EXTENDING GENSTAT CAPABILITY TO ANALYZE RAINFALL DATA USING MARKOV CHAIN MODEL

Jacob Otieno Ong'ala (Mr.)
Center for Applied Research in Mathematical Sciences,
Strathmore University,
Nairobi
E-Mail:jotieno@strathmore.edu

David Ster (Dr.)
[2]Department of Mathematics and Applied Statistics
Maseno University
Kisumu

Roger Stern (Dr.)
Statistical Services
Center, University of Reading
United kingdom

Abstract:

Rainfall is of critical importance for many people particularly those whose livelihoods are dependent on rain fed agriculture. Methods of analysis of daily rainfall records based on Markov chain models have been available for many years and their value is widely recognized. However they are rarely used because of the complexity of their analysis. This paper describes how these models are being made more accessible through a series of specially written procedures and menus in GenStat, a widely available statistics package.

Key Words: Climatic analysis, Generalized Linear Model, GenStat, Markov Model, Rainfall data

1. Introduction:

Many people all over the world have devoted themselves in collecting climatic data for longer period of time but little is done on the analysis (Stern and Coe 1984). Despite a wide range of available statistical software, the effort of climatic data analysis still does not match their collection. Rainfall is one of the important climatic variables in planning and decision making in the agricultural sector particularly in those regions whose livelihood is dependent on rain fed agriculture. For this reason, as extensive understanding of rainfall regime is an important prerequisite in such planning.

Rainfall variable is a stochastic process in nature and therefore they require stochastic models to describe them (Mimikou 1983). Markov Chain is one of the stochastic models that have gained popularity in describing rainfall characteristics since its introduction by (Gabriel and Neumann 1962). They found that the daily rainfall occurrence for the Tel Aviv data successfully fitted using the first-order Markov chain model. Kotegota et al. in their paper (Kottegoda, Natale and Raiteri 2004) also reported that the first order Markov chain model found to fit the observed data in Italy successfully. However, (Wilks 1999) reported that there are cases where first order Markov chain model failed to fit the observed data and therefore higher order Markov chain model was an alternative to improve these inadequacies.

Although a number of powerful statistical packages have the capability to analyze rainfall data using Markov chain models, most of them do not have specialized routines for

doing this. Instat was introduced in the early 1980s as a simple statistical package to help in the teaching of statistics. It was later improved by adding more components with particular interest for processing climatic data (Stern, et al. 2006). Today it is the only available package with a specialized routine accessible for analyzing rainfall data using Markov chain models. Though it is not powerful enough to handle generalized linear model (GLM) (Gallagher and Stern 2009).

The primary objective of this paper is to use GenStat command language to implement a specialized routine for Markov modeling of rainfall data in GenStat Package by creating procedures and making them accessible by creating their dialogues and menus. The improvement of the package to handle Markov modeling of rainfall data will encourage most researchers and other interested parties to utilize climatic data in their work since the procedures will be accessible to perform such analysis. Section 2 provides a theoretical background to Markov modeling and GLM. Section 3 discusses some of the methods used during the implementation. The program itself is discussed in Section 4, and an example is discussed in Section 5. Finally section 6 concludes with a discussion.

2. Background Information:

A two state Markov chain is the commonly used type of Markov model where state is the condition of a day. A day is referred to as wet if rainfall received is greater than a threshold value (a minimum value say 0.85) or dry if the rainfall amount is at most than the threshold value. We will describe the Markov model for rainfall occurrences and amounts.

2. 1. The Generalized Linear Model

The Generalized Linear Model was introduced by (Nelder and Wedderburn 1972). It is used where the response variable neither follows a normal distribution nor have homogenous variances (Payne, et al. 2009). Comparing GLM and Multiple regression models (a form of general linear model) makes its features seen more clearly (Stern and Coe 1982). The expression below can be used to define General linear models:

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

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

$$Y = X\beta + \epsilon \qquad [2]$$

Where Y is a vector of response X is the matrix of explanatory variable (Covariate), β is a vector of unknown parameters (where β are estimated by solving the least-square equations shown in Eq. (3)) and ϵ is a vector of unobservable of errors corresponding to the observation.

$$X'X\hat{\beta} = X'Y \qquad [3]$$

The approach used by (Nelder and Wedderburn 1972) was to describe any given model in terms of its link function and its variance function. The variance function describes the relationship between 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.

Suppose we generalize Eq. (2) with a linear predictor based on the mean of the outcome variable, then the function $g(\mu)$ will be called the link function.

$$g(\mu) = \theta = X\beta \qquad [4]$$

The link function can be inverted as shown in Eq. (5)

$$\mu = g^{-1}(X\beta) \hspace{4cm} [5]$$

Rainfall occurrence y_i take a binomial distribution (it can rain or not rain) with mean μ then its link function is a logit as derived by (Nelder and Wedderburn 1972) and expressed as:

$$g(\mu) = log\left[\frac{\mu}{1-\mu}\right] = X\beta \hspace{3cm} [6]$$

Then the μ can be expressed as:

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

$$[7]$$

2. 2. Markov chain

A Markov chain is a time ordered probabilistic process that goes from one state to another according to some probabilistic transition rules determined by the current state only (Perera et.al, 2002). 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. Markov chain is useful for analyzing events whose likelihood depends on what happened last.

2.2.1.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

$$P_{ij,\tau} = Pr(X_\tau = j | X_{\tau-1} = i) \hspace{3cm} [8]$$

2.2.2.High Order Markov Chain

A Markov chain of order λ is referred to as high order Markov chain if λ greater than1. The probability of a day of the year having a particular state will depend on the states of the λ previous days. The Markov chains of order 2 and 3 satisfy the conditions in Eq. (9) and Eq. (10) respectively.

$$P_{i_2,i_1j,\tau} = Pr(X_\tau = j | X_{\tau-1} = i_1, X_{\tau-2} = i_2, \dots, X_0 = i_0) = Pr(X_\tau = j | X_{\tau-1} = i_1, X_{\tau-2} = i_2)$$
$$[9]$$

Where $P_{i_2,i_1j,\tau}$ is the transition probability of state j in day τ, year n given state i_1 in day $\tau - 1$ and state i_2 in day $\tau - 2$

$$P_{i_3,i_2,i_1j,\tau} = Pr(X_\tau = j | X_{\tau-1} = i_1, X_{\tau-2} = i_2, \dots, X_0 = i_0) = Pr(X_\tau = j | X_{\tau-1} = i_1, X_{\tau-2} = i_2, X_{\tau-3} = i_3))$$
$$[10]$$

Modeling high order Markov chain leads to a high-dimensional space of parameters (Vardi and Ju 1999). 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 climatic data to estimate them. However, (Longhai and Radford 2008) and" (Vardi and Ju 1999). Suggest that the high dimension of parameters can be reduced in such situations.

2.2.3.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 (Stern and Coe 1982). 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 $r_{i,\tau}$ (Stern and Coe 1984) 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 below.

$$r_{i,\tau} = \frac{n_{i1,\tau}}{n_{i1,\tau}+n_{i0,\tau}} \tag{11}$$

Where, $n_{i1,\tau}$ is the number of years with rain on day τ and $n_{i0,\tau}$ is the number of years with no rain on day τ

The random variable $n_{i1,\tau}$ is binomially distributed 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

$$p_{i,\tau} = g(\theta_{i\tau}) \tag{12}$$

Where g() is a logit link function connecting the probabilities $p_{i,\tau}$ to the function $\theta_{i\tau}$ which is linear unknown parameters (Stern and Coe 1984). The model is a generalized linear model since binomial is a member of the exponential family (Nelder and Wedderburn 1972) $p_{i,\tau}$ is therefore expressed as

$$p_{i,\tau} = \frac{exp(\theta_{i\tau})}{1+exp(\theta_{i\tau})} \tag{13}$$

Stern (Stern and Coe 1984) suggested that Fourier analysis may be used to express $\theta_{i\tau}$ as shown below:

$$\theta_{i\tau} = a_{i0} + \sum_{k=1}^{m}[a_{ik}\sin(k\tau') + b_{ik}\cos(k\tau')] \tag{14}$$

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

2.2.4.High Order Markov Chain

A Markov chain of order λ is referred to as high order Markov chain if λ greater than1. 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 Eq. (15) and Eq. (16) respectively.

$$P_{i_2,i_1j,\tau} = Pr(X_\tau = j|X_{\tau-1} = i_1, X_{\tau-2} = i_2, \dots, X_0 = i_0) = Pr(X_\tau = j|X_{\tau-1} = i_1, X_{\tau-2} = i_2) \tag{15}$$

Where $P_{i_2,i_1j,\tau}$ is the transition probability of state j in time τ, given state i_1 in time $\tau - 1$ and state i_2 in time $\tau - 2$

$$P_{i_3,i_2,i_1j,\tau} = Pr(X_\tau = j|X_{\tau-1} = i_1, X_{\tau-2} = i_2, \dots, X_0 = i_0) = Pr(X_\tau = j|X_{\tau-1} = i_1, X_{\tau-2} = i_2, X_{\tau-3} = i_3) \tag{16}$$

Modeling a high order Markov chain leads to a high-dimensional space of parameters (Vardi and Ju 1999). 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, (Longhai and Radford 2008) and (Vardi and Ju 1999) suggest that the high dimension of parameters can be reduced in such situations.

3. Methods

The implementation of this work involves creating four procedure using GenStat command language. These are: '*count*', '*prepare*', '*fitting*' and '*fittingamount*' using the GenStat command language. The count procedure reads the raw data then counts the number of days with a specific state over the years and calculates the amount of rain in the rainy days using Markov model. The prepare procedure in calculates the probability of rain for each day of the year, the fitting procedure fits the probability of rain for each day of the year while *amountfitting* procedure fits the amount of rainfall.

GenStat has a capability of allowing users to create their own menus for newly developed procedures (Gallagher and Stern 2009). Once a procedure has been written, it can be recalled and used in the command interface or its corresponding menu and dialogs built and used in a graphical user interface as described by (Gallagher and Stern 2009).

4. Program
4.1. Procedures

The dialogs and menus for the four procedures are built in GenStat and can be accessed through a newly created menu called '*user*' (any time you add a procedure into GenStat, it will be listed in the menu 'user' See Figure 1). The user menu contain four submenus namely; *Counts and Total From daily data*, *Probability of rain*, *fitting probability* and *fitting amounts* corresponding to '*count*', '*prepare*', '*fitting*' and '*fittingamount*' procedures respectively (see Figure 2 and Figure 3). Once the procedures have been built in GenStat system, they can now be used to analyze climatic data using Markov chain model.
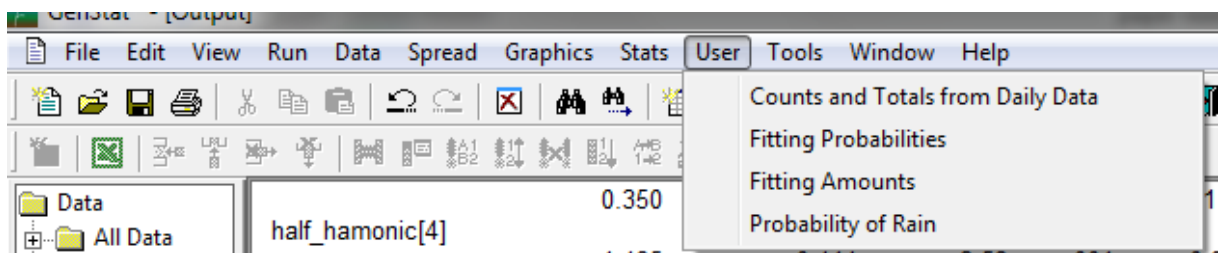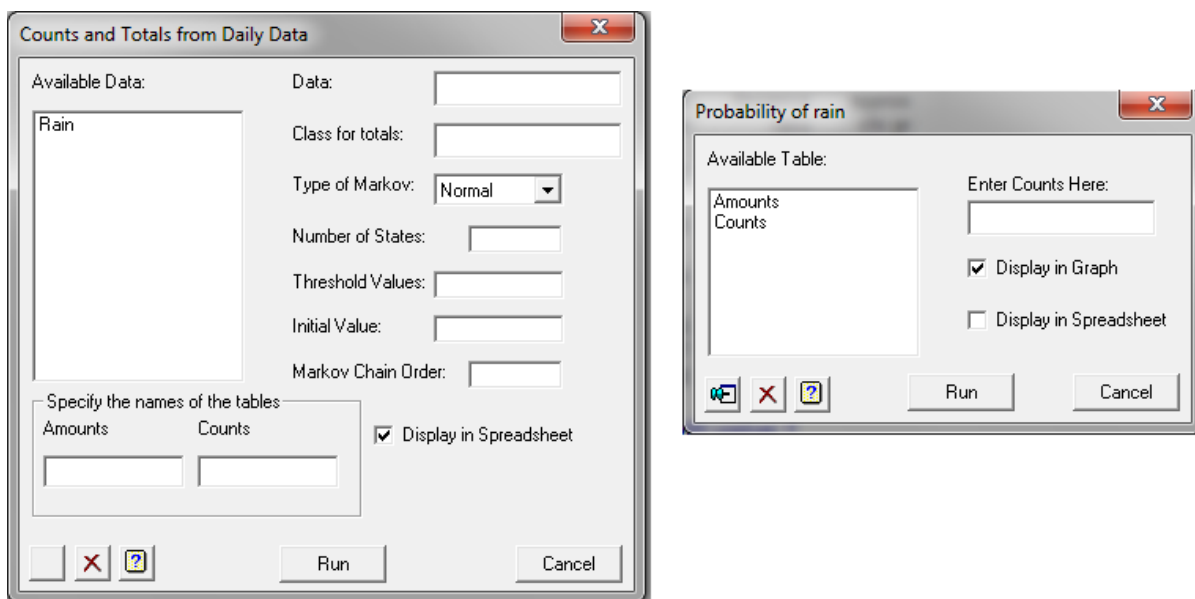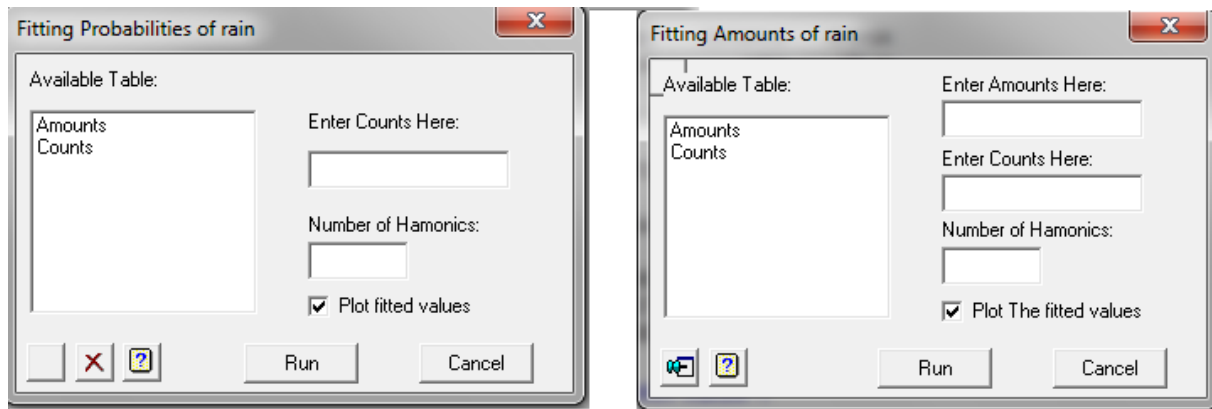

Figure 1: User menu

Figure 2: Counts and Total form Daily Data and Probability of rain dialog boxes

Figure 3: Fitting probability and fitting amounts

4.2. Input data

The input data required is in variate data type in a single column for all the years (rainfall data) and a factor column over which the counts are done. The factor column can be in days, months or weeks of the year. A sample of data required for the procedures is shown in Figure 4

Figure 4: Sample rainfall data

In most cases, the rainfall data may not be given in one stalked column as shown in figure 1, in such a case the user is expected to stalk the data into one column over the years the data is given. This facility is available in GenStat by using ***Spread=> Manipulate=> Stalk*** menu.

4.3. Setting up the analysis

Modeling rainfall data using these newly created procedures are done in two stages. The first stage of the analysis is to determining the rain counts and total of specific days over the years using *count* procedure. Then any the remaining three procedures can follow since they use the results from *count* procedure. It is in this first stage of analysis, where the user specifies the order of chain, the threshold value of rainfall and whether or not the model will be high or normal.

In the second stage the calculation of probabilities is done using the 'Probability of rain' dialog box, the user simply specify the table for the counts and then indicate whether to display the result in a table or a graph, both or none. Then to fitting of the probabilities is done by using a 'fitting probabilities' dialog box where the user specifies the table for counts, number of harmonics used for fitting and whether to plot the fitted values or not. Finally, when fitting amounts, the 'fitting amount' dialog box is used; the amounts and counts tables have to be specified.

5. Example.

We will use the procedures to analyze the rainfall dataset for Samaru1, Nigeria collected from 1930-1940. The data is available in Instat library. It is exported to GenStat spreadsheet and then stalked into one variate of rainfall data and a factor column for year. A factor column for day number with levels 1-366 is created over which the counts will be done.

The analysis in this example will be based on the following categories; two states and a three state, normal orders and high order, Markov on daily basis and Markov summarized to group of days totals (weekly, monthly and 5-days etc).

The analysis starts with the count procedures with the options of two states Markov. This results for count and amount of rainfall in a summary table on a spreadsheet. The command associated with the analysis is shown below.

COUNTS [CLASS=DayYear; HIGH=NO; SPREAD=YES; STATES=2] a=Amount; counts=Count; DATA=Rain

Based on the number of states and order of the Markov chain specified, in the *Counts and Total* dialog box, the *'counts'* procedure counts the number of times that a day of the year is having a specific state-condition for the number of years the data is observed and then calculates the rainfall amounts for rainy days which is defined as the actual amount of rainfall recorded minus the 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 probability plots for the model is obtained by using the prepare procedure and plot a graph for the model shown in the Figure 5. The plot is overcrowded and seems hard to read and distinguish; a better plot can be obtained when the days of the years are summarized into groups. The fitted model for counts and amounts are shown in figure 3 (a) and (b). The results in Figure 6(a) indicate that there is a higher chance of rain between day 150 and day 270 though considering the state condition of previous day (Markov chain of order 1), their is a higher chance of rain given that the previous day was rainy than when the previous day is dry. In Figure 6(b), the highest amount of rainfall given that the previous day being wet was experienced between days 150 to day 230. However, the expected amount of rain will be higher given that in the previous day it had rained.

---

[1] This data set is readily available in Instat library, Instat package is downloadable freely from http://www.ssc.rdg.ac.uk/
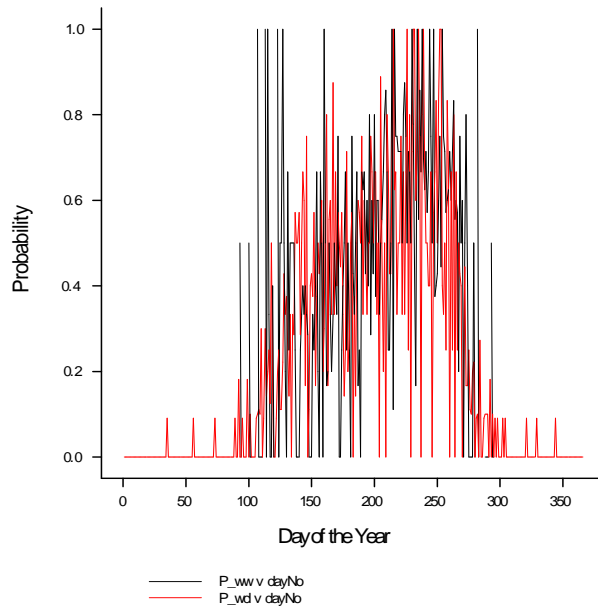
Figure 5: Probability Plot for a two-state Markov Chain Model
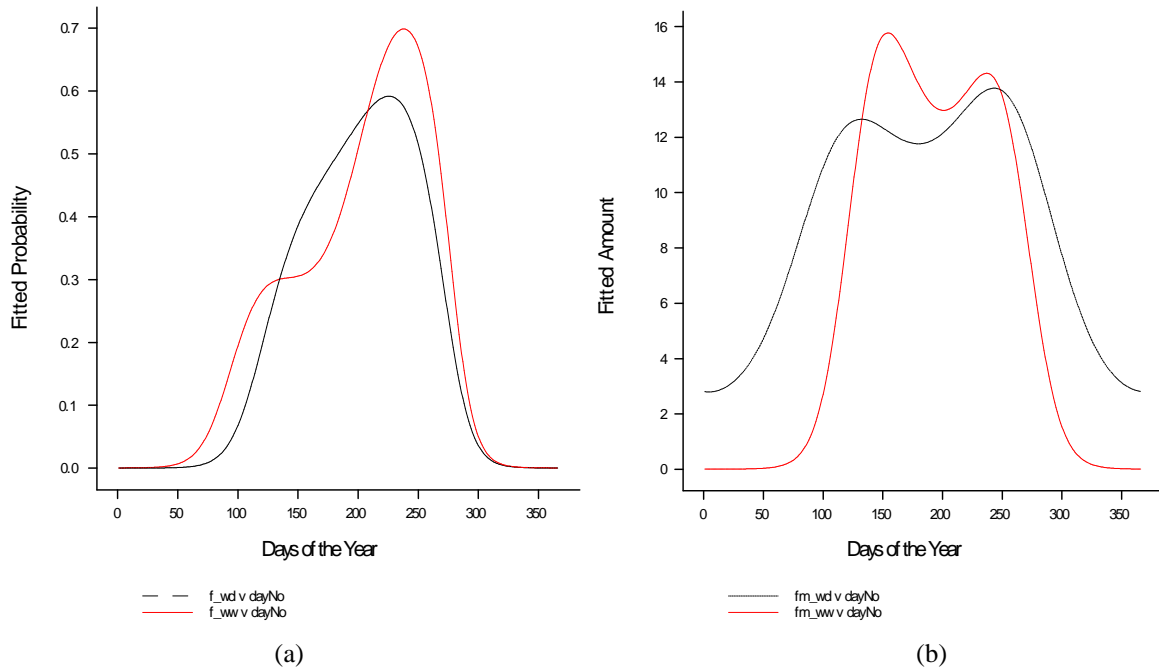
Figure 6: Fitted Probability Plot (a) and Amount (b) for a two-state Markov Chain

The analysis of deviance is also given when the analysis is run (see Table 1) and it suggest that two harmonics is significant (P-Value<0.001) when fitting the model

Table 1: Analysis of deviance

Change	d.f.	deviance	Mean deviance	deviance ratio	Aprox. chi pr
+ half_hamonic[1]	1	55.762	55.762	55.76	<.001
+ half_hamonic[2]	1	19.970	19.970	19.97	<.001
+ half_hamonic[3]	1	11.364	11.364	11.36	<.001
+ half_hamonic[4]	1	18.066	18.066	18.07	<.001
Residual	193	223.368	1.157		
Total	197	328.530	1.668		

6. Summary and Conclusion

The current version of GenStat (version 14) is very powerful in statistical analysis and in particular climatic data analysis with the capability of handling the rainfall data using Markov chain model approach, however this functionality is not accessible a directly that non-statistician can used. In this work therefore, we have presented four GenStat procedures for analyzing rainfall data using Markov Chain model approach. The procedure can now be used directly through the dialogs and menu.

We have illustrated the use of these procedures by applying it to rainfall data for Samaru, Nigeria. The example illustrated here is only a one case (a normal Markov chain model of order 1 with two states) out of other possibilities that the procedure can perform including: Markov chain model of order (0, 2, 3,…n), more than states model and High order Markov chain explained in section 2.2.4.

Future work might include modeling and implementation 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 in implementing all these aspects of climatic analysis.

7. Acknowledgment

8. References

Gabriel, K, R, and J Neumann. "A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv." *Quarterly Journal of Royal Meteorology Society*, 1962.

Gallagher, J, and R Stern. "Analyzing Climatic Data Using Genstat for Windows." University of Reading, 2009.

Gallagher, J, and R Stern. "Analyzing Climatic Data Using Genstat for Windows." University of Reading, 2009.

Guttorp, P. *Stochastic Modellingof scientific data.* Chapman and Hall/CRC, 1995.

Kottegoda, N, T, L Natale, and E Raiteri. "Some considerations of periodicity and persistence in daily rainfalls." *J Hydrology*, 2004.

Longhai, L, and M. Radford. "Compressing Parameters in Bayesian High-order Models with Application to Logistic Sequence Models." *Bayesian Analysis*, 2008.

Mimikou, M. "Daily precipitation occurrences modelling with Markov chain of seasonal order." *Hydrological Sciences*, 1983.

Nelder, J.A, and R.W.M Wedderburn. "Generalized Linear Model." *Roy Statistics Society*, 1972.

Payne, R.W., et al. *Genstat Release 12 Reference Manual, Part 3 Procedure Library.* Edited by R,W Payne. Hemel Hempstead, Oxford: VSN International, 2009.

Perera, H.K.W.I, D.U.J Sonnadara, and D.R. Jayewardene. "Forecasting the Occurrence of Rainfall in Selected Weather Stations in the Wet and Dry Zones of Sri Lanka." *Sri Lankan Journal of Physics*, 2002.

Siriwardena, L, R Srianthan, and T, A McMahon. *Evaluation of two daily rainfall data generation models, .* December 14, 2002.
www.toolkit.net.au./cgibin/webobjects/toolkit.woa/wa/downloadPublications?

Stern, R, D Rijks, I Dale, and J Knock. "Instat Climatic Guide." University of Reading, 2006.

Stern, R, D, and R Coe. "The Use of Rainfall Models in Agricultural Planning." *Agricultural Meteorology*, 1982.

Stern, R.D, and R Coe. "A model fitting analysis of Daily Rainfall Data." *The journal of the Royal Statistical Society*, no. 1 (1984).

Vardi, Y, and W.H Ju. "A Hybrid High-order Markov Chain Model for Computer Intrusion Detection." National Institute of Statistical Sciences, Alexander, 1999.

Wilks, D. S. "Inter annual variability and extreme-value characteristics of several stochastic daily precipitation models." *Agric For Meteorol*, 1999.