SIMPLIFYING THE MARKOV CHAIN ANALYSIS OF RAINFALL DATA USING GENSTAT

BY

JACOB OTIENO ONG'ALA

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE IN APPLIED STATISTICS

DEPARTMENT OF MATHEMATICS AND APPLIED STATISTICS

MASENO UNIVERSITY

© 2011

ABSTRACT

Despite the rapid development of statistical packages, a lot of climatic data still remain unanalyzed due to lack of specialized routines in most of the packages. One package has a climatic menu though with limitation on complex analysis such as generalized linear models. Others can perform the Generalized Linear Model analysis but do not have a specialized menu for analyzing climatic data. There is no statistical package currently available which has a specialized capability to do climatic analysis easily and includes the use of generalized linear models. This study starts the work of creating a specialized menu in GenStat for analyzing climatic data by implementing Markov modeling of rainfall data. Four procedures have been written and corresponding dialogues were created to ease their use. Incorporating a climatic menu into GenStat package will support researchers in agricultural and many other fields that need an analysis of climatic data as part of their work.

iv

CHAPTER 1: INTRODUCTION

1.1: Background information

Recent developments in statistical software have simplified the entire process of data manipulation and analysis. This permits the analysis of large and complicated data sets. Climatic data has been routinely collected in many locations for many years, but a lot of this data remains unanalyzed. In the past, climatic analysis was often left to specialists due to complexities of the data, such as the strong within and between year variability. The advances in the software bring routine analysis of climatic data within reach of non-specialists. This is important due to the current worldwide interest in climate particularly climate change.

Lack of climatic data analysis has made it difficult to understand the difference between climate change and natural climatic variability. Statistical analysis is therefore important as it aims at "identifying the systematic behavior in the data set and also searching for periodic variation, a quasi-periodic variation, a trend, persistence or extreme events in the climate element under analysis" [2]. This will enhance the understanding of whether there is a climate change or climatic variability within a given time frame.

Although a number of statistical packages have the capability to analyze complex statistical problems, most of them do not have specialized routines for analyzing climatic data. Instat was introduced in the early 1980s as a simple statistics package to help in the teaching of statistics. It was later improved by adding more components with particular interest for processing climatic data [21]. It is the only available package specialized for analyzing climatic data.

The analysis of climatic data is easier and user friendly in Instat due to its additional Climatic Menu. Through this menu, most of the climatic analysis can be done e.g. summaries, analysis of events and Markov modeling of rainfall data, etc. Though Instat has the ability to analyze climatic data, it does not include the facilities, such as generalized linear models (GLMs), needed for a full analysis of climatic data. GenStat on the other hand is a powerful package which can handle complex GLM. It is a general statistics package that offers a wide range of high-quality statistical techniques and graphics.

GenStat has a flexible working environment where data can be analyzed using a user-friendly menu-based interface and a powerful command language from which you can create your own command procedures to perform the analyses. The command language permits the user to write his/her own programs that can cover situations where the standard analysis does not give them what they require. "So any GenStat analysis can be used in the construction of a new technique." [15]. Despite the command language, most users are still reluctant in using it.

It is possible to analyze climatic data using GenStat even though it does not have the climate menu as in Instat [6]. Currently, the climatic data analysis using GenStat is currently only possible if you know and understand the direction to take for the solution. This requires more understanding for users who might want to use the package to analyze their climatic data [6]. The existing GenStat commands and procedures can be used to construct other procedures, to simplify their use and make them easily available to most users; in addition, the procedures can be used to form their corresponding menu and dialogues. A climatic menu can therefore be added to GenStat package to make it easier for many users to do their work without having to master the command language.

This study is aimed at starting the implementation of a climatic menu into the GenStat package which will be able to handle high order Markov Chains to be used in modeling rainfall. Chapter 2 of this report discusses the analysis of rainfall data using Instat and Marksim with particular interest on Markov modeling. The extensive use of the GenStat command language and its dialogue/menu interface is discussed in chapter 3. Chapter 4 describes how rainfall data is prepared for fitting in Instat and how the same is implemented in GenStat using the existing and the newly developed procedures. Markov modeling of climatic data and how it is implemented in GenStat is then discussed in chapter five. Finally chapter six discusses how the GenStat procedures are used in a real and large climatic dataset.

1.2: Basic Concepts

1.2.1: The Generalized Linear Model

The Generalized Linear Model (GLM) was introduced by [14]. It is an extension of the General Linear Model to include response variables that follow any probability distribution in the exponential family e.g. Binomial, Poisson, Multinomial, Gamma, Negative Binomial etc. It may be used where the response variable neither follows a Normal distribution nor have homogonous variances [15]. Comparing GLM and Multiple regression models (a form of general linear model) makes its features seen more clearly [19]. The expression below can be used to define General linear models:

$$y_i = \beta_{0i} + \sum_{j=1}^{p} \beta_{ji} x_{ji} + \epsilon_i \qquad (i = 1, ..., n \ and \ E(\epsilon_i) = 0)$$

(1-1)

These set of n equations can be written in the form of a compact model as shown below.

$$Y = X\beta + \epsilon$$

(1-2)

where

a) Y is a vector of response variable and is a standard linear model meeting the Gauss-Markov condition and can be expressed as shown in equation 1-3.

$$E(Y) = \theta = X\beta$$

(1-3)

b) is the matrix explanatory variable (Covariate),

c) β is a vector of unknown parameters (where β are estimated by solving the least-square equations(1-4) and ϵ is a vector of unobservable of errors corresponding to the observation.

$$X'X\widehat{\beta} = X'Y$$

(1-4)

The approach used by [14] was to describe any given model in terms of its link function and its variance function. The variance function describes the relationship between

3

the mean and the variance of the dependent variable to allow for a proper calculation of the variance under non-normal conditions while the link function describes the non-linear relationship between the mean of the dependent variable and the linear right hand side.

GLMs are used for regression modeling for non-normal data. Suppose we generalize equation (1-3) with a linear predictor based on the mean of the outcome variable, then the function $g(\mu)$ will be called the link function.

$$g(\mu) = \theta = X\beta$$

(1-5)

The link function can be inverted as shown in equation (1-6)

$$\mu = g^{-1}(X\beta)$$

(1-6)

If y_i is binomially distributed with mean μ then the link function is logit as derived by [14] and expressed as:

$$g(\mu) = log\left[\frac{\mu}{1-\mu}\right] = X\beta$$

(1-7)

Then μ can be expressed as:

$$\mu = \frac{exp\,(\theta)}{1 + exp\,(\theta)}$$

(1-8)

1.2.2: Markov Chain

A Markov Chain is a time ordered probabilistic process consisting of a finite number of states and some known probabilities p_{ij}, where p_{ij} is the probability of moving from state i to state j [10]. That is, the probability at some point of time τ being in a certain state is conditioned on the states of the previous time, where the number of previous periods is termed as the order of the Chain. A Markov Chain is useful for analyzing events whose likelihood depends on what happened last.

4

1.2.3: The Markov Chain of first order

In the first-order Markov Chain, the current state is dependent solely on the state of the immediate previous period and the chance that a process is in state j at time τ given that it was in state i at time $\tau - 1$ is represented by transitional probability P_{ij} which is expressed as follows

$$P_{ij,\tau} = Pr(X_\tau = j | X_{\tau-1} = i)$$

(1-9)

P_{ij} can be estimated as shown below [16]:

$$P_{ij} = \frac{q_{ij}}{\sum_{j=1}^{c} q_{ij}}; \qquad i, j = 0, 1, 2, 3, \dots C$$

(1-10)

Where q_{ij} = historical frequency of transition from state i to state j and

C = the maximum number of states

1.2.4: High Order Markov Chain

A Markov Chain of order λ is referred to as high order Markov Chain if λ greater than 1. The probability that on time τ will have a particular state depends on the states of the previous time $\tau - \lambda$. For example the Markov Chains of order 2 and 3 satisfy the conditions in equation (1-11) and (1-12) respectively.

$$P_{i_2,i_1 j,\tau} = Pr(X_\tau = j | X_{\tau-1} = i_1, X_{\tau-2} = i_2, \dots, X_0 = i_0)$$
$$= Pr(X_\tau = j | X_{\tau-1} = i_1, X_{\tau-2} = i_2)$$

(1-11)

Where $P_{i_2,i_1 j,\tau}$ is the transition probability of state j in time τ, given state i_1 in time $\tau - 1$ and state i_2 in time $\tau - 2$

$$P_{i_3,i_2,i_1 j,\tau} = Pr(X_\tau = j | X_{\tau-1} = i_1, X_{\tau-2} = i_2, \dots, X_0 = i_0)$$
$$= Pr(X_\tau = j | X_{\tau-1} = i_1, X_{\tau-2} = i_2, X_{\tau-3} = i_3)$$

(1-12)

5

The equation (1-11) mean that it doesn't matter what stat the day $\tau - 3, \tau - 4$, and so on if you are referring to order 2.

Modeling a high order Markov Chain leads to a high-dimensional space of parameters [22]. A higher order Markov Chain say of order 2 with two states will have four parameters, order 3 will have eight parameters order 4 will have sixteen parameters and order λ will have 2^λ parameters. Increasing the number of states increases the number of parameter in each order. Such models may not be accurate in situation where there may be no sufficient data to estimate them. However, the papers [12] and [22] suggest that the high dimension of parameters can be reduced in such situations.

1.2.5: Modeling Rainfall data using Markov Chain

Since the work by [4] proposed Markov Chain for modeling the sequence of wet and dry days, it gained popularity in representing daily precipitation occurrence processes [13]. This has been contributed by the flexibility of the Markov Chain model and the ease of estimating the parameters and obtaining the fitted model without resorting to simulation [19].

The occurrence of a wet state in day τ is dependent of what state that occurred in the previous day's τ-1, τ-2, τ-3.... The random variable X_τ represents the occurrence or non-occurrence of rain on day τ such that

$$X_\tau = \begin{cases} 0 \ if \ day \ \tau \ is \ dry \\ 1 \ if \ day \ \tau \ is \ wet \end{cases}$$

$$(1\text{-}13)$$

1.2.6: Fitting a first order Markov Model to rainfall data

The first order model assumes that the probability of rain occurring on any day depends only on whether it did or did not rain on the previous day. To fit this model, the parameter for transition probability $p_{i,\tau}$ is estimated over the year [18]. The $P_{i,\tau}$ is the probability of rain in day τ given state i (for $i = 0,1$) in day $\tau - 1$. The estimate of $p_{i,\tau}$ is given by $r_{i,\tau}$ [19] which is the proportion of years with state i in their day $\tau - 1$ that had rain in their day τ. The $r_{i,\tau}$ is expressed as shown in equation (1-14).

$$r_{i,\tau} = \frac{n_{i1,\tau}}{n_{i1,\tau} + n_{i0,\tau}}$$

$$(1\text{-}14)$$

6

where

$n_{i1,\tau}$ is the number of years with rain on day τ

$n_{i0,\tau}$ is the number of years with no rain on day τ

The random variable $n_{i,\tau}$ takes a binomial distribution with the probability of success being $p_{i,\tau}$ and $(n_{i1,\tau} + n_{i0,\tau})$ is the number of trials. Therefore the model used is

$$p_{i,\tau} = g(\theta_{i\tau})$$

$$(1\text{-}15)$$

where $g(.)$ is a logit link function connecting the probabilities $p_{i,\tau}$ to the function $\theta_{i\tau}$ which is a linear unknown parameters [19]. The model is a generalized linear model since binomial is a member of the exponential family [14]. $p_{i,\tau}$ is therefore expressed as

$$p_{i,\tau} = \frac{exp\,(\theta_{i\tau})}{1 + exp\,(\theta_{i\tau})}$$

$$(1\text{-}16)$$

The work by [19] suggested that Fourier analysis may be used to express $\theta_{i\tau}$ as shown below:

$$\theta_{i\tau} = a_{i0} + \sum_{k=1}^{m} [a_{ik}\sin(k\tau') + b_{ik}\cos(k\tau')]$$

$$(1\text{-}17)$$

where $\tau' = 2\pi\tau/366$ and m is the number of harmonics

1.3: Statement of the problem

Markov Chain models have been in existence in the past to describe rainfall occurrence data [18]. Several advances are made in improving the Markov model fitting to rainfall data for example; the use of higher order Markov Chains [19] and finding a suitable Markov order [1]. However little has been done to make the modeling of rainfall data using Markov Chains accessible to users of statistical packages.

A powerful statistical package with specialized routine for analyzing climatic data is needed to handle such models. The introduction of Instat package was meant to simplify the

7

modeling of rainfall data and make it accessible to user though it cannot handle complex analyses like the Generalized Linear Models.

GenStat is another powerful package with the capability of climatic data analysis [6] but it doesn't have a climatic menu. Despite its potential and availability of a guide for climatic analyses, it is still not easy to use it for climatic analysis due to lack of a specialized climatic menu.

1.4: Objective of the study

Analysis in GenStat can be done through the command language if it cannot be achieved through dialogues and menus. It is possible to build the commands into a procedure that can perform a particular task. These procedures can then be used to develop special menus and dialogues. The main objective of this study is to start the implementation of an extension of the Instat climatic menu into the GenStat package through creating procedures, then developing them into menus.

The specific objectives are;

- Create a set of procedures in GenStat to support Markov Chain modeling of rainfall data.
- To add a menu and dialogs for each procedure to facilitate their use.
- Demonstrate the use of these procedures by analyzing daily rainfall data for Katumani, Kenya.

1.5: Significance of the study

Simplifying daily rainfall data analysis will enable more researchers to be able to do the analysis and hence include it in their work. Non specialists will have an opportunity to perform simple climatic analysis on their data without necessarily consulting the data specialists. This is intended to encourage more researchers to utilize climatic data in their research. The Markov modeling analysis will incorporate the Generalized Linear Model that is readily available in GenStat. This new facility will permit the comprehensive analysis of daily rainfall data to easily produce quality reports.

8

Chapter 2:   Literature review

The study on the sequence of daily rainfall occurrence was started by [4]. They assumed that the probability of rainfall on any day depended only on whether the previous day was wet or dry, that is whether rainfall did or did not occur. Given the event on the previous day, then, the probability of rainfall is assumed independent of events of further preceding days. They classified days as wet or dry according to whether there had or had not been recorded at least 0.1mm of precipitation in the 24-hour from 8.00 a.m. to 8.00 a.m. the following day. They found that the daily rainfall occurrence for the Tel Aviv data was successfully fitted with the first-order stationary Markov Chain model. The model parameters were the two conditional probabilities:

$$p_1 = p(Wet\ day/previous\ day\ wet\ )$$

$$(2\text{-}1)$$

$$p_0 = p(Wet\ day/previous\ day\ dry)$$

$$(2\text{-}2)$$

Given the Markov Chain and estimates of its two basic probabilities, they derived various properties of rainfall occurrence patterns. These included probabilities of rainfall i days after a wet or dry day. A wet spell of length k was defined as a sequence of k wet days preceded and followed by dry days. Dry spells were defined correspondingly. Weather cycles were defined as combinations of a wet spell and an adjacent dry spell. The distribution of spell lengths was found to be geometric. They then derived probabilities of a wet spell of length k and dry spell of length m, probability of a weather cycle of n days and the probability of exactly s wet days among n following a wet day. For large n, the distribution of the number of wet days tends to normality. The fit of the model was tested on data of daily rainfall in Tel Aviv [4] for the 27 rainy seasons 1923/24 – 1949/50. Chi-square tests of goodness of fit showed no significant deviations from the actual observations on rainfall so the model was said to hold.

Rainfall usually occurs in seasons of various lengths. They did not have any further coefficients of periodicity, persistence or any other parameters. The model was not suggested in terms of actual rainfall occurrence but as a statistical description of the observations. In

9

their paper [11], they reported that the first order of the Markov Chain model was found to fit the observed data in Italy successfully. This model was based on the assumption that there was a dependency of the daily rainfall occurrence to that of the previous day.

In the analysis of Samaru, Nigeria rainfall [17], he modeled the occurrence of rainfall as a first order Markov Chain. The probability of rainfall was not constant within each period of interest instead the probability of rainfall is a function of the time of the year. At the beginning and end of the season, the probability of rain is greater, if there was rain on the previous day, than if the previous day was dry. In the middle of the season, the probability of rain appears to be approximately the same irrespective of the state of the previous day. A linear logistic model was fitted to the data. However, to estimate the transition probabilities in the Markov Chain model, lots of data is required hence statistical software to ease the analysis.

A climatic Guide [21] described the use of Instat Software to analyze climatic data. The guide explained step by step how to analyze most of the element of climate. This simplifies the analysis of climatic data due its existing dialogue and menus specialized for analyzing climatic data. However there are instances when it is not sufficient. For example, situation where higher order Markov modeling (of more than order 2) is required, Instat cannot achieve. Alternative software to Instat is Marksim which is based on a stochastic weather generator and uses a third-order Markov process to model daily weather data [8].

The widely used stochastic model in rainfall modeling is the GLM [3] due to the stochastic nature of rainfall [13] this has not been achieved in Instat. GenStat software is a more powerful tool because it can be used to perform the more complex analysis of data especially the GLM. GenStat climatic Guide [5] describes how GenStat software can be used to analyze climatic data.

CHAPTER 3: FITTING MARKOV CHAIN MODELS USING INSTAT AND MARKSIM

3.1: Introduction

Markov models are fitted for rainfall data collected from Katumani Experimental Research Station in Kenya (longitude-1°35′ S, Longitude-37°14′ E, Altitude-1600mm) for the years 1961-2001 (41 years). The data is already available in the Instat library, and is used to fit a zero, first and second order Markov models using Instat package. The same data will be simulated using MarkSim software which also fits a third order Markov model to the data. The two packages (Instat and MarkSim) are used to establish the appropriate order of the Markov model of rainfall data for Katumani Experimental Research Station.

3.2: Instat.

Instat package has a specialized menu for Markov modeling of rainfall data. Using the Climatic Menu in Instat, Markov modeling sub menu involves three stages. These are; data preparation, fitting the model and using the model. Figure 1 below shows the "Climatic => Markov Modeling" Menu in Instat.
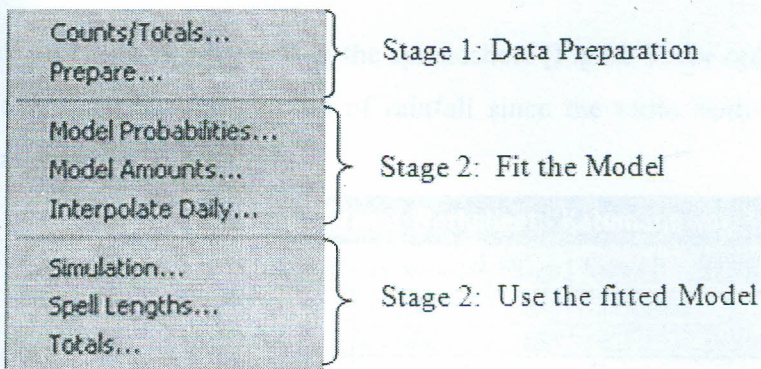

Figure 1: *Markov Modeling Menu*

3.2.1: Data preparation

The data preparation stage summarizes the rainfall data into counts and totals for the model fitting. The Menu "Climatic => Markov Modeling => Counts/Totals" (Figure 1) is used. It is possible to model a zero, first and second order Markov Chains for rainfall occurrences in Instat. When the zero option is selected in the 'Order for Count' section, it

calculates the total number of wet and dry days for each day of the year (two variables only; *dry* and *wet*), first order calculates the number of days with a specific state given the state in the previous day (four variables i.e. '*ww*','*wd*', '*dw*' and '*dd*') while second order calculates number of days with a specific state given the state in the two previous days (eight variables i.e. '*www*', '*wwd*', 'wdw', '*dww*', '*ddw*', '*dwd*','*wdd*', '*ddd*'). Figure 2 is a dialog box in Instat that is used to generate the counts/totals. (We illustrate these stages with a zero order Markov model)
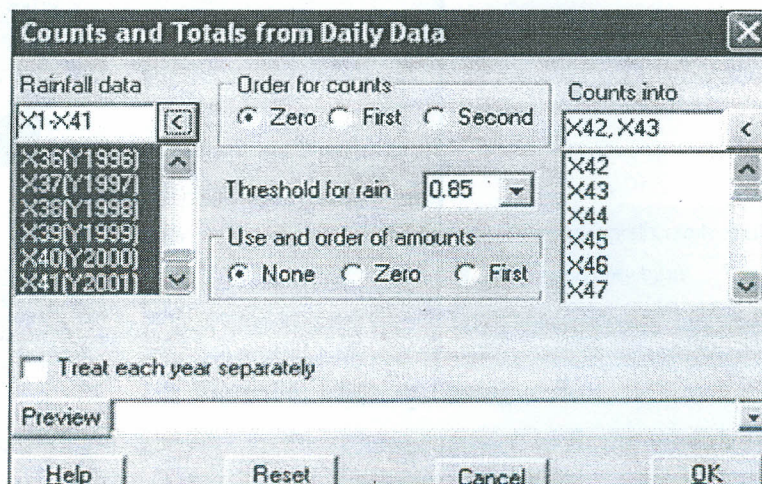

Figure 2: *Counts/Totals Dialogue box*

The result for count is displayed in the spreadsheet (Figure 3) for order zero. Remember the result did not display the amount of rainfall since the radio button *none* for amount was selected as shown in Figure 2.

Figure 3: *Counts of rainy days of order zero.*

The column X42 is renamed as 'dry' and X43 as 'rain'. They contain the number of occasions, in the 41 years, that each day is dry or rainy. For example, 8 years out of 41years had rainfall on day 5 (5th January), while, only 6 out of 41 years received rainfall on the 1st

January. From the counts, the chance of rainfall on a particular day can be calculated, for example using the result in Figure 3, the chance of rain on 1st January. is $\frac{6}{41}$ on 2nd January. $\frac{8}{40}$, on 3rd January $\frac{10}{41}$ etc. These probabilities are calculated using the Menu *"Climate => Markov Modeling => Prepare"*. The dialog (Figure 4) has the option of calculating the summary of rainfall counts and amounts on a daily basis or otherwise.
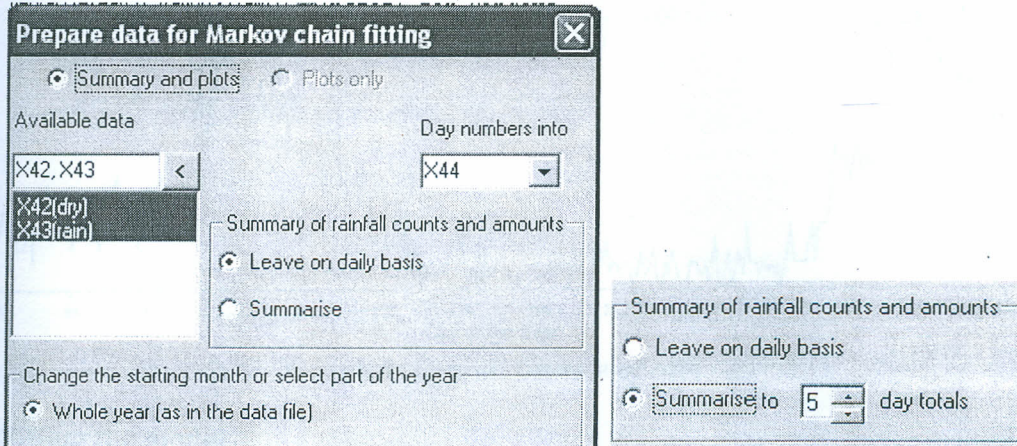
Prepare data for Markov chain fitting

- (•) Summary and plots   ( ) Plots only

Available data   Day numbers into

| X42, X43 | < | | X44 ▾ |

X42(dry)
X43(rain)

Summary of rainfall counts and amounts
- (•) Leave on daily basis
- ( ) Summarise

Summary of rainfall counts and amounts
- ( ) Leave on daily basis
- (•) Summarise to [5 ▲▼] day totals

Change the starting month or select part of the year
- (•) Whole year (as in the data file)

Figure 4: *Prepare Dialogue Box*

The Prepare dialogue box sets up the columns needed to fit the models (Figure 5) and also produces a graph of the chance of rain (Graph 1).

Current Worksheet - Katumani.wor

Columns | Constants | Strings | Labels   Title: Katumani Experimental Research Station - Daily Rainfall (mm) 1961-2001 Lat Lon

	X45	X46	X47	X48	X49	X50	X51	X52	X53	X54	X55	X56
	p_r	circ	ds1	dc1	ds2	dc2	ds3	dc3	ds4	dc4	ds5	dc5
1	0.1463415	-3.124425	716633E-02	-0.9998527	0.0343276	0.9994106	147876E-02	-0.9986741	861474E-02	0.9976432	-0.0857305	-0.9963183
2	0.1951219	-3.107258	-0.0343276	-0.9994106	861474E-02	0.9976432	-0.102821	-0.9946999	0.1369061	0.990584	-0.1708297	-0.9853005
3	0.2439024	-3.090091	147876E-02	-0.9986741	0.102821	0.9946999	-0.1538906	-0.9880879	0.2045521	0.9788557	-0.2546712	-0.9670277
4	0.1707317	-3.072924	861474E-02	-0.9976432	0.1369061	0.990584	-0.2045521	-0.9788557	0.2712339	0.9625134	-0.3366373	-0.9416344
5	0.1951219	-3.055757	-0.0857305	-0.9963183	0.1708297	0.9853005	-0.2546711	-0.9670277	0.3366373	0.9416344	-0.4161247	-0.9093075
6	0.2926829	-3.03859	-0.102821	-0.9946999	0.2045521	0.9788557	-0.3041148	-0.9526354	0.4004539	0.9163169	-0.492548	-0.8702853
7	0.2439024	-3.021422	-0.1198812	-0.9927883	0.2380333	0.971257	-0.3527521	-0.9357168	0.462383	0.8866803	-0.5653447	-0.8248548
8	0.1707317	-3.004255	-0.1369061	-0.990584	0.2712339	0.9625134	-0.4004539	-0.9163169	0.5221326	0.8528643	-0.6339785	-0.7733507
9	0.1707317	-2.987088	-0.1538906	-0.9880879	0.3041148	0.9526354	-0.4470938	-0.8944871	0.5794211	0.8150284	-0.6979442	-0.7161522

Figure 5: *Probability of rainy day or order zero*

The columns set in Figure 5 above X45 shows the probability of rain while X46-X56 shows the figures that are used for fitting the curve.

13

Overall chance of rain
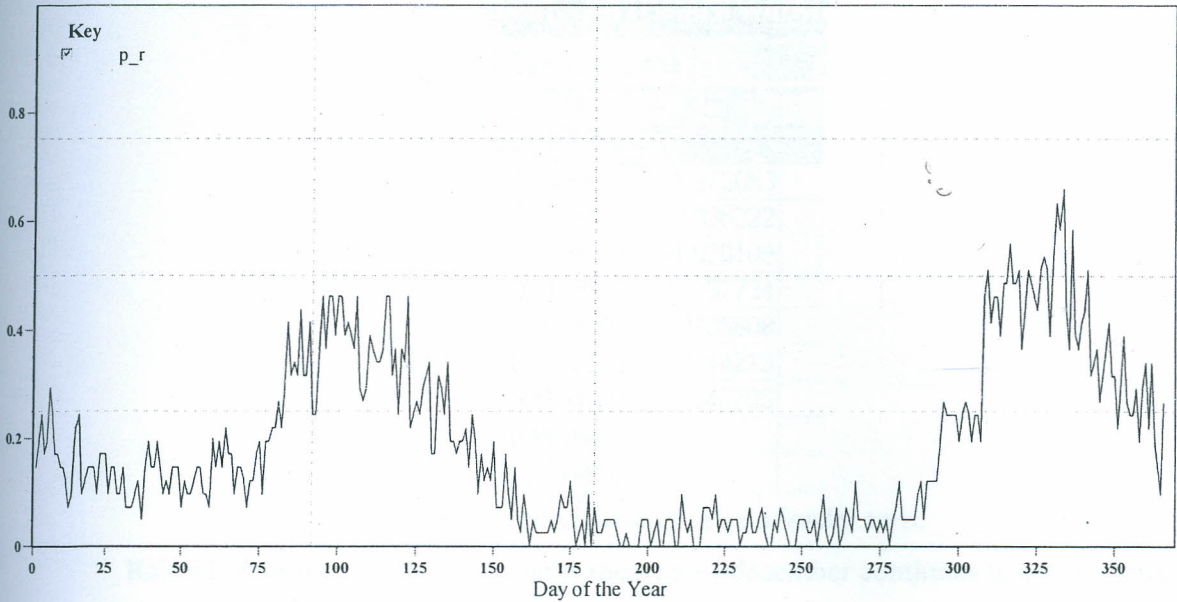
Graph 1: *Overall Chance of rain*

3.2.2: Fitting the model

The second stage of Markov modeling is fitting the Markov model to the rainfall data. It uses the model probabilities dialogue which is obtained through, "*Markov Modeling => Model Probabilities*" menu which opens fitting the probabilities dialoge box (Figure 6) This produces the fitted probabilities, which are given in a column called 'f_r' (Figure 7) and a curve of the fitted probabilities, Graph 2
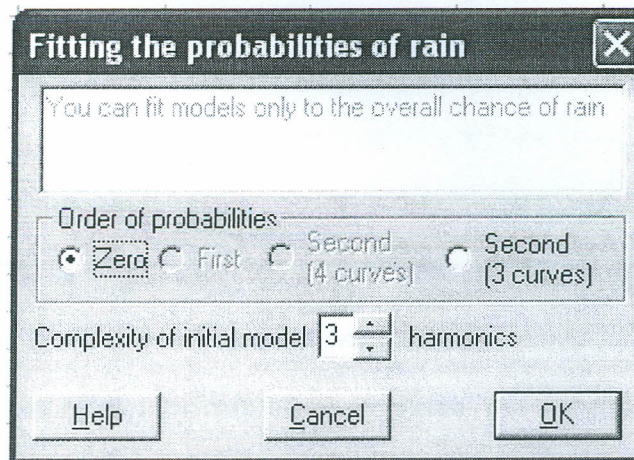

Figure 6: *Fitting Probability Dialogue box*

14

Strings	Labels	Title:
X58	X59	
f_r	estr	
0.2406601	-1.973053	
0.2314891	-0.5935322	
0.2225461	-1.020109	
0.2138521	-0.9957734	
0.2054253	-0.2795608	
0.1972812	0.5144213	
0.1894324	-0.1340795	
0.1818888		
0.174658		

Figure 7: *Fitted Probabilities of order zero.*

Rainfall data is cilic in nature that is the 31st of december continues to 1st of january the following year. There sin and cosine functions are used as indicated in table 5 to compute the fitted values in table 7
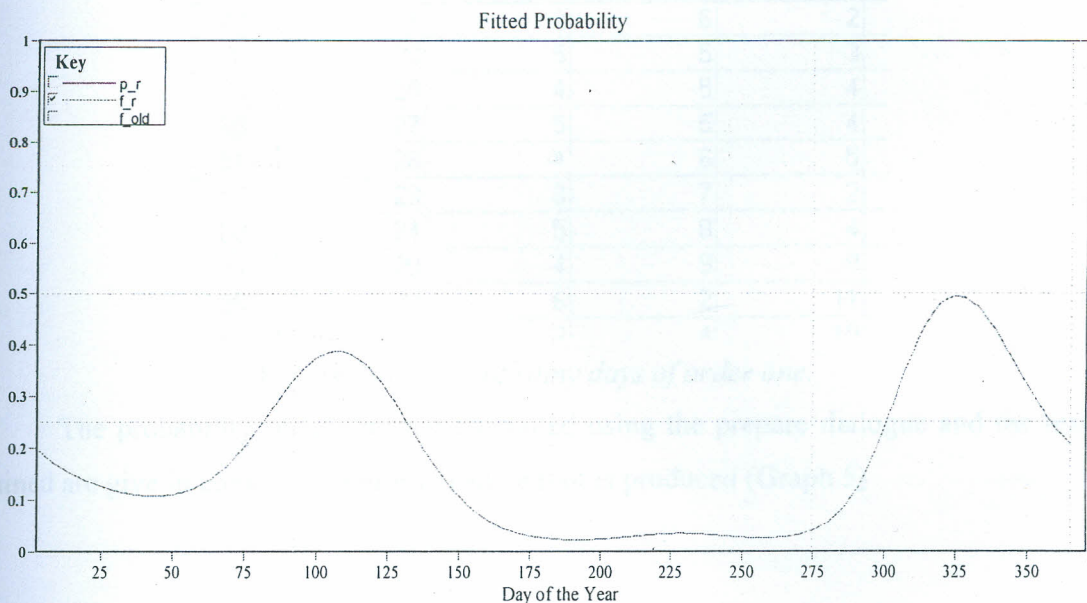
Graph 2: *Fitted Probability curve*

3.2.3: First Order Markov Model for rainfall occurrences

Instat has the ability to fit Markov Chains for counts of up to order two to rainfall data. A first order Markov Chain is modeled by using the dialogue box for Count/Totals is shown below (Figure 8).
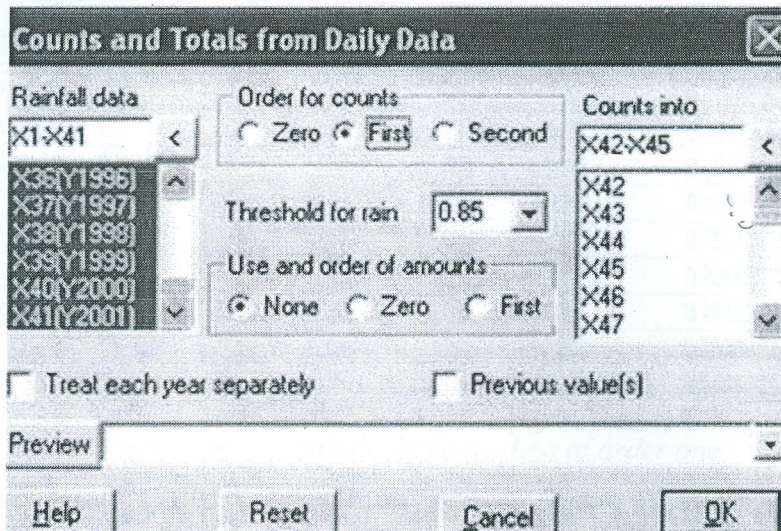
15

Figure 8: *Count/Totals dialogue box for order one.*

The results for the first order is distplayed in figure 9. Four variable are calculated these include *'dd', 'dr', 'rd', 'rr'.*

|	X42	X43	X44	X45
	dd	dr	rd	rr
77	31	2	6	2
78	28	5	5	3
79	28	4	5	4
80	27	5	5	4
81	26	4	6	5
82	23	9	7	2
83	24	5	8	4
84	20	4	9	8
85	22	6	2	11
86	24	3	4	10

Figure 9: *Counts of rainy days of order one.*

The probability of rainfall are computed using the prepare dialogue and the results obtained are give in columns (Figure 10) and a plot is produced (Graph 5)

	X42	X43	X44	X45	X46	X47	X48	X49	X50
	dd	dr	rd	rr	Date	_d	p_rd	_r	p_rr
16	140	20	27	18	79	167	0.1616766	38	0.4736842
17	113	27	30	35	84	143	0.2097902	62	0.5645161
18	95	36	39	35	89	134	0.2910448	71	0.4929577
19	107	29	27	42	94	134	0.2014925	71	0.5915493
20	87	26	30	62	99	117	0.2564103	88	0.7045454
21	87	35	35	48	104	122	0.2868852	83	0.5783132
22	105	34	30	36	109	135	0.2222222	70	0.5142857
23	91	33	37	44	114	128	0.2890625	77	0.5714286
24	100	38	33	34	119	133	0.2481203	72	0.4722222

Figure 10: *Probability of rainy days of order one*

The computation of the probabilities given in Figure 10 is done based on 5 days totals . Date 79 contains the totals for date 77 to date 81 i.e., 27 in rd column is computed by adding all the rd's from date 77 to date 81. Hence P_rd= (rd / _d)= 27/167 =0.161.

First Order and overall chance of rain
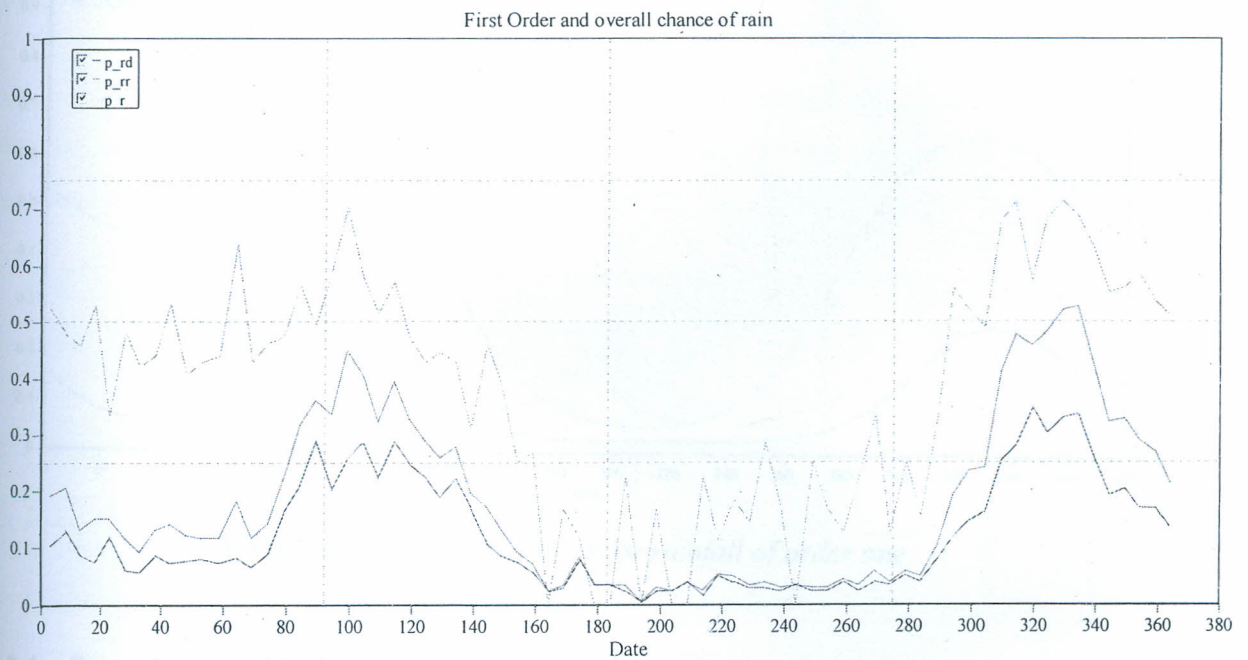
Graph 3: *Probability curves for order two*

The 'Fitting the probability' dialog box is used as shown in Figure 11 is used to fit the probabilities of rainfall. It corresponding results is shown in the
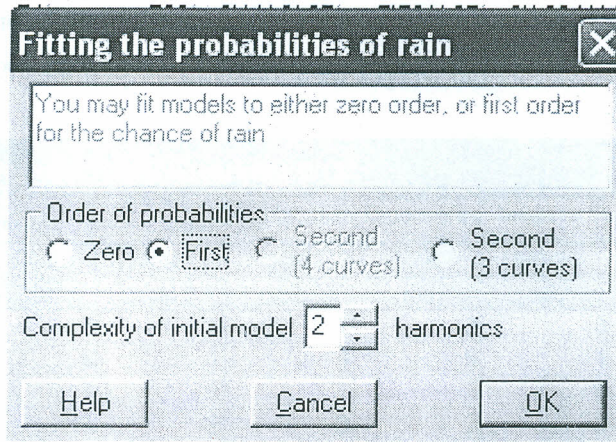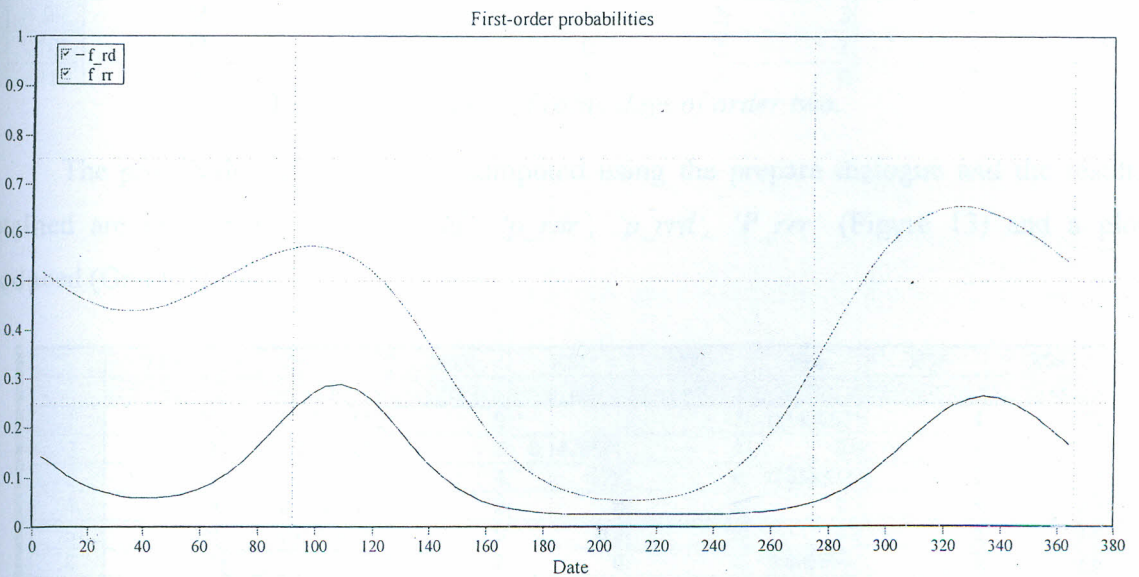
17

Fitting the probabilities of rain ✕

You may fit models to either zero order, or first order for the chance of rain

Order of probabilities
○ Zero ● First ○ Second (4 curves) ○ Second (3 curves)

Complexity of initial model 2 ⇕ harmonics

Help   Cancel   OK

Figure 11: *Fitting probabilities dialogue box for order one*

First-order probabilities

Graph 4: *Fitted Probabilities for rainfall of order one*

3.2.4: Second Order Markov Model for rainfall occurrences

To prepare data for second order Markov Chain model, then 'second-Order of Count' radio button is selected from the 'Count and Total from Daily Data' dialogue box. The results for second order gives 8 columns for counts namely '*ddd*', '*ddr*', '*drd*', '*drr*', '*rdd*', '*rdr*', '*rrd*', '*rrr*' (Figure 12). It is pointed out that out of 41 years, 26 years had a dry day on 3rd January given dry on 2nd January given dry on 1st January being (thick lined circle on Figure

18

12). On the other hand, 2 out of 41 years had a rainy day on 7th January given dry on 6th January given rain on 5th January (dotted circle on Figure 12).

	X42	X43	X44	X45	X46	X47	X48	X49
	ddd	ddr	drd	drr	rdd	rdr	rrd	rrr
1	28	0	6	1	2	0	1	3
2	23	6	1	3	5	1	1	1
3	26	3	1	1	3	1	5	1
4	27	2	2	3	2	0	2	3
5	27	4	1	1	2	1	1	4
6	24	2	1	2	7	0	2	3
7	24	1	4	2	2	2	3	3
8	24	5	2	3	1	1	2	3
9	27	2	2	3	2	3	0	2
10	27	4	4	0	2	1	1	2
11	29	4	1	1	2	0	2	2

Figure 12: *Counts of rainy days of order two.*

The probability of rainfall are computed using the prepare dialogue and the results obtained are give in columns *'p_rdd', 'p_rdr', 'p_rrd', 'P_rrr'* (Figure 13) and a plot produced (Graph 5)

	X51	X52	X53	X54	X55	X56	X57	X58
	_dd	p_rdd	_dr	p_rdr	_rd	p_rrd	_rr	p_rrr
1	29	0.06896552	0	*	7	0.1428571	4	0.75
2	27	0.1851852	7	0.1428571	2	0.5	4	0.25
3	29	0.1034483	4	0.25	6	0.8333333	2	0.5
4	29	0.06896552	2	0	4	0.5	6	0.5
5	29	0.06896552	5	0.2	2	0.5	5	0.8
6	31	0.2258064	2	0	3	0.6666667	5	0.6
7	26	0.07692308	3	0.6666667	7	0.4285714	5	0.6
8	25	0.04	6	0.1666667	4	0.5	6	0.5
9	29	0.06896552	5	0.6	2	0	5	0.4
10	29	0.06896552	5	0.2	5	0.2	2	1

Figure 13: *Probability of rainy days of order two*

The results for 2nd order Markov Chain plots two sets of curves; the 1st order and 2nd order probability curves. The curves in Graph 5 do not have clear distinctions and therefore do not present the probabilities well.

19

Second and first order

Key	
☑	──── p_rdd
☑	·········· p_rdr
☑	──── p_rrd
☑	──── p_rrr
☐	·········· p_rd
☐	p_rr

Day Number of the Year
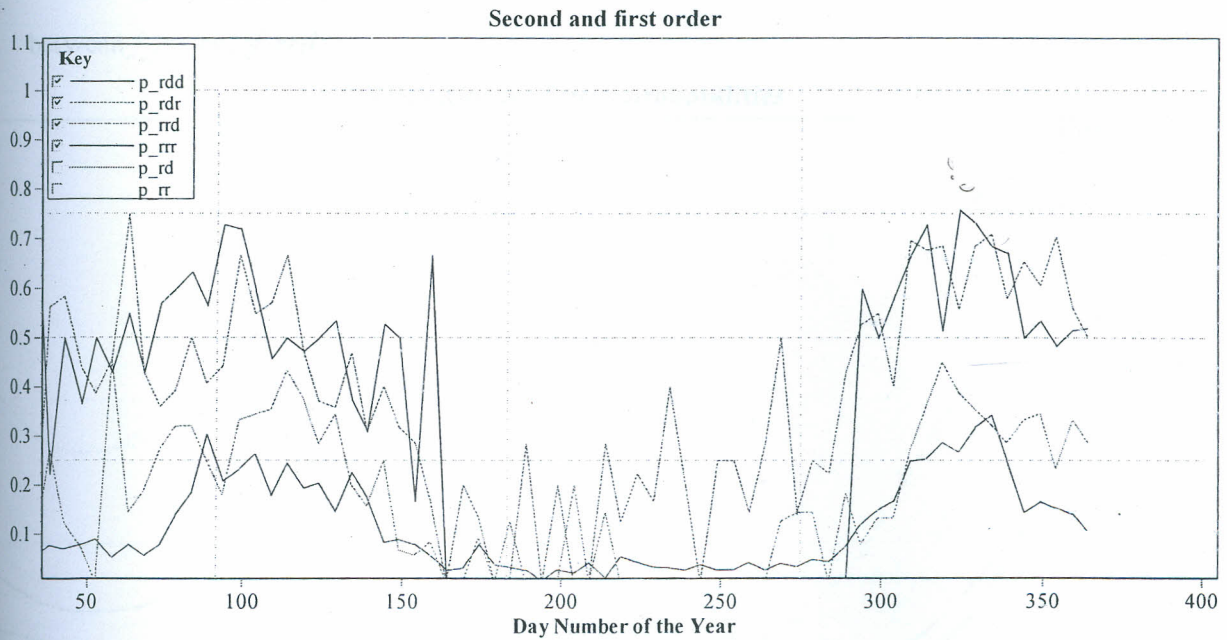
Graph 5: *Probability curves for order two*

The fitted model of order two is now obtained by using 'Fitting Probability dialog box shown in Figure 14. The order of probabilities option to be used at this stage is 'Second (4 curves)' if you want all the curves for rain given the conditions of the two previous days ('*rrr*', '*rrd*', '*rdr*' and '*rdd*').

Fitting the probabilities of rain

You may fit models to either zero order, or first order or second order data for the chance of rain

Order of probabilities
- ○ Zero ○ First ● Second (4 curves) ○ Second (3 curves)

Complexity of initial model [2] harmonics

Help   Cancel   OK
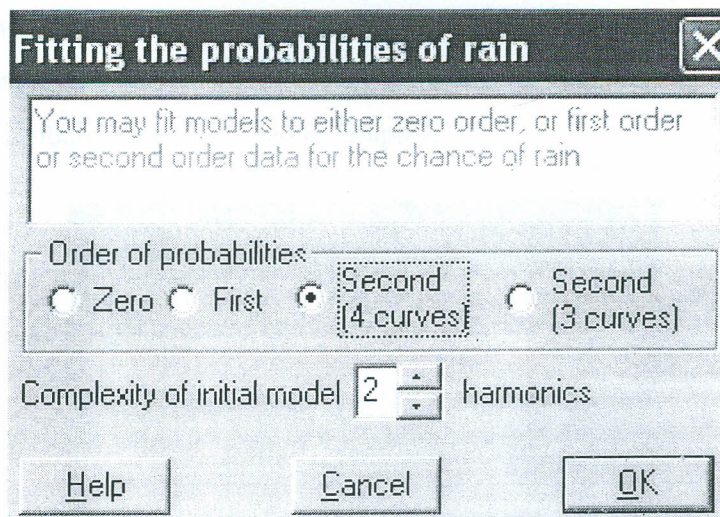
Figure 14: *Fitting probabilities dialogue box for order two with four curves*

20

Graph 6 shows the fitted probabilities of order two. There is no clear distinction between f_rrr and f_rrd.

Second-order Fitted probabilities
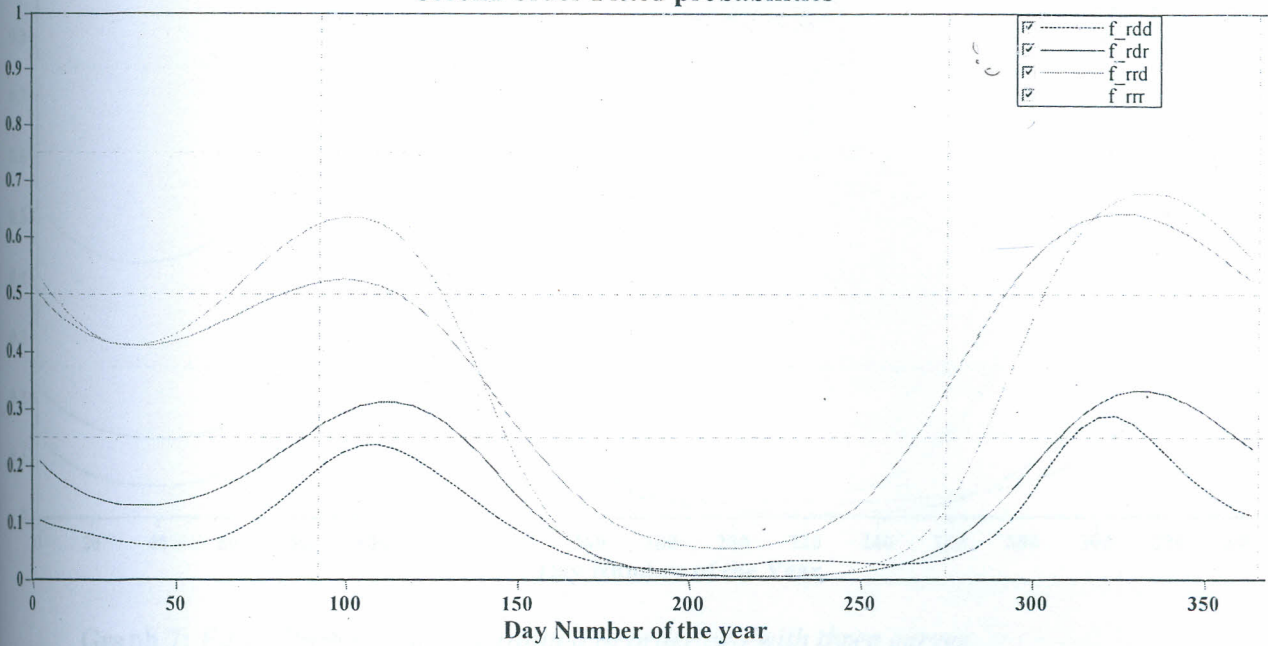
Graph 6: *Fitted Probabilities for rainfall of order two with four curves*

When there is no clear distinction between f_rrr and f_rrd, Instat has an option to model them together by the first order probability f_rr. This option can be seen in the dialog box in Figure 15 as the choice of 'second order with three curves'. This phenomenon is quite common in rainfall data where a rain event seems to "clear the memory" of what happened previously. The fitted probabilities are shown in Graph 7.
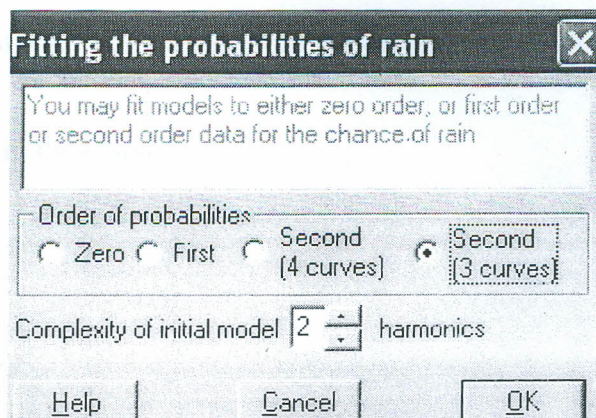
Figure 15: *Fitting probabilities dialogue box for order two with three curves*

21

Second-order probabilities
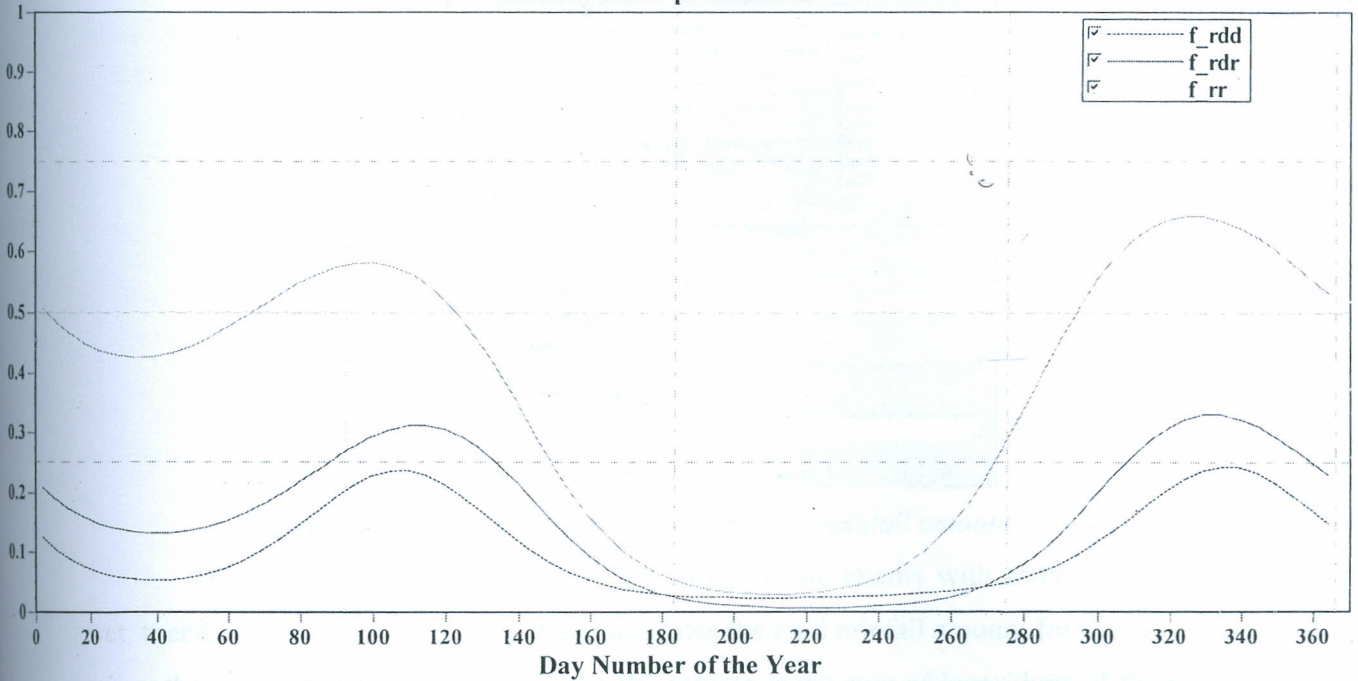
Graph 7: *Fitted Probabilities for rainfall of order two with three curves*

The current version of Instat the Markov command is limited to second-order Markov Chains, however rainfall data for some sites are better modeled using higher order Markov Chains.

3.2.5: Modeling rainfall amounts.

Instat can also model rainfall amounts on rainy days. Modeling the amounts involves the three stages as discussed in Section 3.2: Climatic => Markov Modeling => Counts/Totals menu is used. In the Count and Total dialogue box we choose zero order for amount as shown in

Figure **16** (circle in thick line).

22
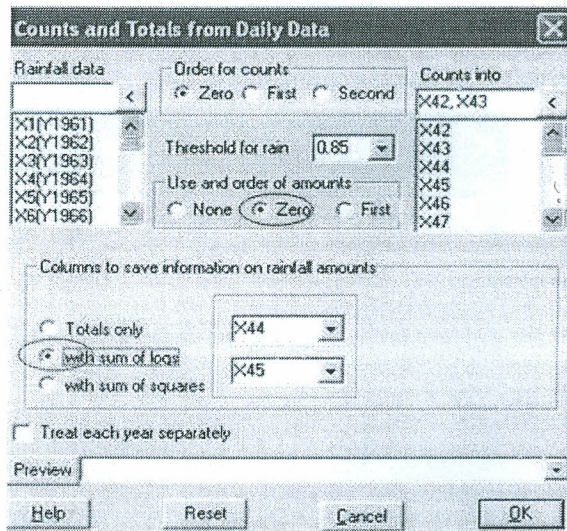
Figure 16: *Count and totals dialogue box for rainfall amount of order zero*

The result after clicking OK button will generate results with three columns; dry, wet, tr and lr (Figure 17). The 'tr' column indicates the total rainfall amount for the rainy days minus the total threshold values while the 'lr' column is the sum of logarithms of the rainfall minus the threshold. The two columns are computed using the equations (3-1) and (3-2)

$$tr_\tau = \sum_{y=1}^{n} (A_\tau - T)_y$$

(3-1)

$$lr_\tau = \sum_{y=1}^{n} \ln(A_\tau - T)_y$$

(3-2)

Where A_τ is the rainfall amount on day τ and T is the threshold value and y is the year number.

	dry	rain	tr	lr
1	35	6	22.2	3.5530
2	33	8	116.7	19.3970
3	31	10	101	16.1049
4	34	7	78.85	13.0649
5	33	8	87.2	13.6700
6	29	12	101	20.4840

Figure 17: *Total rain amounts of order zero.*

The dry and wet columns in the figure above give the number of years in which the day of the year was dry or rainy. The mean amount of rain is obtained by dividing the total tr

23

by the number of rainy days. When fitting the model for amount, Climatic => Markov Modeling => Model Amounts menu is used and it fits the mean rain per rainy day.
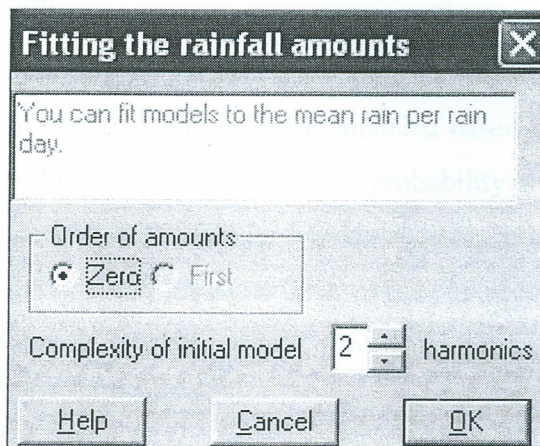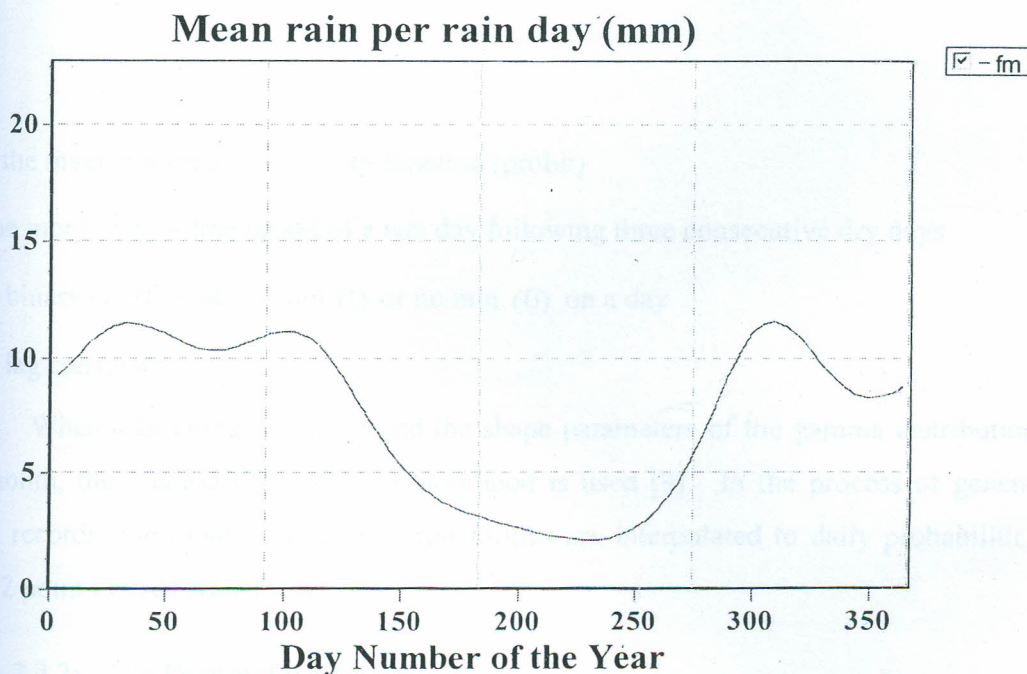

Figure 18: *Fitting rainfall amount for order zero*

Graph 8: *Fitted mean rainfall amounts for order zero.*

The fitted curve is as shown the Graph 8. The current version of Instat can only fit Markov model up to order 1.

3.3: MarkSim

3.3.1: Introduction

MarkSim is a computer tool that generates simulated data for crop modeling and risk assessment [8]. Rainfall in the MarkSim model is modeled based on the third order Markov Chain in two stages. The first stage is transitional probability to determine whether any particular day is wet. Since it is a third order Markov Chain, this wet day will depend on whether there was any rainfall in the previous three days. The second stage of the MarkSim model determines the amount of rainfall. When using MarkSim, a parameter file is created and then used to simulate daily climatic data.

MarkSim defines the probability of day i being wet as

$$P(W_i/a_1 a_2 a_3) = \varphi^{-1}(b_i + a_{i-1}d_1 + a_{i-2}d_2 + a_{i-3}d_3)$$

(3-3)

where:

ϕ^{-1} is the inverse normal probability function (probit)

b_i is the monthly baseline probit of a wet day following three consecutive dry days

a_m are binary coefficients of rain (1) or no rain (0) on a day

d_m are lag constants

When estimating the mean and the shape parameters of the gamma distribution for each month, the method of maximum likelihood is used [9]. In the process of generating rainfall records, the monthly baseline probabilities are interpolated to daily probabilities by using 12 point Fourier transform.

3.3.2: The Parameter file

The information that is required to create a parameter file is the longitude, latitude and elevation of the site of interest. The longitude (1.58333 degrees), latitude (37.2333 degrees) and elevation (1600 meters) for Katumani when entered, MarkSim will now create a parameter file for the simulated data.

25

Figure 19 shows the parameter file created from entering the longitude, latitude and elevation of Katumani into MarkSim.
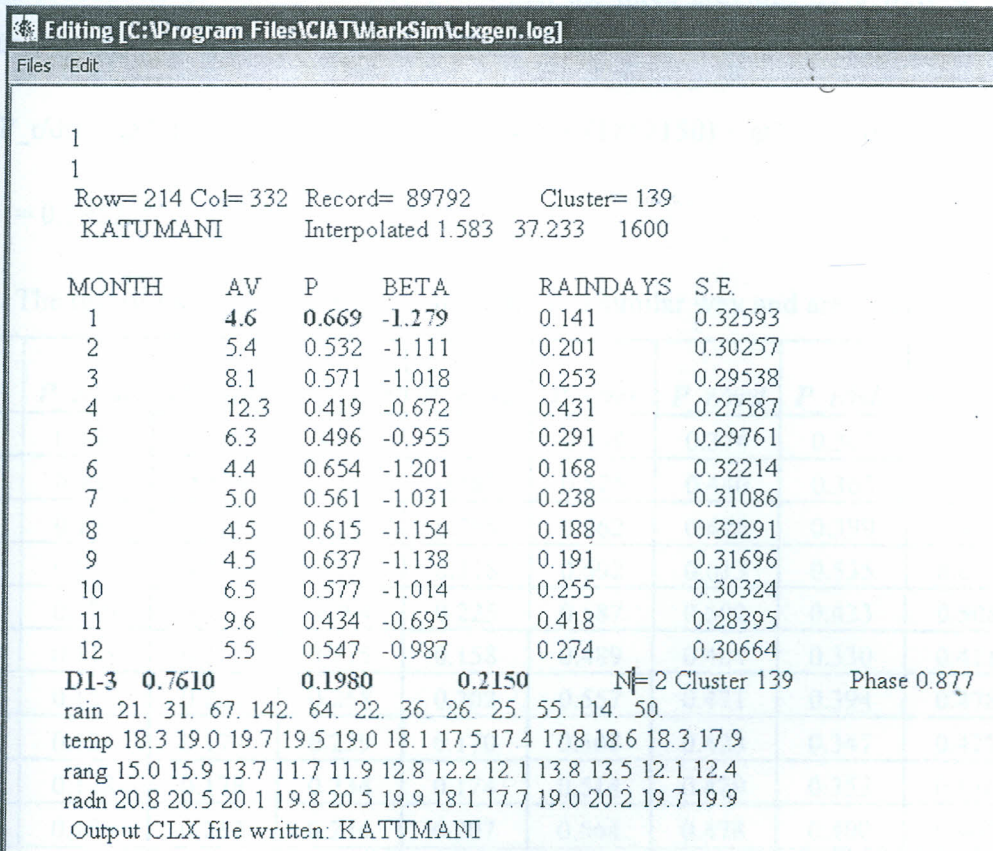
```
Editing [C:\Program Files\CIAT\MarkSim\clxgen.log]
Files   Edit

1
1
Row= 214 Col= 332  Record= 89792        Cluster= 139
KATUMANI           Interpolated 1.583  37.233   1600

MONTH      AV    P      BETA      RAINDAYS  S.E.
  1        4.6   0.669  -1.279    0.141     0.32593
  2        5.4   0.532  -1.111    0.201     0.30257
  3        8.1   0.571  -1.018    0.253     0.29538
  4        12.3  0.419  -0.672    0.431     0.27587
  5        6.3   0.496  -0.955    0.291     0.29761
  6        4.4   0.654  -1.201    0.168     0.32214
  7        5.0   0.561  -1.031    0.238     0.31086
  8        4.5   0.615  -1.154    0.188     0.32291
  9        4.5   0.637  -1.138    0.191     0.31696
 10        6.5   0.577  -1.014    0.255     0.30324
 11        9.6   0.434  -0.695    0.418     0.28395
 12        5.5   0.547  -0.987    0.274     0.30664
D1-3   0.7610          0.1980        0.2150      N= 2 Cluster 139      Phase 0.877
rain 21. 31. 67. 142. 64. 22. 36. 26. 25. 55. 114. 50.
temp 18.3 19.0 19.7 19.5 19.0 18.1 17.5 17.4 17.8 18.6 18.3 17.9
rang 15.0 15.9 13.7 11.7 11.9 12.8 12.2 12.1 13.8 13.5 12.1 12.4
radn 20.8 20.5 20.1 19.8 20.5 19.9 18.1 17.7 19.0 20.2 19.7 19.9
Output CLX file written: KATUMANI
```

Figure 19: *Parameter File for Katumani*

3.3.3: The Baseline Probits and the Lag Parameters

The probability of rain given the state in the three previous days is calculated from the parameter file using (3-3). The monthly baseline probit values show the probability of rain, given that it was dry in the three previous days as $\varphi^{-1}(b_1)$ i.e. when all a_m are 0. The baseline probit value in January was -1.279 (Figure 19: *Parameter File for Katumani*Figure 19) implying that in January the P_r/ddd $= \varphi^{-1}(-1.279)$ or 0.1004.

In a third-order model, eight different rainfall probabilities are considered, '*ddd*', '*ddr*', '*drr*', '*drd*', '*rrr*', '*rrd*', '*rdd*' and '*rdr*', with *r* standing for a rainy day and *d* for a dry day. MarkSim has a third order model for each month, where the eight different probabilities

26

are calculated using the 12 baseline probit (BETA) values and the three lag coefficients (D1-3). MarkSim therefore uses just 15 parameters to calculate the 96 (12 times 8) probabilities of rain. The probability of rain in January given that the three previous days were dry, rainy, rainy, for example is

$$P\_r/drr = \varphi^{-1}(-1.279+(0)0.7610+ (1)0.1980 + (1)0.2150) = \varphi^{-1}(-0.866)$$

$$= 0.1922.$$

The rest of the probabilities are calculated in a similar way and are listed (

	P_r/ddd	P_r/ddr	P_r/drr	P_r/drd	P_r/rrr	P_r/rrd	P_r/rdd	P_r/rdr
Jan	0.100	0.144	0.193	0.140	0.458	0.374	0.302	0.381
Feb	0.133	0.185	0.243	0.181	0.525	0.440	0.363	0.446
Mar	0.154	0.211	0.273	0.206	0.562	0.476	0.399	0.483
Apr	0.251	0.324	0.398	0.318	0.692	0.613	0.535	0.619
May	0.170	0.230	0.294	0.225	0.587	0.502	0.423	0.508
Jun	0.115	0.162	0.215	0.158	0.489	0.404	0.330	0.411
Jul	0.151	0.207	0.268	0.202	0.557	0.471	0.394	0.478
Aug	0.124	0.174	0.229	0.170	0.508	0.423	0.347	0.429
Sep	0.128	0.178	0.234	0.174	0.514	0.429	0.353	0.436
Oct	0.155	0.212	0.274	0.207	0.564	0.478	0.400	0.485
Nov	0.244	0.316	0.389	0.310	0.684	0.604	0.526	0.611
Dec	0.162	0.220	0.283	0.215	0.574	0.489	0.411	0.496

Table 2)

| Months	'ddd'	'ddr'	'drr'	'drd'	'rrr'	'rrd'	'rdd'	'rdr'
Jan	-1.279	-1.064	-0.866	-1.081	-0.105	-0.32	-0.518	-0.303
Feb	-1.111	-0.896	-0.698	-0.913	0.063	-0.152	-0.35	-0.135
Mar	-1.018	-0.803	-0.605	-0.82	0.156	-0.059	-0.257	-0.042
Apr	-0.672	-0.457	-0.259	-0.474	0.502	0.287	0.089	0.304
May	-0.955	-0.74	-0.542	-0.757	0.219	0.004	-0.194	0.021
Jun	-1.201	-0.986	-0.788	-1.003	-0.027	-0.242	-0.44	-0.225
Jul	-1.031	-0.816	-0.618	-0.833	0.143	-0.072	-0.27	-0.055
Aug	-1.154	-0.939	-0.741	-0.956	0.02	-0.195	-0.393	-0.178
Sep	-1.138	-0.923	-0.725	-0.94	0.036	-0.179	-0.377	-0.162
Oct	-1.014	-0.799	-0.601	-0.816	0.16	-0.055	-0.253	-0.038
Nov	-0.695	-0.48	-0.282	-0.497	0.479	0.264	0.066	0.281
Dec	-0.987	-0.772	-0.574	-0.789	0.187	-0.028	-0.226	-0.011

Table 1: *Probit values for the twelve months*

After the probit values have been computed, their corresponding values of probability are then computed. This can be archived well using MS-Excel.

28

	P_r/ddd	P_r/ddr	P_r/drr	P_r/drd	P_r/rrr	P_r/rrd	P_r/rdd	P_r/rdr
Jan	0.100	0.144	0.193	0.140	0.458	0.374	0.302	0.381
Feb	0.133	0.185	0.243	0.181	0.525	0.440	0.363	0.446
Mar	0.154	0.211	0.273	0.206	0.562	0.476	0.399	0.483
Apr	0.251	0.324	0.398	0.318	0.692	0.613	0.535	0.619
May	0.170	0.230	0.294	0.225	0.587	0.502	0.423	0.508
Jun	0.115	0.162	0.215	0.158	0.489	0.404	0.330	0.411
Jul	0.151	0.207	0.268	0.202	0.557	0.471	0.394	0.478
Aug	0.124	0.174	0.229	0.170	0.508	0.423	0.347	0.429
Sep	0.128	0.178	0.234	0.174	0.514	0.429	0.353	0.436
Oct	0.155	0.212	0.274	0.207	0.564	0.478	0.400	0.485
Nov	0.244	0.316	0.389	0.310	0.684	0.604	0.526	0.611
Dec	0.162	0.220	0.283	0.215	0.574	0.489	0.411	0.496

Table 2: *Probability values for the twelve months of Katumani Simulated data*

The value of the lag coefficient represents the extent of separation of the probit values which translates to the difference in the rainfall probabilities. For example, a large first lag coefficient value would represent a large difference between the probability of rain given the previous day was rainy and the probability of rain given that the previous day was dry. A small lag coefficient would represent a small separation of the rainfall probabilities at that coefficient. A very small third lag coefficient would suggest that a second-order model would be appropriate.
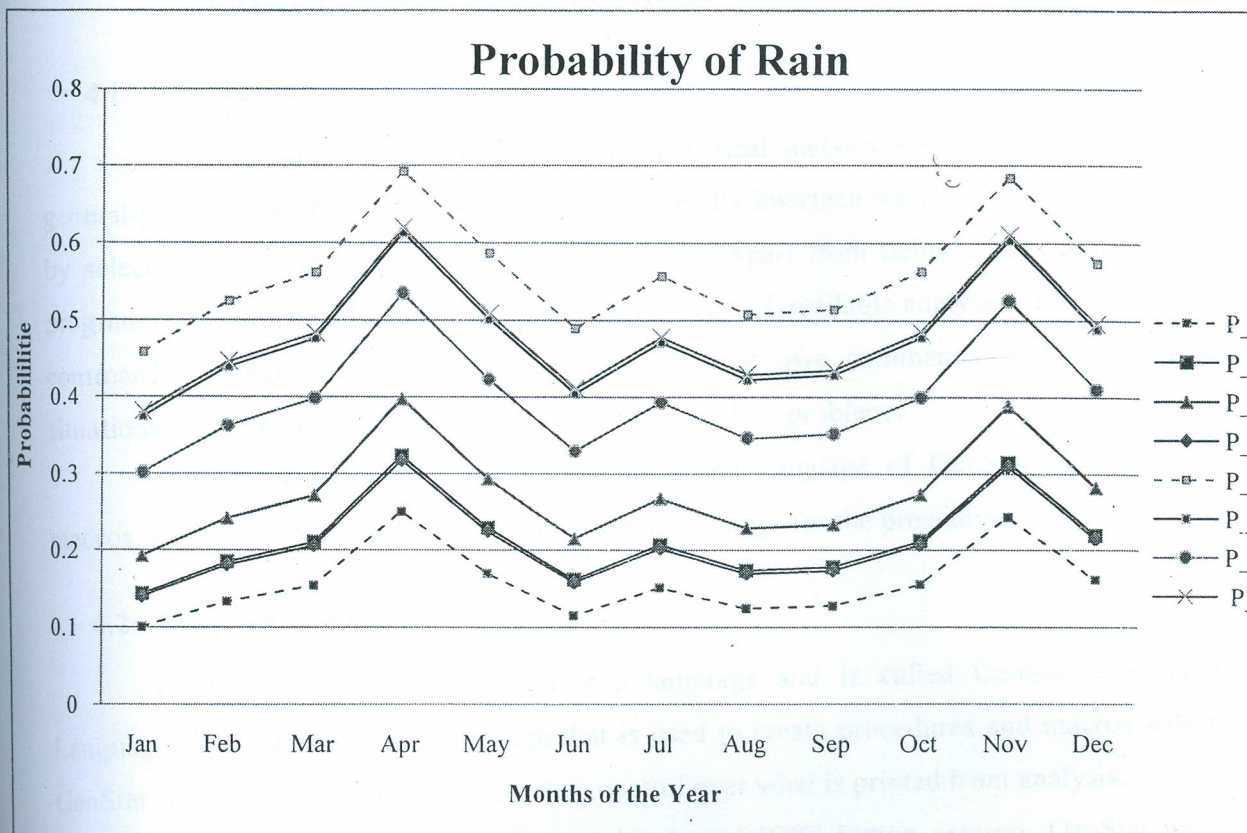
Graph 9: *The Eight Different Probabilities of Rain across the Year*

Graph 9 plots the eight rainfall probabilities over the year. With the large D1 value of 0.5295 there are two distinct groups of probabilities. In the first group the previous day was rainy and in the second group the previous day was dry. The probability of rain is higher when the previous day was rainy. The eight curves on the probit scale would be exactly parallel with the differences in probabilities varying slightly through the year.

Basically, Marksim was designed as computer tool that generates simulated rainfall data amongst other for crop modeling and risk assessment and to model a third order Markov Chain for the simulated rainfall data. The assumption of third order models for different sites is not practical to some sites and therefore Marksim models can only be used well for rainfall data that require third order Markov Chain model. GenStat capability to model any order of Markov Chains is utilized in this work to develop a special routine for doing the same.

CHAPTER 4: GENSTAT COMMAND INTERFACE

4.1: Introduction

GenStat is a computer program for statistical analysis, with all the facilities of a general-purpose statistics package. It has a user friendly interface which makes it easy to use by selecting option from menus and dialogue box. Apart from being a collection of pre-programmed commands for selecting from fixed options of available analyses, it has a flexible command language which enables users to write their own commands and use them in situations where the available commands cannot solve their problems.
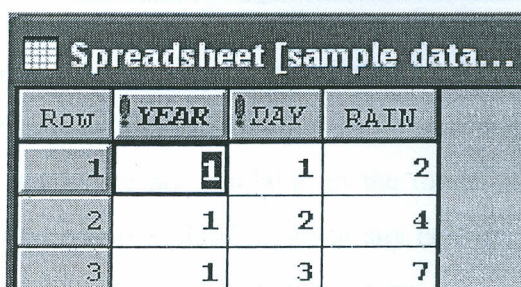
This chapter describes the use of command language of GenStat including the macros, procedures and how to create the menu and dialogs for the procedures.

4.2: Command language

GenStat's own commands forms a language and is called GenStat Command Language. It is this command language that is used to create procedures and macros within GenStat. The commands give you complete control over what is printed from analysis.

Whenever a dialogue box is completed, and [OK] button pressed, GenStat itself produces commands corresponding to the analysis process which are written in the Input Log and sent to the GenStat Server for processing. The results are sent to the output window once the Server has processed the commands. If there were any error in the commands, then the description is sent to the Error Log, as well as to the Output window.

We use a sample data shown in Figure 20 to illustrate the command language.

Figure 20: *Sample rainfall data*

31

To calculate the maximum and minimum rainfall per year 'Start=> Summary Statistics => Summaries of Groups (Variate)' menu is used to obtain the dialogue box selected as shown in Figure 21. The corresponding command for the analysis is generated in the input log window shown in Figure 22.
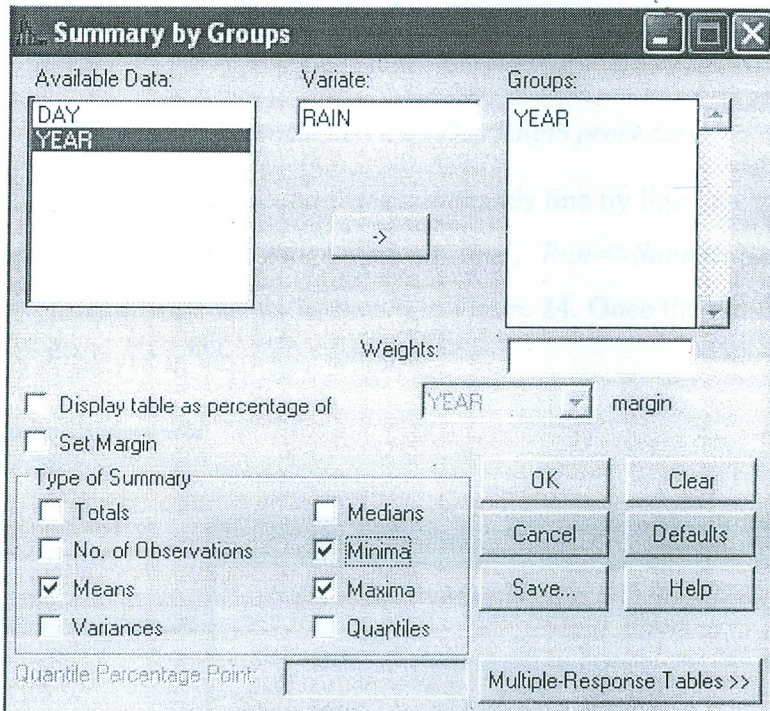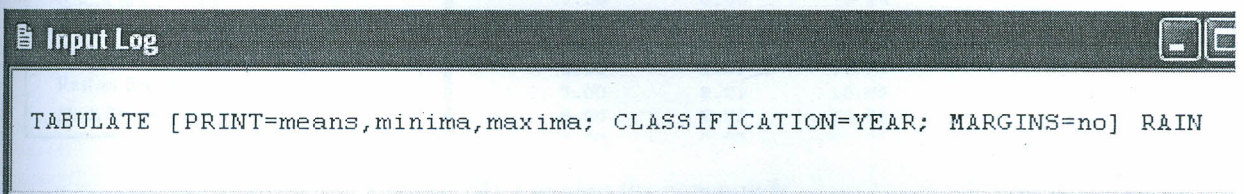

Figure 21: *Dialogue for Summary by Groups*


```
TABULATE [PRINT=means,minima,maxima; CLASSIFICATION=YEAR; MARGINS=no] RAIN
```

Figure 22: *Command corresponding to Summary by Groups Dialogue.*

The GenStat dialogues and menu make use of command without necessarily having to type them. It is important to learn beyond GenStat menu and make use of its commands since there are some facilities that are not available in the menu yet are in command. For example a simple procedure that calculate day lengths at any latitude and any day of the year [6] As a command it is called 'daylength', but there is no equivalent menu option. This command can be used to calculate the day lengths through the command interface.

An example on how 'daylength' is used is shown in Figure 23

32

```
Input Window;1*                                                    _ □ X

PRINT      'Example of how to use procedure DAYLENGTH'
VARIATE    Dayno; !(1...31); DECIMALS=0
DAYLENGTH [LATITUDE=52.205] DAYNUMBER=Dayno; DAYLENGTH=Daylength
CALCULATE Sunset = 12 + Daylength / 2
CALCULATE Sunrise = 12 - Daylength / 2
PRINT      'Sunrise and Sunset times for January at Wellesbourne,UK.'
PRINT      Dayno,Sunrise,Sunset; DECIMALS=2
```

Figure 23: *Example that uses Day length procedure*

The user has an option of submitting these commands line by line or a group of lines or the whole window by using the '*Run=>Submit line*', *Run=>Submit Selection*' and *Run=>Submit window*' menus respectively as shown in Figure 24. Once the window has been submitted, the result is displayed in the output window.

```
Run  Data  Spread  Graphics  Stats  Optio

  Submit Line            Ctrl+L
  Submit to Current Line   Ctrl+Shift+K
  Submit from Current Line Ctrl+Shift+M      Output
  Submit Selection         Ctrl+M
  Submit Window            Ctrl+W            Sunrise and Sunset times for January at Wellesbourne, UK
  Recycle Window           Ctrl+Shift+W
  Submit Clipboard         Ctrl+K            34  PRINT      Dayno,Sunrise,Sunset; DECIMALS=2

  Submit File...           Ctrl+B                Dayno     Sunrise     Sunset
  Submit R Script...                             1.00       8.09       15.91
  Submit BUGS Script...                          2.00       8.07       15.93
                                                 3.00       8.06       15.94
  Restart Session                                4.00       8.05       15.95
  Restart Server                                 5.00       8.04       15.96
  Flush Server                                   6.00       8.02       15.98
                                                 7.00       8.01       15.99
                                                 8.00       7.99       16.01
                                                 ^ ^^       ^ ^^       ^^ ^^
```

Figure 24: *Running the commands*

4.3: Compacting programs

Compacting programs in GenStat involves storing statements (commands) made by users so that they can refer to them later or use them repeatedly. There are two ways of compacting programs in GenStat namely: Macros and Procedures [15]

4.3.1: Procedures

When statements are compacted using Procedure, they must be self-contained, apart from those explicitly defined as options or parameters of the procedure. That is, all the data

33

structures that they use are accessible only within the procedure. Once you write programs for complicated tasks, GenStat allows you to keep them for future use. You may also wish to use procedures written by other people by simply giving the name of the procedure, and specifying options and parameters as required.

When GenStat meets a statement with a name that it does not recognize as one of the standard GenStat directives, it first checks to see whether you have a procedure of that name already stored in your program. Then it checks it up in any procedure library that you may have attached explicitly to your program, taking these in order of their channel numbers. After locating the required procedure, GenStat reads it in, if necessary, and then executes it. Information is transferred to and from a procedure only by means of its options and parameters.

Below is a new procedure (*Counts*) which counts the number of dry and wet days on a given Markov Chain order. It also calculates the mean amount of rainfall of rainy days.

```
PROCEDURE [restore=case]'Counts'
 OPTION NAME= \
  'STATES' ,      "(I: Scalar giving the number of states of the chain)"\
  'LIMITS' ,      "(I: Scalar or vector giving the values to determine the states)"\
  'ORDER'  ,      "(I: Order of the Markov Chain)"\
  'CLASS'  ,      "(I: Factor over which the data are to be totalled)"\
  'HIGH'   ,      "(I: YES/NO indicating if high order, default is NO)"\
  'LABELS' ,      "(I: YES/NO on whether to label the tables of results, default is YES)"\
  'SPREAD' ,      "(I: YES/NO on whether to display the results in spreadsheets)"\
  'INITIAL';      "(I: Scalar or variate - length =ORDER with initial values)"\
  MODE=4(P),3(T),P; SET=3(yes),no,3(yes),no; DECLARED=8(yes);\
  PRESENT=8(yes) ;\
  DEFAULT=2,!(0.85),1,*,'NO','YES','NO',*;\
  TYPE=!T('scalar'),!T('scalar','variate'),!T('scalar'),!T('factor'),\
  3(!T('text')),!T('scalar','variate');\
  VALUE=*,*,*,*,2(!T('YES','NO')),*,*
 PARAMETER NAME= \
  'DATA', "(I: vector from which spell lengths are to be formed" \
  'COUNTS', "(O: tables containing the frequencies)" \
  'AMOUNTS', "(O: table containing the sum of the amounts)"\
  'CATEGORIES', "(O: variate containing the category of each observation)"\
  'SCOUNTS';     "(O: pointer to tables for frequencies)"\
  MODE=P; SET=yes,4(no); DECLARED=yes,4(no); \
  TYPE=!T('variate'),2(!T('table')),!T('variate'),!T('pointer'); \
  PRESENT=yes,4(no)
```

Figure 25: *Procedure Counts- sec 1*

The procedure *'counts'* in Figure 25 can now be executed through the command interface by referring to the procedure name. The command that makes use of the procedure name is as shown in Figure 26 below.

```
COUNTS [order=2;states=2;class=DayOfYear; spread='YES';\
 high='yes'; labels='yes';limits=!(0.2,0.5);\
INITIAL=!(0,0)]Rain;cat=c;counts=counts;a=amounts
```

Figure 26: *Procedure Counts- sec 2*

In the command above, the options are defined in the brackets [] these are; *order, states, class, spread, High, labels, limits*. While the parameters are defined outside the brackets, these are *Rain, categories counts, amounts*. The rules for using the procedure are identical. Using a procedure does not require the knowledge of how the program inside operates, what data structures it contains, nor what directives it uses so long as you specify the correct options and parameters [15].

4.3.2: Macros

A macro is created when GenStat program is placed into a text. Once a macro is formed, you can use the macro to execute the program that is built within it by typing a pair of hash characters ## followed by the macro name. An example of a macro is shown in the Figure 27

```
text[VALUES='COUNTS [order=2;states=3;class=DayOfYear; spread=YES; high=NO; \
labels=yes;limits=!(0.2,0.5);INITIAL=!(0,0)]Rain;cat=c;counts=counts;a=amounts']trial
## trial
```

Figure 27: *Macro*

The program shown in Figure 26 can be executed within a macro shown in the Figure 27. First the command is defined within the single quotes ('') of the text, then the macro is executed by running command line ##trial. The two method of compacting programs are made use of to built more complex commands.

4.4: Creating Menus and Dialogs

GenStat allow its users to create their own menus for existing procedures. Any menu created is added in the main menu called *user* .Once the procedure is completed, they are saved with the extension *.gpi*, in a GenStat folder called *AddIn* found in *C:\Program files\Gen12ed\AddIn*. The codes are then run using *Run=>Submit* window menu. The procedure is then added in the library using *Tools=>Procedure Library=>Build.* [6]

35
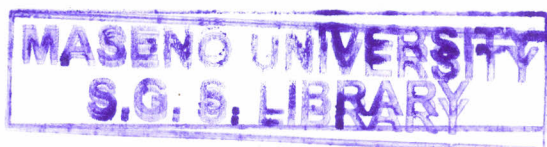
Commands for creating the menu are then created in a new input window. The layout and the content structure are created using the GenStat resource language basic structure is described below.

```
DIALOG   id, width, height
CAPTION   text
PROCEDURE   text
BEGIN

...
... dialogue control (layout and content instruction
---
END
```

DIALOG statement defines the *id* of the dialogue and its size of the dialogue box. CAPTION statement specifies the title that will appear on the dialogue title. The PROCEDURE statement specifies the procedure that will be created in the dialogue. The BEGIN-END is a block of statement that specifies the body structure of the Dialogue box. The file is then saved with extension *.grc*. [6]

The *.grc* file is then attached into GenStat. The layout of the commands used for attaching the *.grc* file is described below.

```
[AddInX]
Type = DIALOG
ProgID = my dialogue id
MenuTitle = name in menu
ResourceFile = dialogue code.grc
```

The underlined text needs to be changed by the user. The command is also created in a new input window then saved with extension *.gad*.

A sample of the dialogue box created in this study is shown in Figure 28. All the commands and procedures are included in the appendix III,IV,V and VI

36

```
DIALOG dialog1, 240, 240
CAPTION "Counts and Totals from Daily Data"
PROCEDURE "COUNTS"
BEGIN
        PUSHBUTTON "&Run",IDRUN,20, 220, 45, 15,BS_RUN
        PUSHBUTTON "&Reset",IDCLEAR,95,220,45,15,BS_CLEAR
        PUSHBUTTON "&Cancel",IDCANCEL,160,220,45,15,BS_CANCEL
        "PUSHBUTTON "&Default",IDDEFAULT,210,180,30,15,BS_DEFAULTS"
        LTEXT "&Available Data:",ITEXT,5,5,55,10
        LISTBOX IDLIST,5,20,90,120,GW_AVAILABLE
        LTEXT "&Data:",ISTATIC,115,5,60,10
        EDITTEXT DATA,115,15,80,14,GW_AVAILABLE | DT_TYPVAR | GW_PARAM | EC_SET
        LTEXT "&Class for totals:",ISTATIC,115,35,,55,10
        EDITTEXT CLASS,115,45,80,14,GW_AVAILABLE | DT_TYPFAC | GW_OPT | EC_SET
        LTEXT "&Type of Markov:",ISTATIC,115,65,,60,10
        COMBOBOX HIGH,175,65,60,40,CB_DROPLIST ! GW_OPT
        SLIST HIGH,"Normal","High"
        RLIST HIGH,"NO","YES"
        LTEXT "&Number of States:",ISTATIC,115,90,,60,10
        EDITTEXT STATES,215,90,20,11, GW_OPT
        LTEXT "&Threshold Values:",ISTATIC,115,105,,60,10
        EDITTEXT LIMITS,200,105,35,11, GW_OPT
        LTEXT "&Initial Value:",ISTATIC,115,120,,60,10
        EDITTEXT INITIAL,200,120,35,11, GW_OPT
        LTEXT "&Markov Chain Order:",ISTATIC,115,135,,80,10
        EDITTEXT ORDER,215,135,20,11, GW_OPT
        GROUPBOX "Save result in", IDSTATIC, 40, 150, 170, 40
        TOGGLEBOX "Amount", a, 55, 165, 40, 14,GW_PARAM,"amount", "NO"
        TOGGLEBOX "Counts", counts, 110, 165, 40, 14,GW_PARAM,"counts", "NO"
        TOGGLEBOX "Class", cat, 155, 165, 40, 14,GW_PARAM,"c", "NO"
        TOGGLEBOX "Display in Spreadsheet", SPREAD, 100, 200, 90, 14,GW_OPT,"YES", "NO"

END
```

Figure 28: *A .grc file for creating Counts and Amount Daily data menu*

```
AddIn1]
Type=DIALOG
ProgID=dialog1
MenuTitle=Counts and Totals from Daily Data
ResourceFile=counts.grc
```

Figure 29: *A .gad file for attaching dialogue command*

37

CHAPTER 5: MARKOV MODELING OF CLIMATIC DATA IN GENSTAT

5.1: Introduction

The potentials of GenStat are exploited to provide routines that can analyze rainfall data using Markov model and make it accessible to users. The First stage involves preparing the data for fitting; the second stage involves fitting the Markov model for the rainfall. Four procedures are written which includes; COUNT, PREPARE, FITTING, AMOUNTFITTING. These procedures are all shown in the appendix III, IV, V and VI

The COUNT procedure was written to read the raw data fed by the user and summarizes it by counting the number of rainy days for each day of the number of the years the data is given. It also calculates the amount of rain in the rainy days. The result of the summaries is then tabulated ready for calculation of the probabilities and fitting. The PREPARE procedure designed to calculates the probability of rain, the FITTING procedure, fits the probability of rain while AMOUNTFITTING fits the amount of rainfall. This chapter provides the documentation of what the procedures that are newly created do.

5.2: COUNT Procedure

COUNT procedure was first written by Roger Stern (unpublished) but in form of macros. This macro was redesigned to procedure for its use in making the menus. It identifies the state condition on each day then TABULATE directive is used to summarize the frequency of each day of the year possessing a specific state over the number of years the data is given.

5.2.1: Options for COUNT Procedure

States – This is a scalar value that defines the number of states, i.e. 2, 3, 4 etc. A state is determined by the amount of the rainfall below or above a specified threshold. The number of state equals the number of limits (threshold) values plus one.

Limits – This is a scalar or a variate that is used to determine the limit of a state, the default value of limit is 0.85. If a day received rainfall amount that is less than this value (0.85), then the day is referred to as a dry day, otherwise it will be a wet day.

Order- It is a scalar value that defines the number of order the Markov model will use to model the data. The values range from zero onwards

Class – It is a factor with 366 values over which the data is to be summed. It corresponds the period that you want that you will use to define the annual seasonality.

High – It specifies whether the Markov order is high or normal as explained in section 1.2.4: of this thesis.

Labels –Specifies whether or not the table of the *counts* will have labels or not (the default is 'yes')

Spread – It is a string specifying whether or not to display the result of counts and amount in a spreadsheet (the default is 'yes').

Initial – It is a scalar or variate that assigns initial value to days before the first data of the data. It is applicable to Markov of order one and above. The initial values are specified to avoid missing data for days before the first day.

Table – It is a string specifying whether or not to display the result of counts and amount in a table structure (the default is 'yes').

5.2.2:   Parameters for Count Procedure

Data- It is an input vector parameter of rainfall data. The rainfall data is given as one variate for all the years.

Counts- It is an output table parameter containing the sum of the frequencies

Amounts – It is an output table containing the sum of the amounts

Categories- An output variate parameter containing the category of each observation

Scounts -   An output pointer parameter to tables for frequencies

5.2.3:   Description of Count Procedure

The COUNTS procedure counts the number of times a day of the year having a specific state-condition for the number of years and also calculates the rainfall amounts for rainy days. Amount of rainfall is defined as the actual amount of rainfall recorded minus the

39

threshold value. That is, if the threshold is 0.85mm and in a specific day, it was recorded that the rainfall was 2mm, then the rainfall amount is 1.15mm.

The frequencies of the days with the same state conditions are organized in a table using TABULATE directive with columns representing the state condition e.g. *dry, wet, dd, dw, ddd, dww*, etc while the rows indicate factor levels for example day number of the year totaling to 366 days (see a sample in **Figure 32** (b)). The amounts of rainfall are also summed over the rainy days. This is further illustrated in the Figure 3o

below. Using the sample data set in Figure 30 (a), the resulted tables that are created when the COUNTS procedure is run with order 1 and threshold of 0.85 are shown in Figure 30 (b)

Spreadsheet ...

Row	Year	DayOf	Rain
1	1	1	2
2	1	2	4
3	1	3	7
4	1	4	0
5	1	5	0
6	2	1	1
7	2	2	6
8	2	3	5
9	2	4	4
10	2	5	0
11	3	1	0
12	3	2	0
13	3	3	0
14	3	4	0
15	3	5	5
16	4	1	0
17	4	2	6
18	4	3	0
19	4	4	0
20	4	5	2

Spreadsheet [L...

Row	Day	dd	dw	wd	ww
1	1	1	1	1	0
2	2	2	1	0	1	2
3	3	3	1	1	0	2
4	4	4	2	1	0	1
5	5	5	1	1	2	0

amounts ...

Row	Day	wd	ww
1	1	0.2	*
2	2	5.2	8.3
3	3	*	10.3
4	4	*	3.1
5	5	5.3	*
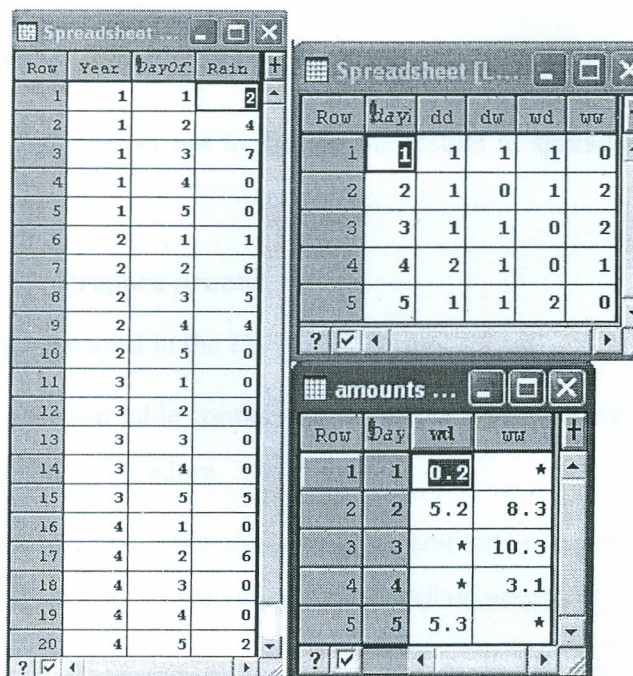
(a)   (b)

Figure 30: *A sample rainfall data and results of count procedures.*

Considering a state condition with wet given wet for day 2, i.e. there are only 2 days that satisfy this condition. The two days had rainfall recorded as 4mm and 6mm. the rainfall amount recorded in the amount table is 8.3 (4mm+6mm−2(0.85mm).

The rainfall amount of a given day of the year with specific state, is calculated using the equation (3-1) and (3-2).

The results of the count procedure are given in a table structure with the number of columns determined by the order of the model.

The table of counts and amounts may be displayed in a spreadsheet by declaring SPREAD=yes. The initial record for rain can be set using the option of INITIAL. The number of values in INITIAL variate must be equal to the ORDER value.

5.3: PREPARE Procedure

The PREPARE procedure converts the input structures (*counts* and *amounts* tables) to variates by making use of VTABLE Procedure [7], this enable manipulation of the result in each columns of the table.

5.3.1: Options for PREPARE procedure

Plot-Specifies whether or not to plot the results, the default is YES

Spread- Specifies whether or not to display the results in spreadsheets, the default is YES

5.3.2: Parameter for Prepare procedure

The parameters that are used in the PREPARE procedure are;

Counts - This is an input table containing the frequencies, it corresponds to the count table resulting from the COUNT procedure

Amounts – This is an input table containing the amounts of rainfall. It correspond to the amount table resulted from the COUNT procedure described above.

5.3.3: Description of Prepare Procedure

The PREPARE procedure reads data from the results created in COUNTS procedure; these are *counts* and *amounts*. The input structures are converted into variate   then the probabilities are calculated by applying the fomula below.

$$P_{ijk\,...} = \frac{f_{ijk\,...}}{\sum_{i=1}^{C} f_{ijk\,...}} \quad i,j,k = 1,2,3,...,C$$

(5-1)

41

Where

C= maximum number of states

$P_{ijk...}$ = The probability of the present state is i given that the state in the previous day is j and day before yesterday is k and so on up to the number of order. If order zero is used then we refer to P_i, for order one (P_{ij}), for order two (P_{ijk}) and so on.

$f_{ijk...}$ - The number of days that posses the condition that the present state is i given that the state in the previous day is j and day before yesterday is k and so on up to the number of order

$\sum_{i=1}^{C} f_{ijk...}$ - The total number of days with the previous day having state j and the day before the having state k regardless of the state in the present day.

The probabilities may be saved, displayed in the spreadsheet when SPREAD=yes or plotted by the DGRAPH directive when PLOT=yes. The high order Markov can be obtained when HIGH=yes which is only applicable when ORDER $\geq$ 2. The order value used in the COUNTS procedure must be maintained in PREPARE procedure

5.4: FITTING command procedures.

Fitting Procedure converts the counts and amounts table into variate using Vtable procedure [7]. The variates are then used to fit the probabilities of rain. The result may be displayed in a graph or saved.

5.4.1: Options for FITTING procedure

Harmonics- It is a scalar to represent the number of harmonic required for the model fitting, default is 3

Spread- Specifies whether or not to display the results in spreadsheets, the default is NO

5.4.2: Parameter for FITTING procedure

The parameters that are used in the FITTING procedure are;

Counts -Tables containing the frequencies, it correspond to the count table resulted from the count procedure

42

Amounts – This is an input table containing the amounts of rainfall. It correspond to the amount table resulted from the count procedure described above.

5.4.3: Description of Fitting Procedure

The FITTING parameter is set to read data structure of type table. (Counts and amounts tables from COUNTS procedure). The input structures are converted into variate then fitted. The variates forms the Y values while the *dayNo* factor is assigned to the X-values when plotting the fitted parameters.

5.5: AMOUNTFITTING command procedures

AMOUNTFITTING Procedure converts the counts and amounts table into variate using VTABLE procedure [7]. The variates are then used to fit the amounts of rain. The result may be displayed in a graph or not.

5.5.1: Options for AMOUNTFITTING procedure

Harmonics- It is a scalar to represent the number of harmonic required for the model fitting, default is 3

Spread- Specifies whether or not to display the results in spreadsheets, the default is NO

5.5.2: Parameter for FittingAmount procedure

The parameters that are used in the AMOUNTFITTING procedure are;

Counts -Tables containing the frequencies, it correspond to the count table resulted from the count procedure

Amounts – This is an input table containing the amounts of rainfall. It correspond to the amount table resulted from the count procedure described above.

5.5.3: Description of FittingAmount Procedure

The AMOUNTFITTING parameters are set to read data structure of type table. (counts and amounts tables from COUNTS procedure). The input structures are converted into variate then fitted. The variates forms the Y values while the *dayNo* factor is assigned to the X-values when plotting the fitted amounts.

43

CHAPTER 6: RESULTS AND DISCUSSION

6.1: Application of the model to Katumani Data

This chapter discusses how the newly developed procedures are used in all the aspects of the options specified. The daily rainfall data collected from 1961- 2001 at Katumani Experiments research Station in Kenya is used to test the procedure. The data is imported from Instat library into GenStat through excel. Alternatively it can be exported to GenStat from Instat window.

6.2: Description of the data

The data is for rainfall recorded in mm for 41 years, the columns represents the years while the row represent the day of the year running from 1 to 366. Part of the screen shot showing the data set in GenStat is in the Appendix I. There are 10 leap years and 31 non-leap years. The maximum of maximum annual rainfall was received in 1963 (186.9 mm) while the minimum of maximum annual rainfall was received in 1987 (26.7 mm) (see Appendix II).

6.3: Organizing the data

The counts procedure reads only two variates; Rain and a factor column over which the count is to be done i.e. day of the year, months of the year, weeks of the year etc. In this text the factor DayOfYear is used. The data is stalked such that the variate Rain is in one column for all the days and DayOfYear in one column for all the years. The *Spread=> Manipulate=> Stalk* menu is used to achieve this. The 'Stalk Column in Sheet' dialogue box is filled as shown in Figure 31. The number of columns in this case will be the number of years (see Appendix I). The stalked data is shown in Figure 32.

The factor column is changed to a variate by just right clicking the year column then choosing *Convert to variate* option. This result will now show the level values of the column. The *Spread => Calculate => Fill* menu is used to fill the year's values as shown in Figure 33. This is important for the calculation of the *DayOfYear*. The stalked data with year variate is shown in Figure 34.

44
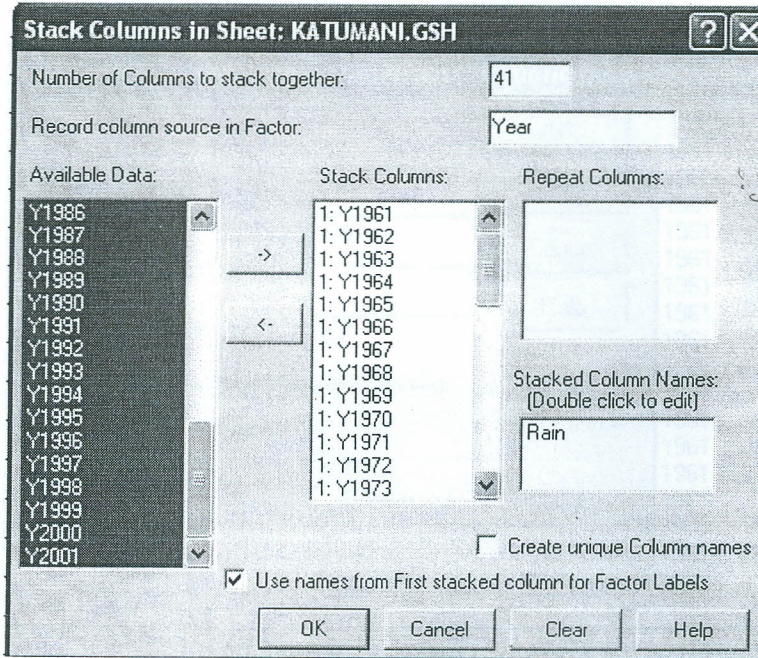
Figure 31: *Stalk column dialogue box*

Figure 32: *Stalked data*

45
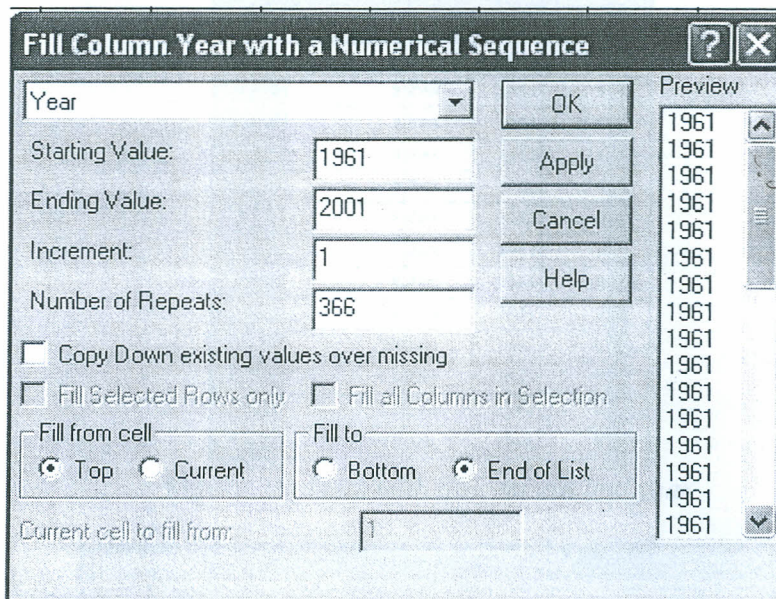
Figure 33: *Fill column Year with Numerical Sequence dialogue box*

Figure 34: *Year in variate*

Create two columns, day of the months and month of the year as shown in the Figure 35. The three columns (day, month and year) are used to calculate the date unless data is entered with date format. Using *Spread=>Calculate=> column,* (Figure 36) the date column is calculated (See Figure 37).

46

Figure 35: *Data with day, month and year*

Figure 36: *Calculate dialogue box for date*

Figure 37: *Date column calculated but with a fault message*

47

The date column is calculated though a fault message is shown, which states that 29th Feb of a non-leap year is invalid, hence no date is created on the 29th Feb. To avoid this fault, the invalid day is eliminated before calculating the dates. To eliminate the invalid data (29th Feb of a non-leap year), you can just delete the rows which are invalid.

Column for *DayOfYear* is calculated using MFRACTION (Date; 1; 1). This always gives 1st March as day 61 in the year instead of NDAYINYEAR (Date; 1) which gives 1st March as the 61st day in a leap year, but 60th day in other years.

The data set ready for the analysis is shown in Figure 38. The *dayofyear* with the 1st march (61st day) highlited.

Row	Month	Year	Date	DayOfYear	Rain
46	Feb	1961	15/02/61	46	0
47	Feb	1961	16/02/61	47	0
48	Feb	1961	17/02/61	48	0
49	Feb	1961	18/02/61	49	0
50	Feb	1961	19/02/61	50	0
51	Feb	1961	20/02/61	51	0
52	Feb	1961	21/02/61	52	0
53	Feb	1961	22/02/61	53	0
54	Feb	1961	23/02/61	54	0
55	Feb	1961	24/02/61	55	0
56	Feb	1961	25/02/61	56	0
57	Feb	1961	26/02/61	57	0
58	Feb	1961	27/02/61	58	0
59	Feb	1961	28/02/61	59	2
60	Mar	1961	01/03/61	61	0
61	Mar	1961	02/03/61	62	0
62	Mar	1961	03/03/61	63	0

Spreadsheet [new.gsh]*

Figure 38: *Data set with 'DayOfYear' and 'Rain' column ready for analysis*

6.4: The User Menu

The four climatic menus corresponding to the procedures are created under the *user* menu in GenStat. Through this menu the procedures can now be accessed and used (Figure

39). Choosing *User=> Count and Totals from daily Data* will open count dialogue box (Figure 42) which correspond to the COUNT Procedure
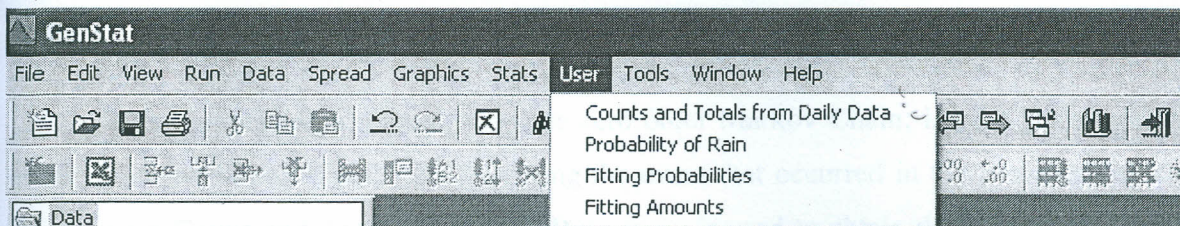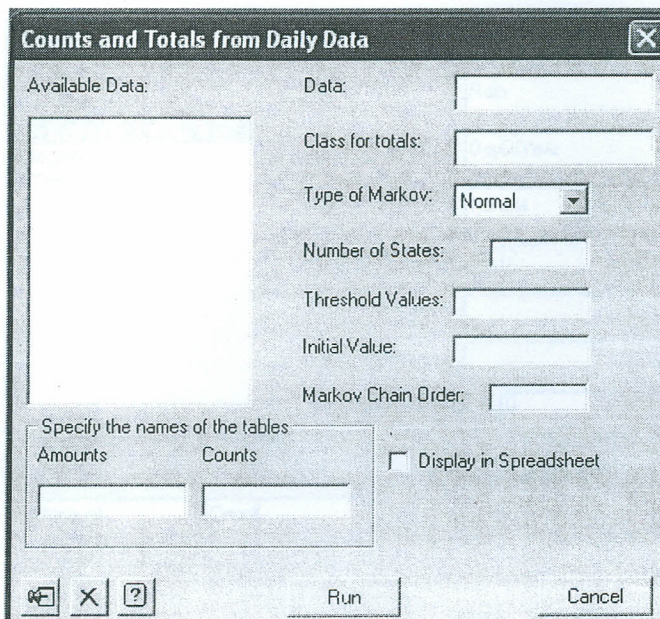

Figure 39: *User Menu*

Figure 40: *Count and Totals from daily Data dialogue box*

The field for 'data' receives the rainfall data and must be set. The 'class for totals' specifies the factor over which the data for counts will be totaled. This can either be the month of the year; day of year etc. 'Type of Markov' field specifies the type of Markov Chain to model, either normal or high. The fields 'number of states', 'threshold value', 'initial value' and 'Markov Chain order', when left blank will use the defaults values.

6.5: Two- State Markov Chain modeling

In a two- State Markov Chain, the states are dry and wet; the threshold value is only one marking the boundary of the two states. Using the default value of threshold option (0.85mm), any value of rainfall greater than 0.85mm is the state of rain while less than or

49

equal to 0.85 is the state of dry. This section discusses Markov Chain modeling of order zero and one.

6.5.1: Order zero

When modeling a normal two-state zero order Markov Chain, i.e. the chance of an event occurring in day τ without considering the event that occurred in the previous day(s), the *User => Count and Totals from daily Data* menu is used to obtain the *Count and Totals from daily Data* dialogue box and then filled as shown below.
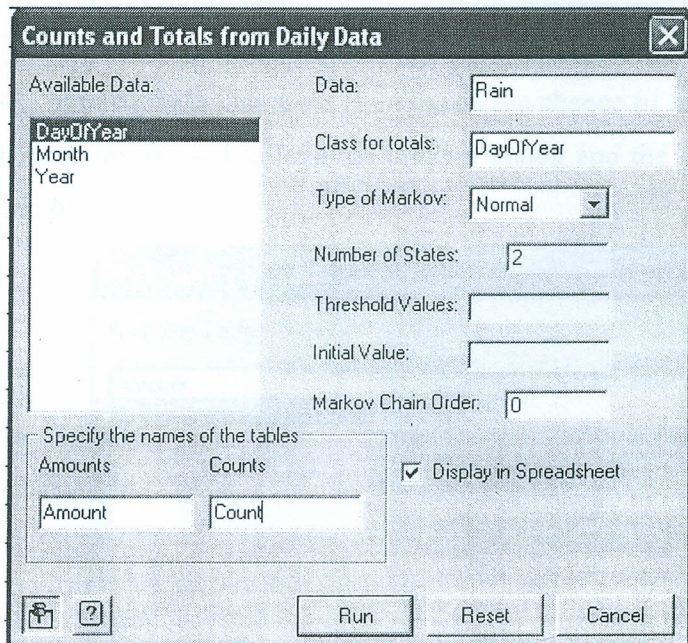

Figure 41: *Dialogue box for Counts and totals from Daily Data (Order Zero)*

The result gives the count and amounts based on order zero (Figure 42 and Figure 43).

Figure 42: *Counts of rainy days (order zero)*

50

Figure 43: *Rainfall amounts (order zero)*

A PREPARE procedure is then used to calculate the chance of rainfall for order zero. The *User => Prepare* menu is used to load the Prepare dialog and the option and parameter set as shown in figure below.

Figure 44: *Dialogue box for Probabilityof rain (Order Zero)*

The results of probability are displayed on a spreadsheet (Figure 45) and a plot of the probabilities (Graph 10) is displayed.

Figure 45: *Probabilities of rainy day form GenStat (order zero)*
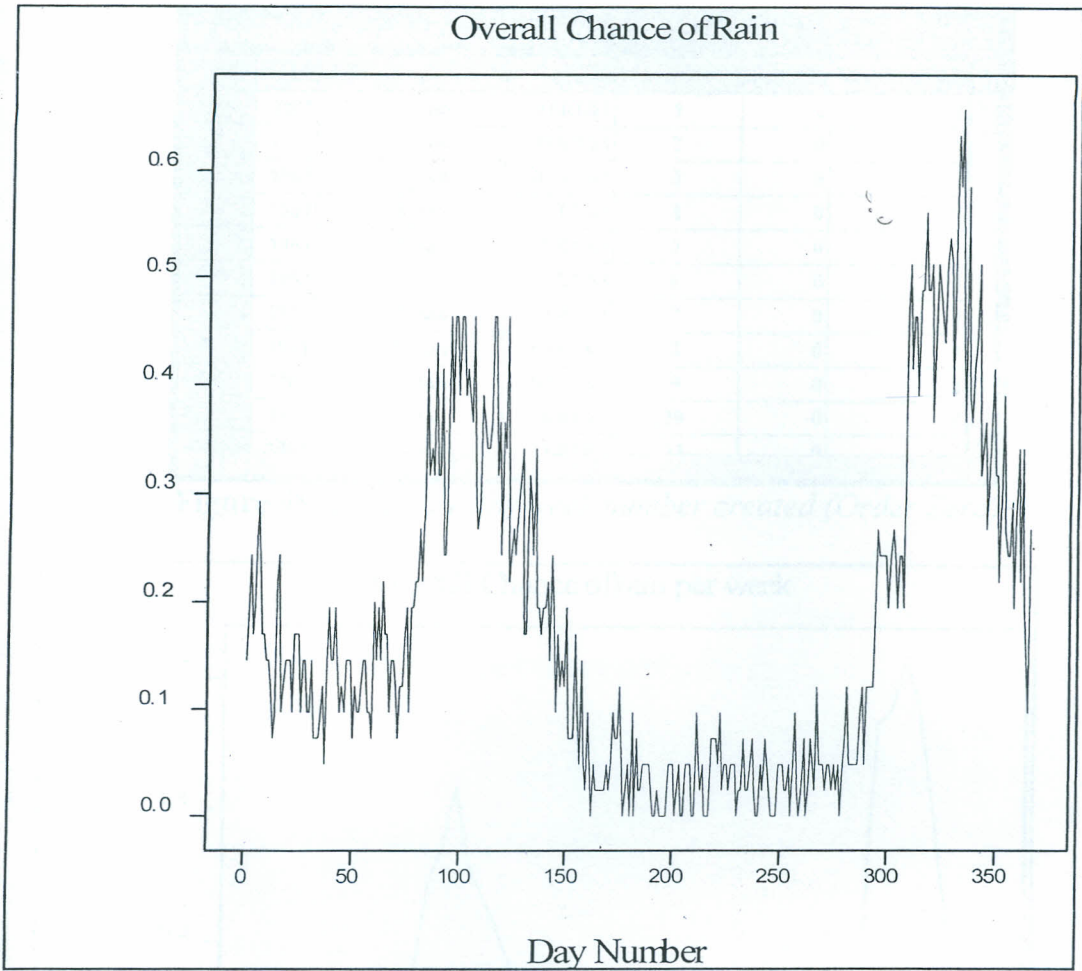
Graph 10: *Probabilities of order zero (GenStat)*

The graph indicates that Katumani has two seasons of rainfall, these are long rains (April-May) and short rains (November). The chance of rain during the short rain period is higher compared to that of long rain period. The same results were depicted when Instat software was used (refer to section 3.2.2:). The day number can be grouped into 5-days totals, week's totals or monthly. To obtain weekly totals, the user will have to create a factor column of week number and use it instead of the DayOfYear (Figure 46). The count totaled in 7 days (weekly) gives a clearer graph (see Graph 11).

52

Figure 46: *Dataset withy week number created (Order Zero)*
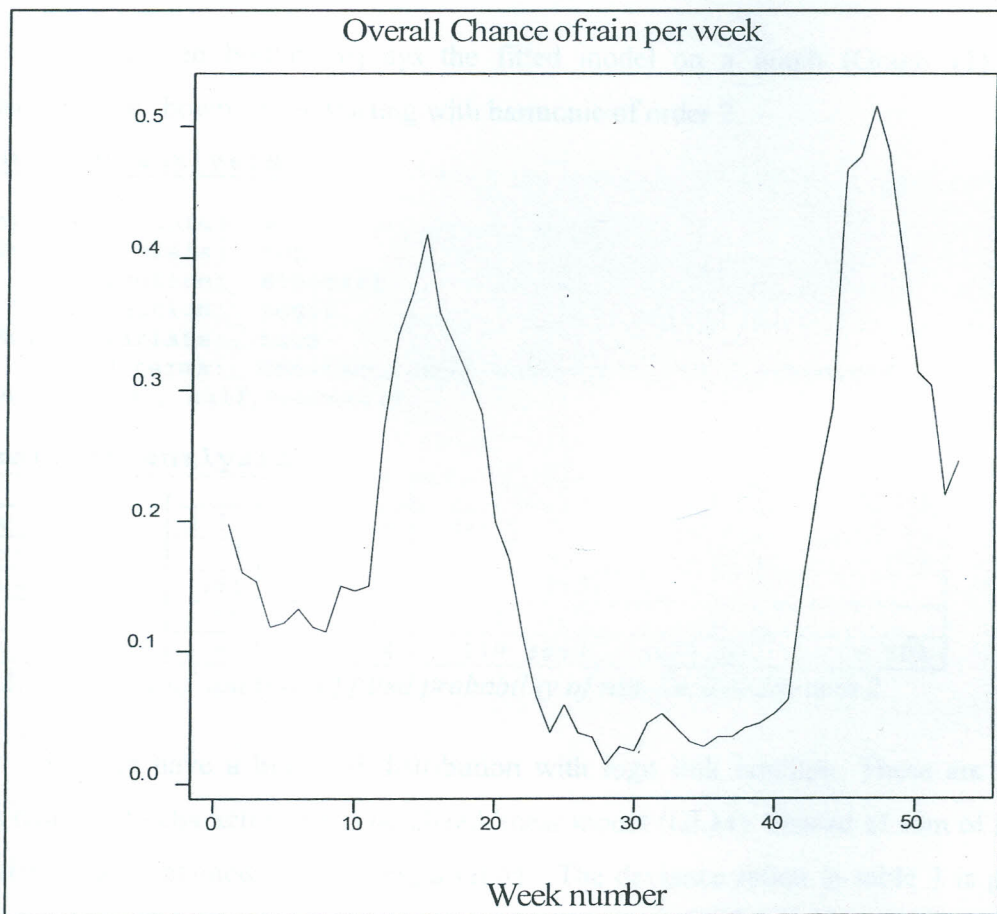
Graph 11: *Overall chance of weekly rainfall of order zero (GenStat)*

The fitted probability is obtained using *User=>Fitting probability of rain* menu which loads 'Fitting Probabilities of rain' dialogue box associated to the FITTING procedure. The options for the procedure are set as; Harmonics=3(Figure 46)
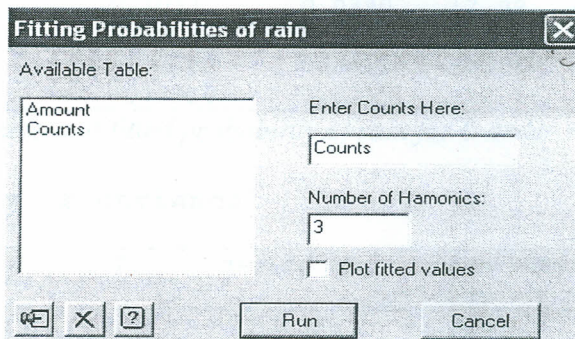

Figure 47: *Dialogue box for fitting Probability of rain (Order Zero)*

Clicking run button displays the fitted model on a graph (Graph 11) and the regression output shown below starting with harmonic of order 2.

Regression analysis

```
   Response variate:   w
   Binomial totals:    tot
      Distribution:    Binomial
     Link function:    Logit
   Weight variate:     tots
     Fitted terms:     Constant, half_hamonic[1], half_hamonic[2],
half_hamonic[3], half_hamonic[4]
```

Summary of analysis

Source	d.f.	deviance	mean deviance	deviance ratio	approx chi pr
Regression	4	1917.8	479.451	479.45	<.001
Residual	361	542.5	1.503		
Total	365	2460.3	6.741		
Change	-1	-119.4	119.364	119.36	<.001

Table 3: *Summary of analysis of fitted probability of rain for 0-order with 2*

The data have a binomial distribution with logit link function. These are the two aspects required to characterize a generalized linear model (GLM). Instead of sum of squares, the deviances are obtained when fitting a GLM. The deviance ration in table 3 is given as 479.45 meaning it is regression mean deviance divide by 1 not 1.503. this is so because the residual for binomial is 1

Estimates of parameters

Parameter		estimate	s.e.	t(*)	t pr.	antilog of estimate
Constant		-1.9904	0.0336	-59.15	<.001	0.1366
half_hamonic[1]	Sin[1]	0.7023	0.0390	17.99	<.001	2.018
half_hamonic[2]	cos[1]	1.2007	0.0468	25.65	<.001	3.323
half_hamonic[3]	Sin[2]	-1.2589	0.0411	-30.60	<.001	0.2840
half_hamonic[4]	cos[2]	-0.3816	0.0355	-10.76	<.001	0.6828

Table 4: *Estimate of parameters of fitted probability of rain for 0-order with 2*

Accumulated analysis of deviance

| Change		d.f.	deviance	mean deviance	deviance ratio	approx chi pr
+ half_hamonic[1]	Sin[1]	1	93.386	93.386	93.39	<.001
+ half_hamonic[2]	cos[1]	1	535.478	535.478	535.48	<.001
+ half_hamonic[3]	Sin[2]	1	1169.576	1169.576	1169.58	<.001
+ half_hamonic[4]	cos[2]	1	119.364	119.364	119.36	<.001
Residual		361	542.511	1.503		
Total		365	2460.315	6.741		

Table 5: *Accumulated analysis of deviance of fitted probability of rain for 0-order with 2 harmonics*

In table 5 above, shows that adding the harmonic of order 2 is significant. When harmonic of order 3 is added the results below is obtained.

Change		d.f.	deviance	mean deviance	deviance ratio	approx chi pr
+ half_hamonic[1]	Sin[1]	1	93.386	93.386	93.39	<.001
+ half_hamonic[2]	cos[1]	1	535.478	535.478	535.48	<.001
+ half_hamonic[3]	Sin[2]	1	1169.576	1169.576	1169.58	<.001
+ half_hamonic[4]	cos[2]	1	119.364	119.364	119.36	<.001
+ half_hamonic[5]	Sin[3]	1	131.654	131.654	131.65	<.001
+ half_hamonic[6]	cos[3]	1	1.028	1.028	1.03	0.311
Residual		359	409.829	1.142		
Total		365	2460.315	6.741		

Table 6: *Accumulated analysis of deviance of fitted probability of rain for 0-order with 3 harmonics*

Fitting the probability of order zero with harmonics of order three is not significant since cos[3] > 0.05. Therefore harmonic of order two is applied and the fitted plot is now drawn and shown in Graph 12.

55
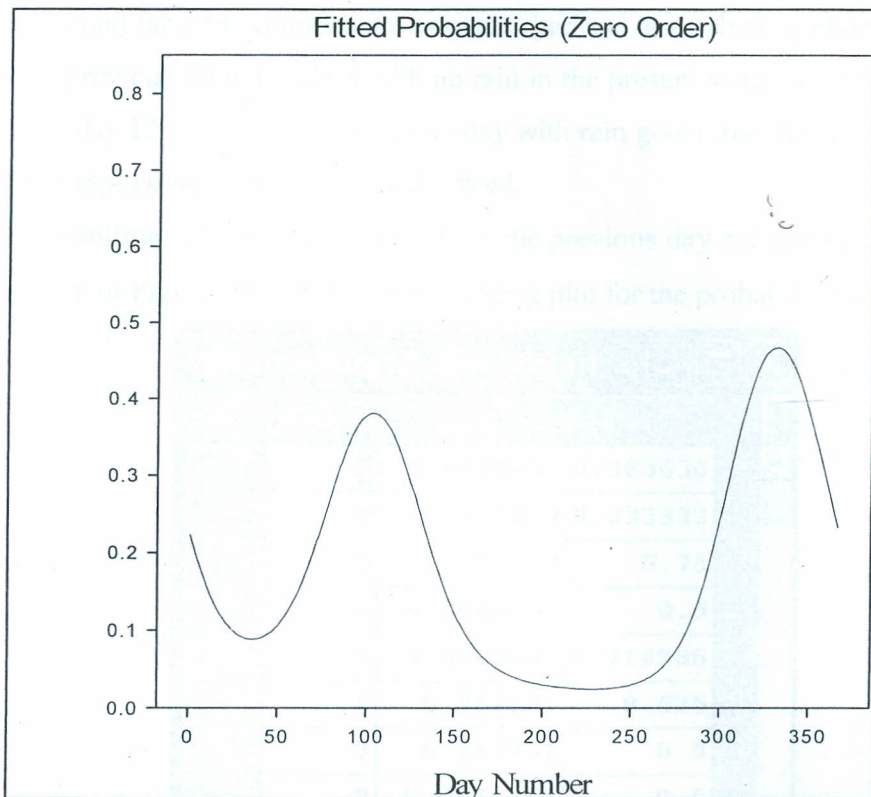
Graph 12: *Fitted Probabilities of order zero (GenStat)*

6.5.2: Order one

A two-state Markov Chain of order one is modeled for the chance of a state in a day given the state condition in the previous day. The dialogue box for counts is filled with the order option equal to one with the rest of the fields the same as in order zero discussed above. The results for counts and amounts for order 1 are displayed in a spreadsheet as shown in Figure 48. (Note: 0,1,2,3 used as column header of 'counts' table corresponds to the states *dd, dw,wd,ww* respectively)

Row	DayOfYear	0	1	2	3
1	1	27	7	2	4
2	2	29	4	6	2
3	3	29	2	4	6

Row	DayOfYear	wd	ww
1	1	4.8	17.4
2	2	88.1	28.6
3	3	20.3	80.7

Figure 48: *First order counts and raindall amounts*

The number of times it was dry on 2nd January given that on 1st January was also dry are 29 out of 41. The value 29 is below *dd* on day 2. On the same row under *wd*, the 6 means that the number of times it was wet on 2nd January given that 1st January was also dry are 6

out of 41. The amount table the columns are only two (*wd* and *ww)* which contain the amounts of rain given the previous state. On days with no rain in the present state their amounts are no calculated. i.e. on day 13th January, there was no day with rain given that the previous day was dry hence the corresponding amount is not calculated.

The probabilities of rain given the state in the previous day calculated by PREPARE procedure is shown in **Figure 49** and the corresponding plot for the probabilities (Graph 13)

| Row	dayNo	P_wd	P_ww
1	1	0.0689655	0.363636
2	2	0.171429	0.333333
3	3	0.121212	0.75
4	4	0.0645161	0.5
5	5	0.0882353	0.714286
6	6	0.212121	0.625
7	7	0.137931	0.5
8	8	0.0645161	0.5
9	9	0.147059	0.285714
10	10	0.0882353	0.428571
11	11	0.0571429	0.666667
12	12	0.114286	0.166667
13	13	0	0.6
14	14	0.0789474	0.333333

Figure 49: *Chance of rain given the state in the previous day*

The values in the table are probabilities of rain given the state in the previous day. For example, the probability of rain on the 1st January given that it was dry in the previous day is 0.06897
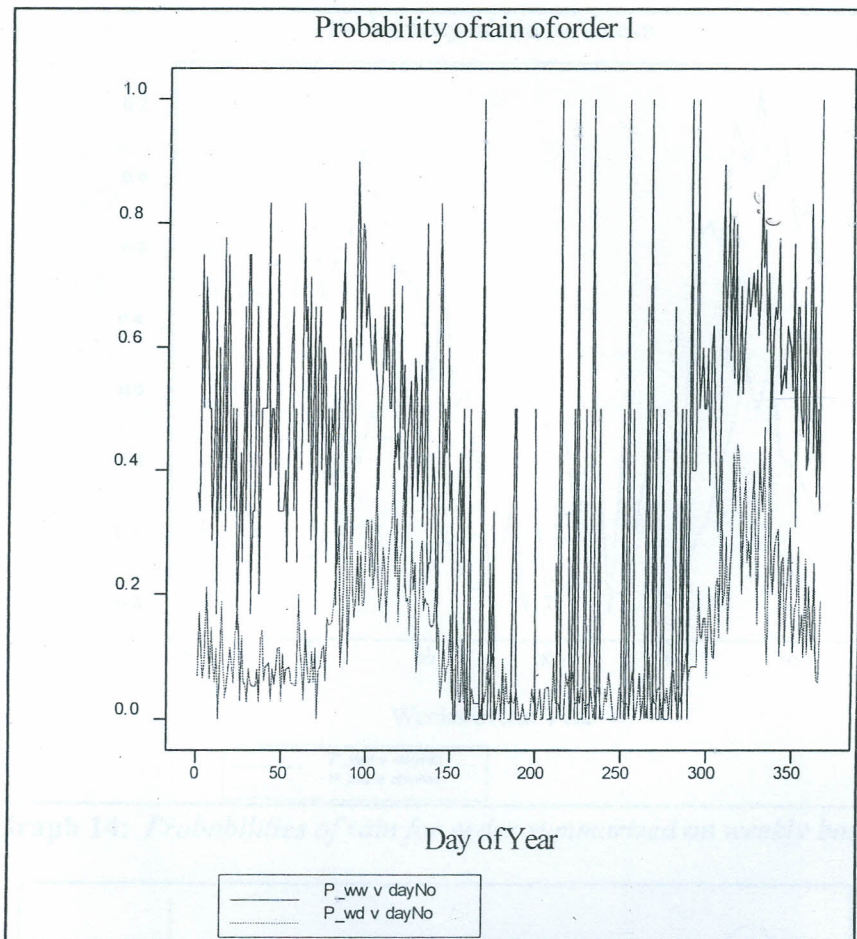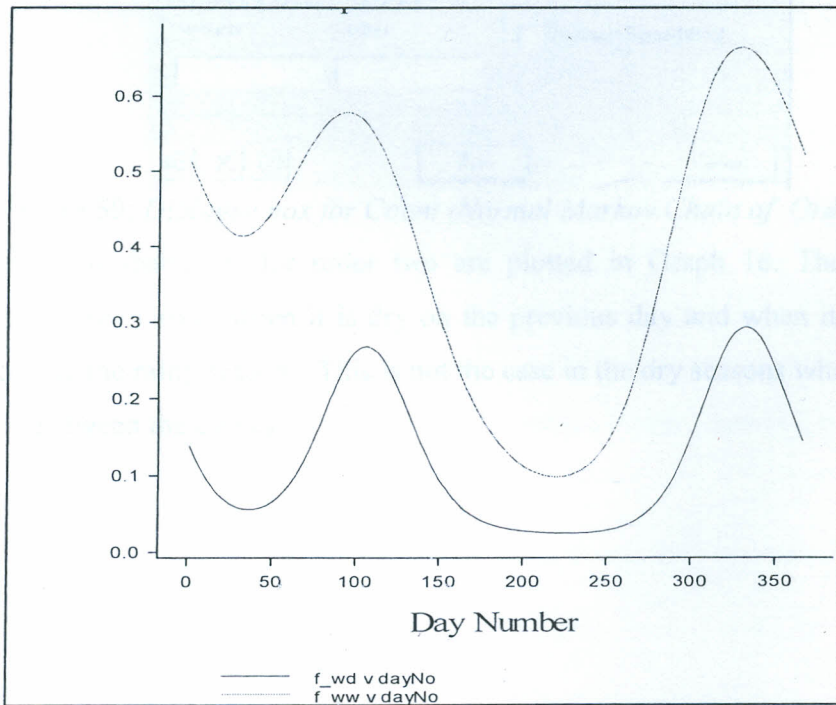
Graph 13: *Probabilities of rain for order one (GenStat)*

When the probabilities are plotted for each day, it is overcrowded and understanding the distinction of the curves can be hard. As before there are solutions for this to give better curves by first grouping the days into 5-day totals and then plot the results (Graph 14), or 10 day, weekly and monthly totals or get the fitted probabilities (Graph 15).
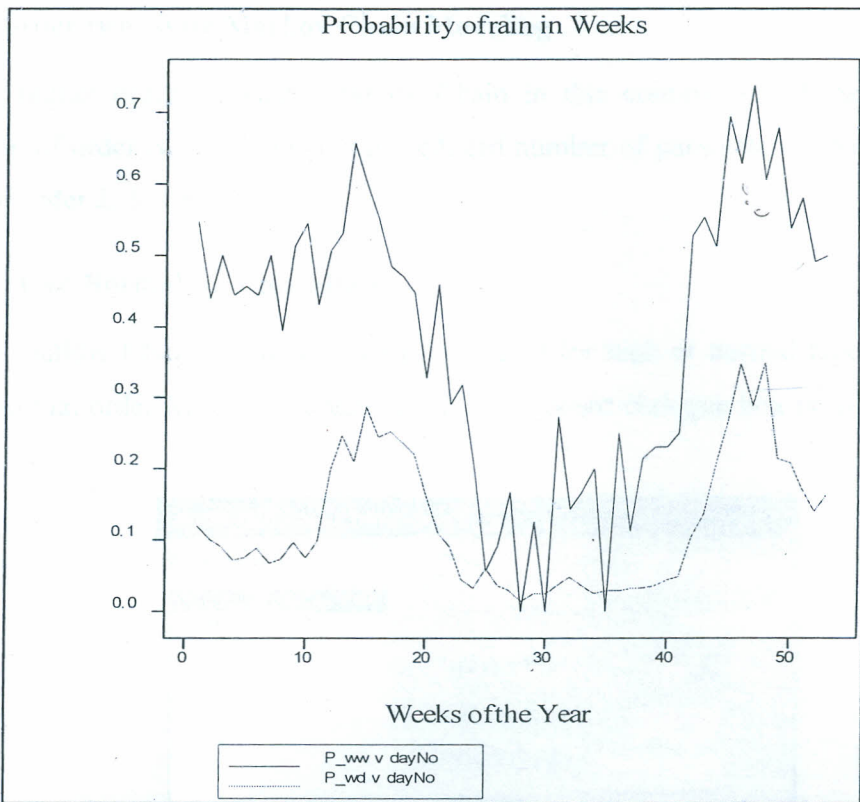
Graph 14: *Probabilities of rain for order summarized on weekly basis*

Graph 15: *Fitted Probabilities of order one (GenStat)*

## 6.6: Higher order two- state Markov Chain Modeling

The higher order two-state Markov Chain in this context would mean modeling Markov Chain of order two and above with reduced number of parameters. We will discuss the results for order 2, 3, 4 and 5.

6.6.1: Order Two Normal Markov Model

The Markov Chain of order 2 can be modeled for high or normal types of Markov Chain. The normal order Markov is obtained when the count dialogue box is filled as shown below.

Figure 50: *Dialogue box for Count (Normal Markov Chain of Order Two)*

The fitted probabilities for order two are plotted in Graph 16. There is a clear distinction between the curves when it is dry on the previous day and when it is wet in the previous day during the rainy seasons. This is not the case in the dry seasons where there is no clear distinction between the curves.

60
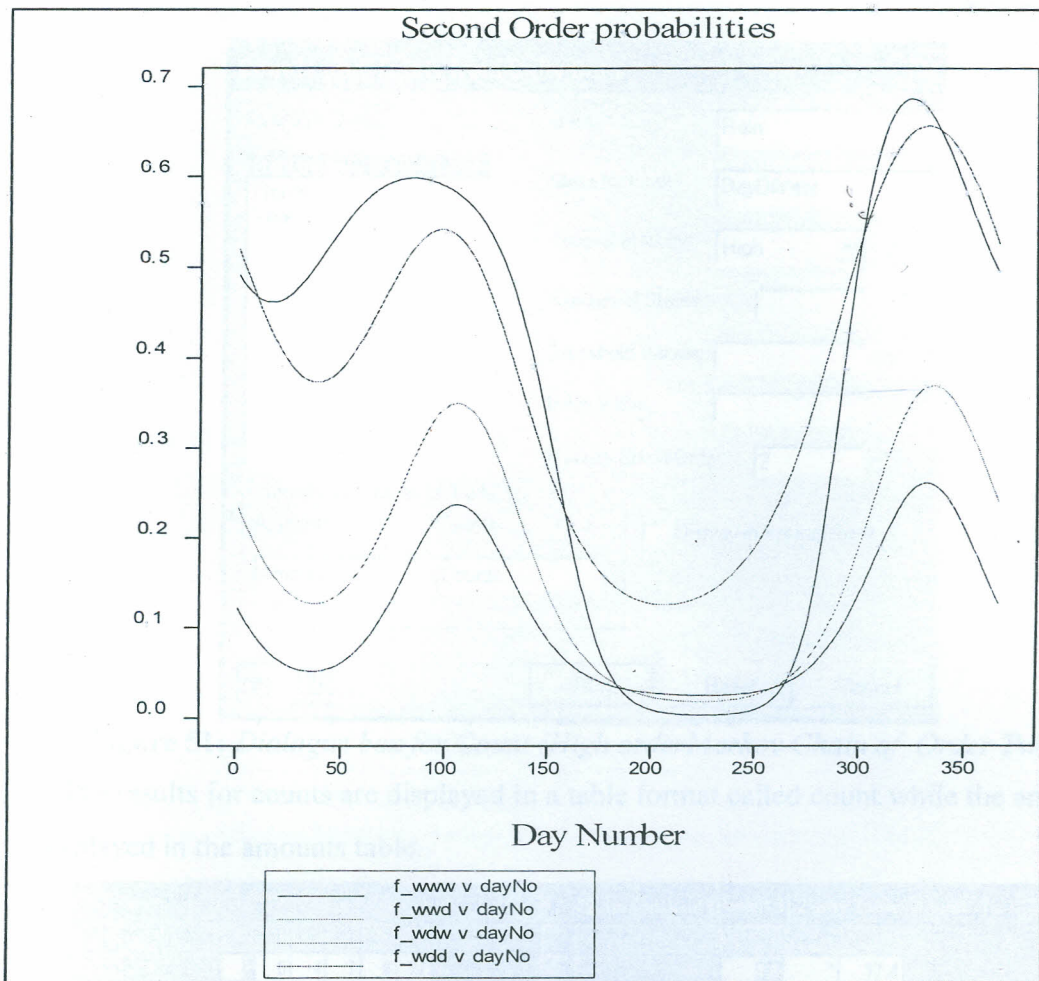
Graph 16: *Fitted Probabilities of rain for order two (Normal Markov)*

6.6.2: Order Two High Markov Model

Considering a high Markov Chain model for order 2, the *www* and *wwd* curves are combined to *ww*. The dialogue box for counts used to obtain a high Markov Chain of order 2 is filled as shown in Figure 51
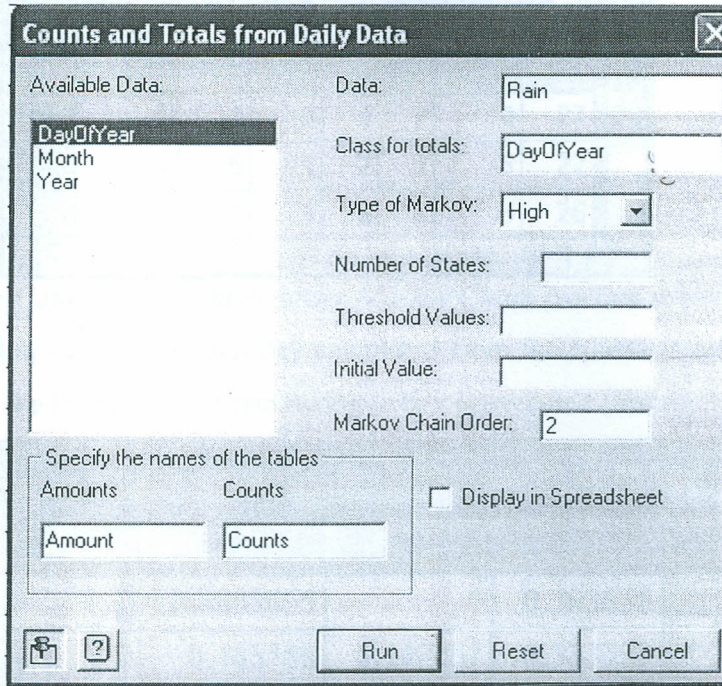
Figure 51: *Dialogue box for Count (High orderMarkov Chain of Order Two)*

The results for counts are displayed in a table format called count while the amounts table is displayed in the amounts table.

Row	DayOfYear	-3	-2	-1	1	2	3
1	1	2	0	4	7	0	27
2	2	5	1	2	4	6	22
3	3	3	1	6	2	3	26
4	4	2	0	5	5	2	27
5	5	2	1	5	2	4	27
6	6	7	0	5	3	2	24
7	7	2	2	6	6	1	24
8	8	1	1	5	5	5	24
9	9	2	3	2	5	2	27

Row	DayOfYear	>1dw	wdw	ww
1	1	4.8	*	17.4
2	2	56.9	31.1	28.6
3	3	19.9	0.4	80.7
4	4	11.0	*	67.8
5	5	13.8	11.5	61.8
6	6	45.8	*	55.1
7	7	26.6	9.9	37.2
8	8	51.9	3.4	59.6
9	9	9.5	76.6	54.3

Figure 52: *Counts and amounts for 2nd -high order Markov Chain(table structure)*

The corresponding variate for table is shown in the figure below. The variate table is obtained when you choose 'Spreadsheet' option in the *Probability of rain dialogue* box. The column headers for counts correspond to the column headers in Figure 53.

Figure 53: *Counts for 2nd -high order Markov Chain (variate structure)*

The probability of rain for each day calculated from the counts is also displayed in the spreadsheet when the Probability of rain dialogue box is executed (see fig 51).

Figure 54: *Probability of rain for 2nd -high order Markov Chain.*

The probability of rain in January 1st given that it was dry in the last two days is 0.06896, while the probability that it rained on January 1st given that it rained in the previous day is 0.3636. The last column in table 54 (P_ww) is used in place of having P_www and P_wwd, which are almost the same when fitted. The fitted probability model is shown in the Graph 17.
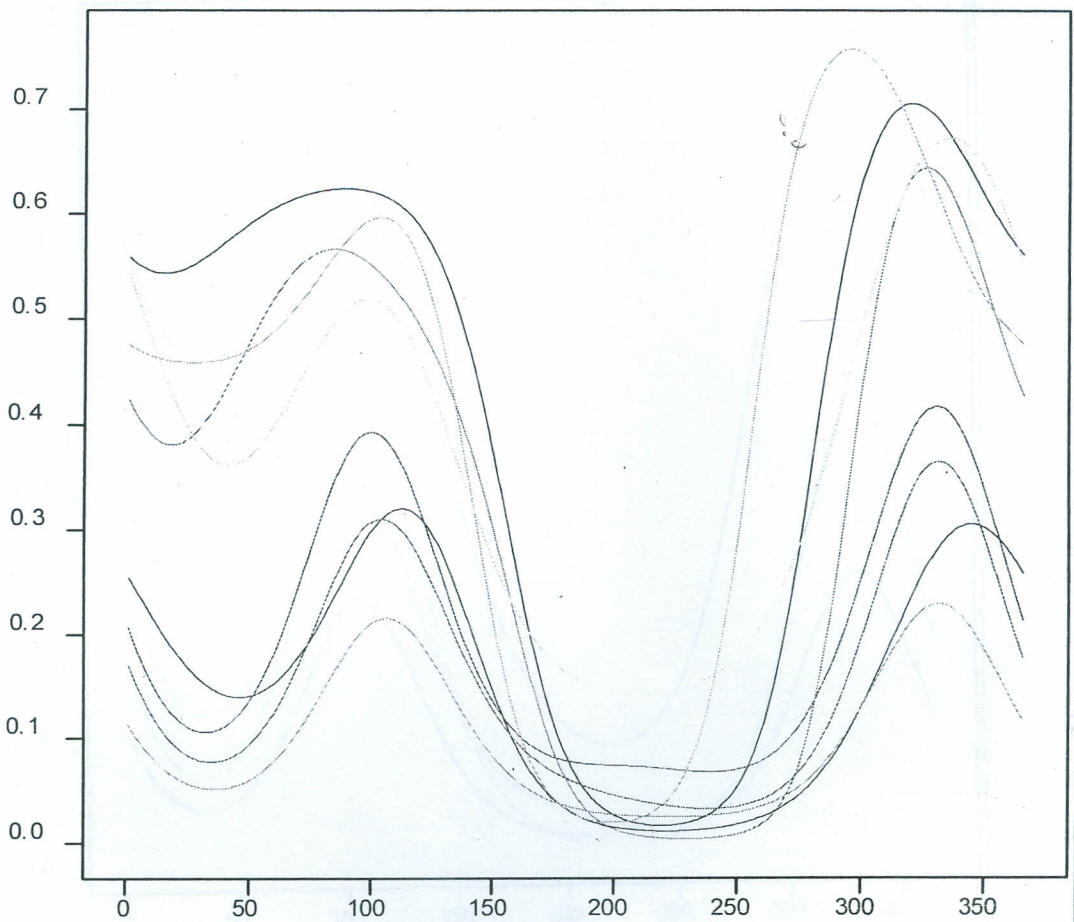
Fitted Probabilities of order 2 (HighOrder)

Day Number

Legend:
- F_w2d v dayNo
- F_dw v dayNo
- F_w v dayNo
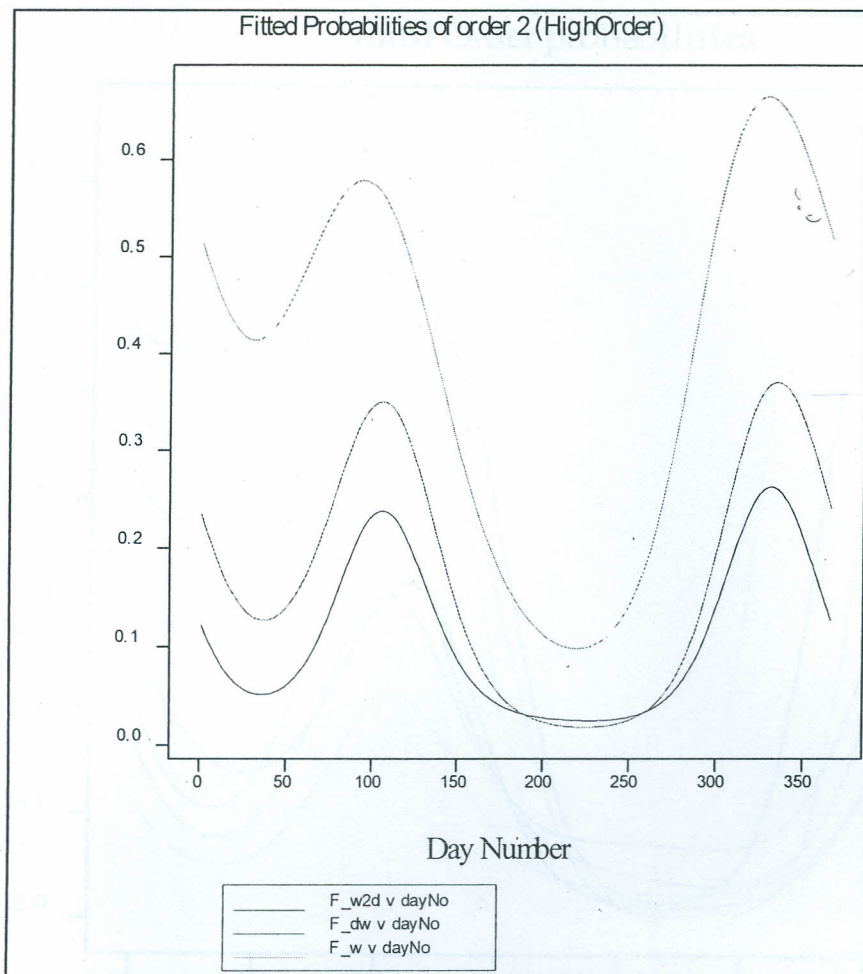
Graph 17: *Fitted Probabilities of order two-High Order (GenStat)*

The fitted probability of rain is high above the probability of dry given rain and that of the probability of wet given two drys in the previous.

6.6.3: Normal and High Markov models for order three and four

The fitted normal and high Markov Chain model for order three are shown in the Graph 18. The fitted model indicates that there is no distinction between the groups of curves when the previous day was dry.
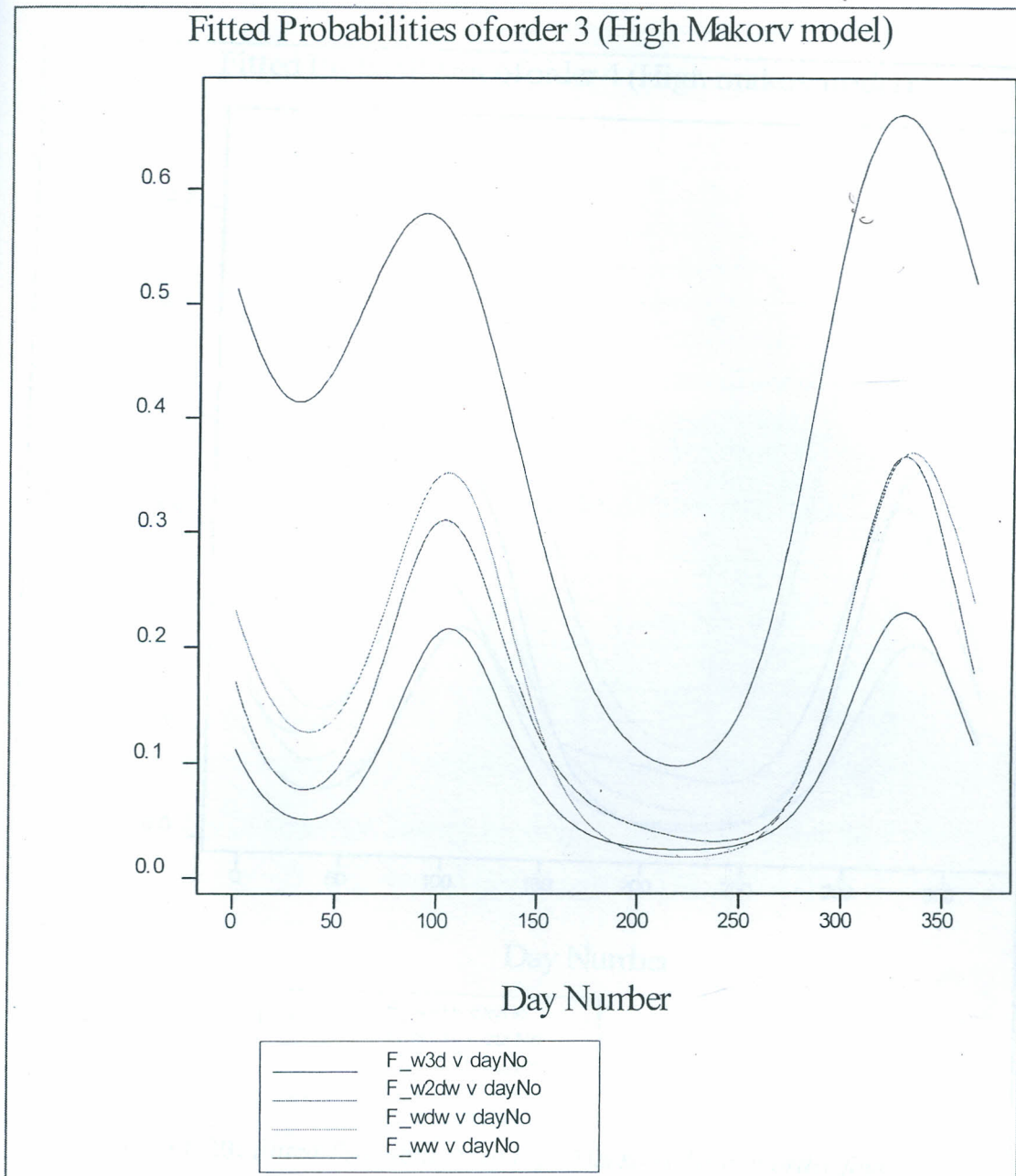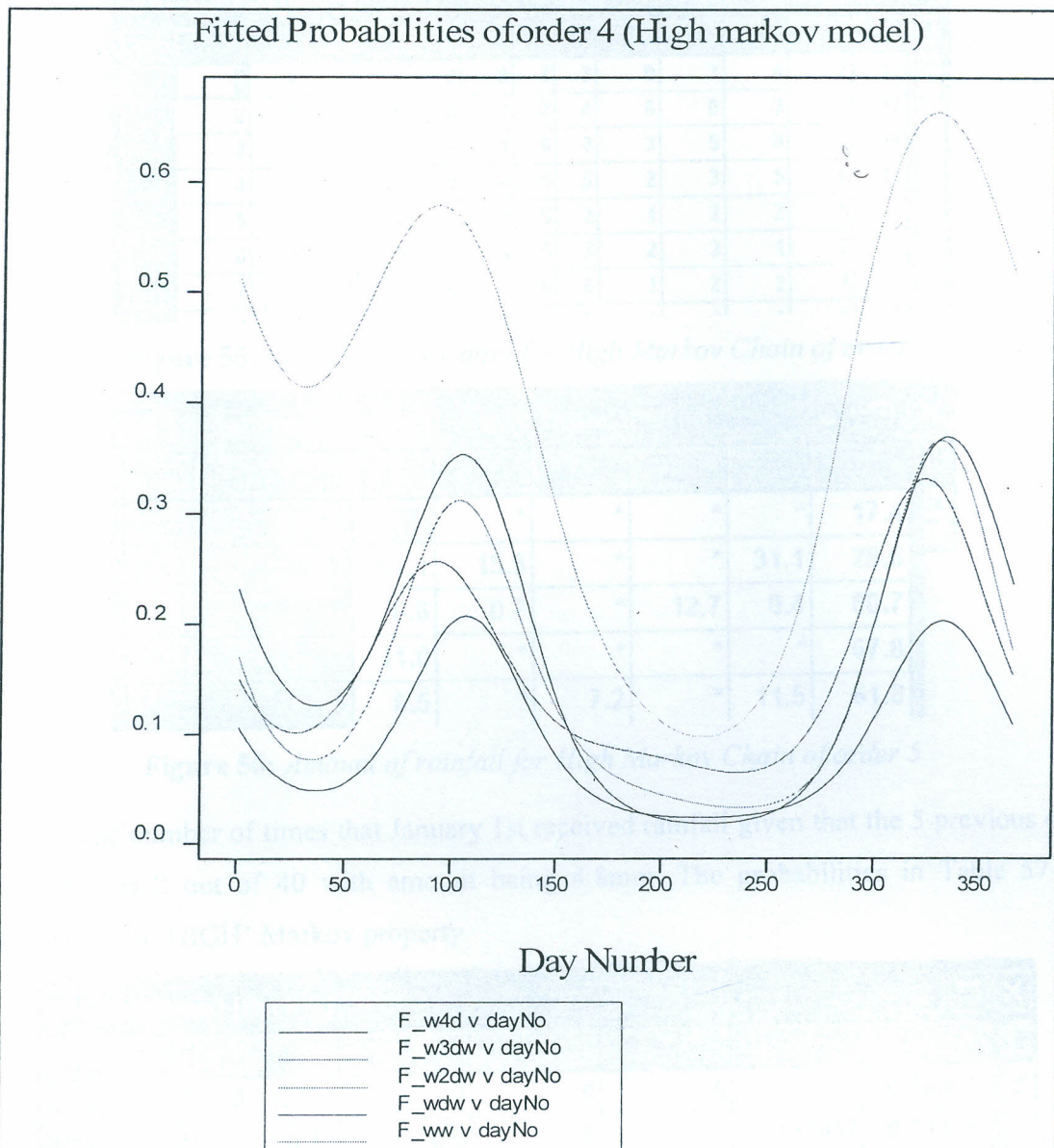
Third Order probabilities

Day number

——————	f_wwww v dayNo
——————	f_wdww v dayNo
——————	f_wwwd v dayNo
··············	f_wdwd v dayNo
——————	f_wwdw v dayNo
– – – – –	f_wddw v dayNo
——————	f_wwdd v dayNo
··············	f_wddd v dayNo

Graph 18: *Fitted Probabilities of normal Markov Chain of order three.*

Graph 19: *Fitted Probabilities of high Markov Chain of order three.*

All the curves in Graph 19 are distinct to each other during the rainy seasons. Using higher orders breaks the distinctions as shown for order 4 (Graph 20) and order 5 (Graph 21)

66

Graph 20: *Fitted Probabilities of high Markov Chain of order four.*

6.6.4: Order five

Order five for a normal Markov will result in so many parameters (64) which might not be significant in Markov modeling. A high order Markov will reduce this number to 10 parameters only. The result obtained for Katumani data is as shown below.

Row	!dayNo	w5d	w4d	w3d	w2d	wdw	ww	dw	_2dw	_3dw	_4dw	_5dw	_6d
1	1	2	0	0	0	0	4	7	0	3	4	7	13
2	2	3	2	0	0	1	2	4	6	0	3	2	17
3	3	1	1	0	1	1	6	2	3	5	0	2	18
4	4	2	0	0	0	0	5	5	2	3	5	0	18
5	5	1	0	1	0	1	5	2	4	2	2	5	17
6	6	5	0	1	1	0	5	3	2	3	1	2	18
7	7	1	0	1	0	2	6	6	1	2	2	1	19

Figure 55: *Counts of rainy days for High Markov Chain of order 5*

| Row	DayOfYear	>4dw	w4dw	w3dw	w2dw	wdw	ww
1	1	4.8	*	*	*	*	17.4
2	2	41.4	15.5	*	*	31.1	28.6
3	3	6.6	0.7	*	12.7	0.4	80.7
4	4	11.0	*	*	*	*	67.8
5	5	6.5	*	7.2	*	11.5	61.8

Figure 56: *Amount of rainfall for High Markov Chain of order 5*

The number of times that January 1st received rainfall given that the 5 previous days were dry was 2 out of 40 with amount being 4.8mm. The probabilities in Table 57 are calculated with 'HIGH' Markov property

Row	!dayNo	P_w5d	P_wd4w	P_wd3w	P_wd2w	P_wdw	P_ww
1	1	0.133333	0	0	0	*	0.363636
2	2	0.15	0.5	0	*	0.142857	0.333333
3	3	0.0526316	0.333333	*	0.166667	0.25	0.75

Figure 57: *Probability of rain for High Markov Chain of order 5*

The probability of rain given five previous days dry is 0.1333.

68
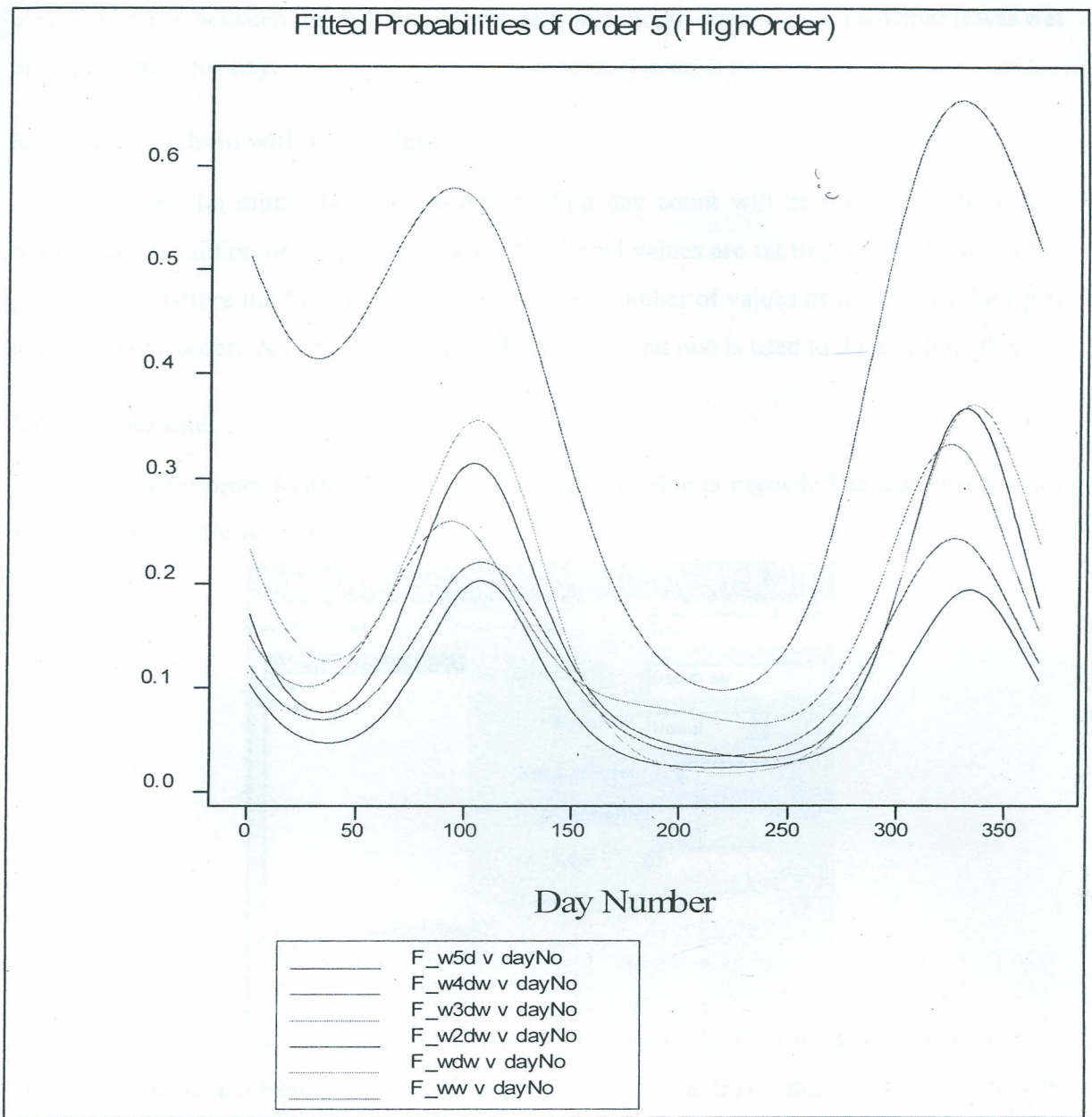
Graph 21: *Fitted Probabilities of high Markov Chain of order five.*

In graph 19, the chance of rain given rain is distinct and high from the other models while the probability of rain given longer dry spell is very low compared to other Chains. The fitted model for rain given that the previous day being dry, has a lot of interactions during dry seasons implying that there is less variations in the number of wet days given dry and less interaction during the wet seasons implying that there is high variation the number of wet days given dry in the previous day. In a nutshell, during the wet season, there seems to be a

clear difference between the numbers of wet days in a month depending on whether it was wet or dry the previous day.

6.7: Markov Chain with Initial data

When the initial data is missing, the first day count will be missing if the model requires the condition of the previous state. The initial values are set to give a memory of the previous day before the first day the data starts. The number of values of initial must be equal to the Markov order. A two-state Markov of order one and two is used to demonstrate this.

6.7.1: Order one

In a 1st order Markov Chain, only one initial value is needed. The dialogue box for count is filled as shown below.
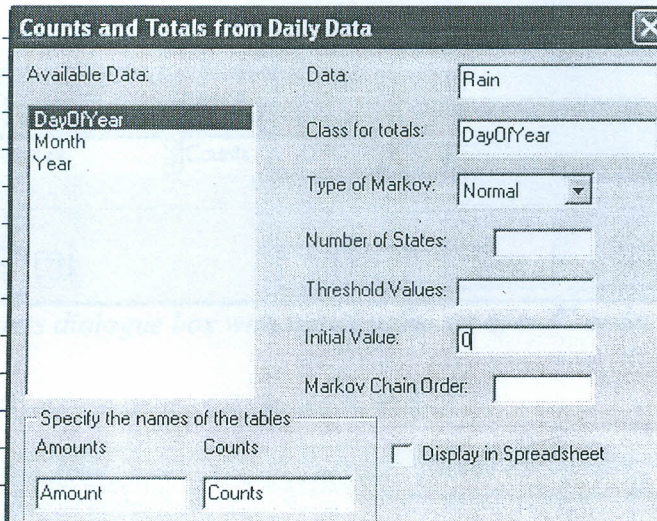

Figure 58: *Counts dialogue box with initial value set of a 1st order Markov Chain*

There is a difference in result on 1st January for a first order Markov Chain with 'initial value' set and 'initial value' not set. The total counts is more (41) in Figure 59 when initial value is set than the total counts (40) in Figure 59 when the initial is not set.

Figure 59: *Counts for Two-state 1st order Markov Chain with initial*

6.7.2: Order Two

When a Markov Chain is of order two it will require two initial values. When the initial values are more than one, they are set in a variate structure. i.e. instead of 0,0 we use !(0,0). The dialogue box for this is filled as shown in Figure 60 below.
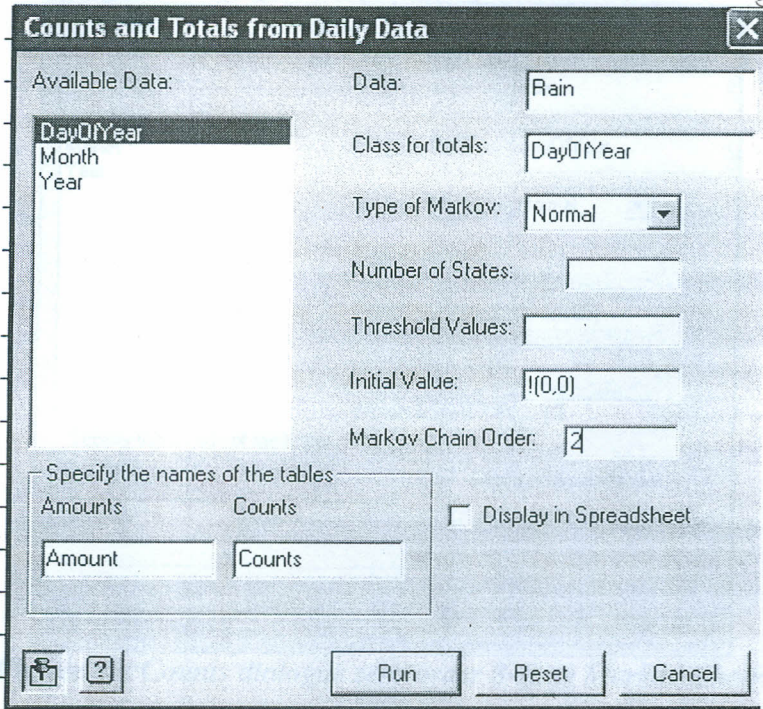

Figure 60: *Counts dialogue box with initial value set of a 2nd order Markov Chain*

The results for this dialogue are shown in Figure 61

Row	DayOfYear	0	1	2	3	4	5	6	7
1	1	28	0	6	1	2	0	1	3
2	2	23	6	1	3	5	1	1	1
3	3	26	3	1	1	3	1	5	1
4	4	27	2	2	3	2	0	2	3
5	5	27	4	1	1	2	1	1	4
6	6	24	2	1	2	7	0	2	3

Row	DayOfYear	wdd	wdw	wwd	www
1	1	2.4	*	7.7	3.2
2	2	11.4	31.1	3.2	25.3
3	3	6.7	0.4	14.2	9.4
4	4	5.5	*	14.9	12.6
5	5	6.9	11.5	16.5	11.3
6	6	6.5	*	11.7	10.6

Figure 61: *Counts and amount table of 2nd order Markov Chain with initial given*

6.8: Three-state Markov Chain Modeling

In certain occasions one may want to group the rain fall data into three states i.e. rain, trace and wet or even more number of states. The model has the capability of modeling

Markov Chains with more than two states. A three state Markov model will have two threshold values which determine the boundaries of each state. This section describes a zero order and a one order state Markov model. To obtain a zero order Markov model the dialogue box for counts can be filled as shown in the Figure 62 below.
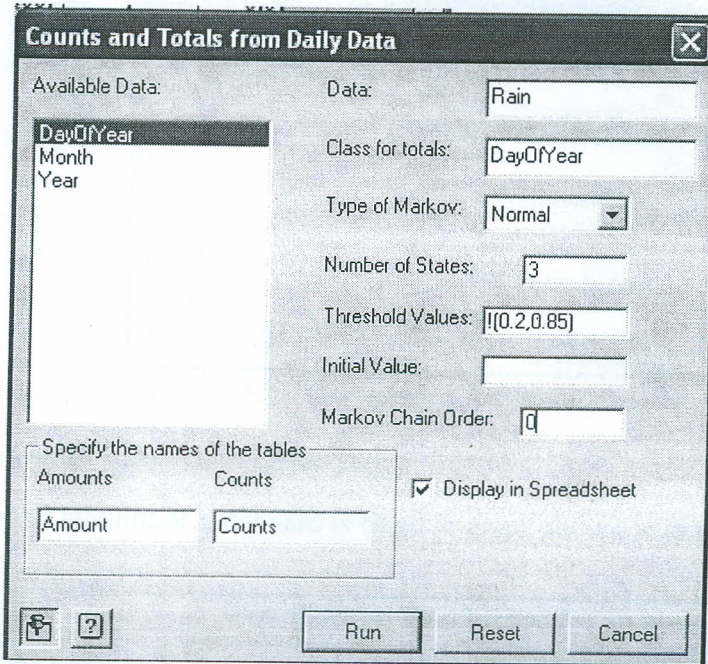
Figure 62: *Counts dialogue zero order 3-state Markov Chain*

6.9: Modeling Rainfall Amount

When fitting rainfall amount, The *User=>Fitting amount* is used, and the dialogue box for fitting amounts is as shown in the Figure 63 below.
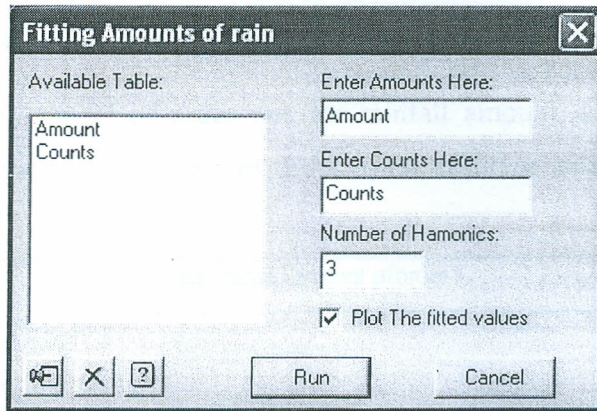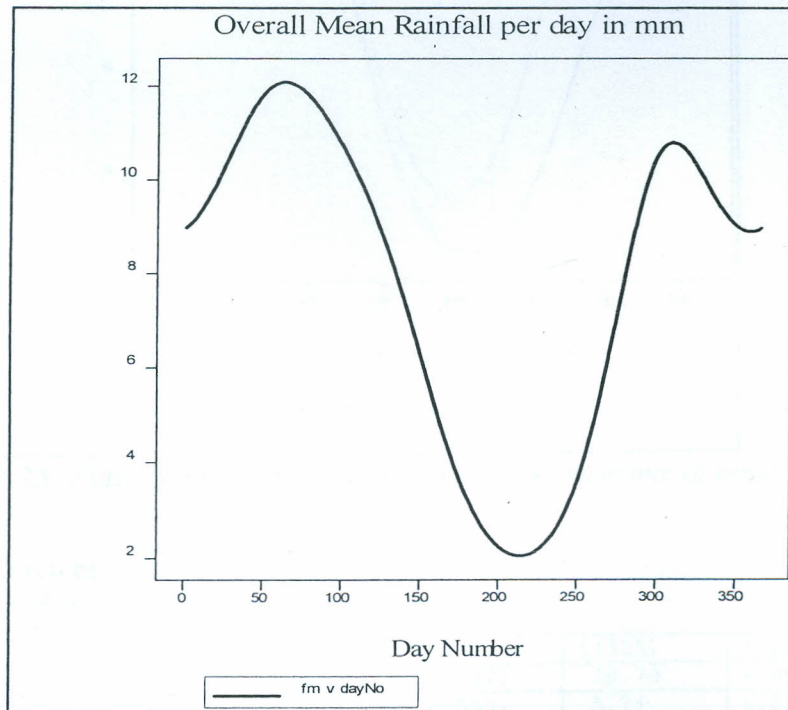

Figure 63: *Dialogue box for fitting Amount*

6.9.1: Modeling rainfall Amount of order zero

The model for rainfall amount of order zero is fitted in graph shown in graph 22

Graph 22: *Fitted Mean rainfall of order zero (GenStat)*

73

The result in Graph 22 indicates that the overall mean rainfall in the long season is higher than in the short season. The mean annual rainfall, result (Graph 26) is realistic based on the conditions that Katumani experienced during the period when the data was collected.

6.9.2: Modeling rainfall Amount of order one

When the Markov model of order one for rainfall amount is fitted starting with harmonic of order 2 followed by harmonic of order 3, the result is shown in the Graph 23 below.
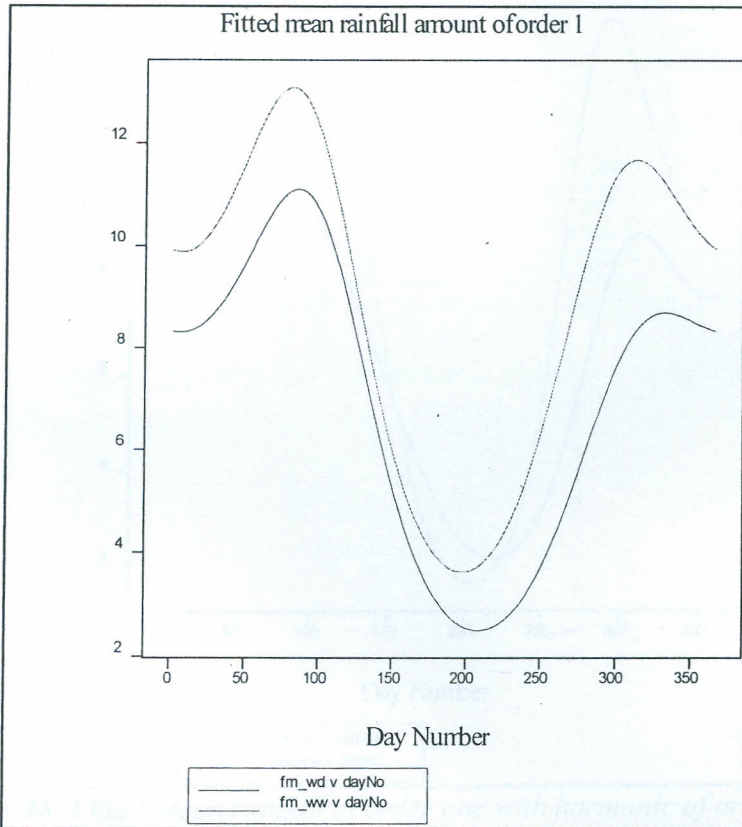
Graph 23: *Fitted Mean rainfall of order one with harmonic of order 2*

Estimates of parameters
Response variate: w_d

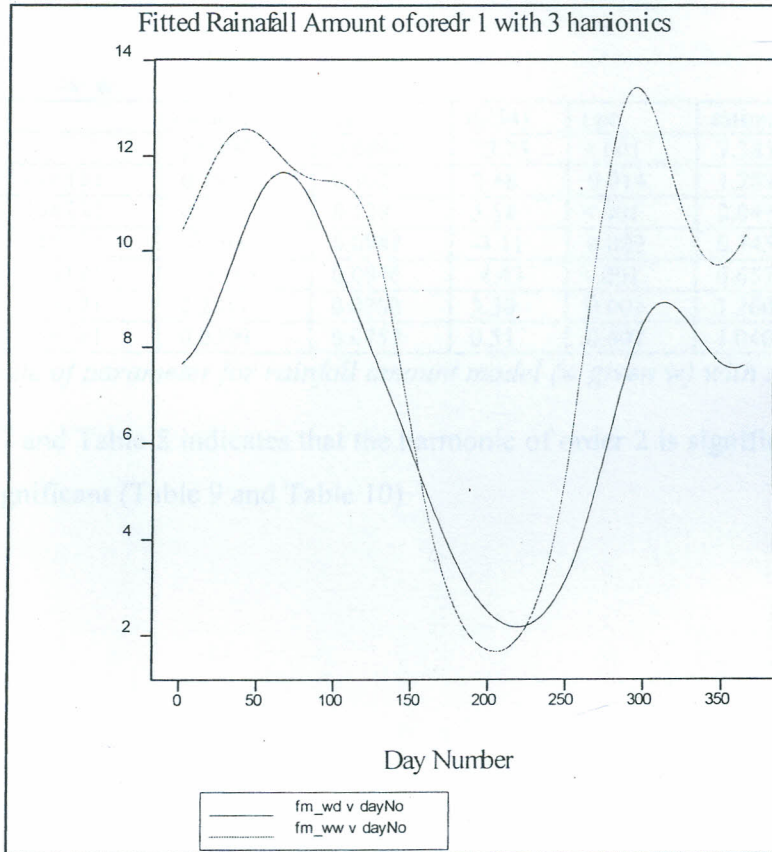| Parameter		estimate	s.e.	t(325)	t pr.
Constant		1.8291	0.0472	38.79	<.001
half_hamonic[1]	Sin[1]	0.3452	0.0601	5.74	<.001
half_hamonic[2]	Cos[1]	0.5191	0.0662	7.85	<.001
half_hamonic[3]	Sin[2]	-0.1965	0.0601	-3.27	0.001
half_hamonic[4]	Cos[2]	-0.2273	0.0608	-3.74	<.001

Table 7: *Estimate of parameter for rainfall amount model (w given d) with 2 harmonics*

74

Response variate: w_w

Parameter		estimate	s.e.	t(256)	t pr.
Constant		2.1022	0.0805	26.13	<.001
half_hamonic[1]	Sin[1]	0.1678	0.0818	2.05	0.041
half_hamonic[2]	Cos[1]	0.478	0.105	4.54	<.001
half_hamonic[3]	Sin[2]	-0.1301	0.0787	-1.65	0.099
half_hamonic[4]	Cos[2]	-0.2827	0.0674	-4.19	<.001

Table 8: *Estimate of parameter for rainfall amount model (w given w) with 2 harmonics*

Graph 24: *Fitted Mean rainfall of order one with harmonic of order 3*

75

Response variate: w_d

Parameter		estimate	s.e.	t(323)	t pr.	estimate
Constant		1.8267	0.0469	38.91	<.001	6.214
half_hamonic[1]	Sin[1]	0.3939	0.0639	6.17	<.001	1.483
half_hamonic[2]	Cos[1]	0.5287	0.0689	7.67	<.001	1.697
half_hamonic[3]	Sin[2]	-0.2301	0.0615	-3.74	<.001	0.7944
half_hamonic[4]	Cos[2]	-0.2121	0.0648	-3.27	0.001	0.8089
half_hamonic[5]	Sin[3]	0.0748	0.0597	1.25	0.211	1.078
half_hamonic[6]	Cos[3]	-0.1121	0.0609	-1.84	0.067	0.8940

Table 9: *Estimate of parameter for rainfall amount model (w given d) with 3 harmonics*

Response variate: w_w

Parameter		estimate	s.e.	t(254)	t pr.	estimate
Constant		1.9994	0.0842	23.75	<.001	7.385
half_hamonic[1]	Sin[1]	0.254	0.102	2.48	0.014	1.289
half_hamonic[2]	Cos[1]	0.717	0.129	5.58	<.001	2.049
half_hamonic[3]	Sin[2]	-0.2933	0.0942	-3.11	0.002	0.7458
half_hamonic[4]	Cos[2]	-0.4191	0.0946	-4.43	<.001	0.6576
half_hamonic[5]	Sin[3]	0.2311	0.0700	3.30	0.001	1.260
half_hamonic[6]	Cos[3]	0.0390	0.0757	0.51	0.607	1.040

Table 10: *Estimate of parameter for rainfall amount model (w given w) with 3 harmonics*

Table 7 and Table 8 indicates that the harmonic of order 2 is significant while that of order 3 is not significant (Table 9 and Table 10)
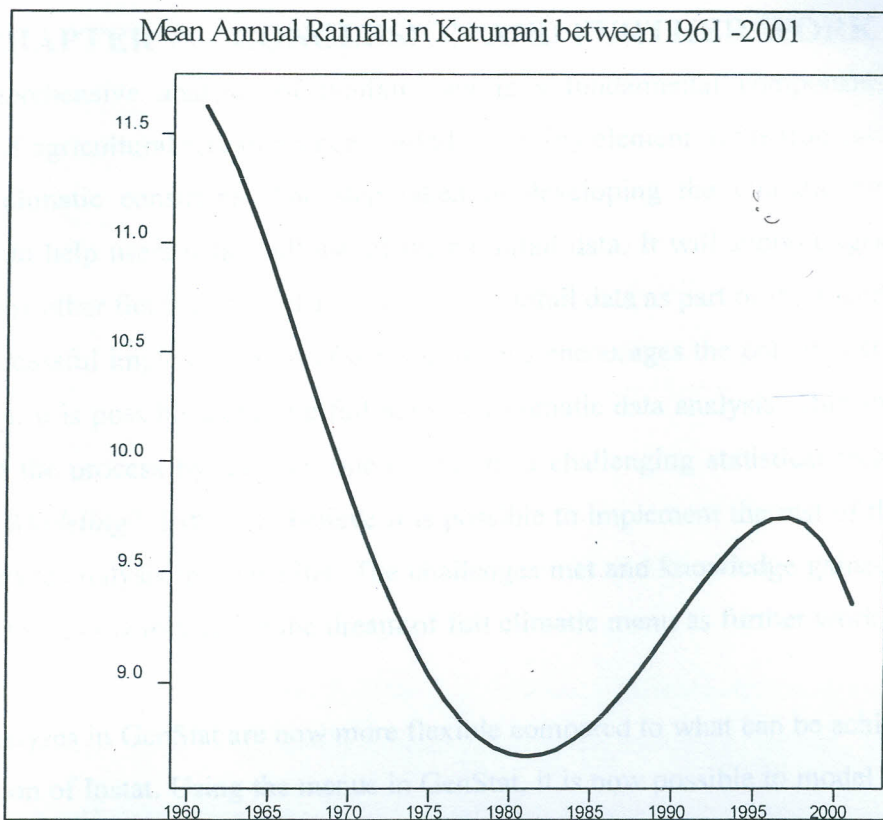
6.9.3: Modeling High Markov model of order two for Rainfall Amount

Graph 25 shows fitted mean rainfall amount of high Markov Chain model of order 2
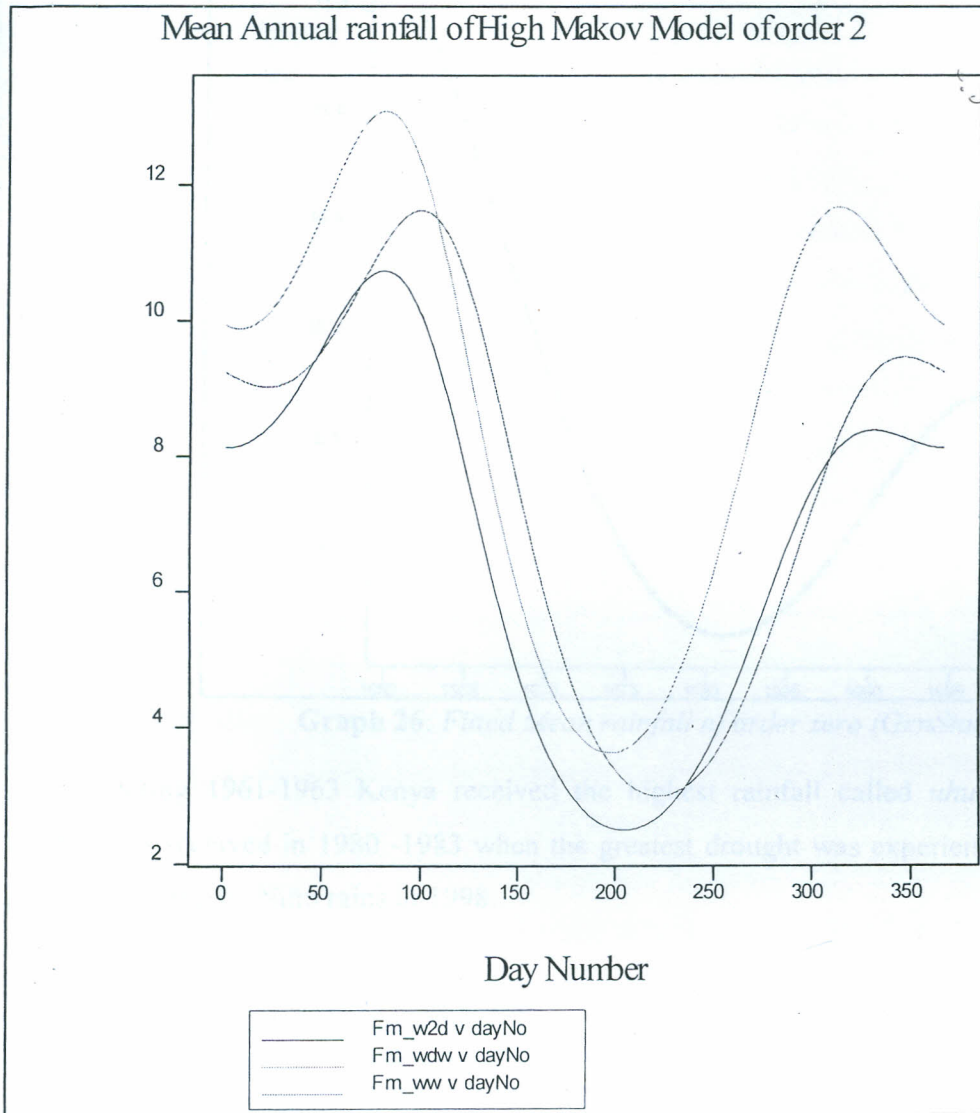
Graph 25: *Fitted Mean rainfall of High Markov of order two*

Order two normal Markov fits better than order highs for Katumani data.

Graph 26: *Fitted Mean rainfall of order zero (GenStat)*

During 1961-1963 Kenya received the highest rainfall called *uhuru*[Y]. The lowest rainfall was received in 1980 -1983 when the greatest drought was experienced. The second peak was during El-Niño rains in 1998.

[Y] Uhuru means Independence (The heavy rains were experience during the period of Kenyan independence.)

CHAPTER 7:   CONCLUSION AND FURTHER WORK

A comprehensive analysis of rainfall data is a fundamental component in the planning phase of agricultural research since rainfall is the key element in the tropical regions amongst other climatic constrains. The step taken in developing the climatic routine is therefore meant to help users make full use of their rainfall data. It will support agricultural research and many other fields that need an analysis of rainfall data as part of their work.

The successful implementation of the four menus encourages the potential statistical programmers that it is possible to have a full menu for climatic data analysis. This study has just jumpstarted the process by tackling one of the most challenging statistical techniques, "*Markov Chain Modeling*". I strongly believe it is possible to implement the rest of the work involved in climatic analysis into GenStat. The challenges met and knowledge gained in this work highly motivates me to achieve the dream of full climatic menu as further work beyond this study

The analyses in GenStat are now more flexible compared to what can be achieved by the current version of Instat. Using the menus in GenStat, it is now possible to model Markov Chain of higher order.

This work has simplified the modeling of rainfall data using GenStat since through the menus, the routines can now be accessed very easily and any other statisticians who are not programmers can still make use of. In addition, the analysis done through these routines can now be relied on since the portray the true distribution of rainfall data. The result from this analysis can therefore be used in publication.

The area of immediate address include modeling climatic events, crop performance index analysis, summaries of climatic data, time series analysis, and temperature analysis etc. For a full utilization of the package in handling climatic data, it is important to look forward to implementing all these aspects of climatic analysis.

The journey is not ending at GenStat only; we are looking forward to collaborate with other researchers who are interested in the same field e.g. of implementing the same in other powerful statistical software e.g. R.

REFERENCES

[1]   **Avik, G., Deepanwita, G., & Dashgupta, S. (2010)**. H*iher order Markov Chain Model for monsoon rainfall over West Belgal, India*. Indian Journal of Radio and Space Physics , 39, 39-44.

[2]   **Buchdahl, J. (1999)**. *Global Climate Change Student Guide: A review of contemporary and prehistoric global climate change. Atmosphere, Climate & Environment, Information Programme, aric*. Manchester Metropolitan University, (p. 99). Manchester.

[3]   **Chandler, R. E., & Wheater, H. S. (2002)**. *Analysis of rainfall variability using generalized linear models: A case study from the west of Ireland*. Water Resources Research. , 38, 1192.

[4]   **Gabriel, K. R., & Neumann, J. (1961)**. *A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv*. Quarterly Journal of Royal Meteorology Society , 88, 90-95.

[5]   **Gallagher, J., & Stern, R. (2004)**. *Analyzing Climatic Data Using Genstat for Windows*. University of Reading.

[6]   **Gallagher, J., & Stern, R. (2009)**. *Analyzing Climatic Data Using Genstat for Windows*. University of Reading.

[7]   **Goedhart, P. W. (1991)**. *Vtable; GenStat Release 7.1 Reference Manual* (p. 464). VSN International.

[8]   **Jones, P. G., & Thornton, P. K. (1999)**. *Fitting a third-order Markov rainfall model to interpolated*. Agricultural and Forest Meteorology , 97, 213-231.

[9]   **Kendall, L. (2008)**. *Generating Complete Sets of Daily Climatic Data Using a Range of Initial Inputs*. MSc Thesis, University of Reading, Department of Applied Statistics, Reading.

[10]   **Khamsi, M. (1999)**. *Markov Chain*. Retrieved May 23, 2010, from S.O.S. Mathematics CyberBoard. : http://www.sosmath.com/matrix/markov/markov.html

[11]   **Kottegoda, N. T., Natale, L., & Raiteri, E. (2004)**. *S*ome considerations of *periodicity and persistence in daily rainfalls*. J Hydrology , 296, 23–37.

[12] **Longhai, L., & Radford, M. (2008).** *Compressing Parameters in Bayesian High-order Models with Application to Logistic Sequence Models.* Bayesian Analysis , 3, 793-82.

[13] **Mimikou, M. (1983).** D*aily precipitation occurrences modelling with Markov chain of seasonal order.* Hydrological Sciences , 28, 2-6.

[14] **Nelder, J., & Wedderburn, R. (1972).** *Generalized Linear Model.* Roy Statistics Society , A135, 370-384.

[15] **Payne, R., Murray, D., Harding, S.& Baird, D. S. (2009).** *Genstat Release 12 Reference Manual, Part 3 Procedure Library.* (R. Payne, Ed.) Hemel Hempstead, Oxford, UK: VSN International.

[16] **Siriwardena, L., Srianthan, R., & McMahon, T. A. (2002).** *Evaluation of two daily rainfall data generation models, .* Retrieved from www.toolkit.net.au./cgibin

[17] **Stern, R. D. (1979).** *Analysis of daily rainfall at Samaru, Nigeria, using a simple two-part model.* Theoretical and Applied Climatology , Volume 28 (Numbers 1-2 / March, 1980), 123-135.

[18] **Stern, R. D., & Coe, R. (1982).** T*he Use of Rainfall Models in Agricultural Planning.* Agricultural Meteorology , 26, 35-40.

[19] **Stern, R., & Coe, R. (1984).** *A model fitting analysis of Daily Rainfall Data.* The journal of the Royal Statistical Society , 147 (1), 1-34.

[20] **Stern, R., & Coe, R. (1982).** *Fitting Models to Daily Data.* Journal of Applied Meteorology , 12 (7), 1024-1031.

[21] **Stern, R., Rijks, D., Dale, I., & Knock, J. (2006).** *Instat Climatic Guide.* University of Reading.

[22] **Vardi, Y., & Ju, W. (1999).** *A Hybrid High-order Markov Chain Model for Computer Intrusion Detection.* National Institute of Statistical Sciences, Alexander.