainfall (mm) for 9 weeks over growing season of rice crop along with yield for 14 years.

1. For the data up to first 13 years generate \( Z_0, Z_1, Z_2 \), where 

\[
Z_0 = \frac{\sum w X_w}{9}, \quad Z_1 = \frac{\sum w r_w X_w}{\sum r_w^2}, \quad Z_2 = \frac{\sum w r_w^2 X_w}{\sum r_w^2}; \ r_w's \ are \ simple \ correlation \ coefficients \ of \ yield \ with \ rainfall \ in \ different \ weeks \ (w=1, 2, ….,9). \ Tabulate \ Y, Z_0, Z_1, Z_2.
\]

2. Estimate the parameters of forecast model \( Y = \alpha + \beta_0 Z_0 + \beta_1 Z_1 + \beta_2 Z_2 + e \)

(In SPSS, go to Analyse -> Regression -> Linear and choose method as ‘Enter’. )

3. Forecast for the 14th year by calculating \( Z_0 \) by utilizing rainfall over the 9 weeks for the 14th year. Note that \( Z_1 \) and \( Z_2 \) here are calculated by taking the correlation coefficients \( r_w \) (w=1 to 9) from the past data itself. Now, for the 14th year plug the values of \( Z_0, Z_1 \) and \( Z_2 \) in the fitted model obtained in (2) above to get forecast for the 14th year and compare it with the actual value of 786 kg/ha.

2.1.3 Logistic Regression Models

When the response variable is quantitative, the usual theory of MLR analysis holds good. However, situations where the response variable is qualitative are also quite common and occur extensively in statistical applications. Logistic regression is widely used when the response variable is qualitative. Sometimes quantitative information on pests and diseases is not available but is available in qualitative form such as occurrence / non-occurrence, low / high incidence etc. The statistical model preferred for the analysis of such binary (dichotomous) responses is the binary logistic regression model. It can be used to describe the relationship of several independent variables to the binary (say, named 0 & 1) dependent variable. The logistic regression is used for obtaining probabilities of occurrence of the different categories (Misra et al. 2004; Agrawal, et al. 2004).

The model is of the form: 

\[
P(E = 1) = \frac{1}{1 + \exp(-z)} \]

where \( z \) is a function of weather variables. If \( P(E = 1) \geq 0.5 \) then there is more chance of occurrence of disease and if \( P(E = 1) < 0.5 \) then probability of occurrence of disease is minimum. If the experimenter wants to be more stringent, then the cutoff value of 0.5 could be increased to, say, 0.7.

Exercise 2.3: Consider Dataset-3, Misra et al. (2004) used weather data during 1987-97 in Kakori and Malihabad mango (Mangifera indica L.) belt (Lucknow) of Uttar Pradesh to develop logistic regression models for forewarning powdery mildew caused by Oidium mangiferae Berthet and validated the same using data of recent years. The forewarning system thus obtained satisfactorily forewarns with the results obtained comparing well with the observed year-wise responses. The status of the powdery mildew (its epidemic and spread) during 1987-97 are given in the dataset, with the occurrence of the epidemic.
denoted by 1 and 0 otherwise. The variables used were maximum temperature \((X_1)\) and relative humidity \((X_2)\). The model is given by

\[
P (Y=1) = 1/ [1+ \exp \{ - (\beta_0 + \beta_1 x_1 + \beta_2 x_2) \}]
\]

Logistic regression models were developed using the maximum likelihood estimation procedure in SPSS. Consider 1987-96 model based on second week of March average weather data using which forewarning probability is obtained for the year 1997. Feed the 1987-96 data upon \(Y, X_1\) and \(X_2\) in SPSS data editor, then choose Analyse -> Regression -> Binary Logistic. Under the output see for “Variables in the Equation” table and note the parameter estimates corresponding to intercept, \(X_1\) and \(X_2\) as \(\hat{\beta}_0 = -72.47; \hat{\beta}_1 = 1.845; \hat{\beta}_2 = 0.22\). Then the model becomes

\[
P(Y=1) = 1/ \{1+ \exp (-(-72.47 + (1.845* X_1) + ( 0.22* X_2 ))}\}
\]

Plugging in the values \(X_1 = 31.50\) and \(X_2 = 68.29\), of year 1997 it can be seen that \(P(Y=1) = 0.66\). This is the forewarning probability of occurrence of powdery mildew in mango using logistic regression modeling for 1997. The logistic regression model yielded good results. If \(P(Y=1)<0.5\), then probability that epidemic will occur is minimal, otherwise there is more chance of occurrence of epidemic and this can be taken as objective procedure of forewarning the disease. As we were having the information that there was epidemic during the year 1997, it can be seen that the logistic regression model forewarns the actual status correctly.

2.2 Time Series Models

Time series (TS) data refers to observations on a variable that occur in a time sequence. Mostly these observations are collected at equally spaced, discrete time intervals. A basic assumption in any time series analysis/modeling is that some aspects of the past pattern will continue to remain in the future. Also under this set up, the time series process is based usually on past values of the main variable but not on explanatory variables which may affect the variable/ system. So the system acts as a black box and we may only be able to know about ‘what’ will happen rather than ‘why’ it happens. So if time series models are put to use for forecasting purposes, then they are especially applicable in the ‘short term’. Ideally, at least 50 observations are necessary for TS forecasting, however practically some 20 observations are needed.

2.2.1 Exponential Smoothing Models

An important step in analysing TS data is to consider the types of data patterns, so that the models most appropriate to those patterns can be utilized (Makridakis et al. 1998). Four types of time series components can be distinguished. They are

(i) Horizontal – when data values fluctuate around a constant value
(ii) Trend – when there is long term increase or decrease in the data
(iii) Seasonal – when series is influenced by seasonal factor/ recurs on regular periods
(iv) Cyclical – when the data exhibit rises and falls that are not of a fixed period

Note that many data series include combinations of the preceding patterns. After separating out the existing patterns in any time series data, the pattern that remains unidentifiable form the ‘random’ or ‘error’ component. Time plot (data plotted over time)
and seasonal plot (data plotted against individual seasons in which the data were observed) help in visualizing these patterns while exploring the data. A crude yet practical way of decomposing the original data (ignoring cyclical pattern) is to go for a seasonal decomposition either by assuming an additive or multiplicative model viz.

\[ Y_t = T_t + S_t + E_t \] or \[ Y_t = T_t \cdot S_t \cdot E_t, \]

where \( Y_t \) - Original TS data
\( T_t \) - Trend component
\( S_t \) – Seasonal component
\( E_t \) – Error/ Irregular component

If the magnitude of a TS varies with the level of the series then one has to go for a multiplicative model else an additive model. This decomposition may enable one to study the TS components separately or will allow workers to de-trend or to do seasonal adjustments if needed for further analysis.

2.2.1.1 Simple Exponential Smoothing (SES) Models

Let the time series data be denoted by \( Y_1, Y_2, \ldots, Y_t \). Suppose we wish to forecast the next value of our time series \( Y_{t+1} \) that is yet to be observed with forecast for \( Y_t \) denoted by \( F_t \). Then the forecast \( F_{t+1} \) is based on weighting the most recent observation \( Y_t \) with a weight value \( \alpha \) and weighting the most recent forecast \( F_t \) with a weight of \((1-\alpha)\) where \( \alpha \) is a smoothing constant/ weight between 0 and 1. Thus the forecast for the period \( t+1 \) is given by

\[ F_{t+1} = F_t + \alpha(Y_t - F_t) \]

Note that the choice of \( \alpha \) has considerable impact on the forecast. A large value of \( \alpha \) (say 0.9) gives very little smoothing in the forecast, whereas a small value of \( \alpha \) (say 0.1) gives considerable smoothing. Alternatively, one can choose \( \alpha \) from a grid of values (say \( \alpha = 0.1, 0.2, \ldots, 0.9 \)) and choose the value that yields the smallest MSE value.

If the above model is expanded recursively then \( F_{t+1} \) will come out to be a function of \( \alpha \), past \( Y_t \) values and \( F_t \). So, having known values of \( \alpha \) and past values of \( Y_t \), our point of concern relates to initializing the value of \( F_1 \). Because the weight attached to this user-defined \( F_1 \) is minimal, its effect on \( F_{t+1} \) is negligible.

**Exercise 2.4:** For the Dataset-4, find the simple exponential smoothing one-step-ahead forecasts for the period 2000-01 to 2003-04 by using the recursive model \( F_{t+1} = \alpha Y_t + (1+\alpha)F_t \) for \( t = 1, 2, 3, \ldots \) upon the years 1970-71 to 1999-2000 with \( F_1 = Y_1 \) and \( \alpha = 0.1 \).

In SPSS, go to Analyse -> Time series -> Exponential smoothing model. In the pop up menu choose radio option button ‘simple’. A large value of \( \alpha \)(say 0.9) gives very little smoothing in the forecast, whereas a small value of \( \alpha \)(say 0.1) gives considerable smoothing. Alternatively, one can choose \( \alpha \) from a grid of values (say \( \alpha = 0.1, 0.2, \ldots, 0.9 \)) and choose the value that yields the smallest MSE value. Another point of concern relates to the initializing value \( F_1 \). One method of initialization is to use the first observed value \( Y_1 \) as the first forecast (\( F_1 = Y_1 \)) and then proceed. Another possibility (“custom option”)
would be to average the first four or five values in the data set and use this as the initial forecast.

### 2.2.1.2 Double Exponential Smoothing (Holt) Models
This is to allow forecasting data with trends. The forecast for Holt’s linear exponential smoothing is found by having two more equations to SES model to deal with – one for level and one for trend. The smoothing parameters (weights) \( \alpha \) and \( \beta \) can be chosen from a grid of values (say, each combination of \( \alpha = 0.1, 0.2, ..., 0.9 \) and \( \beta = 0.1, 0.2, ..., 0.9 \)) and then select the combination of \( \alpha \) and \( \beta \) which correspond to the lowest MSE.

In SPSS, the weights \( \alpha \) and \( \beta \) can be chosen from a grid of values (say, each combination of \( \alpha = 0.1, 0.2, ..., 0.9 \) and \( \beta = 0.1, 0.2, ..., 0.9 \)) and then select the combination of \( \alpha \) and \( \beta \) which correspond to the lowest MSE.

### 2.2.1.3 Triple Exponential Smoothing (Winters) Models
This method is recommended when seasonality exists in the time series data. This method is based on three smoothing equations – one for the level, one for trend, and one for seasonality. It is similar to Holt’s method, with one additional equation to deal with seasonality. In fact there are two different Winters methods depending on whether seasonality is modeled in an additive or multiplicative way.

In SPSS, periodicity can be assigned to data by defining “Dates”. In SPSS, choose, Data -> Define Dates -> Cases Are: - defines the time interval used to generate dates; First Case Is:- Defines the starting date value, which is assigned to the first case. Sequential values, based on the time interval, are assigned to subsequent cases; Periodicity at higher level:- Indicates the repetitive cyclical variation, such as the number of months in a year or the number of days in a week. The value displayed indicates the maximum value you can enter. A new numeric variable is created for each component that is used to define the date.

### 2.2.2 ARIMA Models
A TS is said to be stationary if its underlying generating process is based on a constant mean and constant variance with its autocorrelation function (ACF) essentially constant through time (Pankratz, 1983). Thus, if we consider different subsets of a realization (time series ‘sample’) the different subsets will typically have means, variances and autocorrelation functions that do not differ significantly. A statistical test for stationarity is the most widely used Dickey Fuller test. To carry out the test, estimate by OLS the regression model:

\[
y'_t = \phi y'_{t-1} + b_1 y'_{t-2} + ... + b_p y'_{t-p} \]

where \( y'_t \) denotes the differenced series \( (y_t - y_{t-1}) \). The number of terms in the regression, \( p \), is usually set to be about 3. Then if \( \phi \) is nearly zero the original series \( y_t \) needs differencing. And if \( \phi < 0 \) then \( y_t \) is already stationary.

Autocorrelation refers to the way the observations in a time series are related to each other and is measured by the simple correlation between current observation \( (Y_t) \) and observation from \( p \) periods before the current one \( (Y_{t-p}) \). It ranges from \(-1\) to \(+1\). The maximum number of useful autocorrelations are roughly \( n/4 \) where \( n \) is the number of periods upon which information on \( y_t \) is available. Partial autocorrelations are used to measure the degree of association between \( y_t \) and \( y_{t-p} \) when the y-effects at other time lags
1, 2, …, p-1 are removed. Note that usually up to order 2 for p, d, or q are sufficient for developing a good model in practice.

Theoretical ACFs and PACFs (Autocorrelations versus lags) are available for the various models chosen. Thus compare the correlograms (plot of sample ACFs versus lags) with these theoretical ACF/PACFs, to find a reasonably good match and tentatively select one or more ARIMA models. The general characteristics of theoretical ACFs and PACFs are as follows:- (here ‘spike’ represents the line at various lags in the plot with length equal to magnitude of autocorrelations)

<table>
<thead>
<tr>
<th>Model</th>
<th>ACF</th>
<th>PACF</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>Spikes decay towards zero</td>
<td>Spikes cutoff to zero</td>
</tr>
<tr>
<td>MA</td>
<td>Spikes cutoff to zero</td>
<td>Spikes decay to zero</td>
</tr>
<tr>
<td>ARMA</td>
<td>Spikes decay to zero</td>
<td>Spikes decay to zero</td>
</tr>
</tbody>
</table>

In general, an ARIMA model is characterized by the notation ARIMA (p, d, q) where, p, d and q denote orders of auto-regression, integration (differencing) and moving average respectively. In ARIMA parlance, TS is a linear function of past actual values and random shocks. For instance, given a time series process \( \{y_t\} \), a first order auto-regressive process is denoted by ARIMA(1,0,0) or simply AR(1) and is given by

\[
y_t = \mu + \phi_1 y_{t-1} + \epsilon_t
\]

and a first order moving average process is denoted by ARIMA (0,0,1) or simply MA(1) and is given by

\[
y_t = \mu - \theta_1 \epsilon_{t-1} + \epsilon_t
\]

Alternatively, the model ultimately derived, may be a mixture of these processes and of higher orders as well. Thus a stationary ARMA (p, q) process is defined by the equation

\[
y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} + \ldots - \theta_q \epsilon_{t-q} + \epsilon_t
\]

where \( \epsilon_t \)'s are independently and normally distributed with zero mean and constant variance \( \sigma^2 \) for \( t = 1, 2, \ldots, n \). Note here that the values of p and q, in practice lie between 0 and 3.

For development of seasonal ARIMA models, identification of relevant models and inclusion of suitable seasonal variables are necessary for seasonal modeling and their applications, say, forecasting production of crops. Seasonal forecasts of production of principal crops are of greater utility for planners, administrators and researchers alike. Agricultural seasons vary significantly among the states of India. For example, Tamil Nadu has unique three-season cropping pattern for Paddy crop whereas two-season paddy rules elsewhere in the country. Thus seasonal forecasts of crop production can also be made using seasonal ARIMA models.

The basic stages involved in developing ARIMA models are now discussed.

(i) Identification Stage
The foremost step in the process of modeling is to check for the stationarity of the series, as the estimation procedures are available only for stationary series. There are two kinds
of stationarity, viz., stationarity in ‘mean’ and stationarity in ‘variance’. A cursory look at
the graph of the data and structure of autocorrelation and partial correlation coefficients
may provide clues for the presence of stationarity. Another way of checking for
stationarity is to fit a first order autoregressive model for the raw data and test whether the
coefficient \( \phi_1 \) is less than one. If the model is found to be non-stationary, stationarity
could be achieved mostly by differencing the series. Or go for a Dickey Fuller test (see
section 4). Stationarity in variance could be achieved by some modes of transformation,
say, log transformation. This is applicable for both seasonal and non-seasonal stationarity.
Thus, if \( X_t \) denotes the original series, the non-seasonal difference of first order is
\( Y_t = X_t - X_{t-1} \) followed by the seasonal differencing (if needed)
\( Z_t = Y_t - Y_{t-s} = (X_t - X_{t-1}) - (X_t - X_{t-s-1}) \). The next step in the
identification process is to find the initial values for the orders of seasonal and non-seasonal
parameters, \( p, q, \) and \( P, Q \). They could be obtained by looking for significant
autocorrelation and partial autocorrelation coefficients. If second order auto correlation coefficient is significant, then an AR(2), or MA(2) or ARMA(2)
model could be tried to start with. This is not a hard and fast rule, as sample
autocorrelation coefficients are poor estimates of population autocorrelation coefficients.
Still they can be used as initial values while the final models are achieved after going
through the stages repeatedly.

(ii) Estimation Stage
At the identification stage one or more models are tentatively chosen that seem to provide
statistically adequate representations of the available data. Then we attempt to obtained
precise estimates of parameters of the model by least squares as advocated by Box and
Jenkins. Standard computer packages like SAS, SPSS etc. are available for finding the
estimates of relevant parameters using iterative procedures.

(iii) Diagnostic Stage
Different models can be obtained for various combinations of AR and MA individually
and collectively. The best model is obtained with following diagnostics:

(a) Low Akaike Information Criteria (AIC)/ Schwarz-Bayesian Criteria (SBC): AIC is
given by \( \text{AIC} = (-2 \log L + 2 m) \) where \( m=p+q+P+Q \) and \( L \) is the likelihood function.
Since \(-2 \log L \) is approximately equal to \( \{ n(1+\log 2\pi) + n \log \sigma^2 \} \) where \( \sigma^2 \) is the
model MSE, AIC can be written as \( \text{AIC} = \{ n(1+\log 2\pi) + n \log \sigma^2 + 2 m \} \) and because
first term in this equation is a constant, it is usually omitted while comparing between
models. As an alternative to AIC, sometimes SBC is also used which is given by \( \text{SBC} = \log \sigma^2 + (m \log n) / n. \)

(b) Non-significance of auto correlations of residuals via Portmonteau tests (Q-tests based
on Chisquare statistics)-Box-Pierce or Ljung-Box texts: After tentative model has
been fitted to the data, it is important to perform diagnostic checks to test the
adequacy of the model and, if need be, to suggest potential improvements. One way to
accomplish this is through the analysis of residuals. It has been found that it is
effective to measure the overall adequacy of the chosen model by examining a
quantity \( Q \) known as Box-Pierce statistic (a function of autocorrelations of residuals)
whose approximate distribution is chi-square and is computed as follows:

\[
Q = n \sum r^2(j)
\]

where summation extends from 1 to \( k \) with \( k \) as the maximum lag considered, \( n \) is the
number of observations in the series, \( r(j) \) is the estimated autocorrelation at lag \( j; k \) can
be any positive integer and is usually around 20. Q follows Chi-square with (k-m)
degrees of freedom where m is the number of parameters estimated in the model. A
modified Q statistic is the Ljung-box statistic which is given by

$$Q = n(n+2) \sum r^2(j) / (n-j)$$

The Q Statistic is compared to critical values from chi-square distribution. If model is
correctly specified, residuals should be uncorrelated and Q should be small (the
probability value should be large). A significant value indicates that the chosen model
does not fit well.

All these stages require considerable care and work and they themselves are not
exhaustive.

**Exercise 2.5:** For the Dataset-5, do the following:
(i) Draw time plot. Perform the following Dickey - Fuller test. Fit the model
$$y_t = \phi y_{t-1} + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3}$$
with 'no intercept' and observe whether $\hat{\phi} \approx 0$ (i.e. data is non- stationary) or $\hat{\phi} < 0$
(i.e. data is stationary). Whether the data needs differencing?
(ii) By drawing correlograms identify the tentative values of $p$ and $q$.
(In SPSS go to Graph -> Time series -> Autocorrelation)
(iii) Obtain parameter estimates of the model upon the years 1970-71 to 1999-2000 by
identifying a series of ARIMA (p,d,q) models (p=0,1,2,3; d obtained in question 1; q =
0,1,2,3) also preserving parsimony that might be useful in describing the time series.
Which of your models is the best according to their AIC values?
(In SPSS go to Analyse -> Time series -> ARIMA model)
(iv) For the models obtained in (3) & (4) perform separately diagnostic tests upon residuals
using (i) Residual ACF (it) Box-Pierce Chi-square test.
(v) Forecast for lead years 2000-01 to 2003-04.(Inside option Save -> ‘Predict through’)

### 2.3 Probabilistic Models

#### 2.3.1 Markov Chain Models

A system which takes the form of a chain of finite stages with a limited number of
possible states (condition classes) within each stage is called a Markov chain if there
exists a case of simple dependence that any state of a particular stage depends directly on
any of the states of the preceding stage. In the crop yield forecasting context, a first order
Markov chain is constructed to provide a forecast distribution of crop yield for the various
crop condition classes (states) at selected phenological (or calendar dates) stages of plant
life. The only assumption under the first order Markov chain (or simply a Markov chain)
set up is that the past (crop) conditions are statistically uninformative for predicting the
future (yield forecasts), after the present (crop) conditions are known (Jain and Agrawal,
1992; Ramasubramanian et al. 2004)

For simplicity, the Markov chain forecast model is discussed by taking the particular case
of sugarcane (a one-year growth period crop) yield forecasting. Let there be six stages
including the final harvest stage, the stages having been determined on the basis of
calendar dates i.e. first stage is 3 months after planting, second stage is 4 months after
planting etc. and sixth stage at harvest. For development of model using the available (say,
for one year) data, firstly construct states within stages on the basis of percentiles upon the data points on biometrical characters, say, average plant height, girth of cane, number of plants per plot etc. at each of the first five stages and upon (only) yield at final stage. For brevity, let us consider, only one biometrical character, say, plant height (X₁). At stage−1, let the percentiles, say, quartiles for the data points upon X₁ (i.e. X₁ at stage−1) be a₁₁, a₁₂, a₁₃. Thus four states viz. X₁₁<=a₁₁; a₁₁<X₁₁<=a₁₂; a₁₂<X₁₁<=a₁₃; X₁₁>a₁₃ (classes denoted as a₁, a₂, a₃, a₄) can be obtained with distributions say, f₁₁, f₁₂, f₁₃, f₁₄. Similarly with quartiles (a₂₁, a₂₂, a₂₃) upon X₁₂ four states within stage−2 as X₁₂<=a₂₁; a₂₁<X₁₂<=a₂₂; a₂₂<X₁₂<=a₂₃; X₁₂>a₂₃ (classes denoted as b₁, b₂, b₃, b₄) with distributions say, f₂₁, f₂₂, f₂₃, f₂₄ can be obtained. Likewise four states of other stages can be defined noting that the final stage will have ten states because more finer percentiles, say, deciles (say, a₆₁, a₆₂, a₆₃, a₆₄, a₆₅, a₆₆, a₆₇, a₆₈, a₆₉) can be used for forming states on the basis of variable Y(yield) with distributions say, f₆₁, f₆₂, f₆₃, f₆₄, f₆₅, f₆₆, f₆₇, f₆₈, f₆₉ as this is the main character under study.

Now compute the transition counts of data points moving from any state of one stage to any of the states of the succeeding stage. For example, to find transition probability matrix (TPM) from stage−1 to stage−2, to begin with, consider the f₁₁ data points in state-1 (i.e. a₁) of stage-1 and count the corresponding data points of X₁₂ that satisfy the state conditions b₁, b₂, b₃, b₄ of stage-2. In this way, the data points which are in state-1 of stage-1 are redistributed into the different states of the next stage-2. Next consider the 36 data points in state-2 (i.e. a₂) of stage-1 and count the corresponding data points of X₁₂ that satisfy the state conditions b₁, b₂, b₃, b₄ of stage-2. The same procedure can be followed for finding other transition counts as well to form a transition frequency matrix say ((ffᵢⱼ)) where i, j=1,2,3,4 for stage-1 to stage-2. Here, it is noted that the data points of any stage tend to remain in the same states on transiting to the next stage but it can also be noted that they will be redistributed to other (usually adjacent) states of the next stage as well. However such phenomenon may become less apparent when more than one variable/transformed data are used. The TPM from stage-1 to stage-2 is finally obtained by dividing each row element of this matrix by its corresponding row sum (these row sums are nothing but distributions at various states of stage-1). This TPM say A₁₂ from stage-1 to stage-2 will be a matrix of order 4. Other TPMs can similarly be obtained. Similarly, TPMs A₂₃, A₃₄, A₄₅ will be of order four and A₅₆ will be of order (4x10). At stage-6, form a (10x1) vector, say, yₘ, on the basis of midpoints of class intervals of yield. Now calculate the predicted yield distributions (PYDs) which will be a (4x1) vector at each stage as given here in the table below.

<table>
<thead>
<tr>
<th>At stage</th>
<th>PYD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A₁₂ A₂₃ A₃₄ A₄₅ A₅₆ yₘ</td>
</tr>
<tr>
<td>2</td>
<td>A₂₃ A₃₄ A₄₅ A₅₆ yₘ</td>
</tr>
<tr>
<td>3</td>
<td>A₃₄ A₄₅ A₅₆ yₘ</td>
</tr>
<tr>
<td>4</td>
<td>A₄₅ A₅₆ yₘ</td>
</tr>
<tr>
<td>5</td>
<td>A₅₆ yₘ</td>
</tr>
</tbody>
</table>

In the year for which yield forecast has to be found out, collect observations upon a sample of plots, information upon variable X₁ only for the same set up of stages 1 through
5. Classify these observations as per states of stages of the model developed already. This will result in 'weights' that fall in various states of any particular stage of the developed model. The mean yield forecast at any particular stage can be obtained as the weighted average of the PYDs (four in number) at that stage, weights (four in number) being the number of observations of the forecast year in different states (in the particular stage) of the developed model. Thus the developed model can be used in practice for crop yield forecasting.

**Exercise 2.6:** Consider Dataset-6-1. Here X1i is the average height in metres at stages i = 1 to 5 & Y is the yield in kg/ha at harvest stage 6 of sugarcane crop. Perform the following tasks:

1. Percentiles of variables: Calculate quartiles for the variables X11, X12, X13, X14 and X15 and deciles for Y (hint: PERCENTILE (<A1:A144>, 0.75) will give third quartile)

2. Formation of states within stages
   (i) Insert a new worksheet and copy the column of X11 in it.
   (ii) Classify the variable X11 into either group number '1' if X11 <= Q1 or '2' if Q1 < X11 <= Q2 or '3' if Q2 < X11 <= Q3 or '4' if X11 > Q3 in four subsequent adjacent columns and '0' otherwise where Q1, Q2 and Q3 are the quartiles of X11. (hint: IF(B1<=Q1,1,0); IF(B1>Q1*ANDB1<Q2,2,0) and so on) Sum these four columns row-wise to get the states as 1,2,3,4. Let this column be named Z1.
   (iii) Classify the variable X12 into four states by following the same procedure as discussed in 2(ii) above but this time using the quartiles of X12. Let the resultant column be named Z2.

3. Computation of TPMs
   (i) Copy the columns Z1 and Z2 obtained in 2(ii) and 2(iii) in a new worksheet.
   (ii) Calculate 4x4 frequency matrix by forming 16 columns and finding column totals.
   For f11 - IF( (C1=1)*AND(C2=1),1,0)
   For f12 - IF( (C1=1)*AND(C2=2),1,0)
   For f13 - IF( (C1=1)*AND(C2=3),1,0)
   For f14 - IF( (C1=1)*AND(C2=4),1,0)
   For f21 - IF( (C1=2)*AND(C2=1),1,0)
   ................
   For f43 - IF( (C1=4)*AND(C2=3),1,0)
   For f44 - IF( (C1=4)*AND(C2=4),1,0)

   and by arranging a matrix {f11,f12,f13,f14; f21,f22,f23,f24; f31,f32,f33,f34; f41,f42,f43,f44;} find row sums of this matrix as f1, f2, f3, f4. Divide first row of this matrix by f1, second row, by f2 etc., to get the TPM of stage-1 to stage-2 i.e. A12.

   (iii) Consider the frequency matrices of stage-i to stage-(i+1) for i = 2,3,4,5 as given below and compute corresponding TPMs A23, A34, A45, A56 by dividing each row by their row totals.

   F23= {30, 7, 0, 1; 8, 21, 5, 1; 0, 7, 23, 6; 0, 1, 9, 25}
   F34 = {30, 7, 1, 0; 6, 23, 7, 0; 0, 7, 19, 11; 0, 0, 9, 24}
   F45 = {29, 6, 1, 0; 8, 21, 6, 2; 1, 8, 23, 4; 0, 0, 5, 30}
   F56 = {10, 7, 10, 5, 2, 1, 1, 0, 0, 2; 4, 5, 2, 6, 6, 3, 3, 3, 3, 0; 0, 1, 2, 3, 5, 6, 7, 6, 3, 2; 1, 1, 0, 1, 4, 4, 5, 8, 11}
4. Means of Predicted Yield Distributions (PYDs)

(i) Find minimum and maximum values of Y and using the deciles obtained in task 1 form the class-frequency table, to calculate the midpoints of the class intervals as a 10x1 vetor, say, Ym.

(ii) Find mean of PYDs as

At stage 5: \( y_{5} = A_{56} \times Ym \)
At stage 4: \( y_{4} = A_{45} \times A_{56} \times Ym \)
At stage 3: \( y_{3} = A_{34} \times A_{45} \times A_{56} \times Ym \)
At stage 2: \( y_{2} = A_{23} \times A_{34} \times A_{45} \times A_{56} \times Ym \)
At stage 1: \( y_{1} = A_{12} \times A_{23} \times A_{34} \times A_{45} \times A_{56} \times Ym \)

5. Mean Yield Forecasting

(i) Consider a new set of data (see 'dataset6-2') of 156 observations upon variables X11, X12, X13, X14, X15 at five stages 1,2,3,4,5 respectively.

(ii) Classify the observations upon variable at stage 1 i.e. X11 as per the 'states' of the stage-1 of the 'fitted' model i.e. according to the quartiles obtained for 'dataset6-1' data obtained in task 1. Count the number of observations falling in state-1 of stage-1, then state-2 of stage-1 and so on and denote these as weights w11, w12, w13, w14.. Likewise compute the weights w21, w22, w23, w24 at stage 2 etc.

(iii) Forecast at each stage i (=1,2,3,4,5), by taking weighted average using the corresponding four weights from 5(ii) and the (4x1) vector ys1 at stage1, ys2 at stage2 etc. from 4(ii) to obtain mean yield forecasts at each stage.

3. Concluding Thoughts

Some of the statistical forecast modeling techniques in agriculture has been discussed. Many other models like models based on Group Method of Data Handling (GMDH), growth models, models based on soft-computing techniques such as fuzzy regression, Artificial Neural Networks (ANNs), state space modeling and forecasting using remotely sensed data have not, however, been included.

References and Suggested Reading

### APPENDIX: Datasets

#### Dataset-1

<table>
<thead>
<tr>
<th>No.</th>
<th>Y</th>
<th>X1</th>
<th>X2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21.94</td>
<td>57</td>
<td>0.75</td>
</tr>
<tr>
<td>2</td>
<td>56.35</td>
<td>104</td>
<td>1.35</td>
</tr>
<tr>
<td>3</td>
<td>67.42</td>
<td>108</td>
<td>1.64</td>
</tr>
<tr>
<td>4</td>
<td>43.16</td>
<td>96</td>
<td>1.26</td>
</tr>
<tr>
<td>5</td>
<td>49.45</td>
<td>94</td>
<td>1.52</td>
</tr>
<tr>
<td>6</td>
<td>56.33</td>
<td>104</td>
<td>1.79</td>
</tr>
<tr>
<td>7</td>
<td>59.75</td>
<td>98</td>
<td>1.82</td>
</tr>
<tr>
<td>8</td>
<td>59.58</td>
<td>98</td>
<td>1.75</td>
</tr>
<tr>
<td>9</td>
<td>52.67</td>
<td>106</td>
<td>1.43</td>
</tr>
<tr>
<td>10</td>
<td>50.97</td>
<td>92</td>
<td>1.38</td>
</tr>
<tr>
<td>11</td>
<td>28.02</td>
<td>52</td>
<td>0.96</td>
</tr>
<tr>
<td>12</td>
<td>42.71</td>
<td>70</td>
<td>1.49</td>
</tr>
<tr>
<td>13</td>
<td>52.50</td>
<td>86</td>
<td>1.24</td>
</tr>
<tr>
<td>14</td>
<td>54.69</td>
<td>110</td>
<td>1.32</td>
</tr>
<tr>
<td>15</td>
<td>59.98</td>
<td>111</td>
<td>1.62</td>
</tr>
<tr>
<td>16</td>
<td>64.87</td>
<td>127</td>
<td>1.58</td>
</tr>
</tbody>
</table>

#### Dataset-2

<table>
<thead>
<tr>
<th>Year</th>
<th>Yield (kg/ha)</th>
<th>Year</th>
<th>Rainfall (in mm) over weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>973</td>
<td></td>
<td>34.6</td>
</tr>
<tr>
<td>2</td>
<td>795</td>
<td></td>
<td>27.4</td>
</tr>
<tr>
<td>3</td>
<td>948</td>
<td></td>
<td>26.7</td>
</tr>
<tr>
<td>4</td>
<td>884</td>
<td></td>
<td>146.1</td>
</tr>
<tr>
<td>5</td>
<td>626</td>
<td></td>
<td>32.0</td>
</tr>
<tr>
<td>6</td>
<td>1023</td>
<td></td>
<td>69.2</td>
</tr>
<tr>
<td>7</td>
<td>1024</td>
<td></td>
<td>61.0</td>
</tr>
<tr>
<td>8</td>
<td>351</td>
<td></td>
<td>39.1</td>
</tr>
<tr>
<td>9</td>
<td>654</td>
<td></td>
<td>34.2</td>
</tr>
<tr>
<td>10</td>
<td>860</td>
<td></td>
<td>38.8</td>
</tr>
<tr>
<td>11</td>
<td>1209</td>
<td></td>
<td>213.6</td>
</tr>
<tr>
<td>12</td>
<td>941</td>
<td></td>
<td>107.9</td>
</tr>
<tr>
<td>13</td>
<td>917</td>
<td></td>
<td>2.5</td>
</tr>
<tr>
<td>14</td>
<td>786</td>
<td></td>
<td>15.0</td>
</tr>
</tbody>
</table>

#### Dataset-3

<table>
<thead>
<tr>
<th>Year</th>
<th>Third week(Y) of March</th>
<th>Average weather data in second week of March</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>X₁</td>
</tr>
<tr>
<td>1987</td>
<td></td>
<td>30.14</td>
</tr>
<tr>
<td>1988</td>
<td></td>
<td>30.66</td>
</tr>
<tr>
<td>1989</td>
<td></td>
<td>26.31</td>
</tr>
<tr>
<td>1990</td>
<td></td>
<td>28.43</td>
</tr>
<tr>
<td>1991</td>
<td></td>
<td>29.57</td>
</tr>
</tbody>
</table>
## Forecasting Techniques in Agriculture

### Dataset-4 and Dataset-5

<table>
<thead>
<tr>
<th>Year</th>
<th>Yield</th>
<th>Year</th>
<th>Yield</th>
<th>Year</th>
<th>Area</th>
<th>Year</th>
<th>Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970-71</td>
<td>224.5</td>
<td>1971-72</td>
<td>334.0</td>
<td>1972-73</td>
<td>343.2</td>
<td>1973-74</td>
<td>302.0</td>
</tr>
<tr>
<td>1971-72</td>
<td>334.0</td>
<td>1972-73</td>
<td>434.2</td>
<td>1973-74</td>
<td>302.0</td>
<td>1974-75</td>
<td>268.2</td>
</tr>
<tr>
<td>1972-73</td>
<td>343.2</td>
<td>1974-75</td>
<td>474.5</td>
<td>1975-76</td>
<td>308.6</td>
<td>1975-77</td>
<td>288.9</td>
</tr>
<tr>
<td>1973-74</td>
<td>302.0</td>
<td>1974-75</td>
<td>474.5</td>
<td>1975-76</td>
<td>308.6</td>
<td>1976-77</td>
<td>288.9</td>
</tr>
<tr>
<td>1974-75</td>
<td>268.2</td>
<td>1975-76</td>
<td>476.4</td>
<td>1976-77</td>
<td>288.9</td>
<td>1977-78</td>
<td>369.7</td>
</tr>
<tr>
<td>1975-76</td>
<td>308.6</td>
<td>1976-77</td>
<td>518.2</td>
<td>1977-78</td>
<td>369.7</td>
<td>1978-79</td>
<td>402.7</td>
</tr>
<tr>
<td>1977-78</td>
<td>369.7</td>
<td>1978-79</td>
<td>606.3</td>
<td>1979-80</td>
<td>386.7</td>
<td>1980-81</td>
<td>357.0</td>
</tr>
<tr>
<td>1978-79</td>
<td>402.7</td>
<td>1979-80</td>
<td>654.2</td>
<td>1980-81</td>
<td>357.0</td>
<td>1981-82</td>
<td>379.7</td>
</tr>
<tr>
<td>1979-80</td>
<td>386.7</td>
<td>1980-81</td>
<td>644.7</td>
<td>1981-82</td>
<td>379.7</td>
<td>1982-83</td>
<td>396.7</td>
</tr>
<tr>
<td>1980-81</td>
<td>357.0</td>
<td>1981-82</td>
<td>640.0</td>
<td>1982-83</td>
<td>396.7</td>
<td>1983-84</td>
<td>416.1</td>
</tr>
<tr>
<td>1981-82</td>
<td>379.7</td>
<td>1982-83</td>
<td>580.0</td>
<td>1983-84</td>
<td>416.1</td>
<td>1984-85</td>
<td>335.0</td>
</tr>
<tr>
<td>1982-83</td>
<td>396.7</td>
<td>1983-84</td>
<td>510.0</td>
<td>1984-85</td>
<td>335.0</td>
<td>1985-86</td>
<td>333.1</td>
</tr>
<tr>
<td>1983-84</td>
<td>416.1</td>
<td>1984-85</td>
<td>510.0</td>
<td>1985-86</td>
<td>333.1</td>
<td>1986-87</td>
<td>364.5</td>
</tr>
<tr>
<td>1984-85</td>
<td>335.0</td>
<td>1985-86</td>
<td>390.0</td>
<td>1986-87</td>
<td>364.5</td>
<td>1987-88</td>
<td>100</td>
</tr>
<tr>
<td>1985-86</td>
<td>333.1</td>
<td>1987-88</td>
<td>100</td>
<td>1988-89</td>
<td>89</td>
<td>1989-90</td>
<td>143</td>
</tr>
<tr>
<td>1986-87</td>
<td>364.5</td>
<td>1988-89</td>
<td>89</td>
<td>1989-90</td>
<td>143</td>
<td>1990-91</td>
<td>170</td>
</tr>
<tr>
<td>1996-97</td>
<td>173</td>
<td>1997-98</td>
<td>95</td>
<td>1999-00</td>
<td>162</td>
<td>1999-00</td>
<td>162</td>
</tr>
<tr>
<td>1997-98</td>
<td>95</td>
<td>1998-99</td>
<td>139</td>
<td>2000-01</td>
<td>100</td>
<td>2000-01</td>
<td>100</td>
</tr>
<tr>
<td>1998-99</td>
<td>139</td>
<td>1999-00</td>
<td>162</td>
<td>2001-02</td>
<td>147</td>
<td>2001-02</td>
<td>147</td>
</tr>
<tr>
<td>1999-00</td>
<td>162</td>
<td>2000-01</td>
<td>100</td>
<td>2002-03</td>
<td>158</td>
<td>2002-03</td>
<td>158</td>
</tr>
<tr>
<td>2000-01</td>
<td>100</td>
<td>2001-02</td>
<td>147</td>
<td>2003-04</td>
<td>189</td>
<td>2003-04</td>
<td>189</td>
</tr>
</tbody>
</table>

### Dataset-6-1

<table>
<thead>
<tr>
<th>Plot No.</th>
<th>X11</th>
<th>X12</th>
<th>X13</th>
<th>X14</th>
<th>X15</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.14</td>
<td>0.32</td>
<td>0.75</td>
<td>1.07</td>
<td>1.34</td>
<td>44.03</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>0.30</td>
<td>0.64</td>
<td>1.25</td>
<td>1.50</td>
<td>1.72</td>
<td>77.68</td>
</tr>
</tbody>
</table>

### Dataset-6-2

<table>
<thead>
<tr>
<th>Plot No.</th>
<th>X11</th>
<th>X12</th>
<th>X13</th>
<th>X14</th>
<th>X15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.38</td>
<td>0.62</td>
<td>0.70</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>156</td>
<td>0.24</td>
<td>0.74</td>
<td>1.16</td>
<td>1.38</td>
<td>1.51</td>
</tr>
</tbody>
</table>