FILLING-IN MISSING MONTHLY STREAM-FLOW DATA FOR RIVERS WITH SEASONAL RUNOFF

BY

NAHID AFZA

A Thesis
Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Civil Engineering
University of Manitoba
Winnipeg, Manitoba

ISBN 0-315-76620-4

FILLING-IN MISSING MONTHLY STREAM-FLOW

DATA FOR RIVERS WITH SEASONAL RUNOFF

BY

NAHID AFZA

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

ABSTRACT

In this thesis the use of the seasonal group characteristics (occurrence of high and low flow groups) in the monthly stream flow data infilling was investigated. Two multivariate monthly flow data infilling models were developed. One model reconstruct a flow group of missing data river by conditioning on the simultaneously observed flow group in the nearby located river. The other model reconstruct a flow group of a missing data river by conditioning on the preceding flow group of the same river. The later model performed very poorly while the first model performed satisfactorily only in cases of a longer period of concurrent data.

Further, in this thesis the scope of the use of seasonal group characteristics (homogeneity characteristic e.g. high flow group contains high flows and vice versa) to extract seasonal samples (homogeneous samples) for the application in the regression models were studied. These samples were found beneficial only in the reconstruction of one seasonal data by inducing larger estimation error in other season. Due to the random variation in the occurrence of the flow groups, the adopted procedure of seasonal segmentation (splitting the year in two periods of high or low flow) assigned few flows to incorrect season thus causing larger estimation error. Thorough investigation of such sample are needed prior to the use for estimation of missing data.

TABLE OF CONTENT

CHAPTER 1

INTRODUCTION

1.0 PROBLEM DEFINITION

In the water resources planning and management practice, adequately long uninterrupted monthly stream-flow data series are needed in order to analyze the sequential properties of the historical data for various purposes, such as, forecasting, synthetic trace generation, determination of yield, capacity and operational policy of storage facilities. Often the existing historical monthly stream-flow record is not only short but also contain one or more gaps. A gap in the monthly stream-flow data series may be caused by the instrumental malfunction during the data measurement, data transmission and storage. A gap can also occur due to the calibration error caused by the occurrence of an extreme event. Such a gap as shown in Fig. 1.1, divides the data series into disjointed sub-series. The gap needs to be bridged by using a suitable data infilling method which considers the complicated nature of monthly stream-flow data.

1.1 COMPLEX NATURE OF MONTHLY FLOW DATA AND DILEMMA IN THE CHOICE OF AN INFILLING MODEL

The monthly stream-flow data is characterized by persistence and cyclic variation in the magnitude of flow. The cyclic variability in the magnitude of monthly flow data [e.g. Fig. 1.2] causes cyclic variation in correlogram [Fig. 1.3].

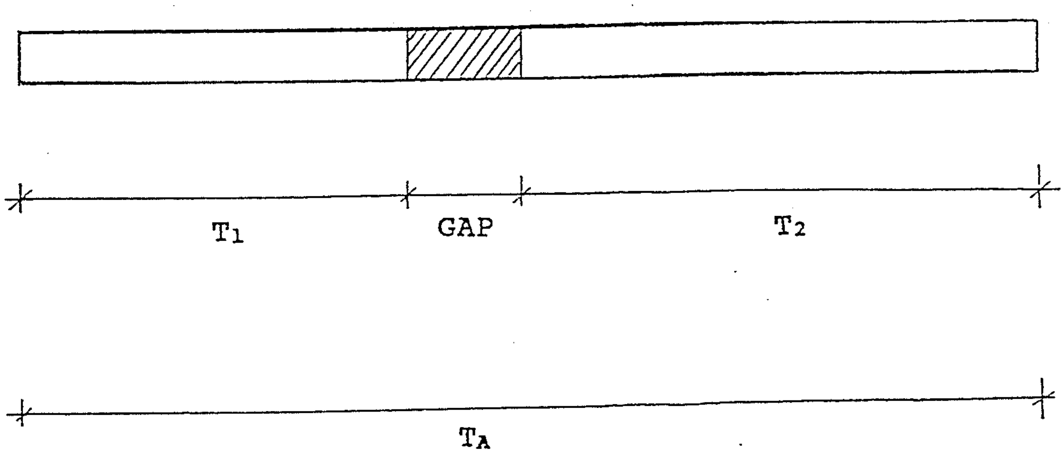


Fig.1.1: Gap in stream-flow data series

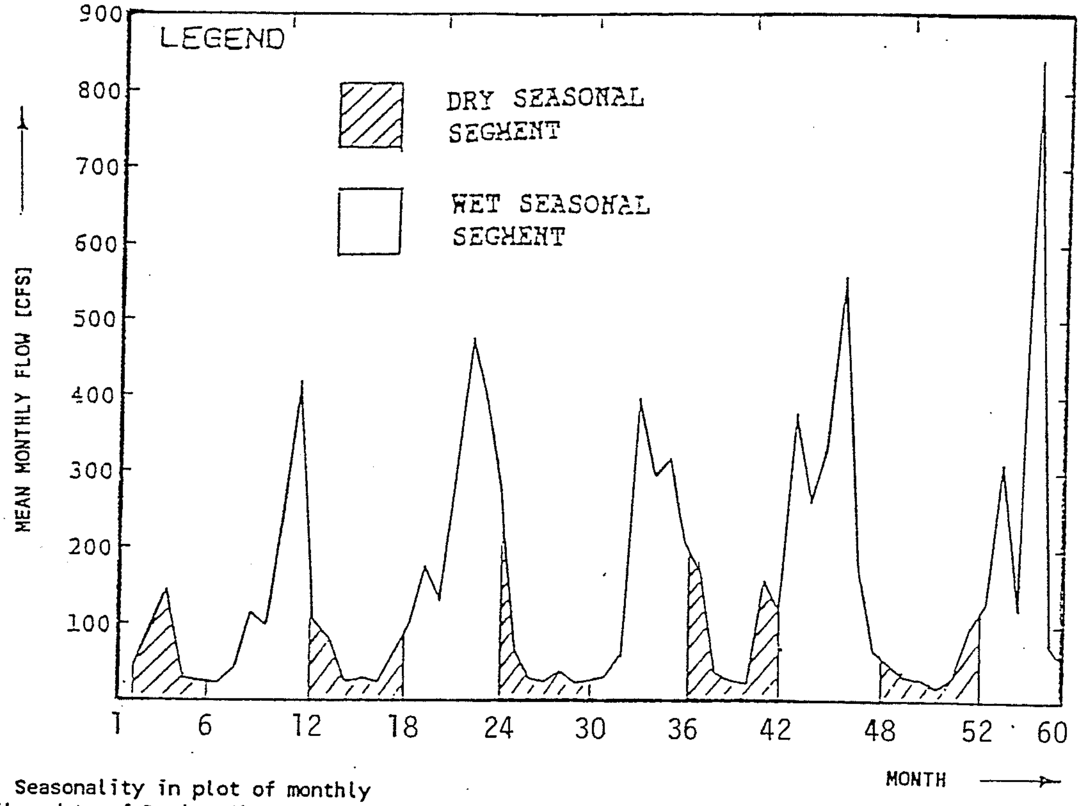


Fig.1.2: Seasonality in plot of monthly stream-flow data of Dunlap Ck. [June 1958 - May 1963]

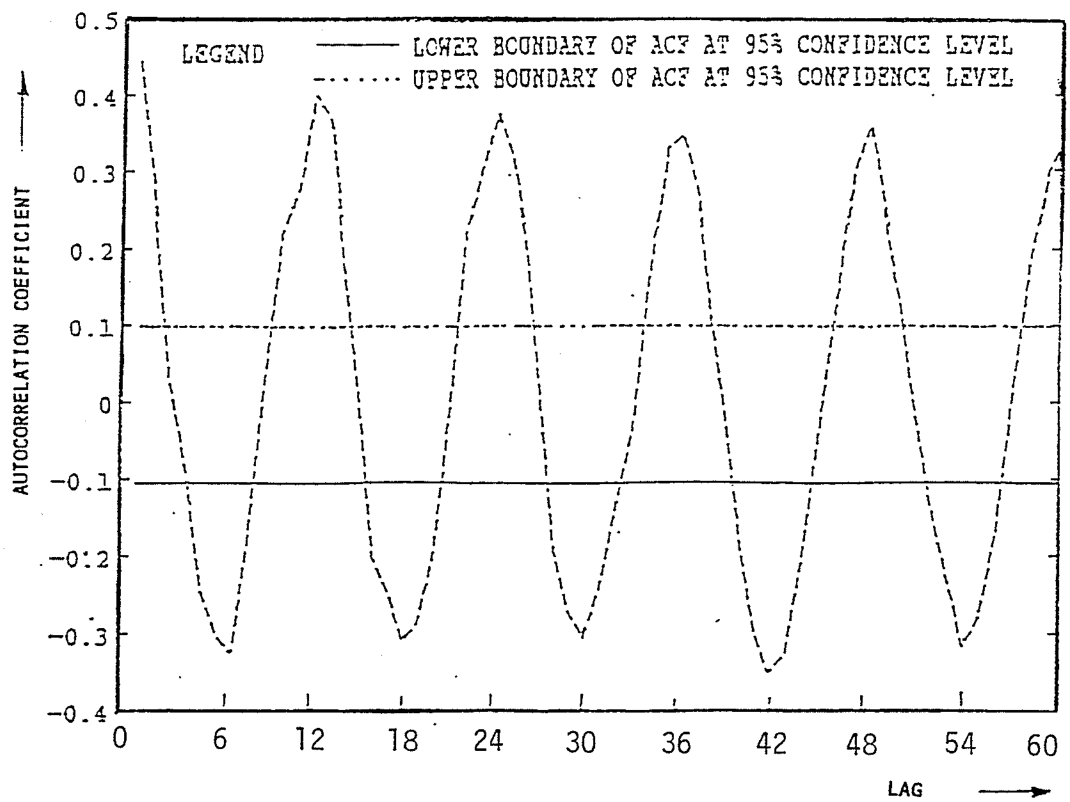


Fig.1.3: Seasonality in correlogram of monthly stream-flow data of Dunlap Ck.

The value of cross correlation coefficient between monthly stream-flow series of two nearby located rivers is often very high. Due to the high cross correlation coefficient, a regression model could be considered to be viable means for infilling of missing data of the short data river by exploiting the concurrent data of a base river (cross river information transfer). But, use of such a model in the serially correlated series produce serially correlated error which is considered as an indicator of possible distortion of a fitted line. When such a distorted line is applied to estimate the missing data of a serially correlated monthly stream-flow data series, the estimated data may differ very much from the actual value of the missing data.

Circumventing such a situation could possibly be achieved by using de-seasonalized data and then fit a AR(1) model on the residual. But the procedure of de-seasonalization is criticized for causing drastic drop in the correlation between {the missing data series and a base series pair} causing substantial amount of loss of information in the cross river information transfer [Harmancioglu and Yevjevich 1987].

Due to the cyclic variation, the monthly stream-flow data belonging to different months may vary substantially and when only one infilling equation is calibrated from all available data, wide variation among the observations of the sample may cause unrealistic estimate of the missing data. A possibility of encompassing the cyclic variation could be the use of 12 different infilling equations for infilling the missing monthly stream-flow data of the short data series belonging to the 12 different months [Hirsh (1979, 1981), Vogel and Stedinger (1985)], but this procedure is not reliable in case of the availability of few years of concurrent observations.

Three major groups of missing stream-flow data filling models are available:

- A group of models use regression type equations on raw data and disregard persistence.
- Another group of models use either a multiple regression or a multivariate regression to infill multi-site data. Both of the models use de-seasonalized data.
- The third group of models incorporate either regional statistics parameters or some physiographic characteristics in the infilling equation and use either raw data or de-seasonalized data.

1.2 PROPOSED MODELS

The persistence and cyclic variation of the monthly stream-flow data offers very little scope to build an infilling model which would simultaneously meet all these constraints. In the hydrological literature, the existing infilling models make some form of compromise among these constraints to obtain an realistic estimate of the missing data. Panu [1978, 1980] considered a completely different approach to encompass serial persistence and cyclic variation. Panu [1978,1980] viewed that the time plot of a monthly data series as a sequence of high and low flow groups. He considered further that the persistence can be embraced in terms of inter-group relationship (of lag-one Markovian nature) between groups and intra-group relationship between the members of the groups.

In this thesis, an effort is made in order to investigate the efficacy of the use of group characteristics in a data infilling model. To pursue the above goals two multivariate models, referred as MULBS and SESTRNALL, (detail specifics are given in chapter three) are proposed. These models encompass the cyclic variation completely and thus reconstruct a segment rather than an element of a time series.

These models consider group relationship which is different from the conventional notion of persistence. The following concept of **seasonal group characteristics** is considered to build the configuration of the proposed models:

- The monthly stream-flow data plot of a river [Fig. 1.2] exhibits periodic occurrence of a **peak and a valley** over a period of one year, these are denoted by **wet and dry**

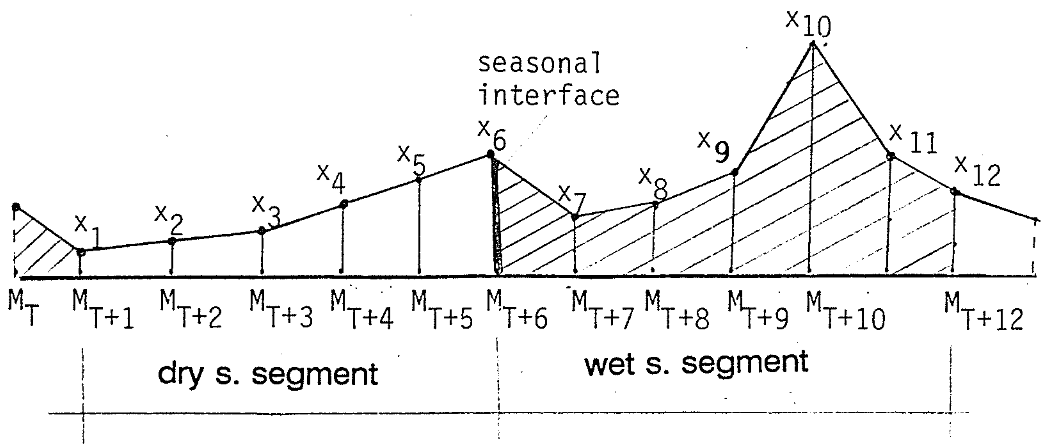
seasonal segment respectively.

- Each of the wet and dry seasonal segments lasts for six months, each can be denoted by a six dimensional vector $X=[x_1 \ldots . x_6]^T$. This vector represents the association of the six monthly flow values within a **group** and the flows belonging to a group are considered to be **similar** (high or low) (Fig: 1.4). This similarity criteria is referred by **intra-seasonal homogeneity.**
- The dry seasonal segments are considered similar to each other but dissimilar from the wet seasonal segment.
- The consecutive six months period (seasonal period) over which the high/low flow persists in the wet/dry seasonal segments are considered as **season** [this definition is different from the definition of a season used in the time series analysis, in the time series, the seasonal length is twelve month]
- The plot of monthly data [Fig.1.3] can be considered as a sequence of seasonal segments [Fig. 3.3.a, b].

RATIONALE BEHIND THE DERIVATION OF THE PROPOSED MODEL

MULBS:

In addition to the above mentioned seasonal group characteristics, following characteristics are considered in the derivation of MULBS model:



Vector representation of Dry seasonal segment $\quad X_D = [\ x_1\ \ x_2\ \ x_3\ \ x_4\ \ x_5\ \ x_6\ ]^T$

Vector representation of Wet seasonal segment $\quad X_W = [\ x_7\ \ x_8\ \ x_9\ \ x_{10}\ \ x_{11}\ \ x_{12}\ ]^T$

x_j = flow on j^{th} month on a year

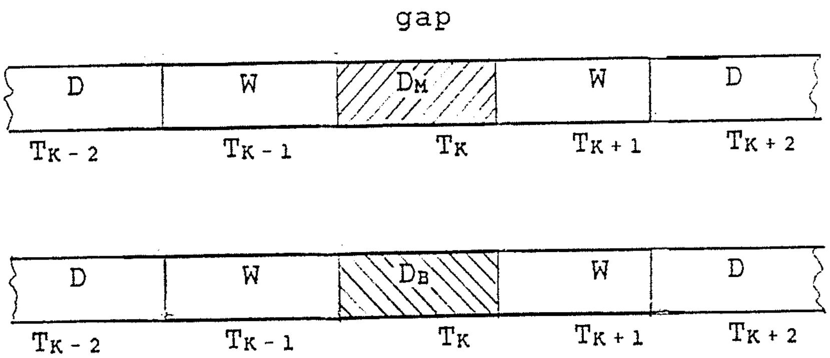
Average dry seasonal group relationship = $COV(X_D)_{6x6}$, when computed from all available dry seasonal segments

Average wet seasonal group relationship = $COV(X_W)_{6x6}$, when computed from all available wet seasonal segments

Fig.1.4: Vector representation of dry and wet seasonal segments of a year.

The simultaneous plot of stream-flow of two or more nearby located rivers [Fig.4.1.2] exhibits almost concurrent occurrence of peak and valley which can be justified by the coincidence of precipitation. This plot suggests a dependence among the simultaneously observed seasonal segments and such a dependence appears to be consistent over time (when there is a high peak in one river, high peak is also exhibited in the other rivers).

Such a concurrence of observation of similar seasonal segments in nearby located river justifies to consider a model which would be able to reconstruct a missing seasonal segment of the short data river conditioned on the simultaneously observed corresponding seasonal segment of the base river. This model reconstruct a specific seasonal segment of the short data series at T_k^{th} seasonal period by conditioning on the corresponding observed seasonal segment of a base river series at T_k^{th} seasonal period [Fig. 1.5]. This model reconstruct a dry seasonal segment of the short data series by conditioning on the existing dry seasonal segment in base river series. Similarly, this model reconstruct a wet seasonal segment of the short data series by conditioning on the existing wet seasonal segment in the base river series. For the reconstruction procedure this model utilizes a multivariate conditional distribution, detail of this procedure is given in chapter three. This model completely disregard any



Short data series A

Base R. data series B

. D= a dry s.segment
. W= a wet s.segment

D_M = missing dry seasonal segment
D_B = observed dry seasonal segment

objective: find:

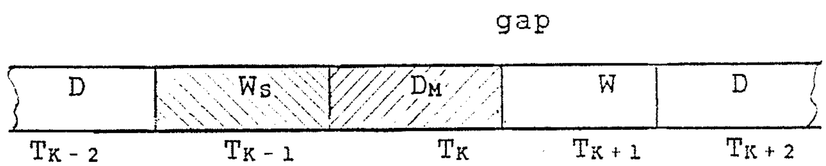
$P(D_M \mid D = D_B)$, derivation of distribution of D_M conditioned on observed D_B

Fig.1.5: Reconstruction of a missing dry seasonal segment D_M by MULBS model

relationship at the interface of adjacent seasonal segments, thus, disregard any relationship being carried over from one seasonal segment to the next seasonal segment (i.e. the model ignores inter-seasonal group relationship).

SESTRNALL MODEL:

BACKGROUND: In the discussion of the scope of application of pattern recognition principle, Panu (1978) proposed a method of analyzing the seasonal group characteristics of monthly flow data of a river and finding a missing data segment by projecting the preceding segment of the same river. This method assumes that the seasonal segments follows a Markovian transition. The transitional probability, cluster configuration of the seasonal segments and distances of the seasonal segments would provide sufficient information to infill a gap segment [these concepts are explained in chapter three]. This method was attempted by Frenette (1988), but the work remained incomplete. This method was initially considered for the thesis, but, this method was modified subsequently. The method proposed by Panu (1978) and the rationale behind the modification is presented in Appendix A.1.2. The modified model (SESTRNALL) considers the inter-seasonal relationship as the infilling basis, but, adopts different operational procedure.



Short data series A

. D= a dry s. segment
. W= a wet s. segment

D_M = missing dry seasonal segment
W_S = observed wet seasonal segment

objective: find

$P(D_M \mid W = W_S)$, derivation of distribution of D_M conditioned on observed W_S

Fig.1.6: Reconstruction of a missing dry seasonal segment D_M by SESTRNALL model

SESTRNALL MODEL: This model reconstruct a specific seasonal segment of the short data series at T^kth seasonal period by conditioning on the preceding seasonal segment of the same series at $T_{k-1}{}^{th}$ seasonal period [Fig. 1.6]. This model reconstruct a dry seasonal segment of the short data series by conditioning on the preceding wet seasonal segment of the short data series. In a similar way, this model reconstruct a wet seasonal segment of the short data series by conditioning on the preceding dry seasonal segment of the same series. For the reconstruction procedure this model utilizes a multivariate conditional distribution, detail of this procedure is given in chapter three. This model takes account of the relationship at the interface of adjacent seasonal segments, thus, considers that the group relationship being carried over from one seasonal segment to the next seasonal segment (i.e. the model accounts for inter-seasonal group relationship). It also considers intra-group relationships.

1.3 THESIS OBJECTIVE:

- Investigation of the efficacy of using seasonal group characteristics in multivariate monthly flow data infilling models and rationale for future use.

.Scope of the use of seasonal group characteristics in extracting samples (based on seasonal homogeneity) for the existing univariate regression type data filling model.

The thesis is organized as outlined below:
In Chapter two a review of the existing infilling models are given to put the proposed models into perspective. The Chapter three develops the mathematical background of the proposed models and it also develops a procedure of sample extraction based on similarity criteria. The Chapter four discusses the application of the proposed models in real world data. The results obtained are analyzed in Chapter five. The Chapter six contains concluding remarks and scope of further research.

CHAPTER 2

LITERATURE REVIEW

2.0 PREFACE

The existing monthly flow data reconstruction models can be broadly classified into three major categories so that the models belonging to a particular category do not differ from each other in three aspects, namely; (i) the way the models handle serial persistence and cyclic variation, (ii) the criteria (**information source**) the models use in the infilling process and (iii) the restriction the models impose on the number of short data series. These categories are: regression type models, multi-site models and models using physiographic factors. Only first two category of models are comparable to a limited extent to the proposed models .

2.1 EXISTING MODELS

REGRESSION TYPE MODELS :These models disregard serial persistence and cyclic variation in monthly data, permit infilling of gap at a single site and incorporate cross-river information transfer for infilling process. The least square regression, regression with noise [Hirsh (1982)] and maintenance of variance extension (MOVE) [Hirsh(1982), Vogel and Stedinger (1985)] belong to this category. These models are basically meant for annual flow data augmentation, hence, do not take account of serial persistence and cyclic variation. Their structural configurations can be compared by considering the following hypothetical case:

Let

$x_1 \quad x_2 \quad x_3 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \; x_{(n1+n2)}$

$y_1 \quad y_2 \quad y_3 \ldots\ldots\ldots\ldots \; y_{n1}$

represent a **n1 + n2** period long stream-flow data sequence **X** of a base river and **n1** period long stream- flow data sequence **Y** of the river with missing data. The **n2** period long gap in series **Y** is reconstructed by both the regression and the MOVE models by means of regressing series **Y** on series **X**

Least square regression equation without a noise term is:

$$\hat{y}_i = \bar{y}_1 + b\,(x_i - \bar{x}_1) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [2.1]$$

where,

$\bar{x}_1$ and $\bar{y}_1$ respectively are the sample estimates of mean of n1 period of concurrent data of series X and Y respectively,

b is the estimate of least square regression parameter, $\hat{y}_i$ and x_i are respectively the estimates of missing data of series Y at i^{th} period and the concurrent observed data of series X.

The regression equation with noise is:

$$\hat{y}_i = \bar{y}_1 + b\,(x_i - \bar{x}_1) + \alpha \,\Theta\, \sqrt{(1 - r^2)}\, s_{y1}\, e_i \ldots\ldots\ldots [2.2]$$

where,

e_i = noise at period i

α = unbiasing factor computed by equating $E(s_y^2) = \sigma_y^2$

Θ = Noise indicator i.e. $\theta = 0$ when no noise is added

$\theta = 1$ when noise is added

r = estimate of correlation coefficient

s^2_{y1} = estimate of variance of n1 period of existing observations of the series Y

When estimates of missing data are computed by Equation [2.2], the estimates of an augmented series mean and variance would be unbiased. These augmented series estimators are known as Matalas and Jacobs's unbiased estimators.

In stream-flow data reconstruction, mathematical formulation of *MOVE* models were done by excluding the noise term from the regression equation [2.2], yet, preserving some desired characteristics of the augmented series (i.e. Matalas and Jacobs's unbiased estimators of the mean and the variance of the augmented series). The noise term was needed to be excluded in order to produce an unique estimate of the missing data [Hirsh (1982)]. Four MOVE models were developed i.e. MOVE.1 and MOVE.2 by Hirsh (1982) , MOVE.3 and MOVE.4 by Vogel and Stedinger (1985).

Vogel and Stedinger (1985) presented the comparison among various MOVE models in detail, which is summarized below :

$$\hat{y}_i = a + b\, x_i \qquad [2.3]$$

All the MOVE models follow the general configuration given by Equation [2.3] and in each case, the parameters are derived on the basis of the desired characteristics to be preserved , namely:

In MOVE.1, the regression parameters are derived so that a **n1 + n2** long generated sequence of **Y** series by Equation [2.3] reproduces the historical mean and variance of **Y** series.

In MOVE.2, the regression parameters are derived so that a **n1 + n2** long generated sequence of **Y** reproduces Matalas-Jacobs's unbiased estimators of mean and variance of augmented series.

In MOVE.3, the regression parameters are derived so that a **n1** long historical data sequence together with **n2** long data sequence generated by Equation [2.3] reproduces Matalas-Jacobs's estimators of mean and variance of the augmented series.

In MOVE.4, the regression parameters are derived so that a **n1** long historical data sequence together with **n2** long data sequence generated by Equation [2.3] reproduces Vogel and Stedinger's minimum variance estimators of mean and variance of the augmented series. The minimum variance estimators of mean and variance of the augmented series were developed by Vogel and Stedinger(1985) by using a linear combination of the corresponding parameter estimators of the observed short data series and the Matalas and Jacobs's unbiased estimators of the augmented series as given below:

$$\hat{\mu}^{*}_{y} = (1 - \Theta_1)\ \bar{y}_1 + \Theta_1\ \hat{\mu}_y \qquad \text{.....................[2.4]}$$

$$\hat{\sigma}^{*2}_{y} = (1 - \Theta_2)\ s^2_{y1} + \Theta_2\ \hat{\sigma}^2_y$$

where

$\hat{\mu}_y$ = Matalas and Jacobs's estimator of the mean of the augmented series

$\hat{\sigma}^2_y$ = Matalas and Jacobs's estimator of the variance of the augmented series

The parameters Θ_1 and Θ_2 are computed by minimizing the variance of mean and the variance estimators given in Equation [2.4].

All these models violate the underlying model assumption of serially independent X and Y series. Such an use would induce auto-correlated error, which apart from giving an incorrect estimate of the missing data, may also cause consistent bias thus producing a serious distortion in the sequential properties. The proponent of these models [Hirsh (1979, 1982), Vogel and Stedinger (1985)] are quiet aware of this problem. They rationalize such an use by the fact that the data series of nearby located rivers have similar serial correlation properties, regression type equations would be able to map these characteristics from base river to the river with missing data, thus, the serial correlation structure of the river with missing data will not be distorted by such an infilling.

The MOVE models (MOVE.2, MOVE.3, MOVE.4) claim that they maintain the variance of the Y series (a time series), but in reality, they maintain the variance of an independent series [Equation 2.2].

Hirsh (1982) found the performance of the MOVE.1 and MOVE.2 to be superior to the regression equations, particularly with respect to the bias in higher order statistics. The performance of MOVE.3 and MOVE.4 in infilling missing data has not been found in the hydrological literature.

<u>MULTI-SITE MODELS</u>

These models consider the persistence and cyclic variation explicitly, permit infilling of gap at multiple sites and incorporate cross-river information transfer for infilling process. Among the two models belonging to this category, one model use multiple regression equation [Young et al. (1970)] and another model use multivariate regression [Kottegoda and Elgy (1977)] as infilling equation belong to this category.

The operational procedure of the model using multiple regression, [Young et. al. (1970)] consists of arranging normal standardized stream-flow data in matrix format such that the sites (represented by rows) be arranged top to bottom of the matrix in descending order of the bulk of available data [Fig. 2.1]. The procedure furthermore, consists of using a linear predictor relationship given by Equation [2.5] to infill the gap from right to left (on the row), top to bottom of the data matrix thus making optimal use of existing and or infilled data. For any site k, in case of existing endpoints i.e. $y_{k(T-1)}$ and $y_{k(T+1)}$, the infilled estimate of $y_{k(T)}$ need to be adjusted with respect to these existing points so that the infilled data would comply with the assumed underlying AR(1) process of the data series.

$$\hat{y}_{si} = \sum_{k=1}^{s} a_{sk}\, y_{k(i+1)} + \sum_{k=1}^{s-1} b_{si}\, y_{ki} + t_{si}(1-R^2_s)^{1/2} \quad ..[2.5]$$

where,

$\hat{y}_{si}$ = estimate of infilled data of s^{th} site at i^{th} period

$y_{k(i+1)}$ = existing or infilled data of k^{th} site at $(i+1)^{th}$ period

y_{ki} = data of k^{th} site at i^{th} period

b_{si}, a_{sk} = least square regression parameters

t_{si} = noise ~ N(0,1)

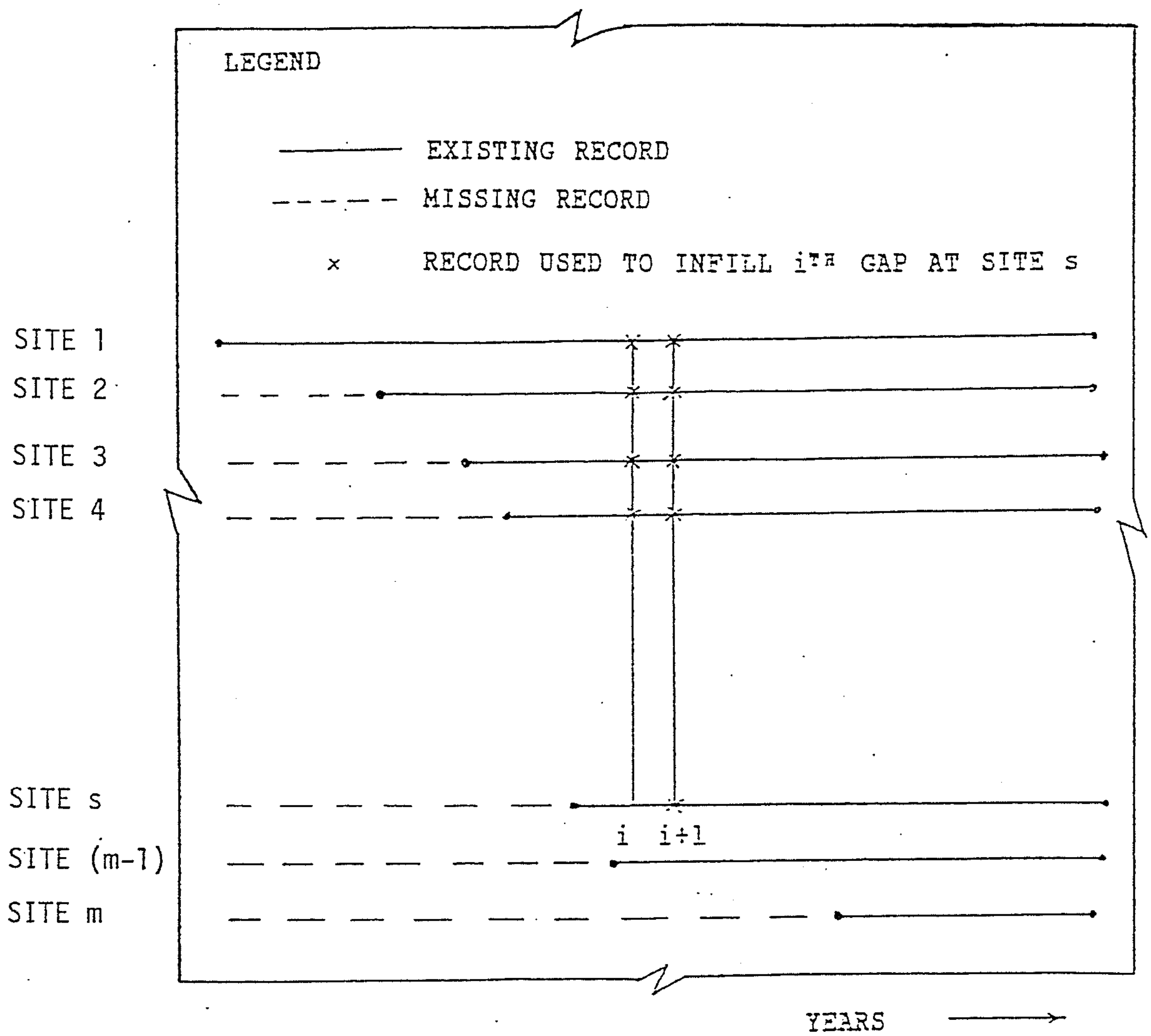


Fig. 2.1. Gap reconstruction at multiple sites

A multivariate multi-site AR(1) model with the following configuration was considered [Kottegoda and Elgy (1977)] for infilling multi-site monthly stream-flow data, there is

$$\mathbf{y}_{t+1} = \mathbf{A} \cdot \begin{vmatrix} \mathbf{y}_t \\ \mathbf{x}_{t+1} \\ \mathbf{x}_t \end{vmatrix} + \mathbf{B}\,\mathbf{\Omega}_{t+1} \quad \text{.......[2.5]}$$

where,

$\mathbf{y}_{t+1}$ = vector of missing data estimates of predicted variable at p sites on $(t+1)^{th}$ period

$\mathbf{y}_t$ = vector of observed values of predicted variable at p sites on t^{th} period

$\mathbf{x}_{t+1}$ = vector of observed values of predictor variable at n sites on $(t+1)^{th}$ period

$\mathbf{x}_t$ = vector of observed values of predictor variable at n sites on t^{th} period

$\mathbf{\Omega}_{t+1}$ = p- dimensional noise vector at $(t+1)^{th}$ period

A,B = parameter matrices

Both of these models adopt standardization to ensure second order stationarity, but this standardization procedure was found inadequate for ensuring stationarity in auto-correlation structure of the series [Bras and Iturbe (1975)].

De-seasonalization (i.e. standardization) procedure also reduces the zero-lag cross correlation coefficient between the rivers [Harmanchioglu and Yevjevich (1987)]

Due to the dependence among the rivers, use of more than one predictor river may not contribute to the marginal information gain and unnecessarily complicate the matrix operation procedure. Young et al. (1970) also considered such a situation and suggested to stop the incorporation of base river when multi-collinearity situation arises.

MODELS UTILIZING PHYSIOGRAPHIC FEATURES

These models permit infilling of gap at a single site and incorporate cross-river information transfer in conjunction with some physiographic features such as the drainage area ratio, regional statistics, distance between { the short data river and a base river} for infilling process. Some of the models pay consideration to the cyclic variation while others ignore this issue [Hirsh (1979,), Kottegoda and Elgy (1977)]. These models are suitable in principle for data augmentation in stations with very few data or no data at all rather than infilling purposes.

2.2 DISCUSSION

By comparing the existing models discussed so far with the proposed models (MULBS, SESTRNALL), the followings are observed:

The proposed models differ very much from the conventional data filling models in terms of the underlying concept. The proposed models are intended to reconstruct a shape feature of the time plot such as, a peak or a valley, in its integrated form, while the existing models can be considered to reconstruct the shape pixel by pixel.

The proposed models take full account of cyclic variation in that they consider each month as a distinct element of a seasonal segment.

The proposed models consider group relationship which can be concieved as the relationship among each of the six elements constituting the seasonal segments. For a particular type of seasonal segment, the average relationship is given by a (6x6) symmetric covariance matrix. **Considering such a relationship as an intra-seasonal persistence, is rather vague from the typical hydrological view point of persistence, hence, it is appropriate to consider such a relationship as intra-seasonal group relationship.**

Among the proposed models, the MULBS consider intra-seasonal group relationship but disregards inter-seasonal group relationship, while SESTRNALL model considers both intra-seasonal and inter-seasonal group relationship.

The proposed model are developed in order to investigate whether or not the seasonal group characteristic can be utilized for monthly flow data infilling by means of multivariate models, hence, these models do not claim their superiority over any of the existing models. With respect to the underlying concept, these models lack any kind of similarity with the existing models. The proposed models are empirical in nature and based upon some assumptions, such as, MULBS model assumes that there is a relationship between the simultaneously observed seasonal segments of a {the river with missing data, a base river} pair. It furthermore assumes that this relationship is consistent over time. On the other hand the SESTRNALL model assumes that there is a substantial degree of inter-seasonal dependence and consistence of the dependence [Panu (1978)]. These are some assumptions which need to be verified by a proper scheme derived in the next chapter.

In this thesis the concept of seasonal group characteristics is used to investigate the scope for finding samples on the basis of seasonal homogeneity criteria for the regression type of models. The proponent of the regression type models have expressed concern about the large dispersion in the sample when only one infilling equation is calibrated from all available data. The only remedy could have been the choice of 12 different equations for infilling data belonging to 12 different months, But this implies reduction of the sample size by a factor of 12, which may lead to unreliable estimates of parameters. These researchers suggested to make a compromise between choosing 12 equations for 12 different months or to make two to four seasonal equation. In this thesis, scope of the concept of seasonal group characteristics in extracting such samples is studied. The intent herein is to study the prospects of seasonal group characteristics in order to extract a sample consisting of the concurrent observations of {short data river and a base river} such that the concurrent observations in the sample do not show wide dispersion among each other (**homogeneity**) and are similar to the missing observation and the simultaneous observation in a base river (**similarity**). It is hereby hypothesised that for a missing observation, if an infilling equation is calibrated from a sample with homogeneity and similarity property, the estimate of the missing data would be more accurate than the estimate computed by an infilling equation calibrated form a sample consisting of all available data (**heterogenous sample**). Three sampling scenarios are considered, one of which use clustering concept. These sampling scenarios are given in Chapter three.

In the chapter three, concept and selection of such sampling criteria are discussed.

CHAPTER 3

DEVELOPMENT OF MULTIVARIATE INFILLING MODELS AND DERIVATION OF SAMPLING SCENARIOS FOR REGRESSION AND MOVE.4 MODELS

3.0 OUTLINE OF VARIOUS SUB-PROCEDURES

The proposed multivariate models assume that any section of the monthly flow data plot of over a period of one year can be segmented into two or more different types of seasonal segments corresponding to the yearly low and high flow group characteristics of the river. It is thereby necessary to recognize, analyze such groups in the data, determine the average number of different flow groups over a year period followed by a suitable **segmentation** process. The evidence of groups in the monthly flow data is discussed in section 3.1. Procedure to determine the average number of such groups is also discussed in this section. An imperial algorithm for seasonal segmentation is given in section 3.2.

Among the proposed models, MULBS assumes the presence of considerate level of consistency in the simultaneous occurrence of seasonal segments of certain level of severity in the {river with missing data, a base river} pair [i.e. when there is a 'very high' peak at a seasonal period T, a very high peak is expected concurrently in a base river and such concurrent occurrence of 'very high' peaks is consistent over time]. SESTRNALL model on the other hand assumes consistency in the seasonal transition. These model assumptions are verified by an empirical scheme that uses a combination of sub-clustering and entropy concepts. The sub-clustering concepts needs the hyper-space representation of seasonal segments which is discussed in section 3.3. The assessment of consistency is done in chapter 5.

It has been mentioned in the previous chapter that three sampling scenarios would be considered for the univariate models. One of these sampling scenarios uses derivation of **selected seasonal segments** by an empirical scheme. In this scheme, the seasonally segmented time series of { the river with missing data, a base river} pair are replaced by the sequences of class membership indices of the seasonal segments which are derived by sub-clustering procedure. Juxtaposition of the class-membership index sequences of {the river with missing data, a base river} pair, allows to extract a set of concurrently observed seasonal segments which are similar to a missing seasonal segment at a seasonal period T_k and the concurrently observed seasonal segment at a base river.

For the infilling of any element j of a missing seasonal segment at T_k, an regression type equation can be considered to be calibrated from the sample consisting of all the elements of the selected seasonal segments pertaining to a missing segment at a period T_k. A procedure for extraction of the selected seasonal segments and development of various sampling scenarios respectively are discussed in section 3.4 and section 3.5.

The development of general statistical basis of multivariate models is done in section 3.6, which is utilized in the development of the statistical configuration of MULBS model in section 3.7 and that of SESTRNALL MODEL in section 3.8.

3.1 ON RECOGNITION OF GROUPS IN THE FLOW DATA AND THE DETERMINATION OF AVERAGE NUMBER OF SUCH GROUPS PER YEAR

Seasonal group characteristics is discernable in various representation of data. Groups are either visible or conceivable in the : raw data, correlogram and time plot of the data.

Monthly stream-flow data series printed from the data bank of any standard hydrological agency e.g. USGS, can be considered as a data matrix consisting of rows and columns representing the year and the month respectively pertaining to the collected data. The data along the row shows seasonal characteristics such as relatively high or relatively low flow persisting consecutively over fixed number of months.

The group behaviour is visible in the correlogram of the stream-flow data. The auto-correlation coefficient can be considered as a measure of similarity between the data at a certain lag [Romesburg (1974)]. The correlogram of monthly stream-flow data displays periodic changes in similarity (in slope and in magnitude). The correlogram of monthly stream-flow data of a river with two seasons per year shows a peak at every even multiple of six months and a valley at every odd integer multiple of six months indicating twelve months periodicity. This indicates the presence of two different types of flows. Each type of flow lasts for six months subsequently followed by a complete reversal of characteristics [Panu (1978), Panu and Unny III(1980)].

Seasonal group behaviour of the stream-flow data is further highlighted in the time waveform plot (plot) of the stream-flow data [Fig. 1.2]. Time waveform of stream-flow is a shape representation of stream-flow in time continuum, characterized by the periodic recurrence of peaks and valleys within any year.

On one hand, seasonal group characteristics can be considered as a shape feature in terms of time wave-form representation, on the other hand, seasonal group characteristics can be considered as an attribute of exhibiting similar flows over some fixed months of any year. However different these descriptions may be, they describe one and the same feature of the physical system, namely, a set of consecutive months having high and low flows or relatively high, high, low and relatively low flow scenario depending upon the number of seasons per year. Due to the random variation of precipitation and other climatological factors, in the real world situation , seasonal distinction of the raw data is not very straight forward.

Based on the similarity criteria associated with the correlogram, the average number of seasons per year can be obtained from the correlogram of the data. After determination of average numbers of seasons per year, the association of months to the season or segmentation of the time wave form is done by an empirical algorithm explained in next section.

3.2 SEASONAL SEGMENTATION

An empirical algorithm, in accordance with the definition of season from clustering point of view, is proposed here { see example , Appendix A.2.1}.

Let the correlogram of monthly stream-flow series show **w** seasons per year and let, each season last for **m** months. According to this algorithm, the twelve stream-flow data corresponding to the 1st year of the data matrix are ranked in ascending order of magnitude, the months corresponding to the first m lowest flows are assigned to the seasonal group **k=1**, the months corresponding to the second m lowest flows are assigned to the group **k=2** and so on. Subsequently, the months corresponding to the m highest flows are assigned to the group **k=w**. This procedure is repeated for each year of the available data. For any group **k** , for any month **j**, the total number of assignment of j^{th} month to k^{th} group is counted as follows:

$$n_{k,j} = \sum_{i=1}^{N} z_{i,j,f} \quad \text{.....................[3.1]}$$

where,

$n_{k,j}$ = total number of times j^{th} month is assigned to group k

$z_{i,j,f}$ = value of counter f corresponding to assignment of j^{th} month on i^{th} year to any group

for computation of the assignment of jth month to kth group:

$z_{i,j,f} = 0$ when $f \neq k$
$z_{i,j,f} = 1$ when $f = k$

w = the total number of seasons

N = total number of years of data

The j^{th} month can be conceptually considered to be assigned to a seasonal group k with whom it has been assigned maximum number of times for that month. The assignment of a month to group k is given below:

$$j \in k \quad : \quad \max\{\, n_{k,j} \,\} \;,\; k = 1 \ldots\ldots w \qquad \ldots\ldots\ldots \; [3.2]$$

For each of the twelve months, the total number of assignment to each of the seasonal groups (k=1....w) are counted. For any season, a continuous chain of m months is expected to show maximum number of assignments to that particular season. One can thus infer which chain of m months should be assigned to which season. Prior knowledge about the drainage basin in terms of the time of occurrence of the peak and low flow can be considered as additional aid pertinent to the seasonal segmentation process.

The correlogram and the segmentation algorithm enables the division of continuous time wave-form into **seasonal segment.** For a particular season, the seasonal segments can be grouped or clustered by imposing some criteria. They can be further sub-clustered by imposing some other finer criteria.

3.3 HYPER-SPACE REPRESENTATION AND CLUSTERING OF THE SEASONAL SEGMENTS

A seasonal segment (m months long) can be considered as a *m*-dimensional object and it can be represented by a *m*-dimensional pattern vector **X** (Equation [3.3]). Such a seasonal segment can be considered as a point in *m*-dimensional hyperspace.

$\mathbf{x}_{ik} = [\ \mathbf{x}_{1k}\ ,\ \mathbf{x}_{2k}\ ,\ .....\mathbf{x}_{mk}\ ]^T$ ....[3.3]

where,

$\mathbf{x}_{ik}$ = pattern vector corresponding to the k^{th} seasonal segment on i^{th} year

$\mathbf{x}_{lk}$ = the stream-flow of k^{th} seasonal segment on l^{th} month , $l= 1....m$ [also denoted as l^{th} element]

In hyper-space representation, in case of a river with **w** seasons per year , the points corresponding to all the seasonal segments would constitute **w** different **clusters** $\{C_k\}$ (k=1...w) in a way that the points representing the seasonal segments belonging to a common season k , would lie in a common cluster C_k [Fig. 3.1].

Any seasonal cluster C_k can be further sub-divided into q **sub-clusters** depending upon the degree of refinement in similarity criteria imposed on the membership of a common sub-cluster [Fig.3.2]. For a given seasonal cluster, by means of sub-clustering, one can thus group seasonal segments by imposing more rigorous similarity criteria, i.e., sub-clustering of wet seasonal segments could mean screening of very high peaks from the comparatively less severe ones.

Many kinds of clustering algorithms and diversified nature of similarity or dissimilarity **metric** are available in standard

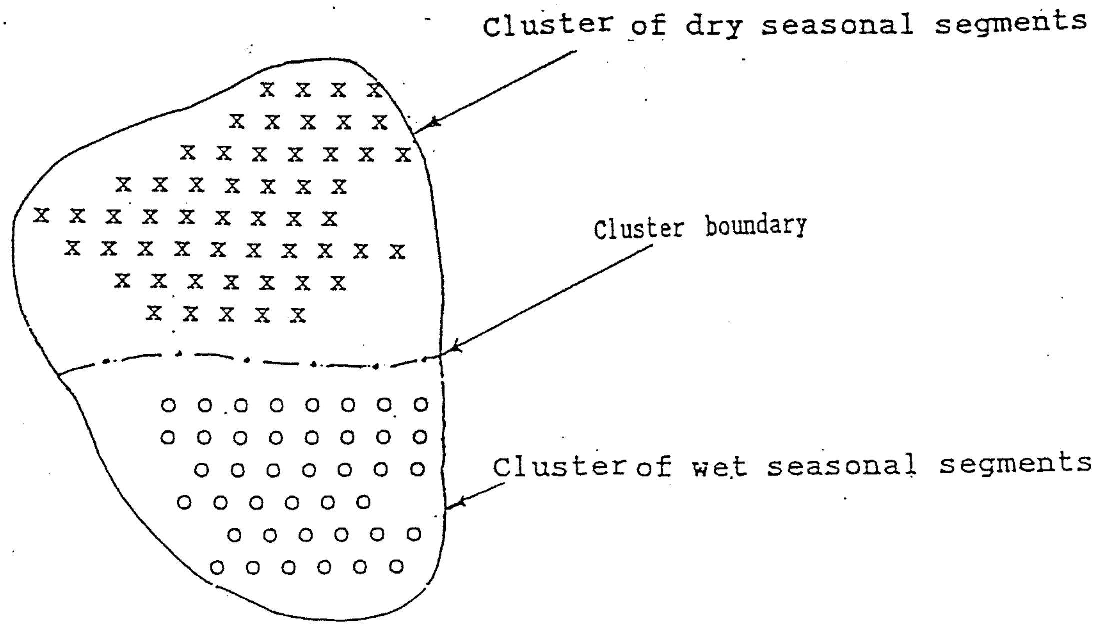


Fig. 3.1: Hyperspace representation of seasonal segments
[Asssuming two seasons per year]

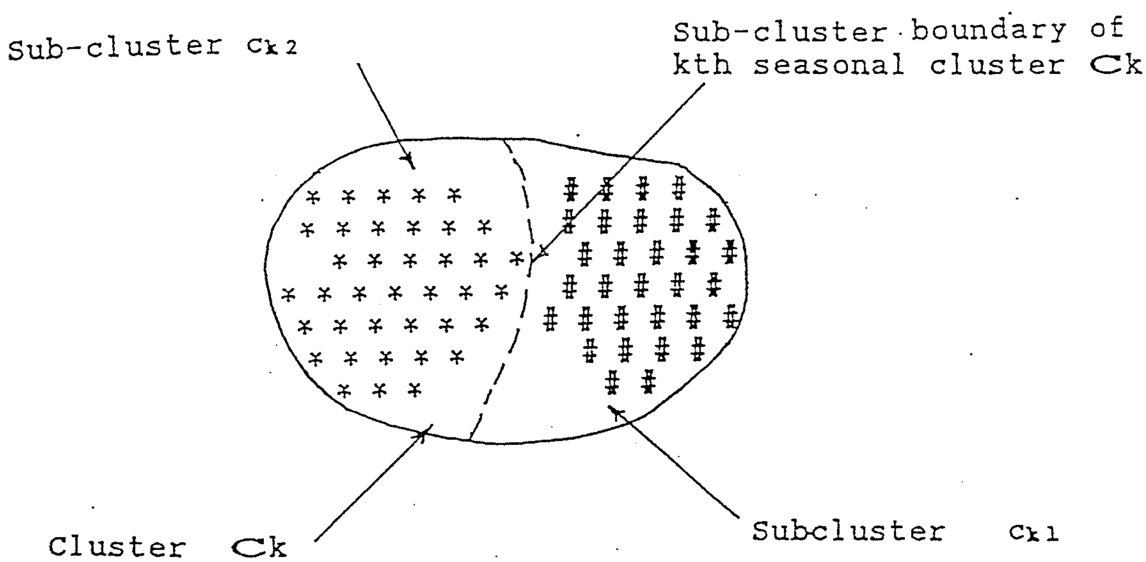


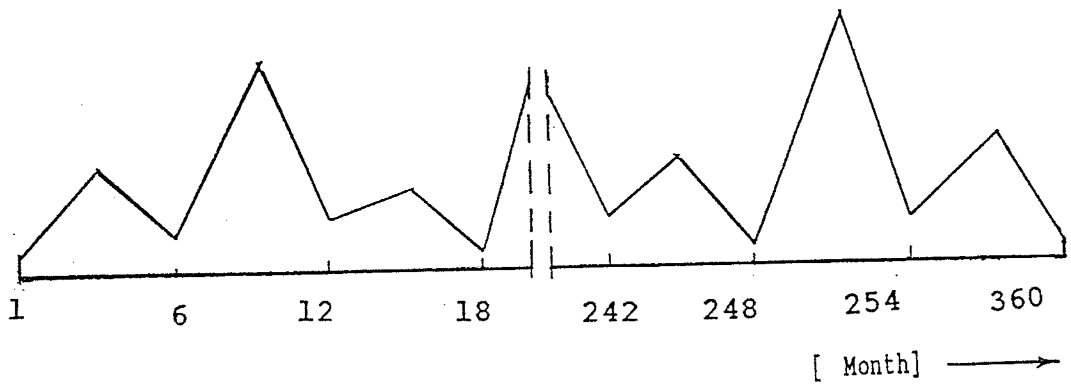
Fig. 3.2: Sub- clustering of kth seasonal cluster
[Assuming two sub-clusters per seasonal cluster]

cluster analysis and pattern recognition text books. In this thesis, a ***k* -mean algorithm** using **Euclidean distance** as **dissimilarity metric** is considered for the sub-clustering of seasonal segment. Additional detail of k- means algorithm is given in Appendix A.1.1 .

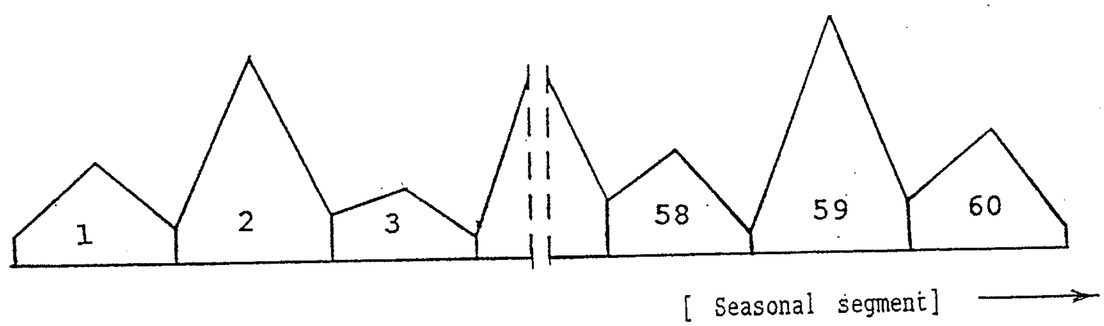
3.4. EXTRACTION OF SELECTED SEASONAL SEGMENTS

One can replace the seasonally segmented time wave-form by the sequence of class-membership indices obtained by the sub-clustering process. Such a **sequence of seasonal class-membership indices** represents the time plot of the seasonal status or degree of severity of the seasonal segments [Fig. 3.3].

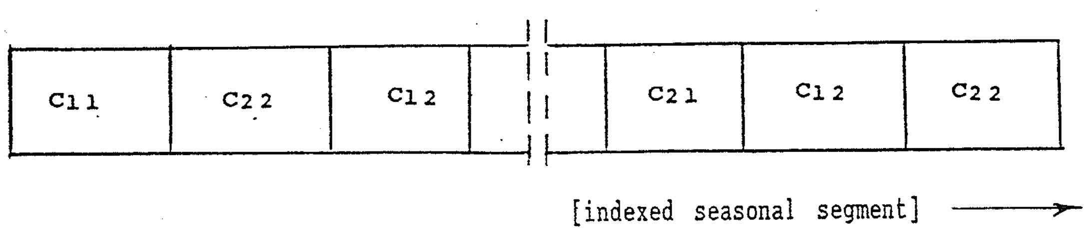
Let Y and X respectively denote the sequences of seasonal class-membership indices of the river with short data and base river [Fig.3.4]. On basis of the similarity of the climate, the rivers of same or nearby basins can be considered to have same number of seasons per year and the same seasonal segmentation pattern, i.e., equal number of seasons per year and association of the same months to the same season). In a pair of nearby rivers, consistency in the simultaneous occurrence of seasonal segments of certain degree of severity can be expected. This consistency can be assessed by computing the probability of occurrence of a seasonal segment of type c_{Ypk} in series Y conditioned on the simultaneous



(a) Time wave-form of stream-flow data [30 years of record]



(b) Sequence of seasonally segmented time wave-form



(c) Replacement of seasonally segmented time wave-form by class-membership index sequence

Fig. 3.3: (a) Time wave-form , (b) seasonally segmented time wave-form and (c) class-membership index sequence. [Assuming 30 years of monthly data, two seasons per year and two sub-clusters per seasonal cluster]

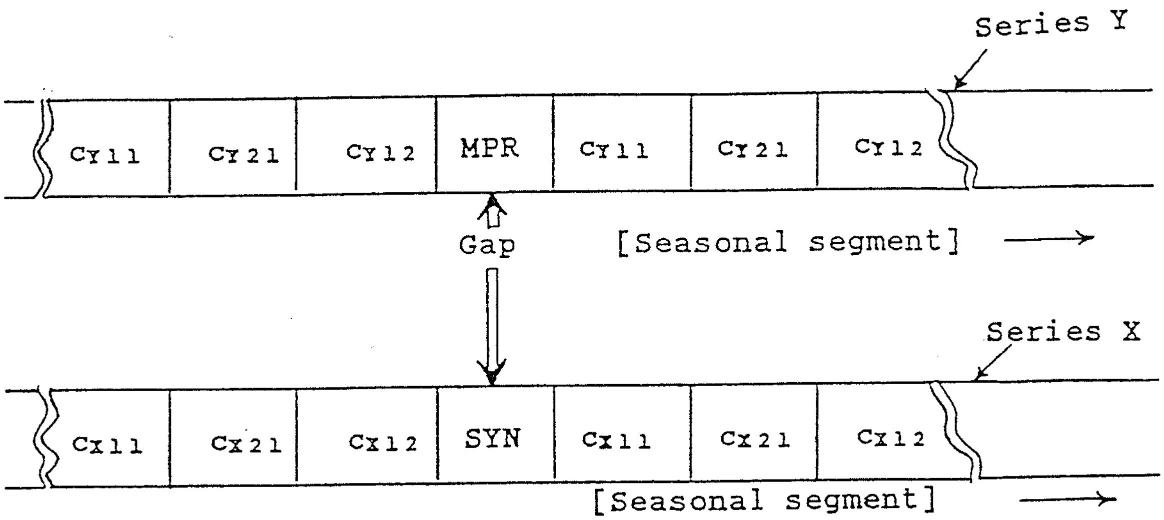


Fig. 3.4: Juxtaposition of sequences of seasonal class-membership indices of X and Y

occurrence of a seasonal segment of type c_{Xqk} in series X as given below:

$$P(c_{Ypk} \mid c_{Xqk}) = \frac{P(c_{Ypk}, c_{Xqk})}{P(c_{Xqk})} \qquad \ldots[3.4]$$

$$= \frac{(N_{XYok} / N_{XYpk})}{(N_{Xok} / N_{Xpk})}$$

where,

p = a sub-cluster index of series Y corresponding to season k [(p=1....k_l),k_l is the total number of sub-clusters of k^{th} seasonal cluster]

q = a sub-cluster index of series X corresponding to season k [(p=1....k_f),k_f is the total number of sub-clusters of k^{th} seasonal cluster]

N_{XYok} = number of times the seasonal segment of type c_{Ypk} in series Y is simultaneously observed with seasonal segment of type c_{Xqk} in series X

N_{XYpk} = number of times the seasonal segment of types c_{Ypk} and c_{Xqk} could possibly simultaneously occur in X and Y

N_{Xok} = number of times the seasonal segment of type c_{Xqk} is observed in series X

N_{Xpk} = number of times the seasonal segment of type c_{Xqk} could possibly occur in series X

In case, when the series X and Y have same number of seasons per year, same seasonal association of the months and same number of sub-clusters per season, N_{XYpk} and N_{Xpk} are equal. Thus, Equation [3.4] reduces to Equation [3.5].

$$P(c_{Ypk} \mid c_{Xqk}) = (N_{XYok} / N_{Xok}) \ldots\ldots\ldots\ldots [3.5]$$

Let there be a missing seasonal segment in season k at T^{th} seasonal period in series Y and let **SYN** represent the class-membership index of corresponding simultaneously observed seasonal segment in series X [Fig. 3.4]. One can search the conditional probability table and select the most probable class-membership index c_{Ypk} of the candidate missing segment in Y by satisfying following constraint:

$$\mathrm{Max}\{ P(c_{Ypk} \mid SYN)\} \ldots\ldots\ldots\ldots [3.6]$$

The sub-cluster p satisfying Equation 3.6 is considered as the most probable class- membership index of the missing seasonal segment and is denoted by **MPR.** One can now search for the years in which seasonal segments of type **SYN** and seasonal segments of type **MPR** are simultaneously observed in series X and Y respectively. Such a set of simultaneously observed seasonal segments is referred as **selected seasonal segments** { example given in Appendix A.2.5 }.

3.5 DEVELOPMENT OF SAMPLING SCENARIOS FOR REGRESSION AND MOVE.4 MODELS

Conventionally, the regression and MOVE.4 models compute parameters of the infilling equations from the sample consisting of all available data, thus, disregard the heterogeneity of the data belonging to different seasons. Three different sampling scenarios are developed for each of the regression and MOVE.4 models. General configuration of the regression and MOVE models are discussed in Chapter two and the detail description of the least square regression and MOVE.4 model considered for various sampling scenarios are presented in Appendix A.1.4 and Appendix A.1.3. The three different sampling scenarios are:

1.) Indiscriminately chosen sample

One infilling equation is calibrated from sample consisting of all available data. The regression and MOVE.4 models under such sampling scenario are denoted by REG and AMOVE .

2.) Seasonal sampling

Separate infilling equations are calibrated for infilling of missing data belonging to separate seasons. For infilling of the missing data belonging to a particular season **k**, the sample is chosen from the elements of all simultaneously observed $\mathbf{k}^{th}$ seasonal segments of { river with missing data, base river pair}. The regression and MOVE.4 models under such sampling scenario are denoted by SREG and SMOVE .

3.) Selected seasonal sampling

Separate infilling equations are calibrated for infilling of missing data belonging to separate seasons. For infilling of the missing data belonging to a particular missing seasonal segment, the sample is chosen from the elements of selected seasonal segments corresponding to the gap segment. The regression and MOVE.4 models under such a sampling scenario are denoted by SSREG and SSMOVE .

These sampling scenarios are more elaborately explained in an example in Appendix A.2.6.

3.6 DEVELOPMENT OF STATISTICAL BASIS OF MULTIVARIATE MODELS

Two multivariate infilling models are considered. Each of them computes the parameters of the conditional distribution of the missing seasonal segment. They differ from each other by the nature of the conditioning variable. One model conditions on the observed seasonal segment in the base river while the other model conditions on the observed or reconstructed seasonal segment preceding the gap segment of the river with missing data. When the predictor and predicted seasonal segments jointly follow a multivariate normal distribution, then the reconstructed seasonal segment is considered to have a multivariate normal distribution. The mean vector and covariance matrix of the reconstructed seasonal segment can be considered as sufficient statistics to describe the configuration of the distribution [Johnson and Wichern (1988)]. Each of the multivariate models considered here has the same following statistical basis .

Let each of $\mathbf{X}_1$ and $\mathbf{X}_2$ represents a **p** variate random vector and let $\mathbf{X} = [\ \mathbf{X}_1\ ,\ \mathbf{X}_2\ ]^T$ be distributed as $N_{2p}\ (\boldsymbol{\mu},\ \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right] \quad , \quad \boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_1 \\ \hline \boldsymbol{\mu}_2 \end{array}\right] \quad , \text{ and } \quad |\boldsymbol{\Sigma}_{22}| > 0$$

Then, the conditional distribution of $\mathbf{X}_1$, given $\mathbf{X}_2 = \mathbf{x}_2$, is multivariate normal with

$$\textbf{Mean} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\, \boldsymbol{\Sigma}_{22}^{-1} (\ \mathbf{x}_2 - \boldsymbol{\mu}_2\ )$$

and .....[3.7]

$$\textbf{Covariance} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\, \boldsymbol{\Sigma}_{22}^{-1}\, \boldsymbol{\Sigma}_{21}$$

where,

Σ_{11} = (p*p) matrix containing the elements of covariance matrix of X_1

Σ_{22} = (p*p) matrix containing the elements of covariance matrix of X_2

$\Sigma_{12} = \Sigma_{21}^{T}$ = (p*p) matrix containing the cross covariance terms between the elements of X_1 and X_2

μ_1 = (p*1) mean vector of X_1

μ_2 = (p*1) mean vector of X_2

The seasonal gap can be infilled either by the estimate of corresponding conditional mean or by a randomly generated seasonal segment by using a multivariate random number generator specified with the conditional configuration as generation basis. In this thesis, estimates of conditional mean are considered as the estimate of missing seasonal segment.

3.7 MULTIVARIATE INFILLING MODEL CONDITIONING ON OBSERVED SEASONAL SEGMENT IN BASE RIVER [MULBS]

The **MULBS** model reconstructs the missing seasonal shape feature, such as a peak or a valley of the time wave form of the river with missing data by conditioning on the corresponding shape feature observed in the base river. This model computes the mean and covariance of a missing seasonal segment of the river with missing data at any seasonal period by conditioning on the observed seasonal segment in base river. Let **A** and **B** respectively represent the seasonally segmented series of the river with missing data and the base river. Let these rivers have same number of seasons per year and same association of months to seasons. Let the vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ represent seasonal segments of **A** and **B** concurrently observed on **k**th season on $\mathbf{i}^{th}$ seasonal period. The vector $\mathbf{X} = [\ \mathbf{X}_1\ ,\ \mathbf{X}_2\ ]^T$ represents the **simultaneously observed seasonal segments.** Let there be a gap on **k**th season at **T**th seasonal period of the seasonally segmented time-wave form of series **A**

(Fig.3.5). Because the vector $\mathbf{X_2} = \mathbf{x_2}$ is observed in the series **B** , hence, the configuration of the distribution of missing seasonal segment conditioned on observed $\mathbf{x_2}$ can be obtained by Equation 3.7. For any seasonal gap on $\mathbf{k}^{th}$ season, all the concurrent observations of the random vectors $\mathbf{X_1}$ and $\mathbf{X_2}$ on $\mathbf{k}^{th}$ season are considered as predicted and predictor vectors. Joint normality of $\mathbf{X_1}$ and $\mathbf{X_2}$ is the requisite precondition to be met for this model. Simultaneously observed seasonal segments and the sampling scenario for the MULBS model are explained in detail in Appendices A.2.3 and A.2.7.

3.8. MULTIVARIATE INFILLING MODEL CONDITIONING ON OBSERVED OR RECONSTRUCTED SEASONAL SEGMENT PRECEDING THE GAP OF THE RIVER WITH MISSING DATA [SESTRNALL]

The **SESTRNALL** model reconstructs the missing seasonal shape feature of the time wave-form by conditioning on the reconstructed or observed preceding shape feature of the river with missing data. Let $\mathbf{X_1}$ and $\mathbf{X_2}$ respectively denote the $\mathbf{k}^{th}$ seasonal segment on $\mathbf{i}^{th}$ seasonal period and $\mathbf{(k-1)}^{th}$ seasonal segment on $\mathbf{(i-1)}^{th}$ seasonal period. Let the vector $\mathbf{X=[X1,X2]}^T$ denote the **transitional seasonal segment** corresponding to the transition of $\mathbf{(k-1)}^{th}$ season to $\mathbf{k}^{th}$ season. Let there be a seasonal gap on $\mathbf{T}^{th}$ seasonal period corresponding to $\mathbf{k}^{th}$ season in the seasonally segmented time series **A** of the river with

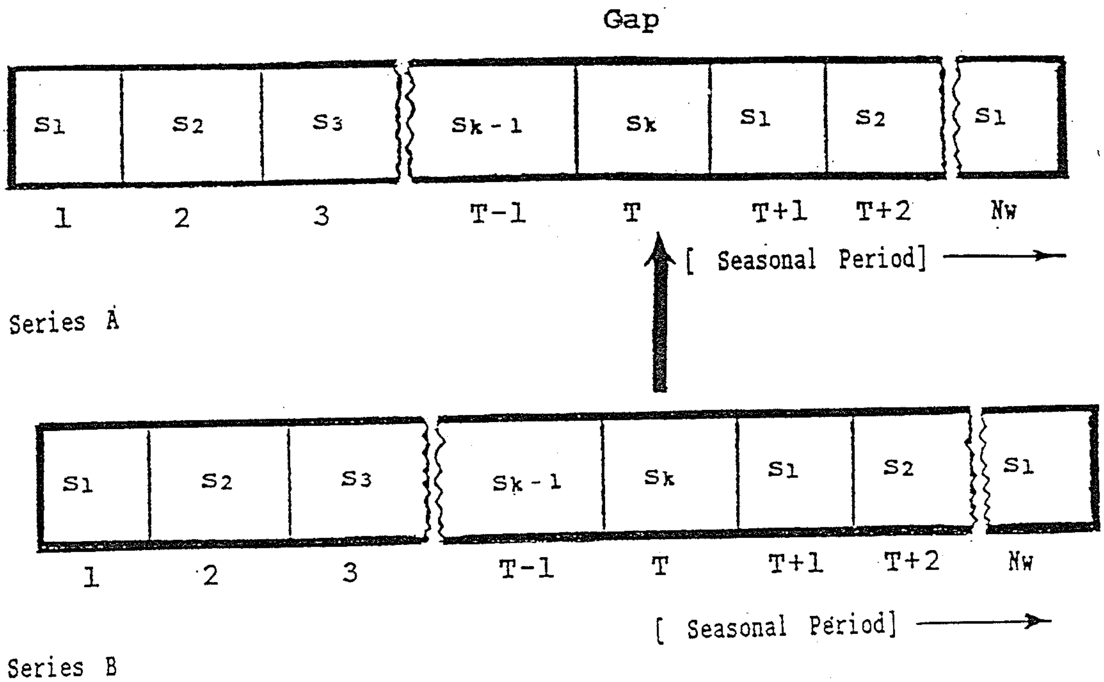


Fig. 3.5:
Seasonally segmented time wave-form of stream-flow series A and B

LEGEND

Direction of information transfer

N = Total number of years of data
w = Total number of seasons per year
s_i = i^{th} seasonal segment on any year [$i \le w$]

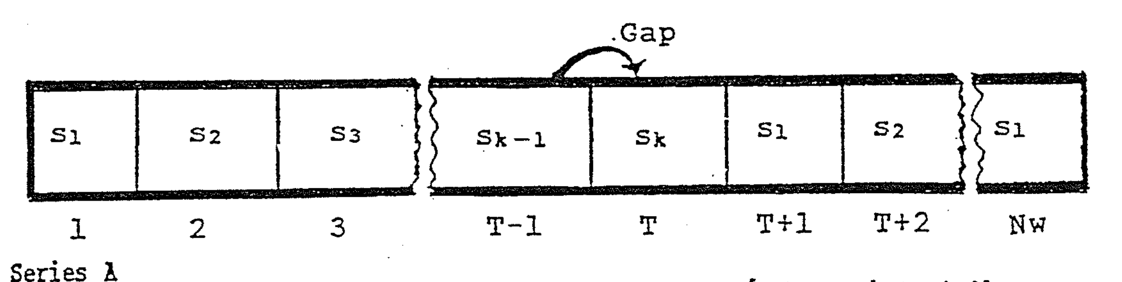


Fig. 3.6:
Seasonally segmented time wave-form of stream-flow series A

missing data (Fig.3.6). Let the seasonal segment at $(T-1)^{th}$ seasonal period corresponding to $(k-1)^{th}$ season be either observed or reconstructed , hence known. Corresponding to a k^{th} seasonal gap, all the observed seasonal segments of k^{th} and $(k-1)^{th}$ seasonal segments are denoted as predicted and predictor vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively. Since the value of $\mathbf{X}_2 = \mathbf{x}_2$ is known, the parameters of the distribution of the missing seasonal segment $\mathbf{X}_1$ by conditioning on the observed or reconstructed vector corresponding to preceding seasonal segment $\mathbf{X}_2$ can be derived from Equation [3.7] . The model parameters of Equation [3.7] are estimated from the existing observations corresponding to **(k-1)**th and **k**th transitional seasonal segments. Transitional seasonal segments and the sampling scenario of SESTRNALL model are given in Appendix A.2.4 and A.2.8.

CHAPTER 4

APPLICATION OF THE MODELS

4.0 OUTLINE OF APPLICATION PROCEDURE

Application of the models developed in chapter three is carried out by reconstructing a year period of monthly stream-flow data of three different rivers in three different watersheds. In case of application of the models incorporating base river information, one river is considered as a river with missing data and whose missing data are infilled by two to three different base rivers. Such a choice of considering one missing data river and multiple base rivers allows to do the followings:

- For a particular base river, for each of the regression and MOVE.4 models, comparison of infilling quality under various sampling scenarios.
- For a particular model[MULBS, regression and MOVE.4 models under various sampling scenarios] , comparison of infilling by various base rivers.

One set of closely located rivers is chosen which is referred as **river cluster** . One member of the river cluster is treated as the river with missing data $\mathbf{R_A}$ and the remaining members are treated as a **base river set** $[\{R_{Bj}\}, j=1..n_B]$ with $\mathbf{n_B}$ being the total number of base rivers. **SESTRNALL** model is applied to information are applied to the pairs of the rivers with missing data and base rivers $[\{R_A, R_{Bj}\}, j=1...n_B]$.

At least three nearby located rivers are needed to be considered to form a river cluster so that one can be treated as river with missing data and the other two rivers can be treated as base rivers. Continuous and reasonably longer period ($\geq$ 30 years) of natural data is desired for each member of the corresponding river cluster.

Three clusters of canadian rivers were initially considered. One of them is rejected in view of meager data in one case [rivers of Dease basin], and another because of urbanization concentration [rivers of Sydenheim basin]. The third river cluster, consisting of the rivers of Lillooet basin, was chosen in spite of the inadequate data (23 years). This river cluster is referred as cluster **LILL** (named after the abbreviation of Lillooet basin).

U.S. Geological survey published data on West Virginia basins shows that the rivers of this area possess the characteristics for considering as members of the river cluster. Hirsh (1979,1982) chose seven rivers from this zone for the comparative study of some of the existing data infilling models. The same seven rivers are considered here, but these rivers are considered in two different clusters due their relative locations. Each of the members of these river clusters has very long period of natural data. These two clusters are referred as **UPB** cluster and **LB** cluster in accordance with their geographical location , namely the members of the UPB cluster are geographically located relatively above the members of the LB cluster.

All the models under study (except REG and AMOVE), need seasonally segmented time series. In order to enable the computational operation, it is therefore convenient to restructure the data matrix so that it's beginning coincides with the beginning month of any of the seasons determined by the segmentation process {Section 3.1}.

In order to facilitate comparison among commensurate models as well as to facilitate comparison of the quality of infilling among various base rivers , the data matrix comprising of same period of data as well as gap over same period is considered. So the same restructured or **slid data matrix** is used for all the models.

The procedure of seasonal segmentation and sliding of data matrix is explained in Appendix A.2.1. The essential features, such as, seasonal segment, simultaneously observed seasonal segment, transitional seasonal segment, selected seasonal segment are described in Chapter three and explained in detail in Appendix A.2.

For the REG and AMOVE models, infilling Equations A.1.4 and A.1.3 are calibrated from the available sample without imposing any selection criteria. For SREG and SMOVE models, infilling Equations A.1.4 and A.1.3 are calibrated from the seasonal samples. For SSREG and SSMOVE models, infilling Equations A.1.4 and A.1.3 are calibrated from the seasonal selected sample. The sampling scenarios are explained in Appendix A.2.6.

For all the three scenarios of the regression models, normality of residual is ensured. At first the respective model is applied to the natural data, and normality of error is tested by both Chi-square goodness of fit and normal plot. In case of non-normality, the data is transformed to satisfy the normal error criteria. In all the three scenarios of the MOVE models, log transformed data is considered [Vogel-Stedinger (1985), Hirsh (1982)].

For the MULBS and SESTRNALL models, the selection of samples is explained in Appendix A.2.7 and in Appendix A.2.8. For each of these multivariate models, normality of the joint variate **X** [equation 3.7] is an essential prerequisite. For the SESTRNALL model, the multivariate normality of the transitional seasonal segments and for the MULBS model, the multivariate normality of the simultaneously observed seasonal segments of {base river and the river with missing data} pair are tested by the procedure explained below.

TEST CRITERIA

For a 2m dimensional normal variate, the Mahalanobis distance (M.D) follows a chi-square distribution. The testing of the distribution of the observed M.D. against a Chi-square distribution is considered as a tool for testing of the multivariate normality of the X variates.

Let each of the seasons comprise of same number of months **m** , then **X** can be considered as a **2m** - dimensional vector for each of the multivariate models. In case of **2m**- dimensional normality of **X** , the Mahalanobis distance (M.D.) of **X** from the mean vector , namely, $[(\mathbf{X}-\mu)^T \Sigma^{-1} (\mathbf{X}-\mu)]$ would follow a Chi-square distribution with **2m** degrees of freedom [Johnson and Wichern (1988)].

OPERATIONAL PROCEDURE:

- The M.D. for the observed X are computed and ranked.
- For each rank j
  - The corresponding exceeding probability α_j is computed by : $\alpha_j = [1 -(j-0.5)/n]$; [n= number of observed X variate].
  - from the of Chi-square table, the theoretical

Chi- square value (TV_j) corresponding to D.F.=2m, is read at α_j

- The observed $M.D._j$'s and the corresponding theoretical Chi- square values are plotted at the same α_j

The curve joining these theoretical values is denoted by **theoretical curve.** The curve joining the observed values is denoted by **observed curve.** The observed curve is visually compared against the theoretical curve. The rationale of such a procedure lies in the fact that it provides a tool for observing the deviation of the observed M.D. from the theoretical Chi-square values at any exceeding probability level α. The test of goodness of fit could have been a choice, but it was not selected because of the biasness of the test statistics to the numbers cells [this biasness is magnified in multiple dimension]. On the basis of the deviation between the observed and the theoretical curve, subjective judgement is applied in order to accept or reject the fact that the observed curve is approximately fitting the theoretical curve. In case of such a **satisfactory fit**, the multivariate normality in the raw data is assumed and the raw data is used in the Equation [3.7].

TRANSFORMATION

In case of poor fit of the observed curve, non-normality of X variate is suspected, which is handled by the following procedure:

The marginal normality of each of the elements of the joint variate **X** is tested. If an element x_j is found non-normal, the element is transformed by Bos-Cox transformation. This is done for all non-normal elements. New Mahalanobis distances are computed and the curve fitting exercise is repeated [Fig.4.0] until a satisfactory fit is obtained and the transformation matrix [consisting of the { shift and power} of the 2m elements] which provides such a fit is chosen to transform the non-normal data to normal data. Equation [3.7] now can be applied to the transformed data under the assumption that the normalization transformation has ensured normality in the data. In case of no satisfactory fit, the transformation matrix which provides the minimum deviation between the observed and the theoretical curve, is chosen for the normalization transformation of the data.

The next section deals with the application of the proposed multivariate models as well as the regression and MOVE.4 models under various sampling scenarios for infilling one years of the stream-flow data of the missing data river belonging to each of the three river clusters.

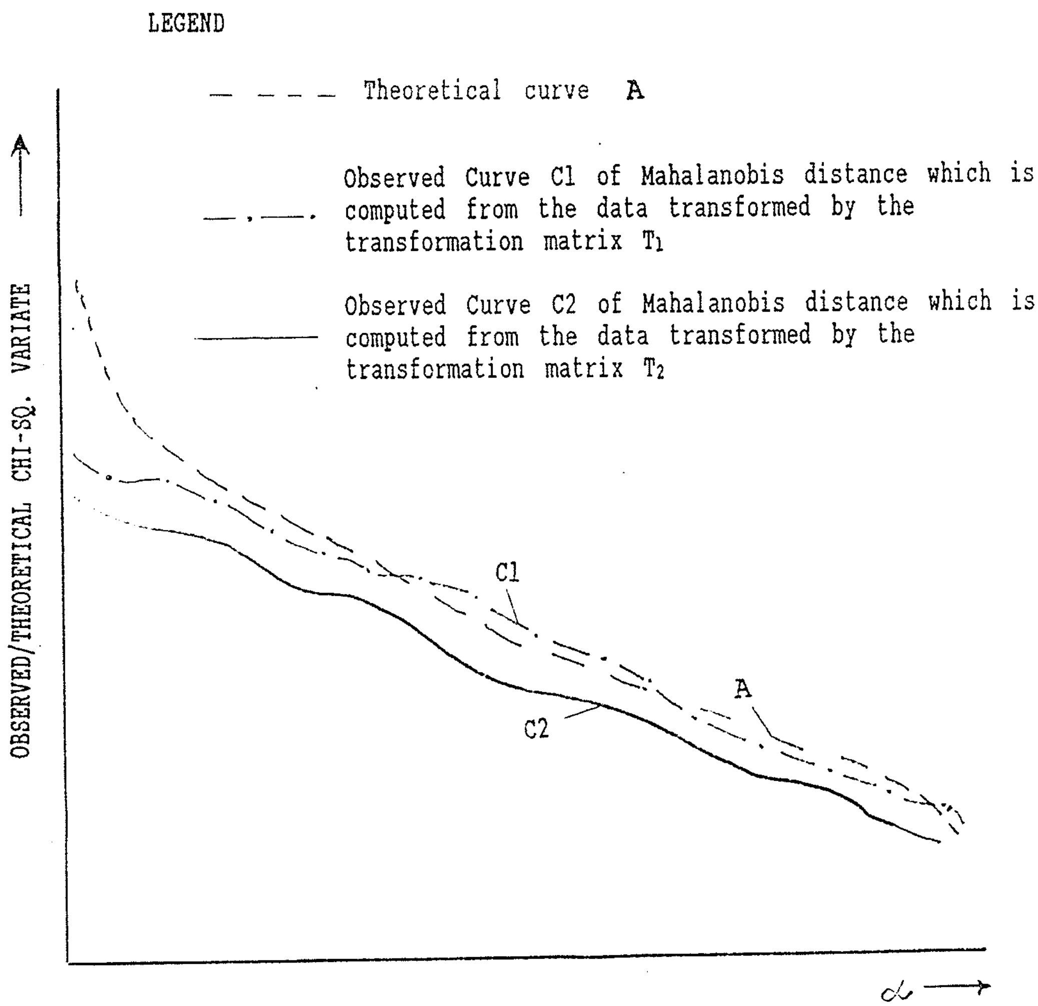


Fig. 4.0: Plots of ranked Mahalanobis distance computed from the data which is transformed by two different transformation Matrices

* Because the curve C1 has less deviation from A than the curve C2, therefore Transformation matrix T1 is a better candidate than the transformation matrix T2 for ensuring multivariate normality

4.1 APPLICATION OF VARIOUS MODELS IN INFILLING MISSING DATA OF RIVERS BELONGING TO UPB CLUSTER

The basic information pertaining to the members of UPB cluster and the location of the members are presented in Table 4.1.1 and in Fig.4.1.1. A plot of a five year period of simultaneously observed data of the member rivers is presented in Fig.4.1.2. At the initial phase, for each of the member rivers, 31 years of data over the period of [Oct. 1958-Sep.1988) are considered. Correlogram analysis [Figs.4.1.3.(a,b,c,d)] shows two seasons per year and the segmentation procedure assigns the periods [June - Nov.] and [Dec.- May] respectively to dry and wet seasons [Table 4.1.2]. The data matrix is slid to begin with the beginning of the dry season. The slid data matrix contains 31 years of data over the period [June 1958 - May 1989]. One year of monthly data of Craig Ck. over the period [June 1986 - May1987] is assumed to be missing. The SESTRNALL model is applied to Craig Ck. and the rest of the models are applied to {Craig Ck., Dunlap Ck.}, {Craig Ck., Johns Ck.} and {Craig Ck., Cowpasture R.} pairs. Multivariate normality is tested for both MULBS and SESTRNALL models. In case of unsatisfactory fit, power transformation of element of the joint variate **X** are done . The final Chi-square plot of the Mahalanobis distances of the observations of **X** variate are presented in Figs. 4.1.4.(1, 2, 3, 4, 5 and 6) and Figs.4.1.5.(1 and 2). The reconstructed data of Craig Ck. corresponding to the different information sources { Johns Ck., Dunlap Ck., Cowpasture R., Craig Ck. } pertinent to data reconstruction are presented in Table 4.1.3.(1, 2, 3 and 4)

Table 4.1.1: Information pertinent to the members of UPB cluster

River*	USGS Station Number	Comment	Latitude	Longitude	Area mi^2	Period of data used in analysis	Missing Period
Craig Creek	02018000	River with missing data	37-39-59	79-54-42	329	June 1958-May 1989 (31 Years)	June 1986-May 1987
Cowpasture River	02016000	Base River	37-47-30	79-45-35	461	June 1958-May 1989 (31 Years)	
Dunlap Creek	02013000	Base River	37-48-10	80-02-50	164	June 1958-May 1989 (31 Years)	
Johns Creek	02017500	Base River	37-30-22	80-06-25	104	June 1958-May 1989 (31 Years)	

* All rivers are tributaries of the Upper James R.

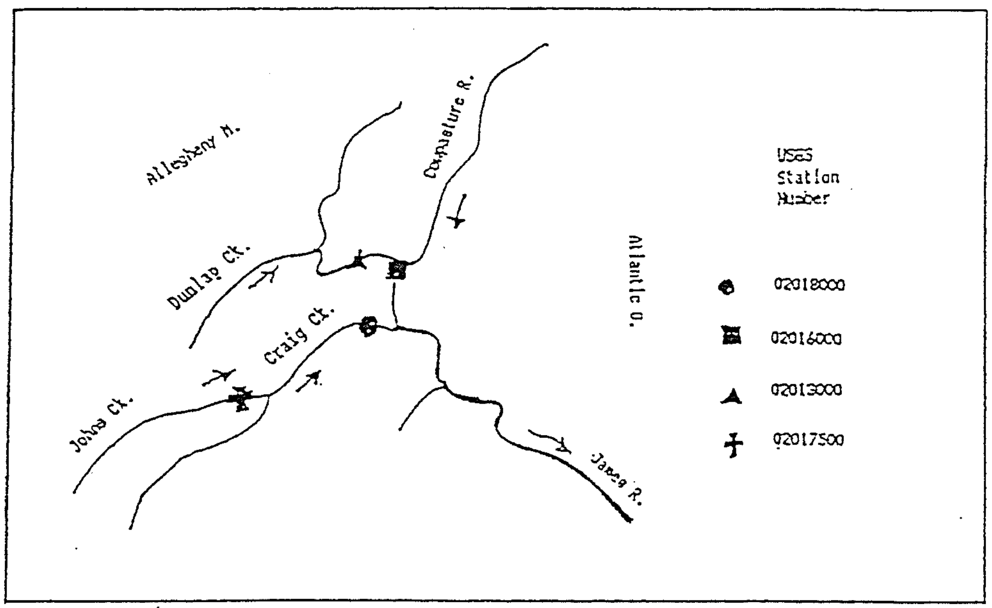


Fig.4.1.1: The location of the members of UPB cluster

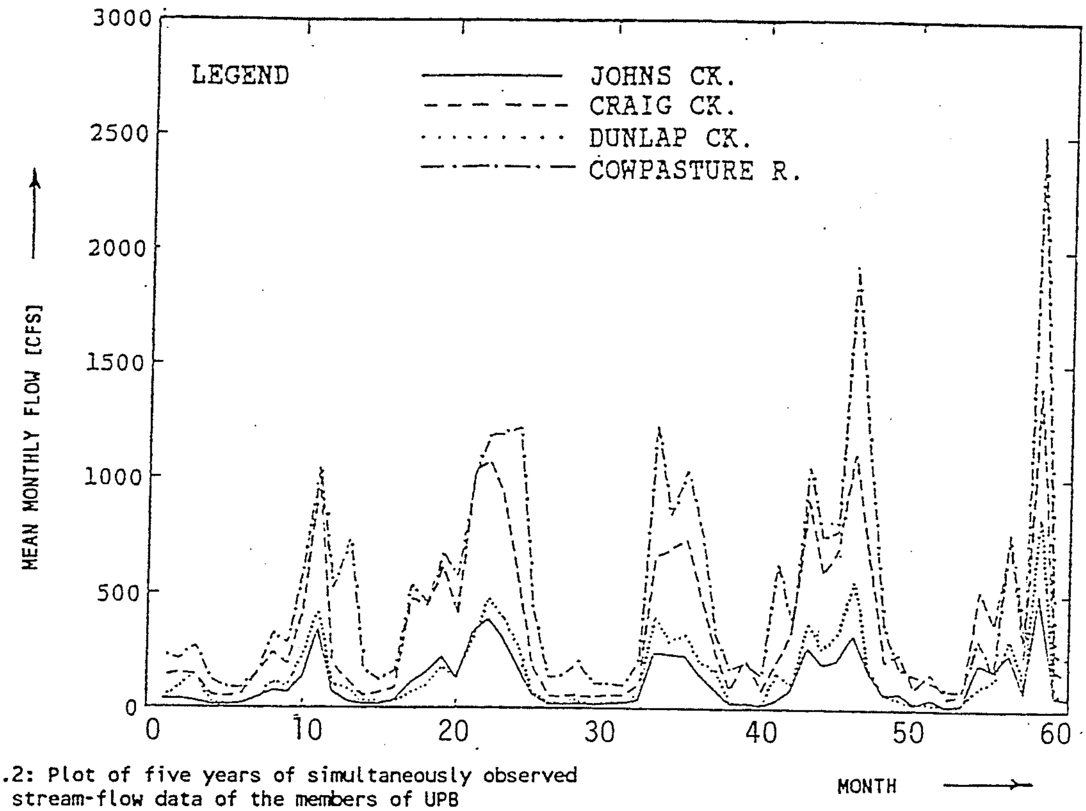


Fig.4.1.2: Plot of five years of simultaneously observed monthly stream-flow data of the members of UPB cluster [June 1958-May 1963]

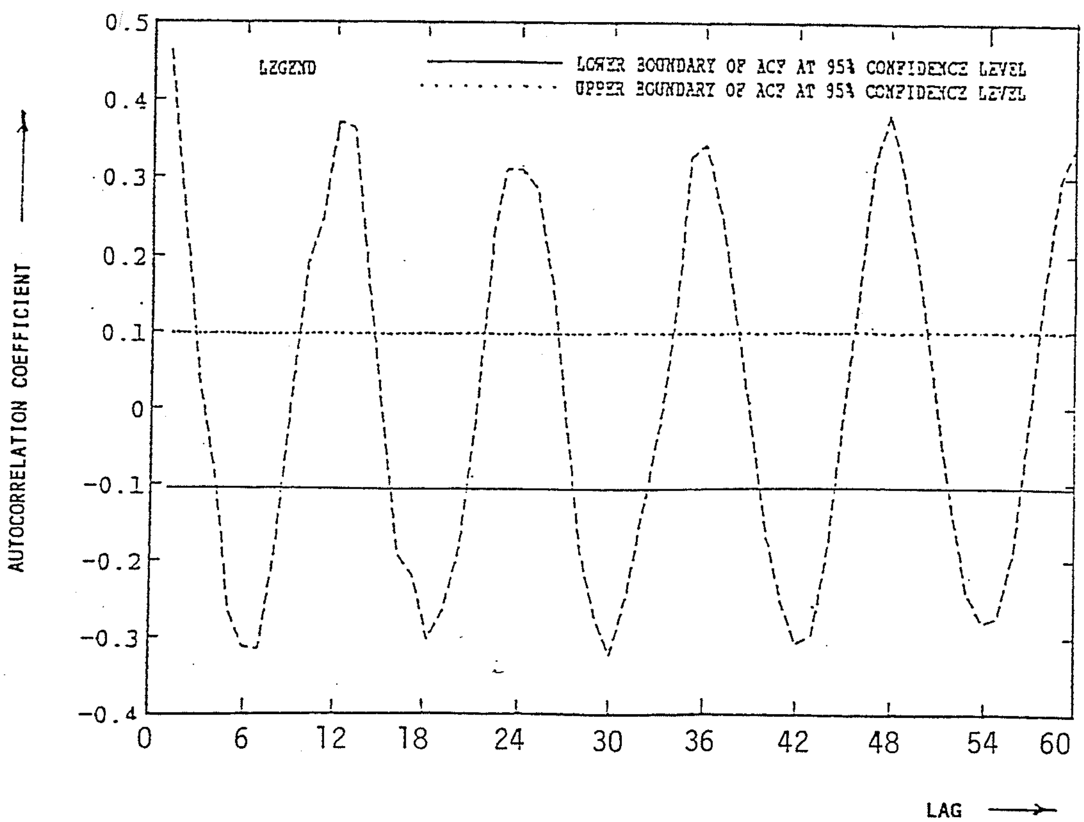


Fig.4.1.3.(a): Correlogram of Craig Ck.

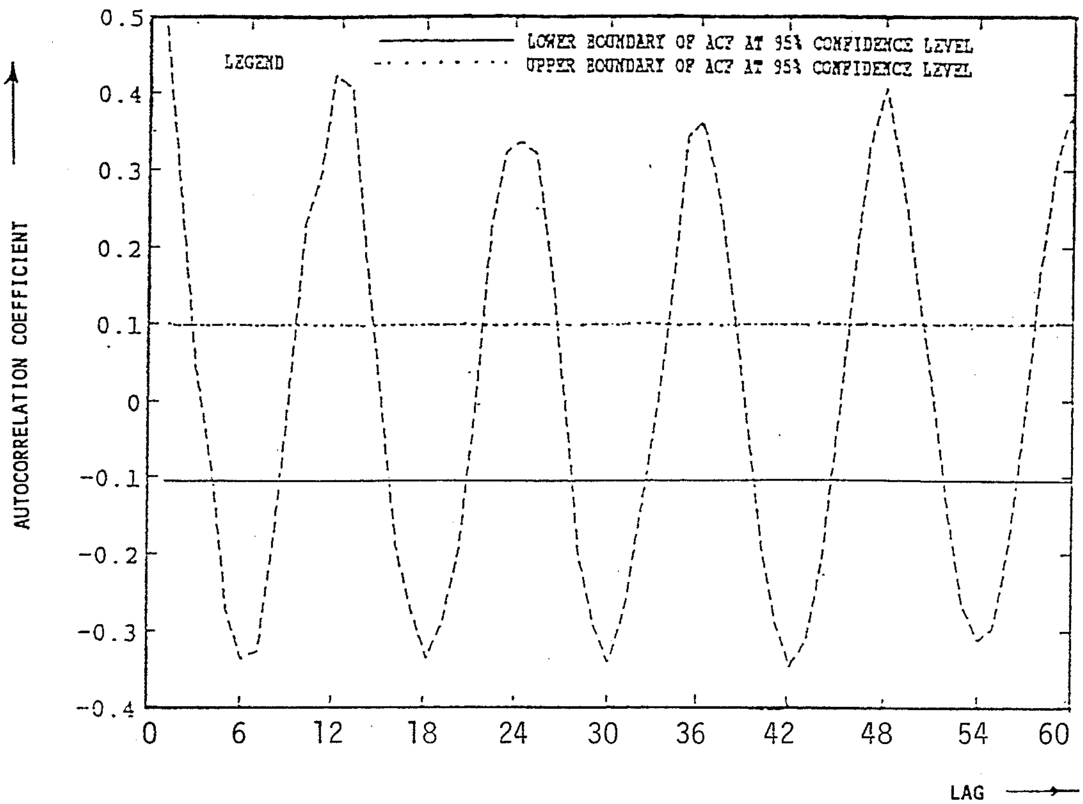


Fig.4.1.3.(b): Correlogram of Johns Ck.

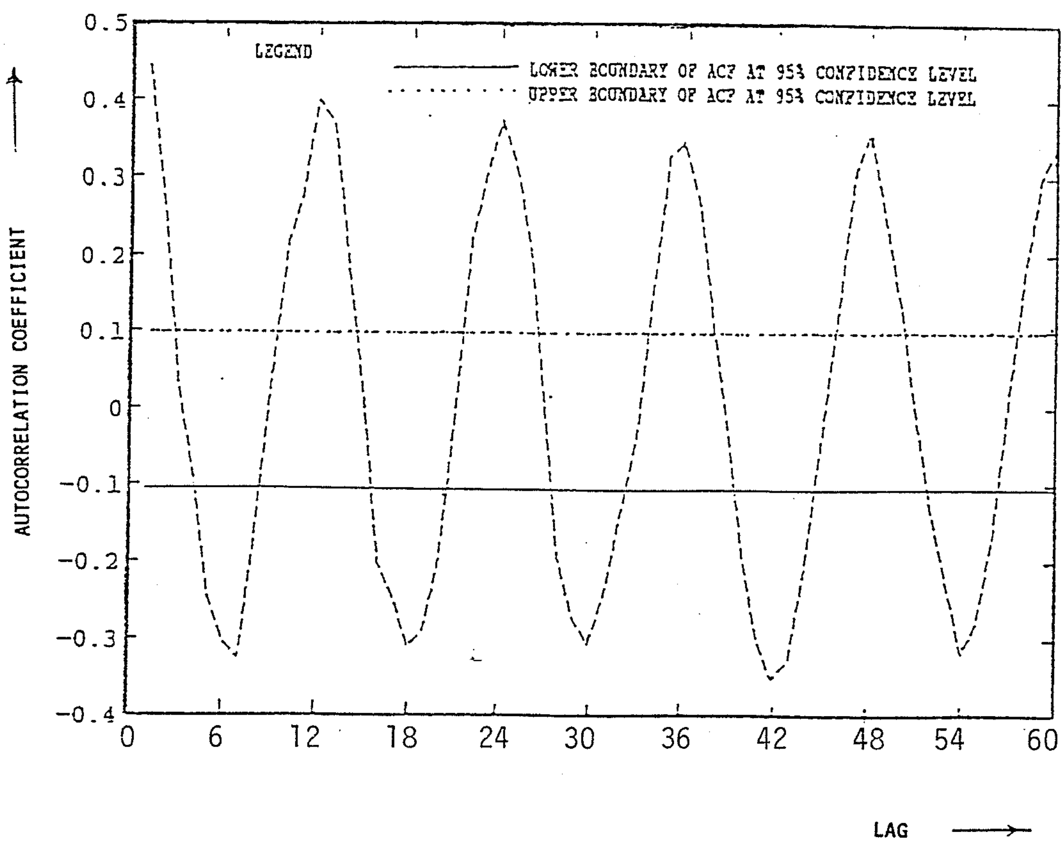


Fig.4.1.3.(c): Correlogram of Dunlap Ck.

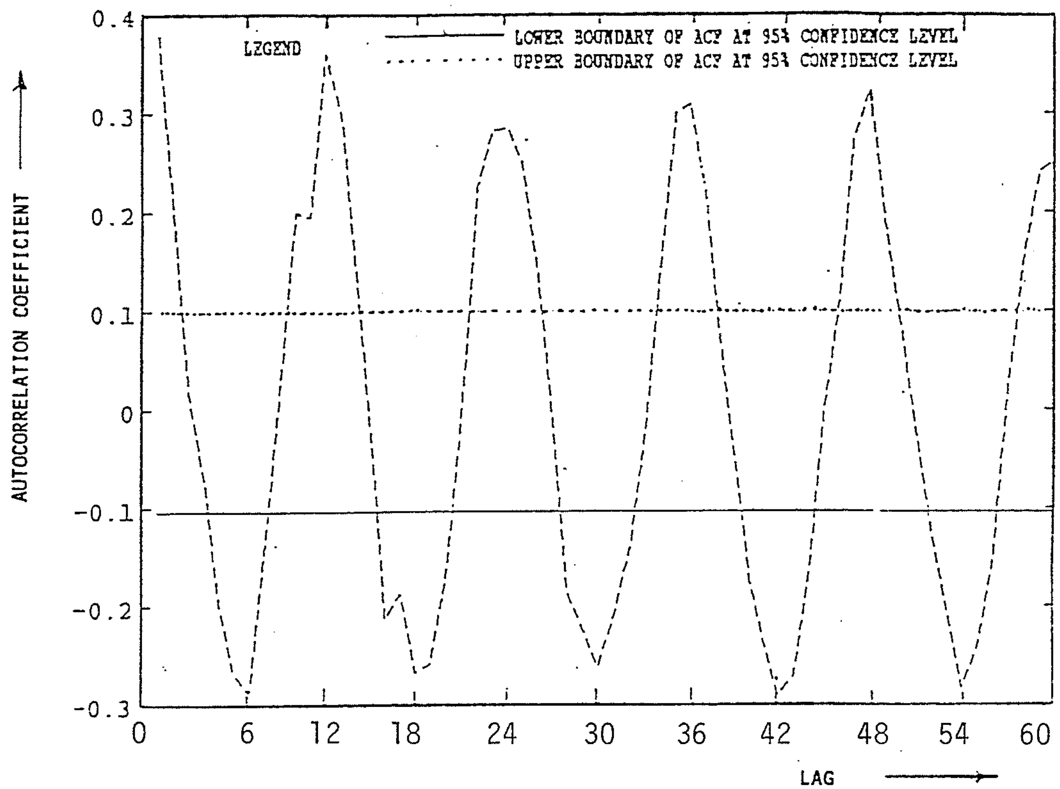


Fig.4.1.3.(d): Correlogram of Cowpasture R.

Table 4.1.2: Seasonal segmentation of the time wave-form of the members of the UPB cluster

River	Number of assignment to seasonal group	Month											
		June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
Johns Creek	Group 1	21	30	29	30	26	23	10	7	2	0	3	5
	Group 2	10	1	2	1	5	8	21	24	29	31	28	26
Dunlap Creek	Group 1	19	29	27	31	27	23	9	6	3	1	4	7
	Group 2	12	2	4	0	4	8	22	25	28	30	27	24
Craig Creek	Group 1	20	29	29	29	24	22	14	7	3	1	3	5
	Group 2	11	2	2	2	7	9	17	24	28	30	28	26
Cowpasture River	Group 1	20	31	27	29	25	18	12	9	3	2	4	6
	Group 2	11	0	4	2	6	13	19	22	28	29	27	25

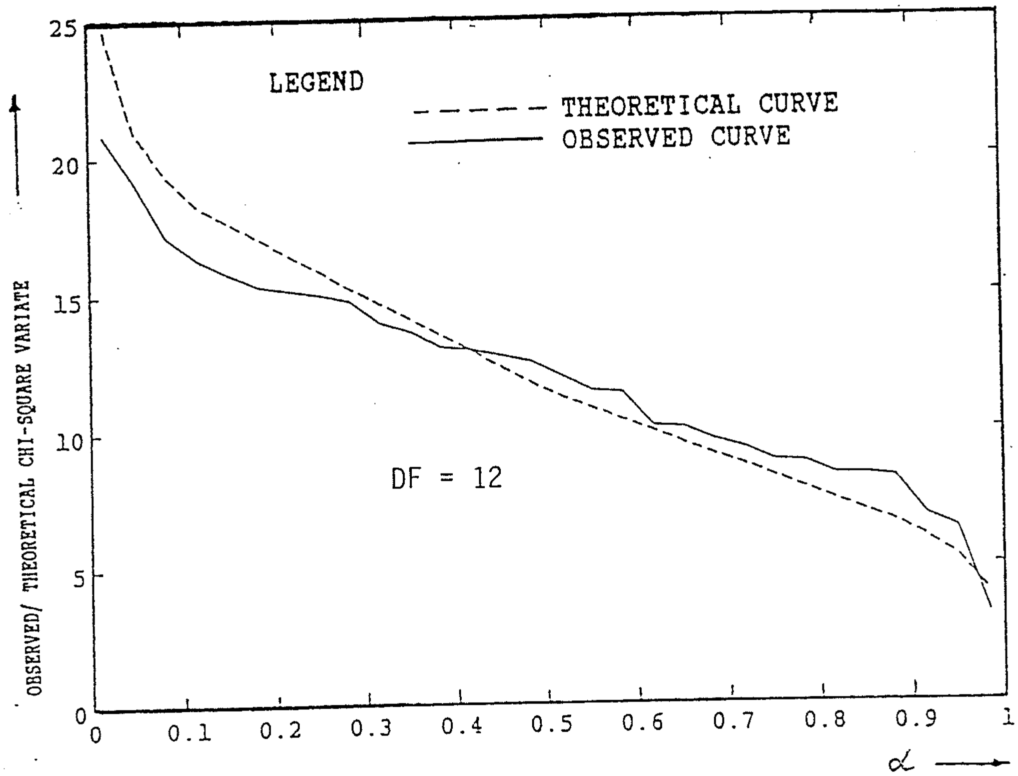


Fig. 4.1.4.1: Chi-square probability plot of Mahalanobis distance of the simultaneously observed dry seasonal segments of (Craig Ck.,Johns Ck.) pair

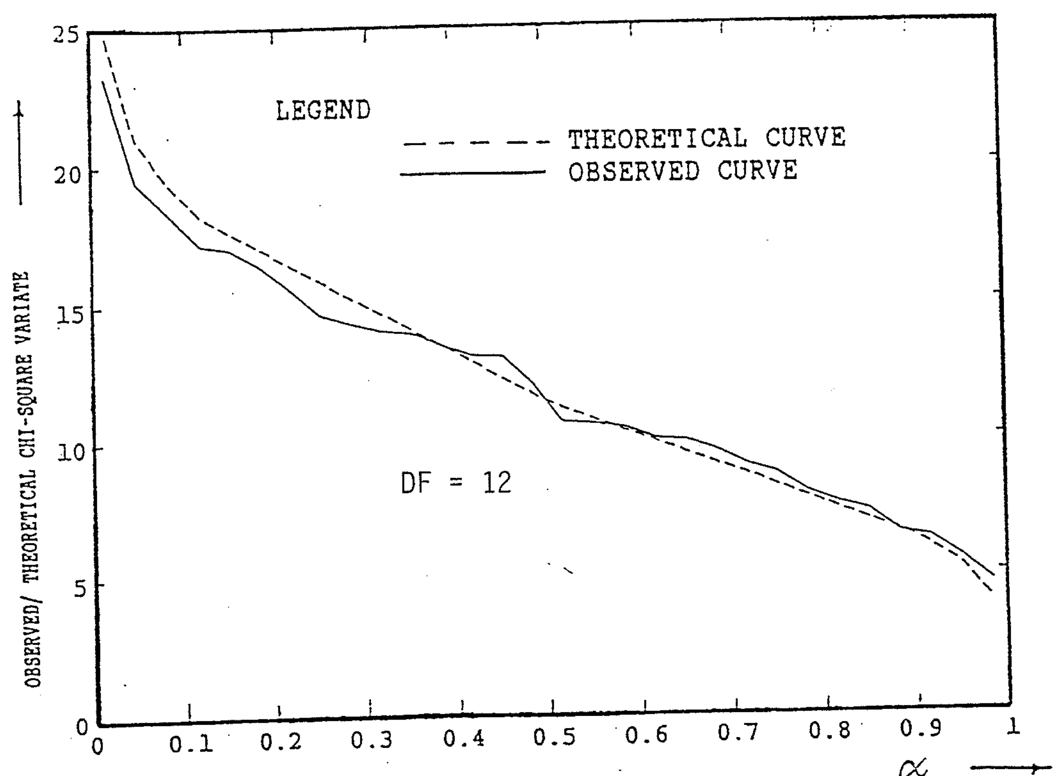


Fig. 4.1.4.2: Chi-square probability plot of Mahalanobis distance of the simultaneously observed wet seasonal segments of(Craig Ck.,Johns Ck.) pair

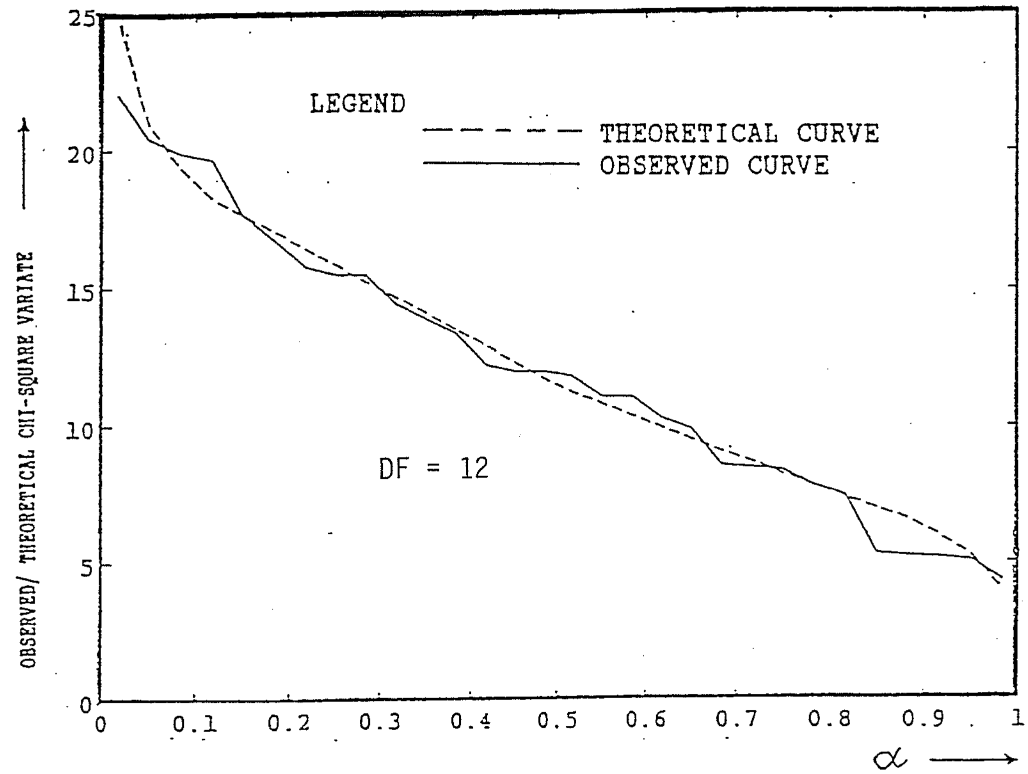


Fig. 4.1.4.3: Chi-square probability plot of Mahalanobis distance of the simultaneously observed dry seasonal segments of {Craig Ck., Dunlap Ck.} pair

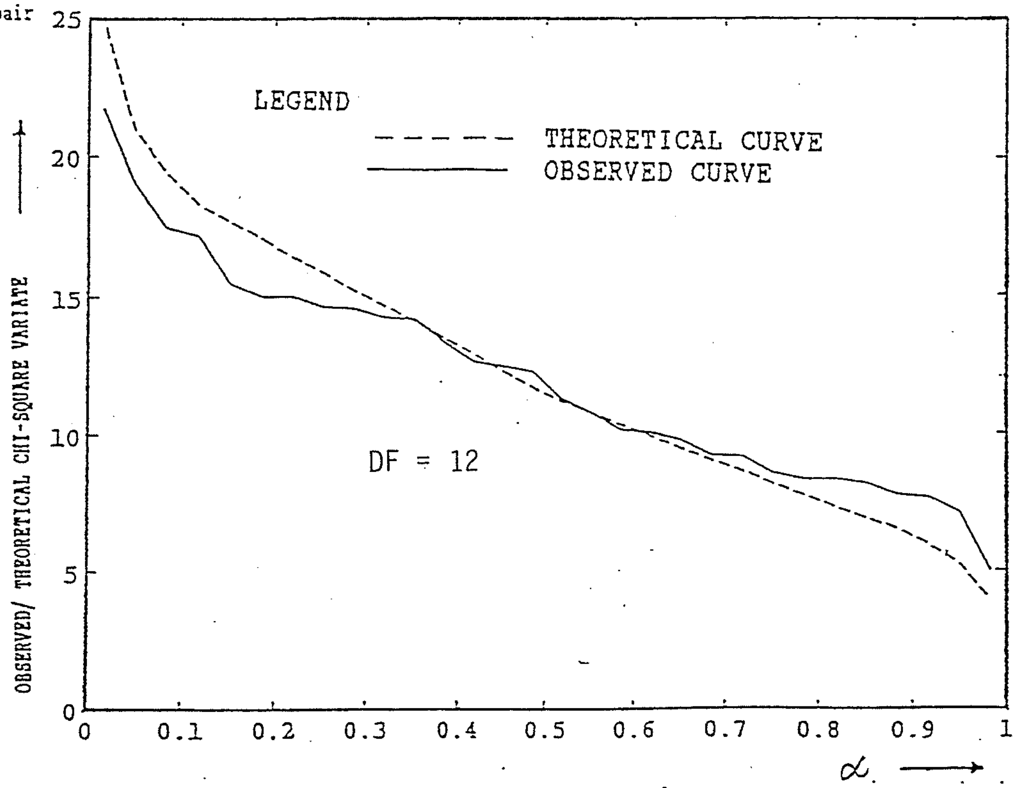


Fig. 4.1.4.4: Chi-square probability plot of Mahalanobis distance of the simultaneously observed wet seasonal segments of {Craig Ck., Dunlap Ck.} pair

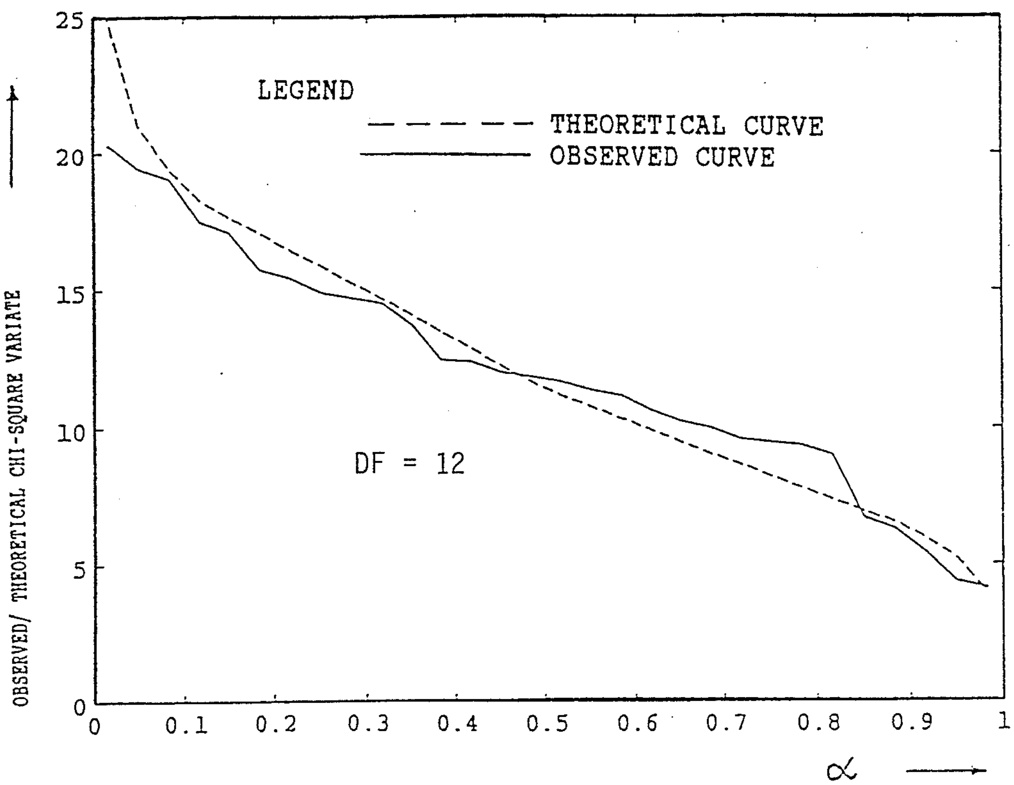


Fig. 4.1.4.5: Chi-square probability plot of Mahalanobis distance of the simultaneously observed dry seasonal segments of (Craig Ck., Cowpasture R.) pair

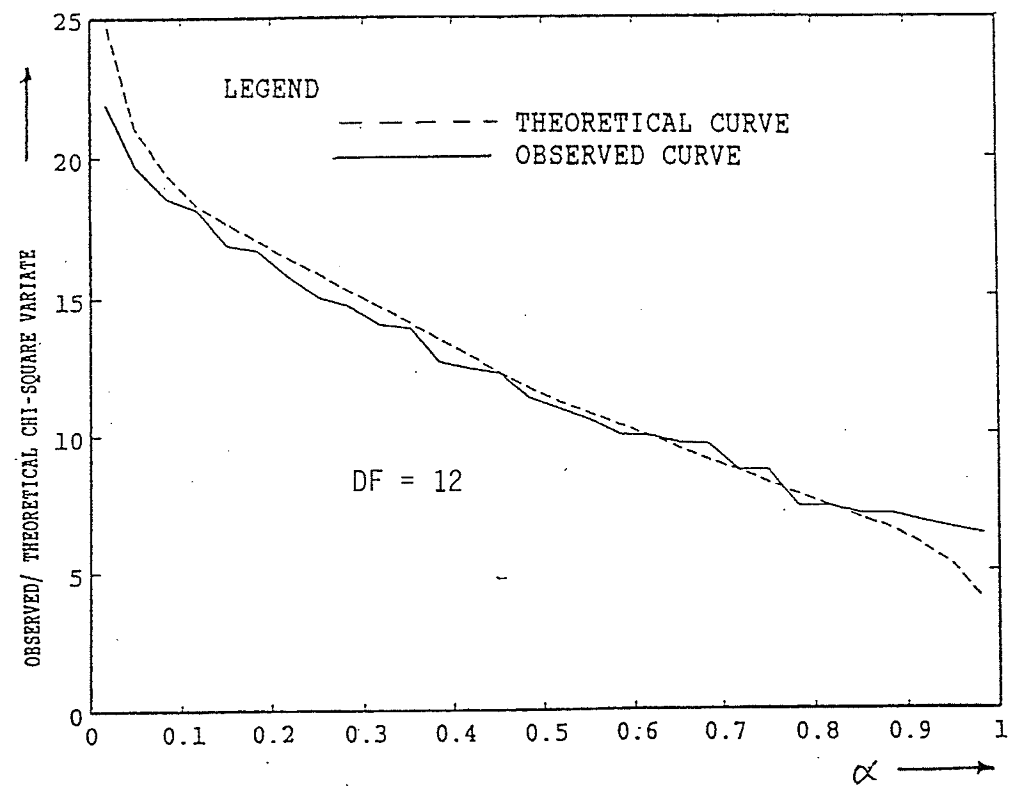


Fig. 4.1.4.6: Chi-square probability plot of Mahalanobis distance of the simultaneously observed wet seasonal segments of (Craig Ck.,Cowpasture R.) pair

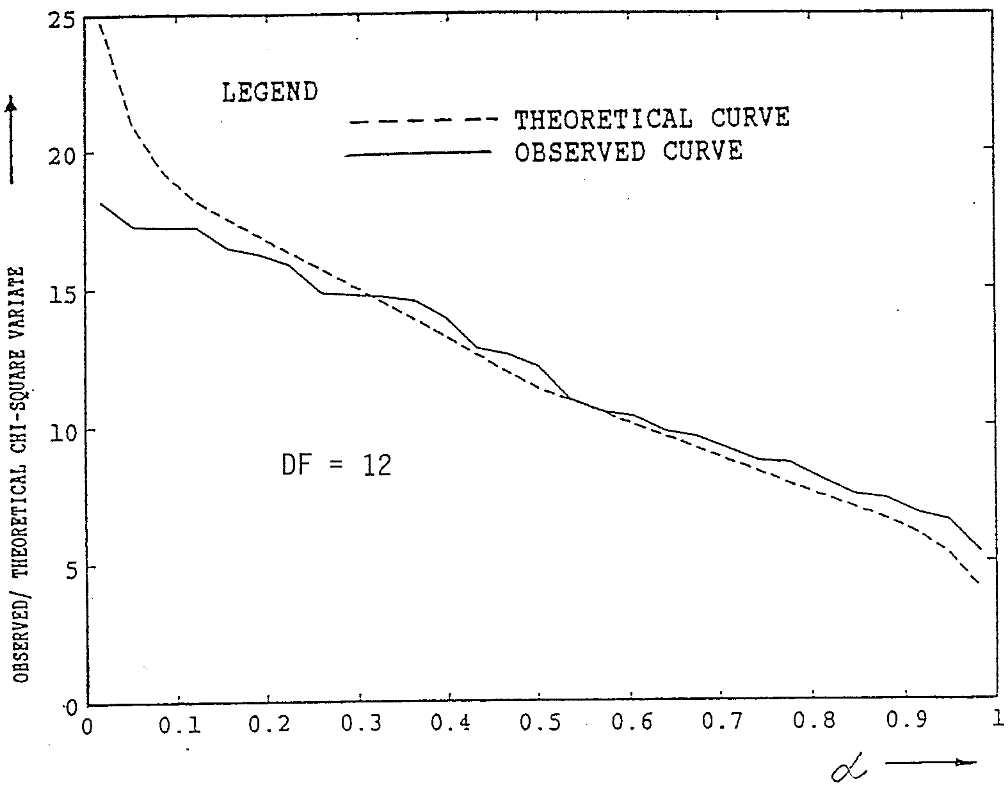


Fig. 4.1.5.1: Chi-square probability plot of Mahalanobis distance of the transitional seasonal segments corresponding to Wet-Dry seasonal transition of Craig Ck.

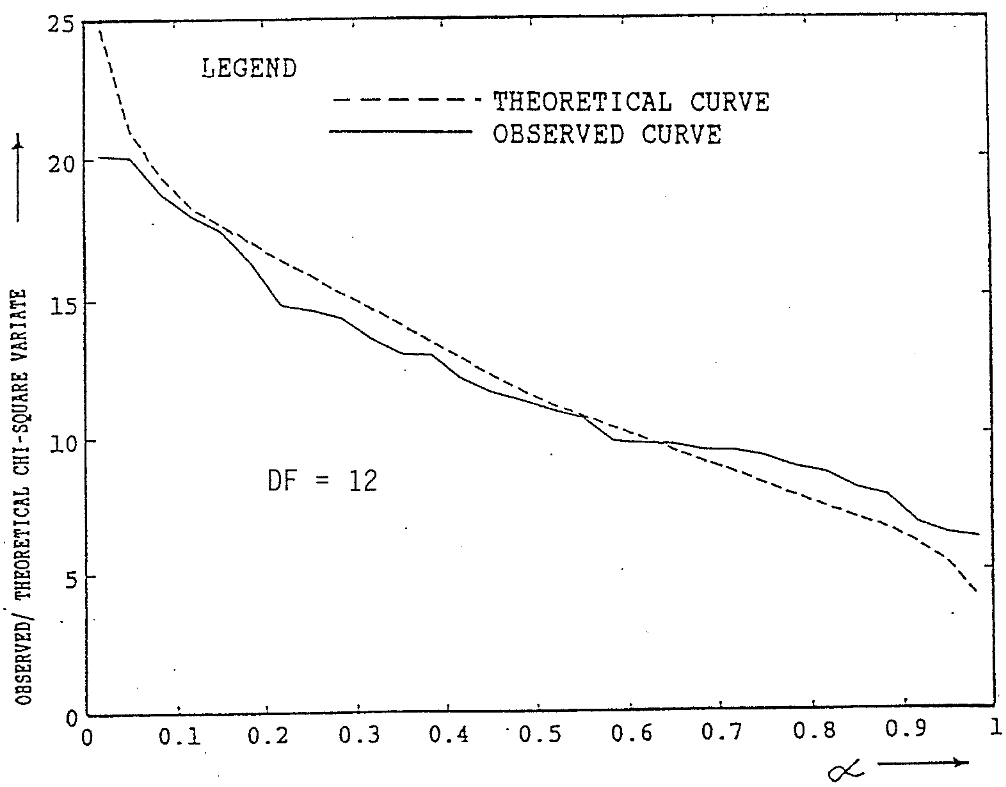


Fig. 4.1.5.2: Chi-square probability plot of Mahalanobis distance of the transitional seasonal segments corresponding to Dry-Wet seasonal transition of Craig Ck.

Table 4.1.3.1 : Infilling of missing monthly mean discharge [CFS] of Craig Creek by various models incorporating Johns Creek information

Infilling Models	Monthly mean Discharge [CFS] for each month of missing period [June1986-May1987]											
	1	2	3	4	5	6	7	8	9	10	11	12
REG	129.8	68.1	74.0	205.8	65.7	305.5	771.9	547.3	690.4	1151.9	2201.7	466.1
SREG	133.6	69.9	75.9	212.3	67.5	315.7	781.9	538.1	692.6	1207.9	2441.9	452.0
SSREG	133.4	70.3	76.4	210.8	67.9	311.9	767.8	530.3	680.9	1181.2	2372.0	446.1
AMOVE	128.2	66.6	72.5	204.6	64.3	305.4	781.5	551.4	697.9	1172.5	2261.1	468.5
SMOVE	134.3	70.1	76.2	213.5	67.7	317.8	782.1	538.2	692.8	1208.2	2442.3	452.1
SSMOVE	134.0	70.5	76.6	211.8	68.2	313.7	765.6	529.1	679.1	1176.9	2360.1	445.3
MULBS	119.3	64.6	70.2	166.8	62.4	310.4	817.5	542.7	711.7	1237.6	2440.5	419.4
OBSERVED	114.0	65.3	68.4	150.0	55.5	222.0	712.0	503.0	655.0	1211.0	2427.0	402.0

Table 4.1.3.2 : Infilling of missing monthly mean discharge [CFS] of Craig Creek by various models incorporating Dunlap Creek information

Infilling Models	Monthly mean Discharge [CFS] for each month of missing period [June1986-May1987]											
	1	2	3	4	5	6	7	8	9	10	11	12
REG	106.7	82.4	107.4	95.3	67.5	186.0	688.4	554.5	753.9	1132.5	1934.5	449.1
SREG	107.0	83.3	107.6	96.0	68.7	183.1	695.0	560.2	762.5	1167.0	2084.8	457.2
SSREG	106.3	83.8	107.0	95.9	69.8	177.3	757.5	624.6	821.5	1181.2	1904.7	517.5
AMOVE	102.3	78.1	103.0	91.0	63.5	182.6	714.0	570.0	785.0	1199.6	2095.7	457.6
SMOVE	108.6	84.4	109.3	97.4	69.6	186.8	701.0	560.1	770.5	1175.5	2049.6	450.0
SSMOVE	107.9	84.9	108.5	97.2	70.6	180.6	751.8	620.6	814.9	1169.0	1879.2	514.8
MULBS	96.3	66.0	90.3	87.8	61.3	198.3	722.1	584.1	763.0	1350.3	2361.7	579.3
OBSERVED	114.0	65.3	68.4	150.0	55.5	222.0	712.0	503.0	655.0	1211.0	2427.0	402.0

Table 4.1.3.3 : Infilling of missing monthly mean discharge [CFS] of Craig Creek by various models incorporating Cowpasture River information

Infilling Models	Monthly mean Discharge [CFS] for each month of missing period [June1986-May1987]											
	1	2	3	4	5	6	7	8	9	10	11	12
REG	131.0	84.2	63.4	74.0	54.3	158.0	602.2	409.5	481.3	840.0	2044.3	482.7
SREG	122.1	81.3	62.5	72.1	54.2	145.1	631.1	441.4	512.7	859.2	1960.1	514.1
SSREG	119.8	81.2	63.2	72.4	55.1	141.2	692.1	520.1	584.4	903.0	1959.8	585.6
AMOVE	126.4	79.2	58.6	69.0	49.8	154.1	634.4	421.9	500.6	902.2	2311.4	502.2
SMOVE	123.2	81.9	63.0	72.7	54.6	146.4	630.5	441.2	512.4	558.0	1955.0	513.8
SSMOVE	120.3	81.6	63.5	72.8	55.4	141.9	667.5	483.9	553.8	881.2	1850.9	555.1
MULBS	119.6	67.6	71.7	56.9	53.9	168.0	608.5	443.8	582.3	963.5	2095.3	369.9
OBSERVED	114.0	65.3	68.4	150.0	55.5	222.0	712.0	503.0	655.0	1211.0	2427.0	402.0

Table 4.1.3.4 : Infilling of missing monthly mean discharge [CFS] of Craig Creek by SZSTRNALL Model

Infilling Models	Monthly mean Discharge [CFS] for each month of missing period [June1986-May1987]											
	1	2	3	4	5	6	7	8	9	10	11	12
INFILLED BY SZSTRNALL MODEL	120.6	68.0	99.2	52.8	61.0	104.5	182.1	330.8	641.0	597.1	380.2	370.4
OBSERVED	114.0	65.3	68.4	150.0	55.5	222.0	712.0	503.0	655.0	1211.0	2427.0	402.0

4.2 APPLICATION OF VARIOUS MODELS IN INFILLING MISSING DATA OF RIVERS BELONGING TO LB CLUSTER

The basic information pertaining to the members of LB cluster and the location of the members are presented in Table 4.2.1 and in Fig.4.2.1. A plot of five years period of simultaneously observed data of the member rivers is presented in Fig.4.2.2. At the initial phase, for each of the member rivers, 31 years of data over the period of [Oct. 1958 - Sep. 1988] are considered. Correlogram analysis [Figs.4.2.3.(a,b,c)] shows two seasons per year and the segmentation procedure assigns the periods [June - Nov.] and [Dec. - May] respectively to dry and wet seasons [Table 4.2.2]. The data matrix is slid to coincide with the beginning of the dry season. The slid data matrix contains 31 years of data over the period [June 1958 - May 1989]. One year of monthly data of Little R. over the period [June 1986 - May 1987] is assumed to be missing. The SESTRNALL model is applied to Little R. and the rest of the models are applied to {Little R., Reed Ck.} and {Little R., Roanoke R. } pairs. Multivariate normality is tested for both MULBS and SESTRNALL models. In case of unsatisfactory fit, power transformation of the element of the joint variate **X** are done. The final Chi-square plots of the Mahalanobis distances of the observations of **X** variate are presented in Figs. 4.2.4.(1, 2, 3 and 4) and Figs.4.2.5.(1 and 2). The reconstructed data of Little R. corresponding to the different information sources { Reed Ck., Roanoke R., Little R. } pertinent to data reconstruction are presented in Table 4.2.3.(1, 2 and 3)

Table 4.2.1: Information pertinent to the members of LB cluster

River	Basin	USGS Station Number	Comment	Latitude	Longitude	Area	Period of data used in analysis	Missing Period
Little River	Kanawaha River Basin	03170000	River with missing data	37-02-15	80-33-25	300	June 1958-May 1989 (31 Years)	June 1986-May 1987
Reed Creek	Kanawaha River Basin	03167000	Base River	36-56-22	80-53-13	247	June 1958-May 1989 (31 Years)	
Roanoke River	Roanoke River Basin	02055000	Base River	37-15-30	79-56-20	395	June 1958-May 1989 (31 Years)	

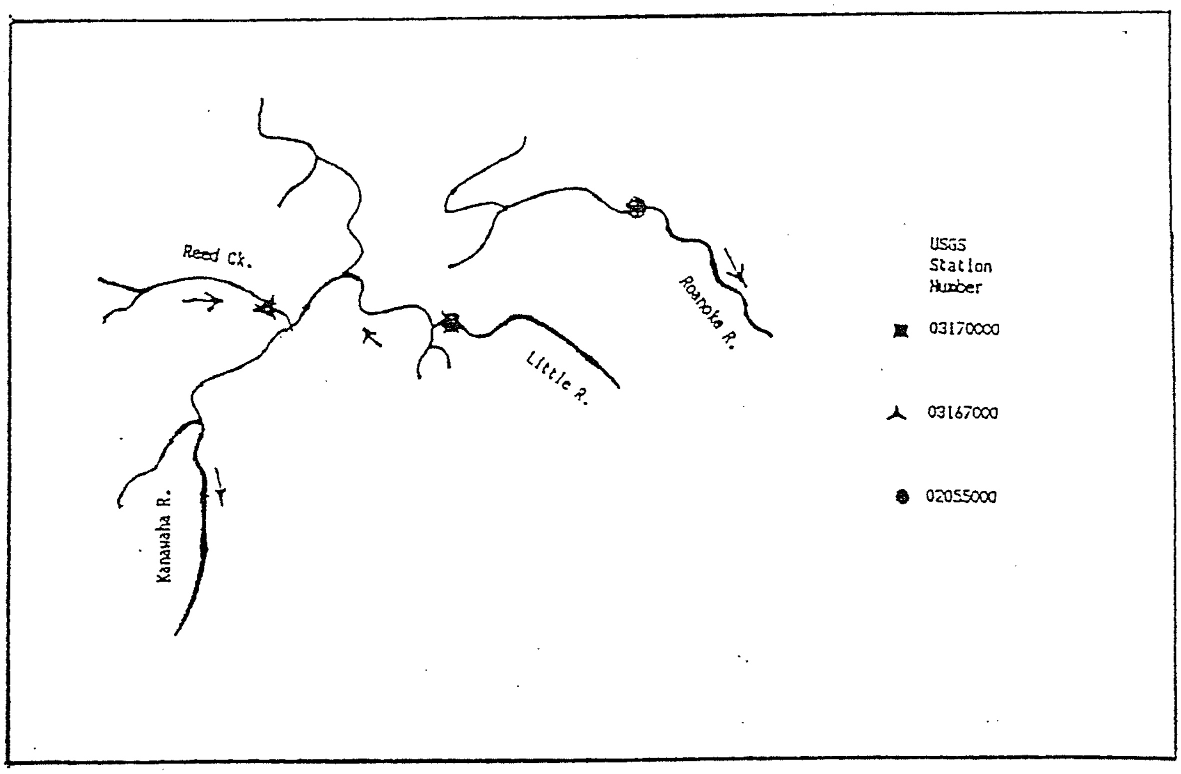


Fig.4.2.1: The location of the members of LB cluster

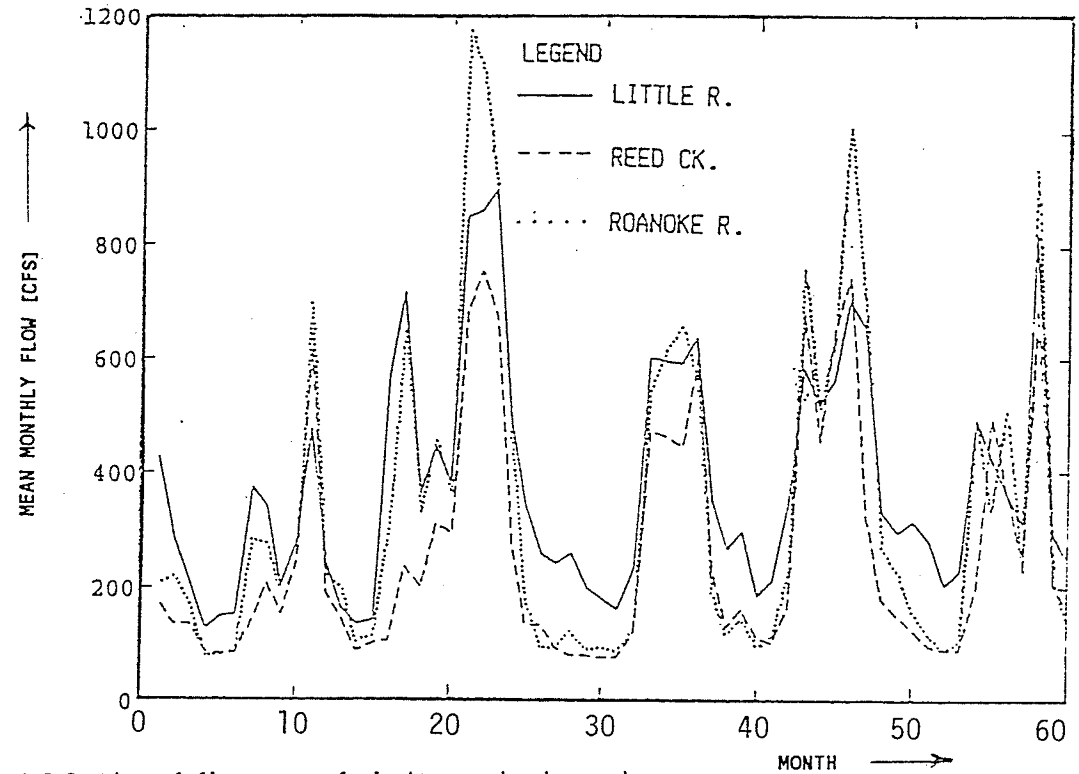


Fig.4.2.2: Plot of five years of simultaneously observed monthly stream-flow data of the members of LB cluster [June 1958-May 1963]

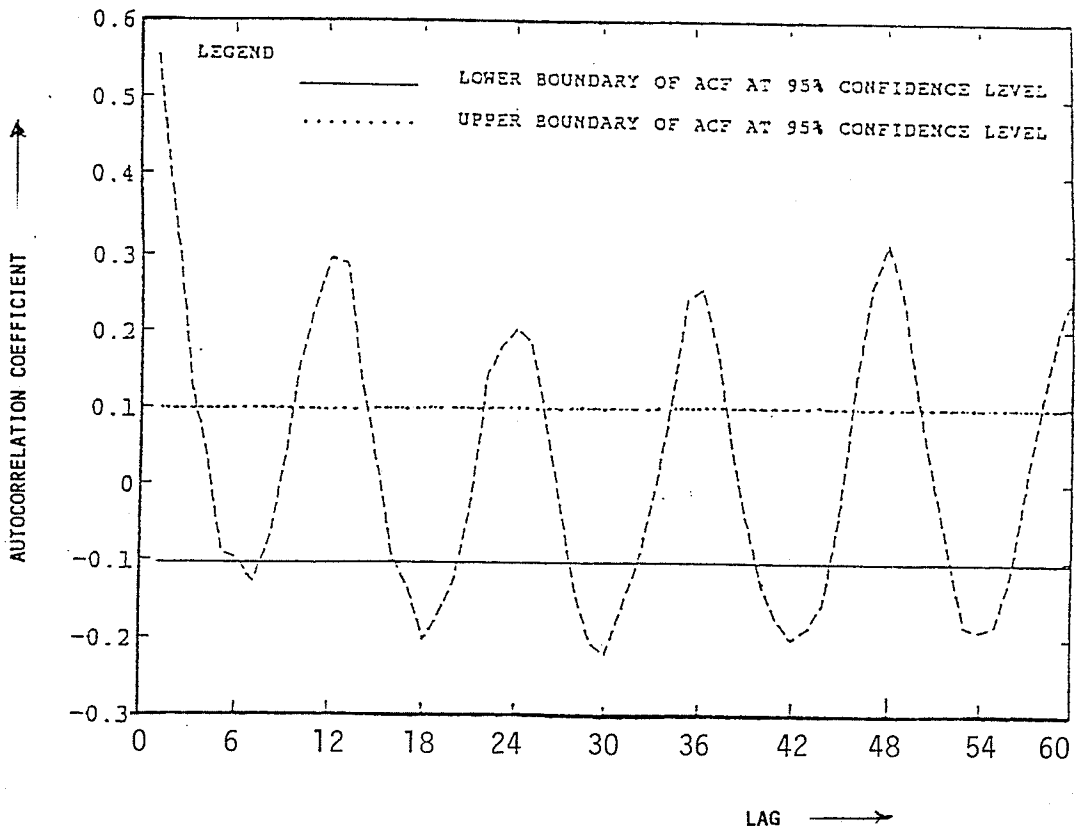


Fig.4.2.3.(a): Correlogram of Little R.

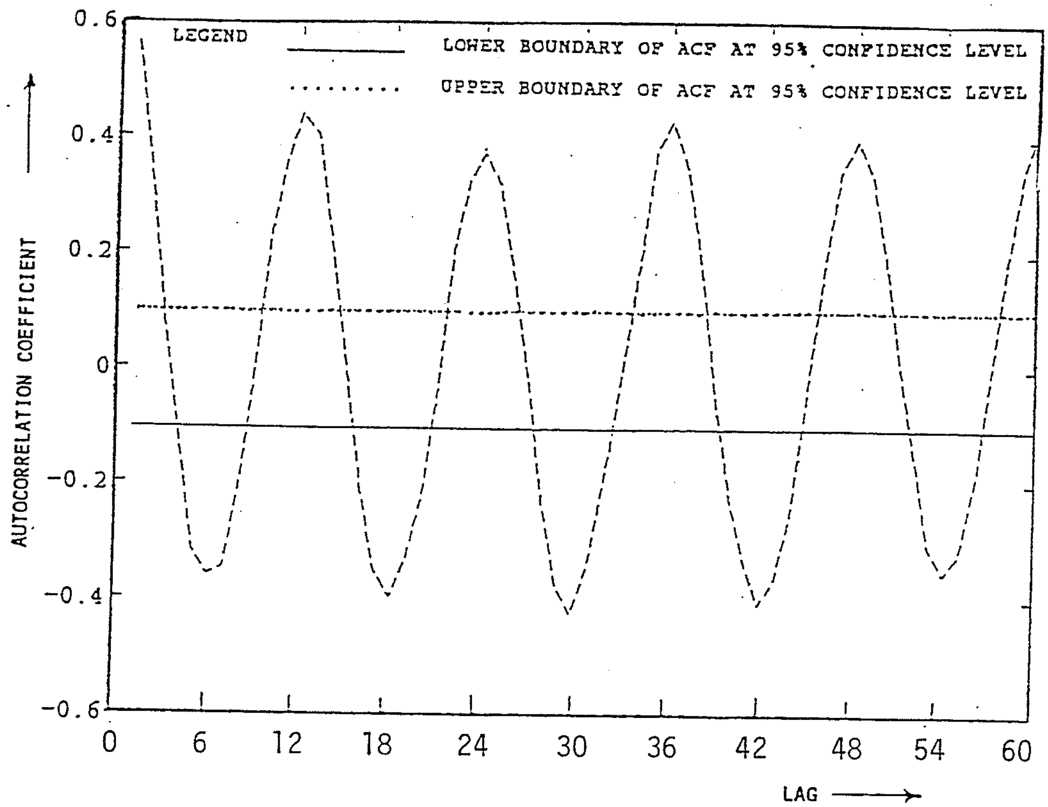


Fig.4.2.3.(b): Correlogram of Reed Ck.

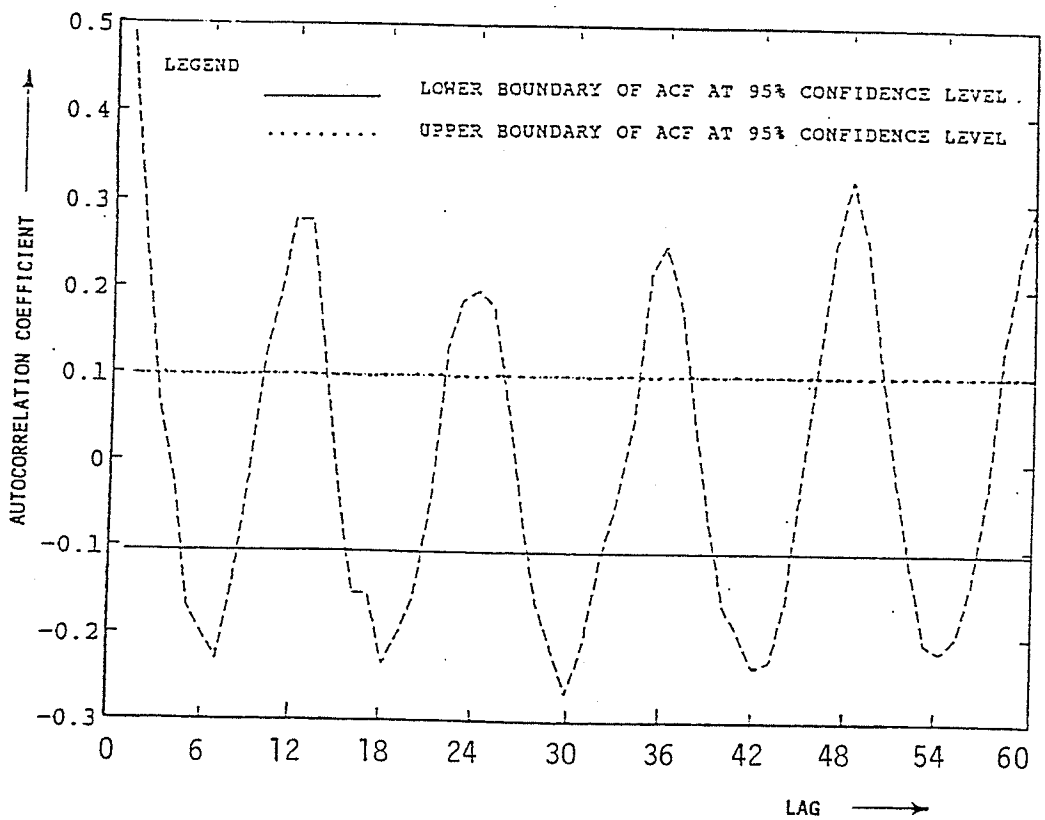


Fig.4.2.3.(c): Correlogram of Roanoke R.

Table 4.2.2: Seasonal Segmentation of the time wave form of members of the LB cluster

River	Number of assignment to seasonal group	Month											
		June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
Roanoke River	Group 1	17	29	27	29	24	24	4	9	4	1	3	5
	Group 2	14	2	4	2	7	7	27	22	27	30	28	26
Reed Creek	Group 1	15	26	30	31	29	24	13	7	2	1	2	6
	Group 2	16	5	1	0	2	7	18	24	29	30	29	25
Little River	Group 1	15	23	28	25	25	22	17	13	4	2	3	9
	Group 2	16	8	3	6	6	9	14	18	27	29	23	22

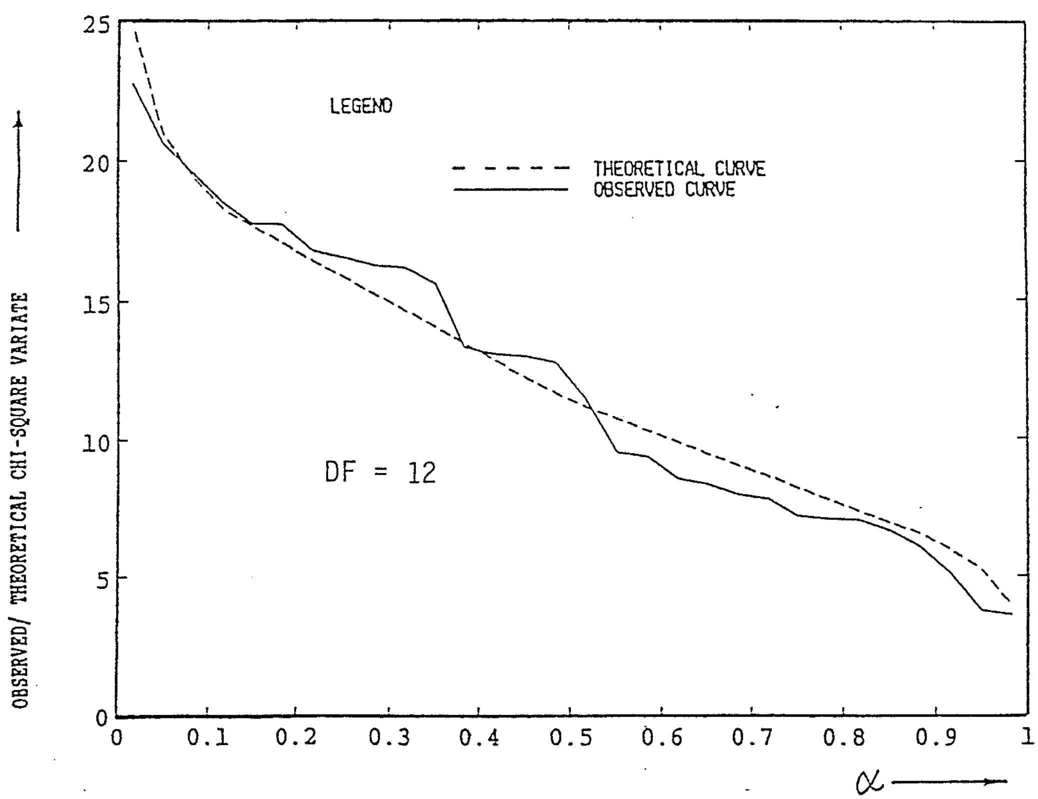


Fig. 4.2.4.1: Chi-square probability plot of Mahalanobis distance of the simultaneously observed dry seasonal segments of {Little R., Reed Ck.} pair

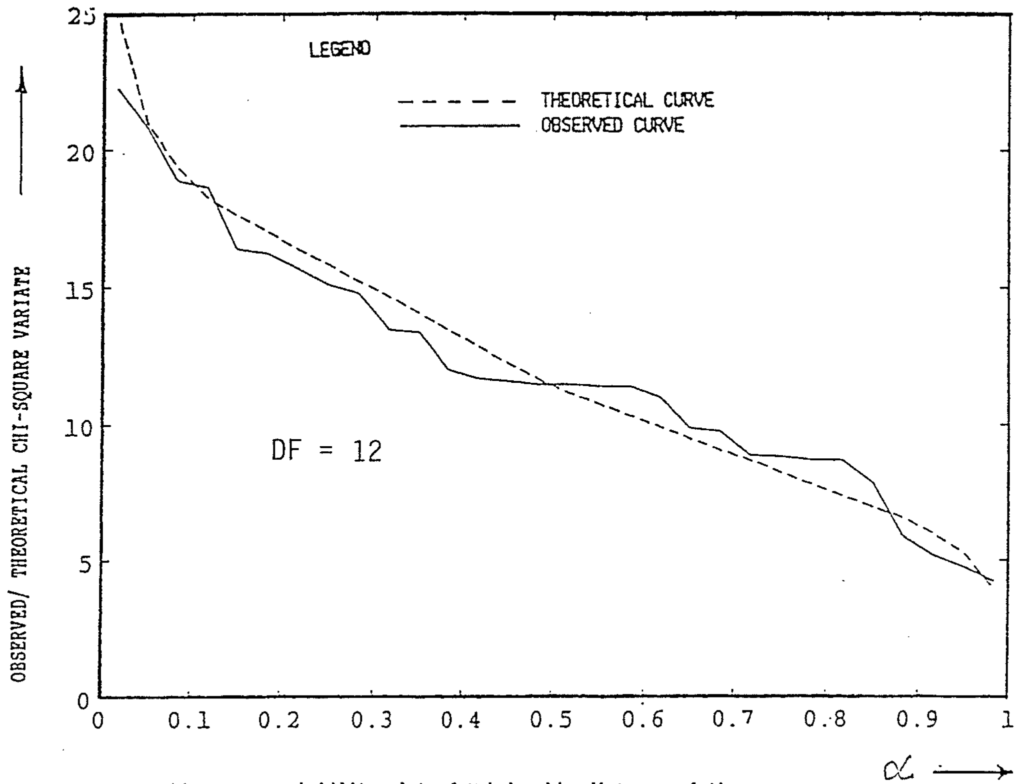


Fig. 4.2.4.2: Chi-square probability plot of Mahalanobis distance of the simultaneously observed wet seasonal segments of {Little R., Reed Ck.} pair

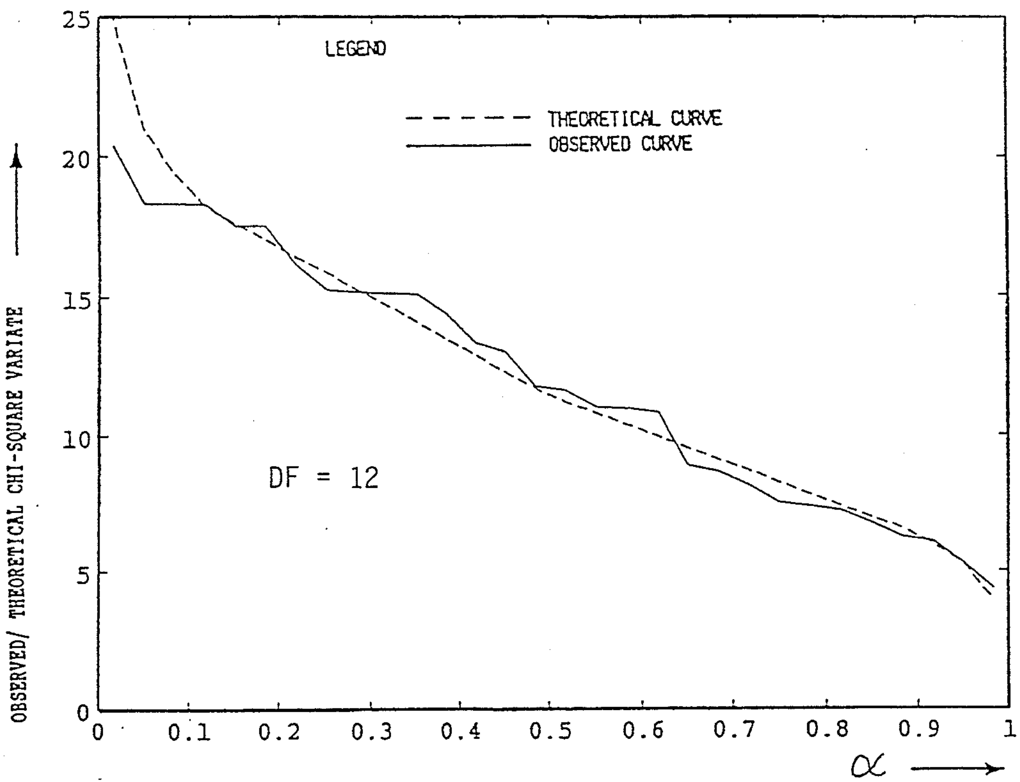


Fig. 4.2.4.3: Chi-square probability plot of Mahalanobis distance of the simultaneously observed dry seasonal segments of {Little R., Roanoke R.} pair

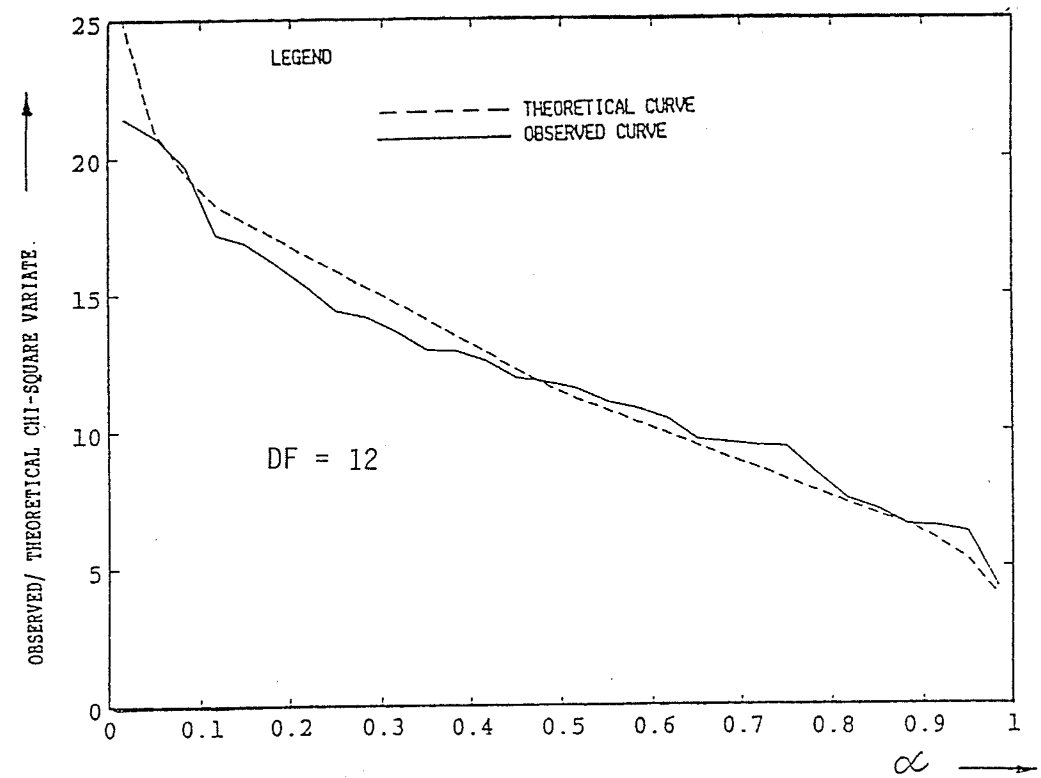


Fig. 4.2.4.4: Chi-square probability plot of Mahalanobis distance of the simultaneously observed wet seasonal segments of {Little R., Roanoke R.} pair

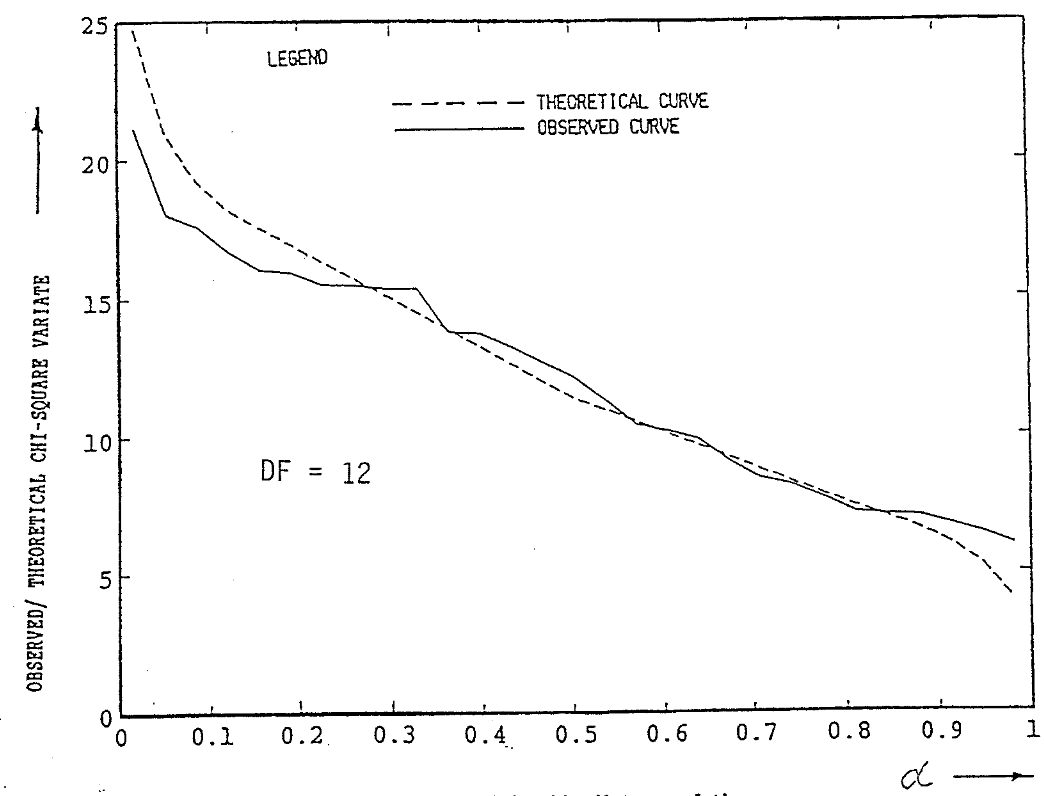


Fig. 4.2.5.1: Chi-square probability plot of Mahalanobis distance of the transitional seasonal segments corresponding to Wet-Dry seasonal transition of Little R.

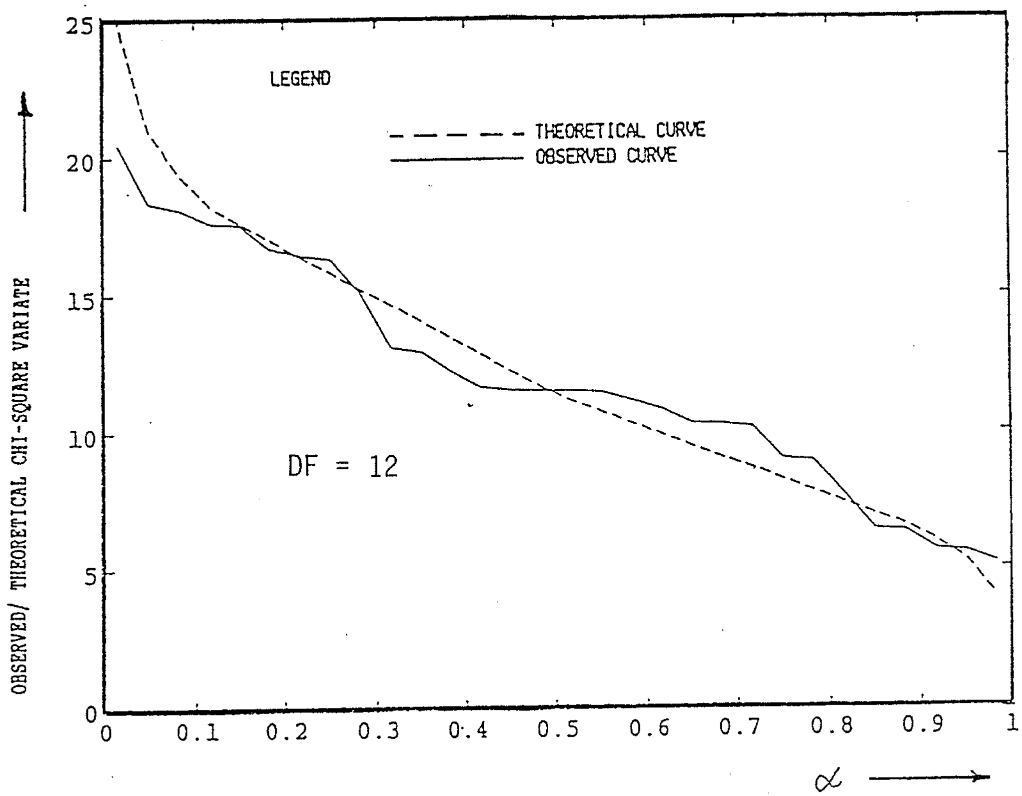


Fig. 4.2.5.2: Chi-square probability plot of Mahalanobis distance of the transitional seasonal segments corresponding to Dry-Wet seasonal transition of Little R.

Table 4.2.3.1 : Infilling of missing monthly mean discharge [CFS] of Little River by various models incorporating Reed Ck. information

Infilling Models	Monthly mean Discharge [CFS] for each month of missing period [June1986-May1987]											
	1	2	3	4	5	6	7	8	9	10	11	12
REG	227.1	164.4	158.4	199.5	150.7	228.2	417.7	417.0	563.5	790.8	988.7	425.2
SREG	238.1	161.4	154.4	203.7	145.4	239.4	401.3	400.5	548.3	780.7	985.4	408.8
SSREG	236.1	167.9	161.4	205.9	153.1	237.3	454.8	454.0	605.0	835.6	1033.9	462.5
AMOVE	219.3	153.8	147.7	190.2	139.8	220.4	428.4	427.4	595.0	863.3	1103.4	436.7
SMOVE	237.8	161.5	154.5	203.6	145.5	239.1	400.9	400.2	547.4	778.8	982.5	408.4
SSMOVE	232.6	166.3	160.0	203.3	151.9	233.7	447.0	446.2	591.0	810.8	998.6	454.4
MULBS	194.6	165.9	145.4	327.0	216.1	291.0	523.8	460.4	565.7	959.3	1366.9	654.9
OBSERVED	160.0	112.0	166.0	213.0	143.0	247.0	470.0	354.0	474.0	982.0	1445.0	632.0

Table 4.2.3.2 : Infilling of missing monthly mean discharge [CFS] of Little R. by various models incorporating Roanoke R. information

Infilling Models	Monthly mean Discharge [CFS] for each month of missing period [June1986-May1987]											
	1	2	3	4	5	6	7	8	9	10	11	12
REG	187.3	139.0	229.9	247.1	148.9	285.5	538.7	460.5	544.6	790.3	1206.9	450.1
SREG	191.1	141.1	235.4	253.4	151.4	293.4	536.7	454.3	542.9	806.3	1264.6	443.4
SSREG	192.5	145.4	233.4	249.9	155.1	286.2	568.9	482.0	575.4	852.6	1333.6	470.5
AMOVE	181.3	132.5	224.9	242.7	142.5	282.4	550.5	466.9	556.8	823.5	1285.3	455.3
SMOVE	191.7	141.6	236.1	254.2	151.8	294.3	536.5	454.3	542.7	805.9	1263.3	443.3
SSMOVE	192.1	145.4	238.8	249.1	155.1	285.0	563.8	478.3	570.1	842.2	1312.5	167.0
MULBS	169.6	160.0	192.5	267.2	148.2	292.6	521.0	503.3	534.7	893.3	1652.4	621.4
OBSERVED	160.0	112.0	166.0	213.0	143.0	247.0	470.0	354.0	474.0	982.0	1445.0	632.0

Table 4.2.3.3 : Infilling of missing monthly mean discharge [CFS] of Little River by SESTRNALL Model

Infilling Model	Monthly mean Discharge [CFS] for each month of missing period [June1986-May1987]											
	1	2	3	4	5	6	7	8	9	10	11	12
INFILLED BY SESTRNALL MODEL	208.3	172.8	172.6	153.1	167.5	228.1	237.1	338.0	398.4	467.2	352.4	354.0
OBSERVED	160.0	112.0	166.0	213.0	143.0	247.0	470.0	354.0	474.0	982.0	1445.0	632.0

4.3 APPLICATION OF VARIOUS MODELS IN INFILLING MISSING DATA OF RIVERS BELONGING TO LILL CLUSTER

The basic information pertaining to the members of LILL cluster and the location of the members are presented in Table 4.3.1 and in Fig.4.3.1. A plot of five years period of simultaneously observed data of the member rivers is presented in Fig.4.3.2. At the initial phase, for each of the member rivers, 23 years of data over the period of [Jan. 1925 - Dec.1947] are considered. Correlogram analysis [Figs.4.3.3.(a, b, c)] shows two seasons per year and the segmentation procedure assigns the periods [Nov. - April] and [May - Oct.] respectively to dry and wet seasons [Table 4.3.2]. The data matrix is slid to coincide with the beginning of the dry season. The slid data matrix contains 23 years of data over the period [Nov.1924 - Oct.1947]. One year of monthly data of Green R. over the period [Nov.1944 - Oct.1945] is assumed to be missing. The SESTRNALL model is applied to Green R. and the rest of the models are applied to {Green R., Soo R.} and {Green R., Rutherford Ck. } pairs. Multivariate normality is tested for both MULBS and SESTRNALL models. In case of unsatisfactory fit, power transformation of element of the joint variate **X** was performed. The final Chi-square plot of the Mahalanobis distances of the observations of **X** variate are presented in Figs. 4.3.4.(1, 2, 3 and 4) and Figs.4.3.5.(1 and 2). The fit of multivariate normality in the case of wet seasonal segments of [Green R., Soo R.] pair, both dry and wet seasonal segments of the [Green R., Rutherford Ck.] pair and the transitional seasonal segments of both [wet-dry] and [dry-wet] transition are found to be not very satisfactory. This test is repeated on all possible cases of transformed data and the best fit among them is presented here which (still show wide deviation). This lack of multivariate normality is suspected to be due to inadequate data (22 years). This explanation can not be proven for rivers of LILL cluster. Nevertheless, a hypothetical test is done on the wet seasonal segments of {Craig Ck., Johns Ck.} pair of the UPB cluster by comparing the Chi- square plot by varying the sample size between 22-29. These plots are presented in Appendix A.2 [Fig. A.2.] . These plots clearly show the positive correlation between better fit of the observed curve to the theoretical curve and the corresponding sample size. In fact, a sample size of 29 shows satisfactory fit. The better fit of dry seasonal segments of {Green R., Soo R.} is possibly due to the little variation among the flows of the dry seasonal segments. In case of the wet seasonal segment, for small sample size, variation in the multivariate observations is possibly too intractable to be encompassed by multivariate normal distribution format. To investigate the true reason behind the non-normality in case of small sample is beyond the scope of the thesis. Inspite of this unsatisfactory fit, the proposed multivariate models were applied. The reconstructed data of Green R. corresponding to the different information sources pertinent to data reconstruction are presented in Table 4.3.3.(1, 2, and 3).

Table 4.3.1: Information pertinent to the members of LILL cluster

River *	WSC Station Number	Comment	Latitude	Longitude	Area mi^2	Period of data used in analysis	Missing Period
Green River	08MG003	River with missing data	50-16-55	122-51-05	855	Nov.1924-Oct. 1947 (23 Years)	Nov. 1944-Oct.1945
Soo River	08MG007	Base River	50-13-30	122-53-00	283	Nov.1924-Oct. 1947 (23 Years)	
Rutherford Creek	08MG006	Base River	50-16-00	122-52-20	179	Nov.1924-Oct. 1947 (23 Years)	

* All rivers are tributaries of the Lillooet R.

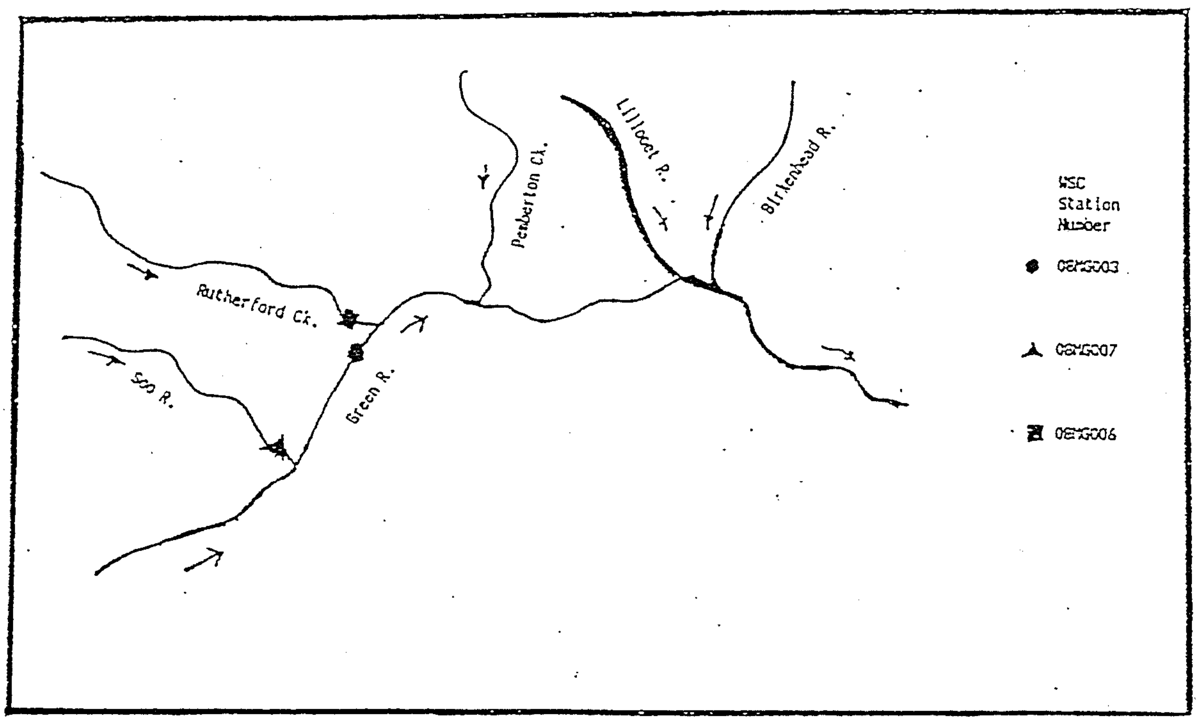


Fig.4.3.1: The location of the members of LILL cluster

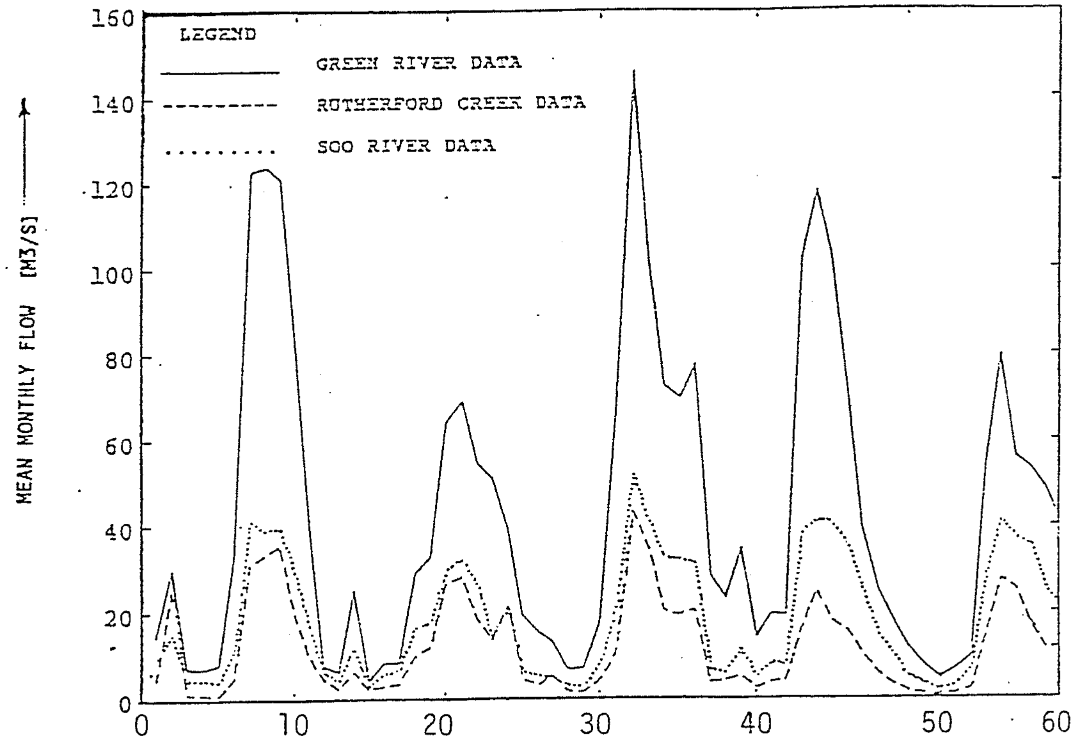


Fig.4.3.2: Plot of five years of simultaneously observed monthly stream-flow data of the members of LILL cluster [Nov.1944 - Oct. 1945]

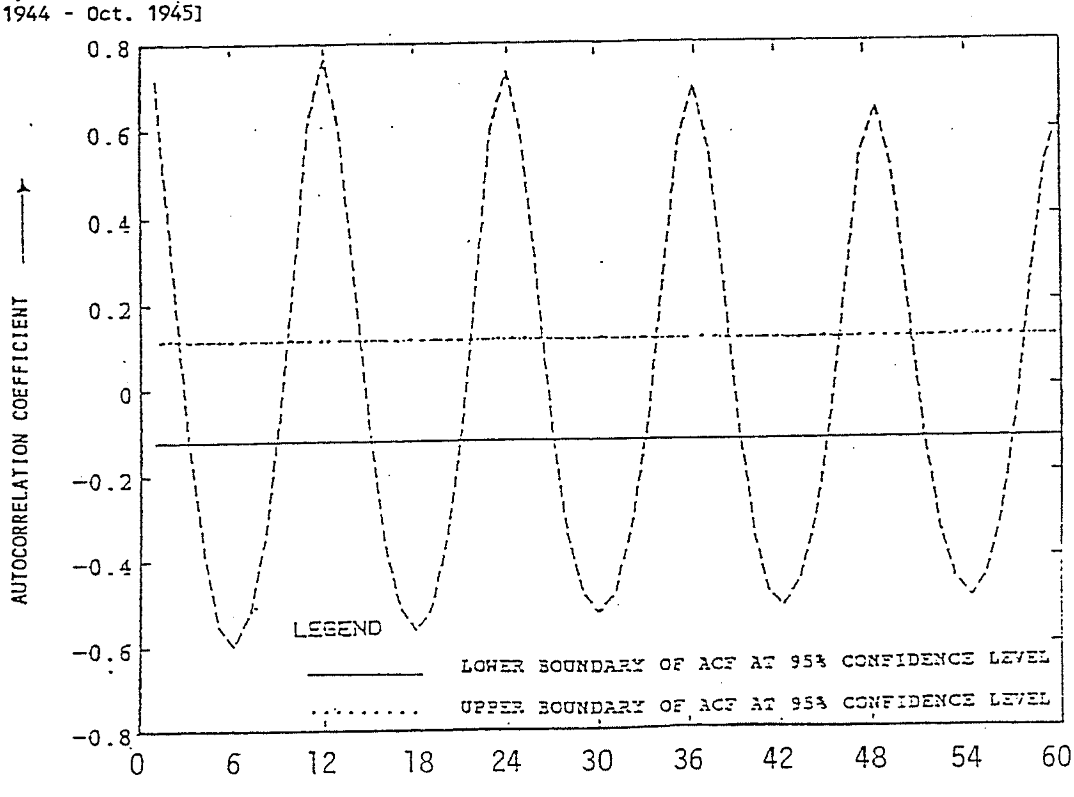


Fig.4.3.3.(a): Correlogram of Green R.

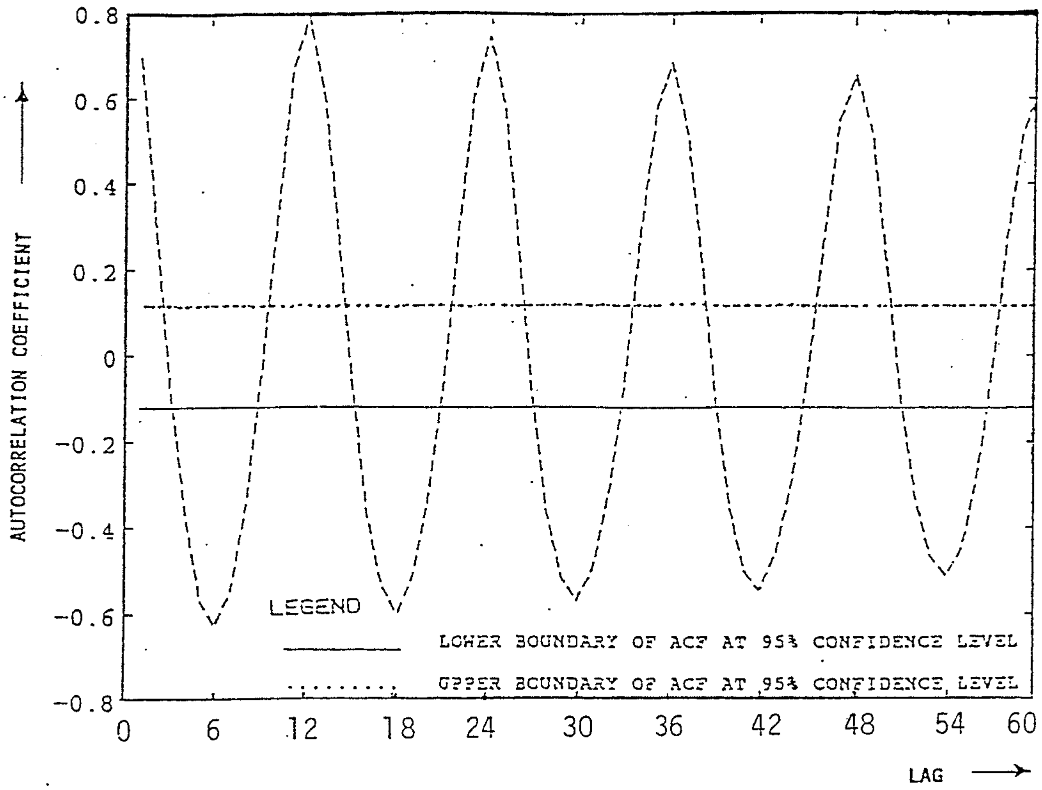


Fig.4.3.3.(b): Correlogram of Soo R.

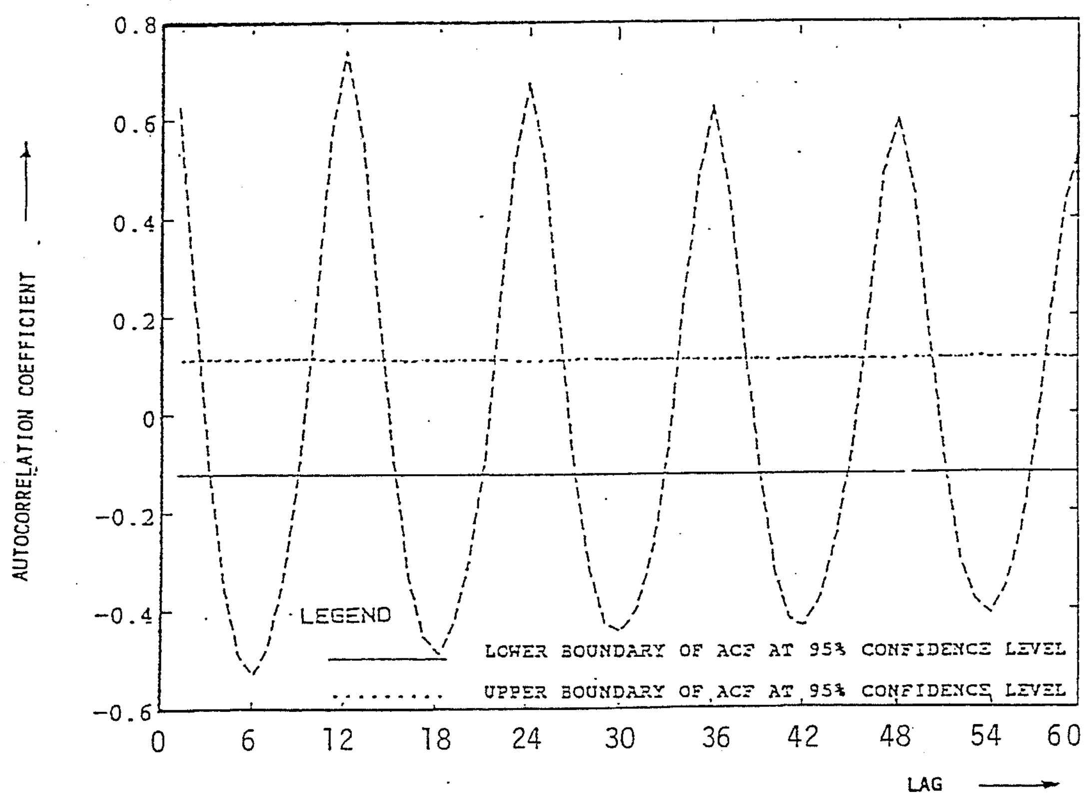


Fig.4.3.3.(c): Correlogram of Rutherford Ck.

Table 4.3.2: Seasonal Segmentation of the time wave form of members of LILL cluster

River	Number of assignment to seasonal group	Month											
		Nov.	Dec.	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sep.	Oct.
Green River	Group 1	18	23	22	23	23	14	0	0	0	0	3	12
	Group 2	5	0	1	0	0	9	23	23	23	23	20	11
Soo River	Group 1	17	21	23	23	23	17	1	0	0	0	4	9
	Group 2	6	2	0	0	0	6	22	23	23	23	19	14
Rutherford Creek	Group 1	19	21	22	23	23	16	1	0	0	0	3	9
	Group 2	4	2	0	0	0	7	22	23	23	23	20	14

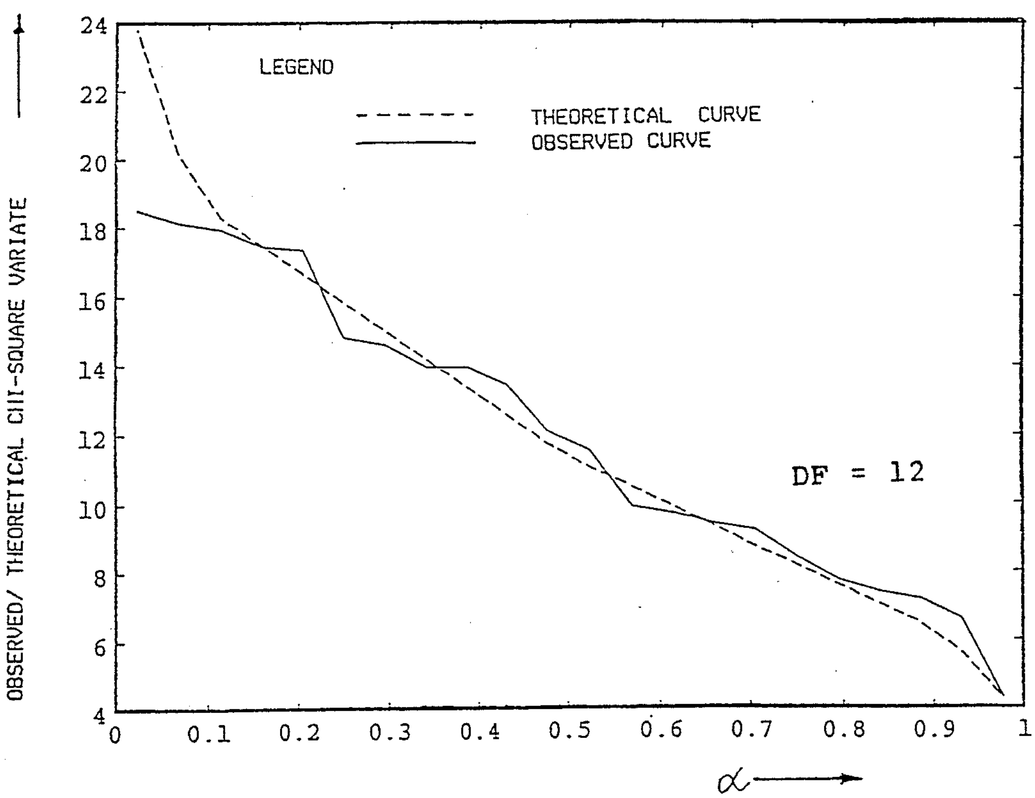


Fig. 4.3.4.1: Chi-square probability plot of Mahalanobis distance of the simultaneously observed dry seasonal segments of {Green R., Soo R.} pair

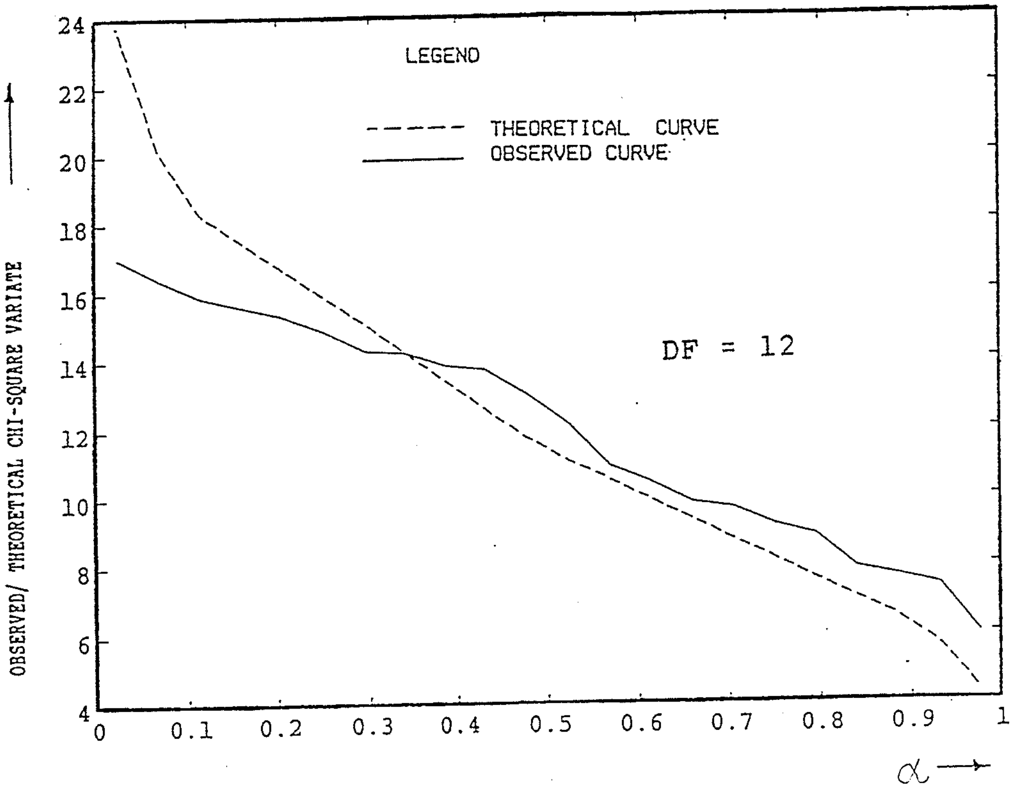


Fig. 4.3.4.2: Chi-square probability plot of Mahalanobis distance of the simultaneously observed wet seasonal segments of {Green R., Soo R.} pair

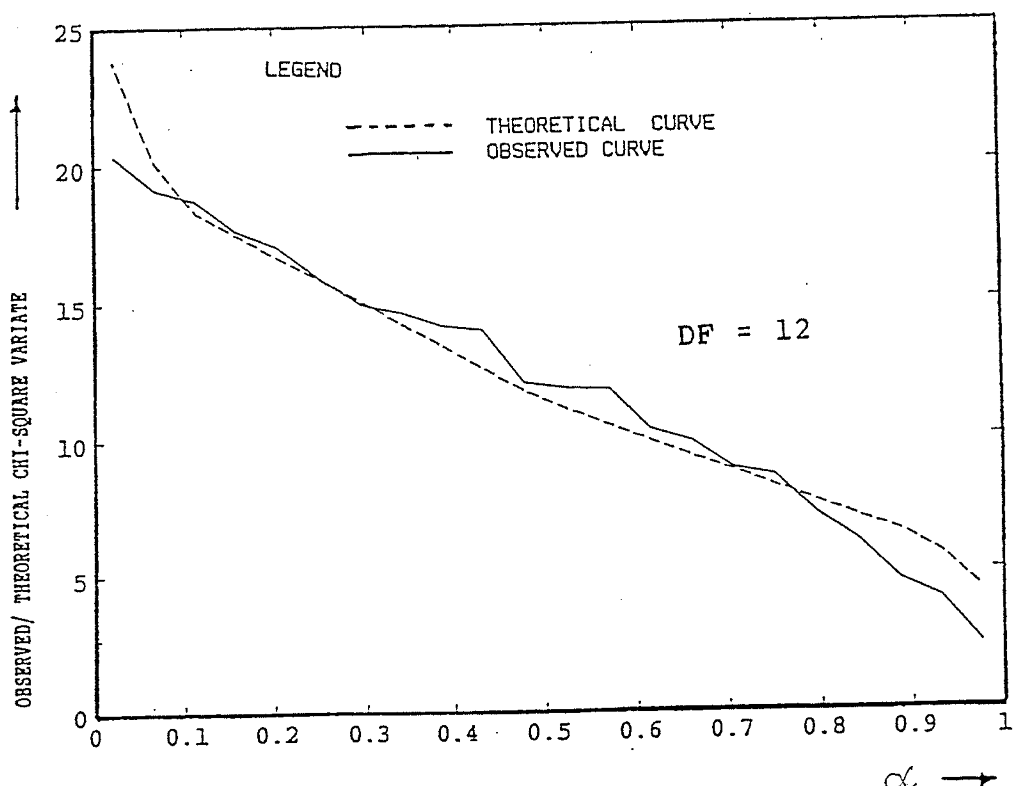


Fig. 4.3.4.3: Chi-square probability plot of Mahalanobis distance of the simultaneously observed dry seasonal segments of {Green R., Rutherford Ck.} pair

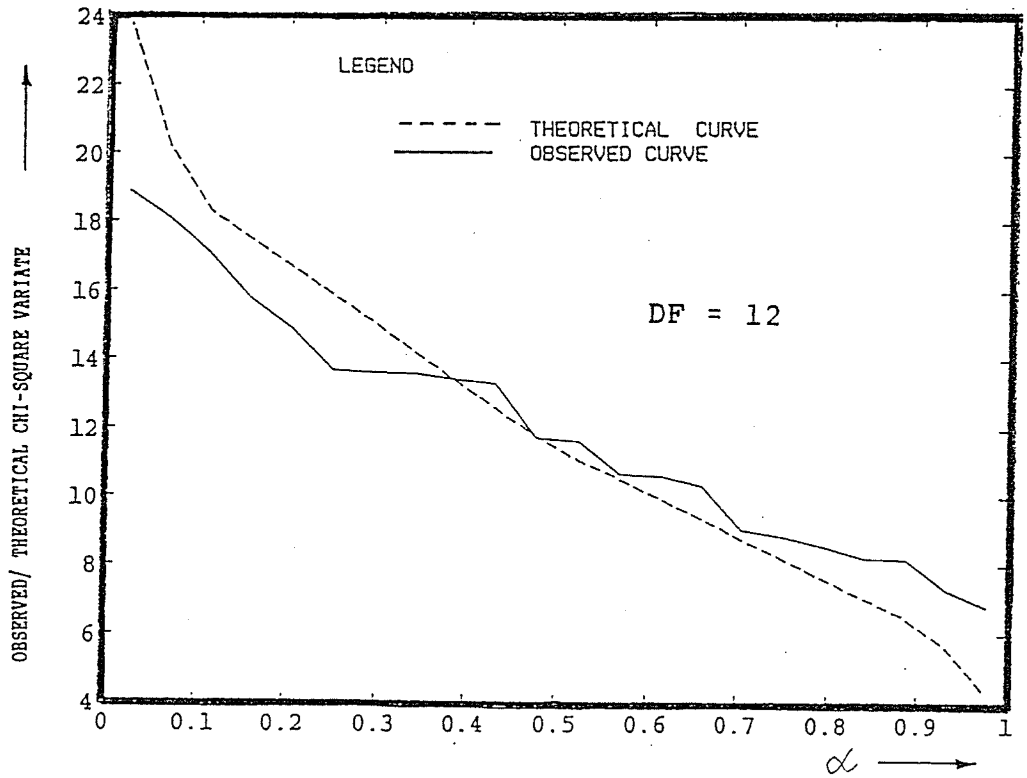


Fig. 4.3.4.4: Chi-square probability plot of Mahalanobis distance of the simultaneously observed wet seasonal segments of {Green R., Rutherford Ck.} pair

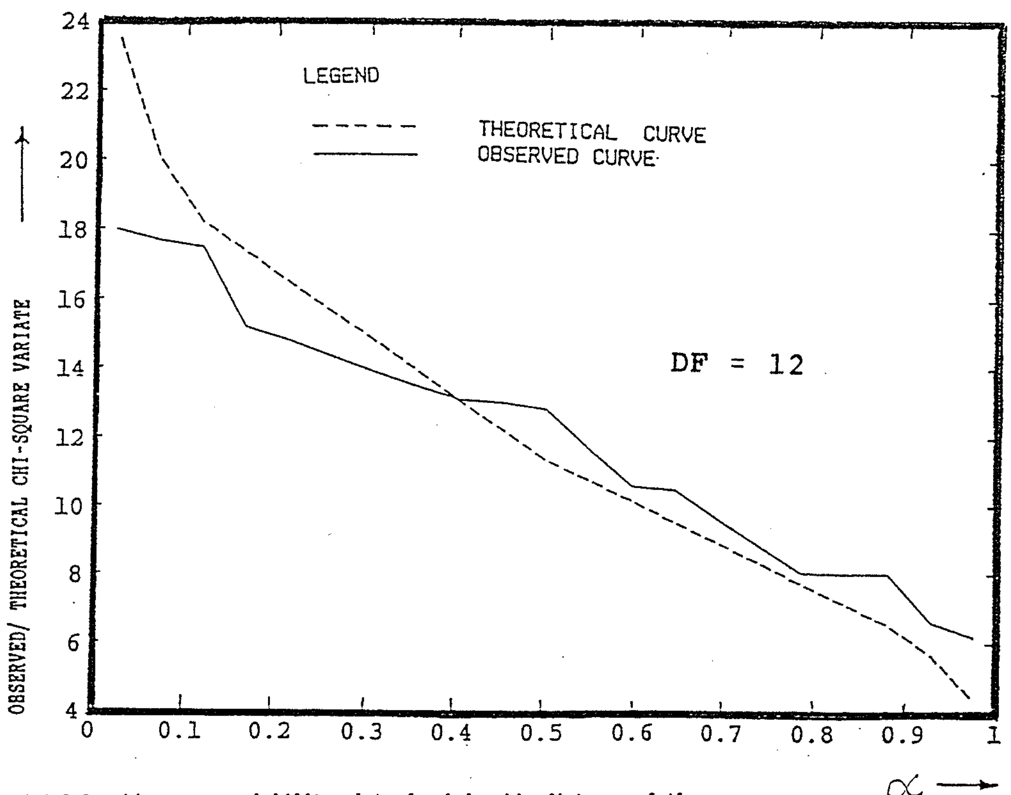


Fig. 4.3.5.1: Chi-square probability plot of Mahalanobis distance of the transitional seasonal segments corresponding to Wet-Dry seasonal transition of Green R.

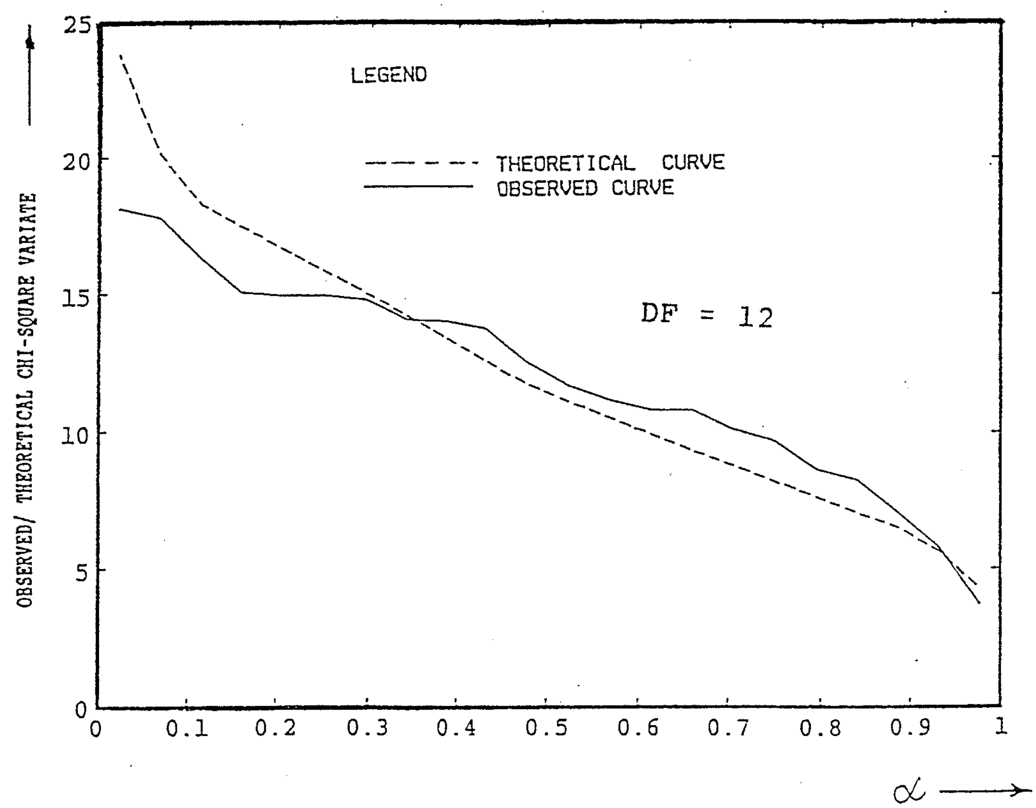


Fig. 4.3.5.2: Chi-square probability plot of Mahalanobis distance of the transitional seasonal segments corresponding to Dry-Wet seasonal transition of Green R.

Table 4.3.3.1 : Infilling of missing monthly mean discharge [m3/s] of Green R. by various models incorporating Soo R. information

Infilling Models	Monthly mean Discharge [m3/s] for each month of missing period [Nov.1944-Oct1945]											
	1	2	3	4	5	6	7	8	9	10	11	12
REG	33.3	18.8	11.6	11.9	8.2	7.8	79.8	109.2	110.4	62.5	24.0	20.5
SREG	31.8	19.0	12.7	12.9	9.7	9.3	78.8	108.6	109.8	61.3	22.6	19.1
SSREG	31.7	19.0	12.1	12.4	8.6	8.2	80.7	114.0	115.3	61.6	21.1	17.6
AMOVE	33.9	19.1	11.6	11.9	7.9	7.5	80.0	108.6	109.7	63.0	24.4	20.8
SMOVE	33.0	19.5	12.3	12.6	8.7	8.2	78.8	108.5	109.8	61.3	22.6	19.1
SSMOVE	33.0	19.5	12.3	12.6	8.7	8.3	80.5	113.6	114.9	61.5	21.1	17.6
MULBS	31.3	22.0	11.2	10.6	9.4	10.6	80.9	124.3	113.0	66.1	36.5	17.3
OBSERVED	35.4	20.3	15.4	13.3	9.0	12.5	80.2	103.0	105.0	72.2	31.7	22.4

Table 4.3.3.2: Infilling of missing monthly mean discharge [m3/s] of Green R. by various models incorporating Rutherford Ck. information

Infilling Models	Monthly mean Discharge [m3/s] for each month of missing period [Nov.1944-Oct.1945]											
	1	2	3	4	5	6	7	8	9	10	11	12
REG	32.3	23.8	20.2	19.5	16.1	17.2	87.1	131.7	129.5	53.7	29.4	24.3
SREG	29.1	21.0	17.6	16.9	13.7	14.8	87.4	126.0	123.9	58.8	37.7	33.3
SSREG	31.2	21.7	17.5	16.7	12.7	14.0	94.0	140.4	137.9	59.5	34.1	28.8
AMOVE	31.1	20.9	16.5	15.7	11.6	12.9	94.4	144.8	142.2	56.3	27.6	21.5
SMOVE	30.9	21.6	17.4	16.6	12.6	13.9	89.8	132.0	129.9	56.4	29.6	23.7
SSMOVE	32.9	22.6	18.0	17.2	12.8	14.3	95.8	147.0	144.3	57.1	27.9	21.7
MULBS	34.7	20.5	18.2	16.3	12.8	17.5	96.8	131.9	104.7	65.1	35.8	26.6
OBSERVED	35.4	20.3	15.4	13.3	9.0	12.5	80.2	103.0	105.0	72.2	31.7	22.4

Table 4.3.3.3: Infilling of missing monthly mean discharge [m3/s] of Green R. by SESTRNALL model

Infilling Model	Monthly mean Discharge [m3/s] for each month of missing period [Nov. 1944 - Oct. 1945]											
	1	2	3	4	5	6	7	8	9	10	11	12
INFILLED BY SESTRNALL MODEL	31.2	21.1	12.9	12.4	10.7	29.8	69.4	111.3	99.6	69.9	51.4	38.1
OBSERVED	35.4	20.3	15.4	13.3	9.0	12.5	80.2	103.0	105.0	72.2	31.7	22.4

CHAPTER 5

RESULTS AND DISCUSSIONS

5.0 ASSESSMENT PROCEDURE

In this Chapter, the results of the applications of various models are evaluated according to the following procedure: A year period of monthly stream-flow data is assumed to be missing which is reconstructed by various models utilizing a range of information sources. Three categories of assessment are made based on following viewpoints:

1. For a particular model, and a particular case of {river with missing data, information source} pair, the infilling quality assessment is made based on purely statistical considerations relevant to the model.

2. For a particular case of {river with missing data, base river} pair, for each of regression and MOVE.4 models, the comparison is made on the basis of the infilling performance under various sampling scenarios.

3. For commensurate models [multivariate models, each of REG, SREG, SSREG, AMOVE, SMOVE and, SSMOVE model], for a particular river with missing data, the comparison is made on the basis of the infilling performance by various information sources.

ASSESSMENT CATEGORY 1

STATISTICAL ASSESSMENT OF MULTIVARIATE MODELS

ASSESSMENT OF CONSISTENCY IN THE RELATIONSHIP

Markovian seasonal transition together with the consistency in the simultaneous occurrence of seasonal segments of a particular severity is considered as information source. The consistent nature of inter seasonal dependence is imposed in SESTRNALL model. In another words, SESTRNALL model assumes that the observation of seasonal segments of a certain severity at any season **(k-1)** at period **(T-1)** can predict the severity of the following seasonal segment at period **T**. MULBS models on the other hand, assumes consistency in simultaneous observation of seasonal segments of certain severity. Entropy in discrete form is used to quantify the consistency.

Entropy is defined as a measure of uncertainty of a system. For a system **X** with $x_1, x_2, \ldots\ldots\ldots x_w$, **w** different possible system states, the probability p_k of occurrence of any system state x_k **(k=1.....w)**, is the only known information. The uncertainty or entropy of the system is given below [Kinchin (1957), Jones (1979)].

$$H = -\sum_{k=1}^{w} P_k \ln P_k \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots [5.1]$$

A sequence of class-membership index [chapter 3.1] representing the sequence of state or severity of the seasonal segment in discrete form is considered as the **system.** Entropy of such a system based on the knowledge of the occurrence probability of the seasonal segment of certain severity ,is given as follows:

$$H_Y = -\sum_{k=1}^{w} \sum_{p=1}^{n_{Yk}} P(c_{Ypk}) \ln P(c_{Ypk}) \quad \ldots\ldots\ldots\ldots [5.2]$$

where,

w = total number of seasons per year

n_{Yk} = total number of sub-clusters in any season k

c_{Ypk} = a particular class-membership index of a seasonal segment observed in season k

For a system following a Markovian transition, the information contained in consistent transition of the seasonal segments would theoretically contribute to the reduction of uncertainty of the system. The entropy of a system following Markovian transition is given below.

$$H_{YM} = \sum_{i=1}^{n} \sum_{j=1}^{n} P_i \, P_{ij} \ln P_{ij} \quad \ldots\ldots\ldots\ldots [5.3]$$

where,

n = total number of system states ($n= k*n_{Yk}$)

P_i = probability of observation of any system state i

P_{ij} = probability of transition of state i to the state j

In case of a system **Y** incorporating cross river information transfer, reduction in entropy of **Y** is expected by conditioning on the base river **X**. Thus, considering the sequences of class-membership index of the series **X** and **Y** as sequences of system state in discrete form, the entropy of **Y** conditioned on **X** is given below.

$$H_{Y|X} = \sum_{k=1}^{w} \sum_{q=1}^{nXk} \sum_{p=1}^{nYk} P(c_{Ypk} \mid c_{Xqk}) \ln P(c_{Ypk} \mid c_{Xqk}) \quad \text{....[5.4]}$$

where,

w = total number of seasons per year

n_{Xk} = total number of sub-cluster of series X in any season k

n_{Yk} = total number of sub-clusters of series Y in any season k

$P(c_{Ypk} \mid c_{Xqk})$ = probability of occurrence of seasonal segment of type c_{Ypk} of series Y conditioned on the simultaneous observation of seasonal segment of type c_{Xqk} of series X

Reliability of the information source is measured by the percentage reduction of system entropy by conditioning on the information source. For SESTRNALL and MULBS models, the percentage reduction in system entropy are computed by Equation 5.5 and Equation 5.6 respectively.

$$H_{RM} = \frac{(H_Y - H_{YM})}{H_Y} \times 100. \quad \text{.............[5.5]}$$

$$H_{RB} = \frac{(H_Y - H_{Y|X})}{H_Y} \times 100. \quad \text{...............[5.6]}$$

The greater the percentage reduction in system entropy, the greater the reliability that can be expected to be associated with the information source.

Multivariate level infilling assessment is done by the Mahalanobis distance of the observed seasonal segment with respect to the corresponding predicted configuration of the missing segment. The Mahalanobis distance is a multivariate analogue of the standard normal deviate **z**. This distance corresponds to the probability contour on which the observed seasonal segment lies with respect to the predicted distributional configuration . The smaller this distance is, the narrower the probability contour (hyper-ellipse) will be and the more accurate will be the quality of prediction [Fig.5.1.0]. This figure shows that any point $\mathbf{X_1}$ lying on contour $\mathbf{P_1}$% has a constant Mahalanobis distance which is less than the distance of any point $\mathbf{X_2}$ lying on the $\mathbf{P_2}$% contour. For a **m** months long seasonal segment, the limiting probability contour corresponding to a certain variate X can be obtained from the relationship given by the following relationship [Johnson and Wichern (1988)]:

Solid ellipsoid of X values satisfying :

$$(X-\mu)^T \Sigma^{-1} (X-\mu) \le \chi^2_m(\alpha) \text{ has a probability of } (1-\alpha) \text{ ...[5.7].}$$

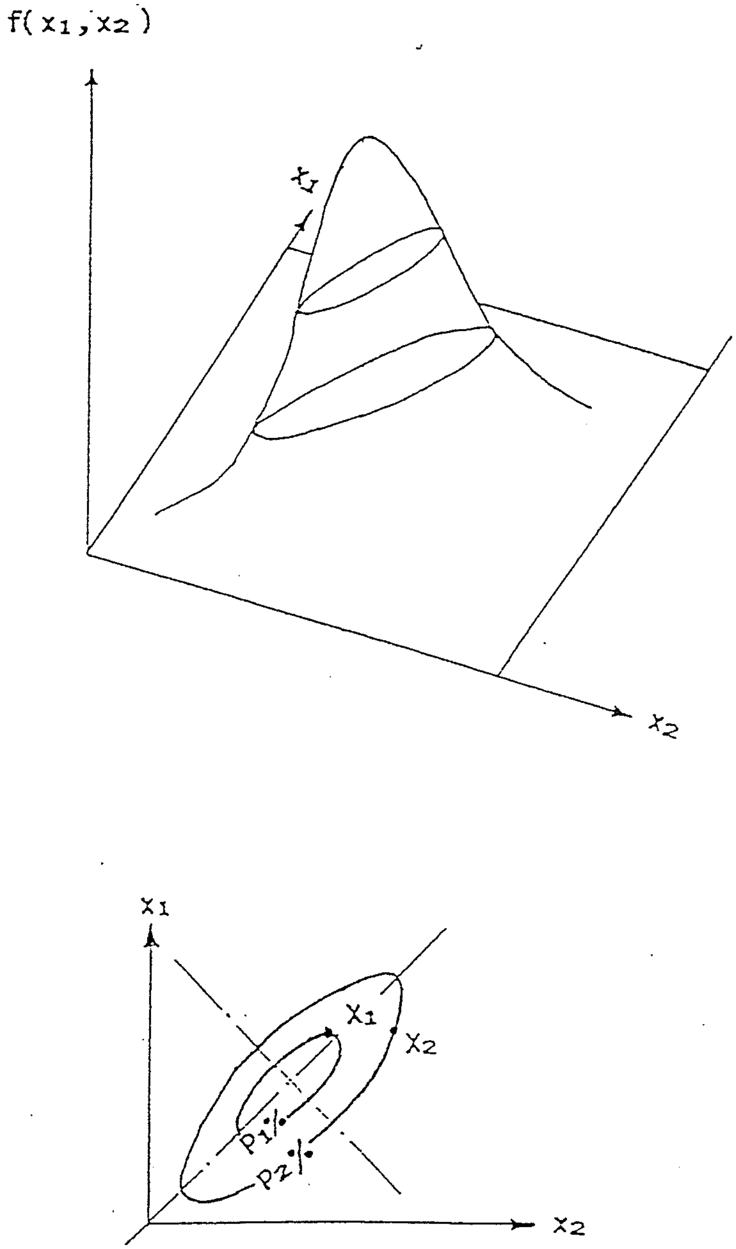


Fig.5.1.0.: Mahalanobis distance and probability Contour in case of of a bivariate normal distribution

where,

μ and Σ are the conditional mean and covariance matrix computed from Equation 3.7

Univariate level assessment is performed by finding the probability band on which each element of the observed seasonal segment lies with respect to the marginal univariate configuration of corresponding predicted element. The band is determined by the elements of the conditional mean vector and the square root of diagonal elements of the conditional covariance matrix Equation 3.7.

STATISTICAL ASSESSMENT OF THE REGRESSION MODELS

For each of the three sampling scenarios of the regression models, standard regression analysis incorporating following assessments:

1. Inference about the regression parameters (t - test)
2. Residual analysis
   - normality of the residual (Chi-Square test of goodness of fit, normal plot)
   - homoscedasticity of the residual (Plot of residual)
   - whiteness of residual (ACF of the residual)
3. Correlation coefficient between the predicted and predictor variables.
4. Quality assessment of prediction is done by finding the

probability band within which the observed element resides corresponding to the configuration of prediction of the corresponding element of the seasonal segment (given by the estimates of the prediction mean and the prediction variance of the corresponding element).

STATISTICAL ASSESSMENT OF THE MOVE.4 MODELS

No procedure of statistical assessment specific to the model is described in the source literature [Vogel- Stedinger (1985)]. No statistical analysis is therefore done for MOVE.4 model under any of the various sampling scenarios.

ASSESSMENT CATEGORY 2

For each case of particular { river with missing data, base river} pair, for each of the MOVE.4 and regression model, the comparison of the quality of infilling under various sampling scenarios are assessed on basis of the following criteria:

a.) Estimation error in the dry season:

$$ed= \{ 1/6(\Sigma((\hat{y}i - yi)/yi)^2 \}^{0.5} \quad [i=1....6]$$

b.) Estimation error in the wet season:

$$ew= \{ 1/6(\Sigma((\hat{y}i - yi)/yi)^2 \}^{0.5} \quad [i=7.....12]$$

c.) Deviation at the peak value:

$$P= (\hat{y}p-yp)$$

d.) Overall estimation error:

$$eo = \{ 1/12(\Sigma((\hat{y}i - yi)/yi)^2 \}^{0.5} \qquad [i=1.....12]$$

For a particular pair of {base river, short data river}, for each of the regression and AMOVE.4 model, the corresponding model varieties are ranked in descending of performance on basis of ed, ew, P and eo.

The plots of the infilled versus the observed data also provides an approximate measures the relative performance of the varieties of regression model for a fixed case of {a base river, the river with missing data} in terms of the deviation of the plot of the infilled data from the plot of the observed data.

ASSESSMENT CATEGORY 3

For a particular river with missing data, comparison of infilling by the SESTRNALL model and MULBS is made. For each of the sampling scenarios of the regression and MOVE.4 model, comparison of quality of infilling among various base rivers are made. Comparison is made on the basis of ed, ew, P and eo.

For a particular model, the relevant information sources are ranked in descending order of model performance measured in terms **ed, ew, P and eo.**

5.1 ASSESSMENT OF QUALITY OF INFILLING OF MISSING DATA IN THE RIVER BELONGING TO THE UPB CLUSTER

ASSESSMENT CATEGORY 1

Table 5.1.1.0 shows that in all cases of MULBS model, the observed seasonal segment lies within the 95% probability contour of prediction. In case of the SESTRNALL model, the observed dry seasonal segment lies within the 95% contour but the wet seasonal segment lies outside the 99.5% contour.

Univariate level assessment in case of MULBS model [Table 5.1.1.(1, 2 and 3)] shows that for the {Craig Ck., Johns Ck.} pair, the observed data corresponding to 1st element of wet seasonal segment lie within the 97.5% contour of marginal prediction while the observed data corresponding to the rest of the elements of both wet and dry seasonal segments lie within the 95% band. In the case of {Craig Ck., Dunlap Ck.} pair, the observed data corresponding to the elements of both dry and wet seasonal segment lie within the 95% prediction band of corresponding element. In case of the {Craig Ck., Cowpasture R.} pair, observed data corresponding to the 4th element of dry seasonal segment lie within 98.6% band and the observed data corresponding to the rest of the elements of both dry and wet seasonal segment lie within the 95% band. Univariate level assessment in case of SESTRNALL model [Table 5.1.1.4] shows that the observed data corresponding to the 1st element of wet seasonal segment lies within the 99% band , the observed data corresponding to the 5th element of the wet seasonal segment lies beyond the 99.9% band while the observed data corresponding to the rest of the elements of both dry and wet seasonal segments lie within the 95% band of prediction.

The entropy reduction varies within the 59% - 78.88% range in case of conditioning on the base river in contrast to a 5% reduction under consideration of Markovian inter seasonal transition [Table 5.1.2].

A summary of the results of the REG, SREG models [Table 5.1.3.(1 and 2)] show that all observed elements are contained within the 95% prediction band. Summary of SSREG model [Table 5.1.3.3] shows that in case of the {Craig Ck., Cowpasture R.} pair, the observed data corresponding to 5th element of the wet seasonal segment is contained within the 99% prediction band while the rest of the observed data are contained within the 95% prediction band. In case of both {Craig Ck., Johns Ck.} and {Craig Ck., Dunlap Ck} pairs, for SSREG model, all the observed data lie within the 95% prediction band.

Table 5.1.1.0 : Distance of Observed seasonal segment of Craig Creek with respect to the predicted conditional configuration by the base R. by MULBS Model and by self Series by SESTRNALL Model(multivariate basis appraisal)

		Season 1			Season 2		
Model	Information Source	Observed Distance	DF	Comment	Observed Distance	DF	Comment
MULBS	Johns Creek	7.88	6	lying within 95% Contour	5.22	6	lying within 95% Contour
MULBS	Dunlap Creek	10.25	6	lying within 95% Contour	12.0	6	lying within 95% Contour
MULBS	Cowpasture River	9.14	6	lying within 95% Contour	2.46	6	lying within 95% Contour
SESTRNALL	Self Series	10.0	6	lying within 95% Contour	20.0	6	lying outside 99.5% Contour

Table 5.1.1.1: Craig Ck. Data infilling by Johns Ck. by MULBS Model: Predicted vs. Observed data (appraisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Mean Predicted	Element Std. Predicted	Observed Element	Comment
	1	0.022	4.780	0.148	4.740	Contained within 95%
	2	0.014	4.170	0.119	4.178	Contained within 95%
	3	0.069	4.250	0.263	4.225	Contained within 95%
1	4	0.064	6.310	0.253	6.090	Contained within 95%
	5	0.031	4.130	0.176	4.016	Contained within 95%
	6	0.023	5.740	0.152	5.403	Contained within 95%
	1	2540.961	817.457	50.408	712.000	Contained within 97.5%
	2	1563.969	542.656	39.547	503.000	Contained within 95%
	3	2294.180	711.687	47.898	655.000	Contained within 95%
2	4	3254.641	1237.556	57.049	1211.000	Contained within 95%
	5	3652.438	2440.499	60.435	2427.000	Contained within 95%
	6	2198.680	419.416	46.890	402.000	Contained within 95%

Table 5.1.1.2: Craig Ck. Data infilling by Dunlap Ck. by MULBS Model: Predicted vs. Observed data (appraisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Mean Predicted	Element Std. Predicted	Observed Element	Comment
	1	0.113	4.567	0.336	4.736	Contained within 95%
	2	0.052	4.190	0.229	4.179	Contained within 95%
	3	0.153	4.503	0.391	4.225	Contained within 95%
1	4	0.877	4.475	0.936	5.011	Contained within 95%
	5	0.120	4.116	0.346	4.016	Contained within 95%
	6	0.074	5.290	0.271	5.403	Contained within 95%
	1	6898.740	722.090	83.059	712.000	Contained within 95%
	2	2371.040	584.140	48.693	503.000	Contained within 95%
	3	16087.520	763.000	126.837	655.000	Contained within 95%
2	4	15926.500	1350.340	126.200	1211.000	Contained within 95%
	5	20996.660	2361.730	144.902	2427.000	Contained within 95%
	6	9672.580	579.300	98.349	402.000	Contained within 95%

Table 5.1.1.3: Craig Ck. Data infilling by Cowpasture R. by MULBS Model: Predicted vs. Observed data (appraisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Mean Predicted	Element Std. Predicted	Observed Element	Comment
	1	0.154	4.784	0.392	4.736	Contained within 95%
	2	0.101	4.214	0.318	4.179	Contained within 95%
	3	0.110	4.273	0.332	4.225	Contained within 95%
1	4	0.946	3.693	0.973	6.090	Contained within 98.6%
	5	0.098	3.986	0.313	4.020	Contained within 95%
	6	0.126	5.124	0.355	5.400	Contained within 95%
	1	0.056	6.411	0.236	6.568	Contained within 95%
	2	0.047	6.096	0.216	6.221	Contained within 95%
	3	0.036	6.367	0.189	6.485	Contained within 95%
2	4	0.035	6.877	0.188	7.099	Contained within 95%
	5	0.063	7.647	0.252	7.794	Contained within 95%
	6	0.148	5.913	0.385	5.997	Contained within 95%

Table 5.1.1.4: Craig Creek data infilling by SESTRNALL Model Predicted vs. observed data of Craig Creek [appraisal on marginal basis]

Season	Monthly Element	Element Variance Predicted	Element Std. Predicted	Element Mean Predicted	Observed Element	Comment
	1	0.4735	0.6881	4.7923	4.7362	Contained within 95%
	2	0.1598	0.3997	4.2193	4.1790	Contained within 95%
	3	0.3291	0.5737	4.5968	4.2254	Contained within 95%
1	4	0.3179	0.5638	3.9655	5.0106	Contained within 95%
	5	0.5048	0.7105	4.1102	5.0160	Contained within 95%
	6	0.6098	0.7809	4.6488	5.4027	Contained within 95%
	1	0.3020	0.5495	5.2043	6.5681	Contained within 99%
	2	0.5083	0.7130	5.8014	6.2206	Contained within 95%
	3	0.2333	0.4830	6.4631	6.4846	Contained within 95%
2	4	0.2524	0.5024	6.3921	7.0992	Contained within 95%
	5	0.2593	0.5092	5.9406	7.7944	Not Contained within 99.9
	6	0.3341	0.5780	5.9146	5.9965	Contained within 95%

Table 5.1.2: Entropy reduction in class-membership index sequence of Craig Ck.

Case	Information source	Marginal Entropy of class-membership index sequence of Craig Ck.	Conditional Entropy of class-membership index sequence of Craig Ck.	Entropy Reduction [%]
Conditional Entropy	Johns Ck.	1.0292	0.2174	78.88
Conditional Entropy	Dunlap Ck.	1.0292	0.3044	70.42
Conditional Entropy	Cowpasture R.	1.0292	0.4218	59.02
Markovian Entropy	Seasonal transition	1.0292	0.9777	5.00

Table 5.1.3.1: Summary of REG model pertinent to Craig Ck. data infilling

Base River	State	computed model							Residual Analysis				Prediction
		b0	b1	tb0	tb1	DF	Comment	R-sq	DF	Comment	Const. var.(e)	State of ACF(e)	Observed vs. prediction interval
Johns Creek	lntd*	1.49	.92	38.61	106.35	358	b0, b1 signi-ficant at 5% level	.97	69	nonnormal at .5% level	satisfied	auto-correlated	all elements contained within 95% interval
Dunlap Creek	lntd*	1.46	0.88	20.84	58.99	358	b0, b1 signi-ficant at 5% level	.91	69	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
Cowpasture River	lntd*	-0.48	1.02	-3.82	48.35	358	b0, b1 signi-ficant at 5% level	.88	69	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval

* lntd : ln transformed data
** nat : raw data

Table 5.1.3.2: Summary of SREG model pertinent to Craig Ck. data infilling

Base River	Season	State	computed model							Residual Analysis					Prediction
			b0	b1	tb0	tb1	DF	Comment	R-sq	Chi-sq(e)	DF	Comment	Const. var.(e)	State of ACF(e)	Observed vs. prediction interval
Johns Creek	1	lntd*	1.51	.93	22.12	49.98	178	b0, b1 significant at 5% level	.93	68	33	nonnormal at .05% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	1.06	1.00	17.11	82.91	178	b0, b1 significant at 5% level	.97	42.8	33	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
Dunlap Creek	1	lntd*	1.57	.85	12.3	26.12	178	b0, b1 significant at 5% level	.79	39	33	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	nat**	78.85	1.87	4.48	33.67	178	b0, b1 significant at 5% level	86	32.6	33	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
Cowpasture River	1	lntd*	-0.13	.94	-.63	24.03	178	b1 is significant at 5% but drop b0	.76	34.4	33	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	0.07	.94	.32	27.6	178	b1 is significant at 5% but drop b0	.81	37	33	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval

* lntd : ln transformed data
** nat : raw data

- Table 5.1.3.3: Summary of SSREG model pertinent to Craig Ck. data infilling

Base River	Season	State	computed model							Residual Analysis				Prediction	
			b0	b1	tb0	tb1	DF	Comment	R-sq	Chi-sq(e)	DF	Comment	Const. var.(e)	State of ACF(e)	Observed vs. prediction interval
Johns Creek	1	lntd*	1.55	.92	17.44	36.24	148	b0, b1 significant at 5% level	.89	48.4	27	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	1.09	.99	7.02	35.46	52	b0, b1 significant at 5% level	.96	13.74	7	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
Dunlap Creek	1	lntd*	1.71	.81	11.47	20.54	154	b0, b1 significant at 5% level	.73	51.85	28	normal at 1% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	2.10	.78	8.69	18.63	52	b0, b1 significant at 5% level	.87	9.70	7	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
Cowpasture River	1	lntd*	0.07	.90	.30	18.68	154	b1 is significant at 5% but drop b0	.69	22.85	28	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	nat**	147.32	0.63	2.82	13.91	58	b0, b1 significant at 5% level	0.77	8.6	9	normal at 5% level	satisfied	auto-correlated	5th element contained within 99% interval while the rest contained within 95% interval

* lntd : ln transformed data
** nat : raw data

ASSESSMENT CATEGORY 2

COMPARISON OF THE REGRESSION MODELS FOR DIFFERENT {RIVER WITH MISSING DATA, BASE R.} PAIRS [Table 5.1.5.a]

In all cases, REG performed best in terms of ed and eo.

In case of the {Johns Ck.Craig Ck.} pair, the SSREG and SREG performed better over REG model in terms of ew and P. In case of {Dunlap Ck., Craig Ck.} pair, SREG performed better over REG only in terms of P. In case of the {Cowpasture R., Craig Ck.} pair, REG performed best with respect to all the four criteria. The plots of the infilled versus observed data are given in Figs.5.1.1.(1,2,3)].

COMPARISON OF MOVE.4 MODELS FOR DIFFERENT {RIVER WITH MISSING DATA, BASE R.} PAIRS

The ranking among the various MOVE.4 models is given in Table 5.1.5.b. The AMOVE performed best in all cases in terms of ed eo. In case of {Johns Ck., Craig Ck,} pair: SSMOVE and SMOVE performed better over AMOVE in terms of ew and P. In case of {Dunlap Ck., Craig Ck,} pair, only SMOVE performed better over AMOVE in terms of ew. Cowpasture R. did not show any beneficial effect with respect to choice of sample, i.e. the models using seasonal sampling criteria did not perform better over the model using heterogenous sample. The plots of the infilled versus observed data are given in Figs.5.1.1.(4,5,6).

Table 5.1.5.a: Table of ranking of performance of varities of regression model (UPB cluster)

River	Model rank in terms of estimation error in dry season [ed]			Model rank in terms of estimation error in wet season [ew]			Model rank in terms of deviation at peak reconstd.- obsvd. [P]			Model rank in terms of overall estimation error [eo]		
Johns Ck.	REG 0.24	SREG 0.27	SSREG 0.27	SSREG 0.06	SREG 0.07	REG 0.10	SREG 14.9	SSREG -55.0	REG -225.3	REG 0.18	SSREG 0.19	SREG 0.20
Dunlap Ck.	REG 0.32	SREG 0.32	SSREG 0.33	REG 0.12	SREG 0.12	SSREG 0.21	SREG -342.2	REG -492.5	SSREG -522.3	REG 0.24	SREG 0.24	SSREG 0.27
Cowpasture R.	REG 0.27	SREG 0.28	SSREG 0.28	SREG 0.21	REG 0.22	SSREG 0.23	REG -382.7	SREG -466.9	SSREG -467.2	REG 0.25	SREG 0.25	SSREG 0.26

Table 5.1.5.b: Table of ranking of performance of varities of MOVE.4 model (UPB cluster)

River	Model rank in terms of estimation error in dry season [ed]			Model rank in terms of estimation error in wet season [ew]			Model rank in terms of deviation at peak reconstd.- obsvd. [P]			Model rank in terms of overall estimation error [eo]		
Johns Ck.	AMOVE 0.23	SSMOVE 0.27	SMOVE 0.28	SSMOVE 0.06	SMOVE 0.07	AMOVE 0.10	SMOVE 15.3	SSMOVE -66.9	AMOVE -165.9	AMOVE 0.18	SMOVE 0.20	SSMOVE 0.20
Dunlap Ck.	AMOVE 0.29	SSMOVE 0.33	SMOVE 0.33	SMOVE 0.12	AMOVE 0.13	SSMOVE 0.20	AMOVE -331.3	SMOVE -377.4	SSMOVE -547.8	AMOVE 0.22	SMOVE 0.25	SSMOVE 0.28
Cowpasture R.	AMOVE 0.28	SMOVE 0.28	SSMOVE 0.28	AMOVE 0.19	SSMOVE 0.23	SMOVE 0.28	AMOVE -115.6	SMOVE -472.0	SSMOVE -576.1	AMOVE 0.24	SSMOVE 0.25	SMOVE 0.28

Table 5.1.4: Zero lag Cross correlation Coefficient between Craig Creek and each of the base rivers

Base River	Cross correlation coefficient corresponding to indiscreminately chosen sample	Cross correlation coefficient corresponding to seasonal sample		Cross correlation coefficient corresponding to selected seasonal sample	
		Season 1	Season 2	Season 1	Season 2
Johns Creek	0.98	0.97	0.99	0.95	0.98
Dunlap Creek	0.95	0.89	0.93	0.86	0.93
Cowpasture River	0.93	0.87	0.90	0.83	0.88

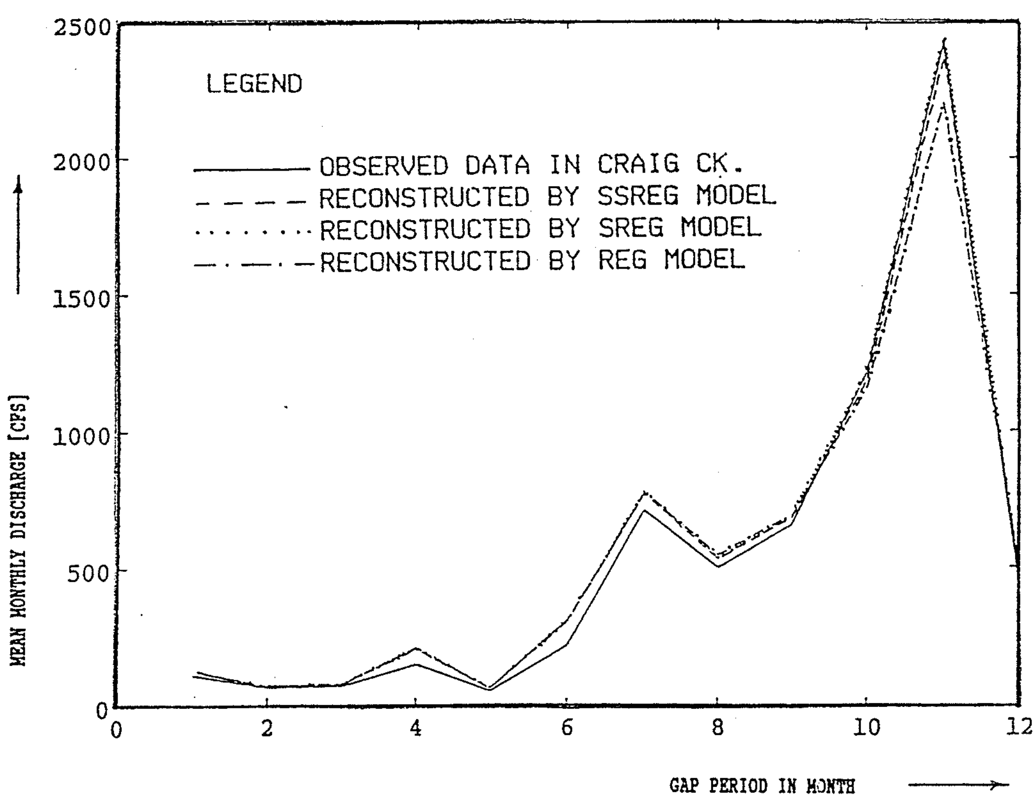


Fig. 5.1.1.1:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by Johns Ck. by REG, SREG, SSREG models (varities of Regression model)

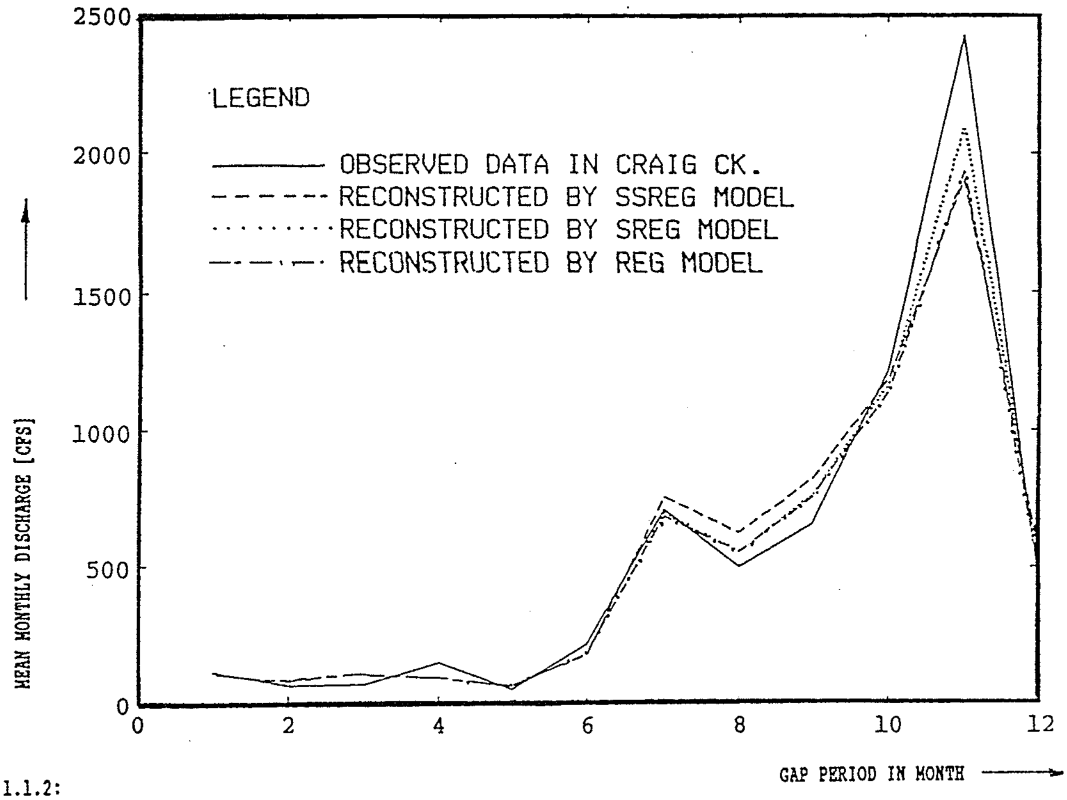


Fig. 5.1.1.2:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by Dunlap Ck. by REG, SREG, SSREG models (varities of Regression model)

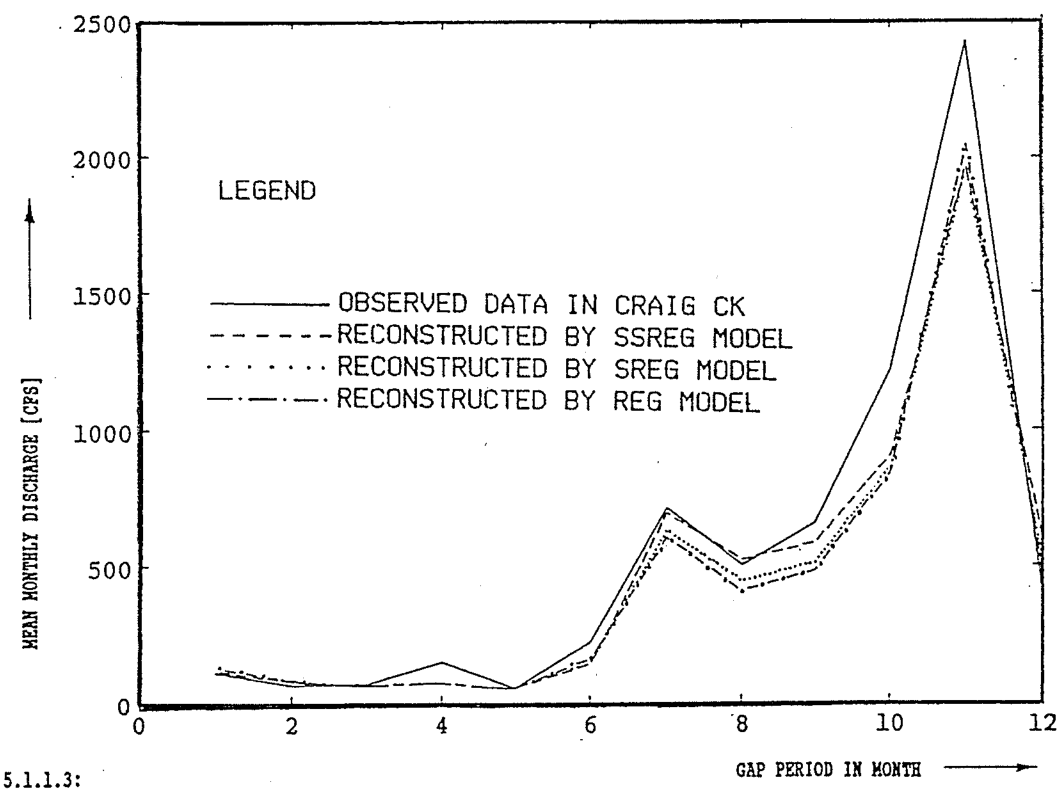


Fig. 5.1.1.3:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by Cowpasture R. by REG, SREG, SSREG models (varities of Regression model)

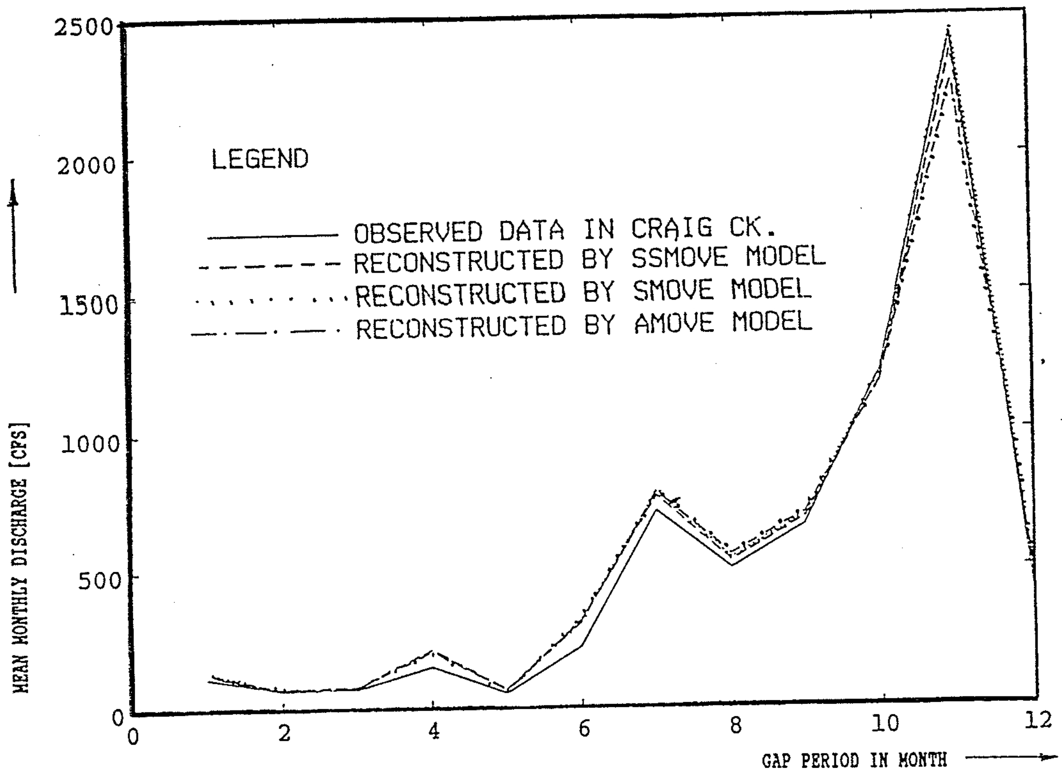


Fig. 5.1.1.4:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by Johns Ck. by AMOVE, SMOVE, SSMOVE models (varities of MOVE.4 model)

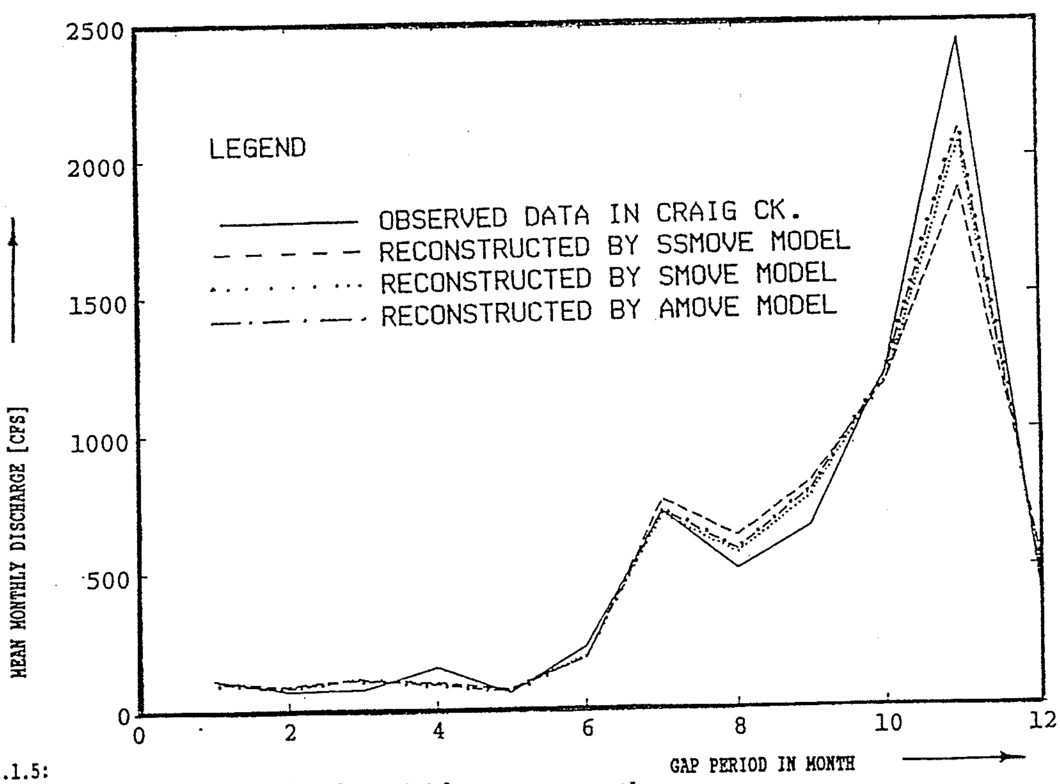


Fig. 5.1.1.5:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by Dunlap Ck. by AMOVE, SMOVE, SSMOVE models (varities of MOVE.4 model)

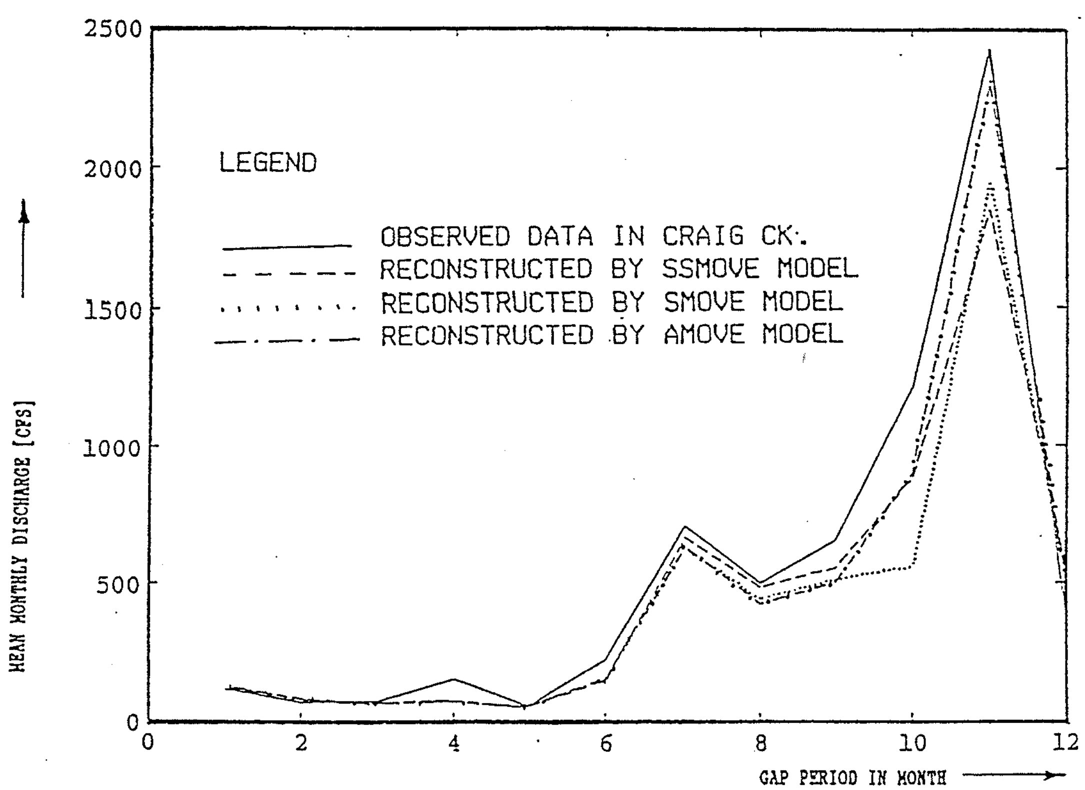


Fig. 5.1.1.6:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by Cowpasture R. by AMOVE, SMOVE, SSMOVE models (varities of MOVE.4 model)

ASSESSMENT CATEGORY 3

MULTIVARIATE MODELS [Table 5.1.6.a]

In all cases, MULBS model performed better than the SESTRNALL model. In case of MULBS model, the quality of reconstruction by incorporating Johns Ck. and Dunlap Ck. information is superior to quality of reconstruction by incorporating Cowpasture R.'s information. The fig. 5.1.2 shows the relative performances among the various sources.

REGRESSION MODEL [Table 5.1.6.b]

Johns Ck. performed best with respect to all the four criteria.

.In case of REG model: Cowpasture R. performed superior to Dunlop Ck. with respect to ed and P. Dunlap Ck. performed superior to Cowpasture R. with respect to ew and eo.

.In case of SREG model: Dunlap Ck. performed superior to Cowpasture R. with respect to ew, P, eo. Cowpasture R. performed superior to Dunlap Ck. with respect to ed.

.In case of SSREG model: Dunlap Ck. performed superior to Cowpasture R. with respect to ew,. Cowpasture R. performed superior to Dunlap Ck. with respect to ed, P and eo.

The figs. 5.1.2.(1,2,3) show the relative performances among the various sources for a particular variety of model.

MOVE MODELS

Johns Ck. performed best with respect to all the four criteria.

.In case of AMOVE model: Cowpasture R. performed better than Dunlap Ck. in terms of ed and P. Dunlap Ck. performed superior to Cowpasture R. with respect to ew and eo. (similar to REG)

.In case of SMOVE model: Dunlap Ck. performed superior to Cowpasture R. with respect to ew, P, eo. Cowpasture R. performed superior to Dunlap Ck. with respect to ed. (similar to REG model).

.In case of SSMOVE model: Dunlap Ck. performed superior to Cowpasture R. with respect to ew, P. Cowpasture R. performed superior to Dunlap Ck. with respect to ed and eo.

The figs. 5.1.1.(4,5,6) show the relative performances among the various sources for a particular variety of model.

Table 5.1.6.a: Table of ranking of performance of various sources for the multivariate models (UPB cluster)

MODEL - -	Source rank in terms of estimation error in dry season [ed]				Source rank in terms of estimation error in wet season [ew]				Source rank in terms of deviation at peak reconstd.- obsvd. [P]				Source rank in terms of overall estimation error [eo]			
	J.Ck.	D.Ck.	C.R.	S.S.	J.Ck.	C.R.	D.Ck.	S.S.	J.Ck.	D.Ck.	C.R.	S.S.	J.Ck.	D.Ck.	C.R.	S.S.
Multiv. M.	0.18	0.23	0.27	0.39	0.08	0.14	0.21	0.52	-13.5	-65.3	-331.7	-2046.8	0.14	0.22	0.22	0.46

Table 5.1.6.b: Table of ranking of performance of various base R. for each sampling scenario, regression model (UPB cluster)

MODEL - -	River rank in terms of estimation error in dry season [ed]			River rank in terms of estimation error in wet season [ew]			River rank in terms of deviation at peak reconstd.- obsvd. [P]			River rank in terms of overall estimation error [eo]		
	J.Ck	C.R	D.Ck.	J.Ck.	D.Ck.	C.R.	J.Ck.	C.R.	D.Ck.	J.Ck.	D.Ck.	C.Ck.
REG	0.24	0.27	0.32	0.10	0.12	0.22	-225.3	-382.7	-492.5	0.18	0.24	0.25
	J.Ck.	C.R.	D.Ck.	J.Ck.	D.Ck.	C.R.	J.Ck.	D.Ck.	C.R.	J.Ck.	D.Ck.	C.R.
SREG	0.27	0.28	0.32	0.07	0.12	0.21	14.9	-342.2	-466.9	0.20	0.24	0.25
	J.Ck.	C.R.	D.Ck.	J.Ck.	D.Ck.	C.R.	J.Ck.	C.R.	D.Ck.	J.Ck.	C.R.	D.Ck.
SSREG	0.27	0.28	0.33	0.06	0.21	0.23	-55.0	-467.2	-522.3	0.19	0.26	0.27

Table 5.1.6.c: Table of ranking of performance of various base R. for each sampling scenario, MOVE.4 model (UPB cluster)

MODEL - -	River rank in terms of estimation error in dry season [ed]			River rank in terms of estimation error in wet season [ew]			River rank in terms of deviation at peak reconstd.- obsvd. [P]			River rank in terms of overall estimation error [eo]		
	J.Ck	C.R	D.Ck.	J.Ck.	D.Ck.	C.R.	J.Ck.	C.R.	D.Ck.	J.Ck.	D.Ck.	C.Ck.
AMOVE	0.23	0.28	0.29	0.10	0.12	0.19	-165.9	-115.6	-331.3	0.18	0.24	0.24
	J.Ck.	C.R.	D.Ck.	J.Ck.	D.Ck.	C.R.	J.Ck.	D.Ck.	C.R.	J.Ck.	D.Ck.	C.R.
SMOVE	0.28	0.28	0.33	0.07	0.12	0.28	15.3	48.0	111.8	0.20	0.25	0.28
	J.Ck.	C.R.	D.Ck.	J.Ck.	D.Ck.	C.R.	J.Ck.	D.Ck.	C.R.	J.Ck.	C.R.	D.Ck.
SSMOVE	0.27	0.28	0.33	0.06	0.20	0.23	-66.9	-547.8	576.1	0.20	0.25	0.28

J.Ck= Johns Ck.
D.Ck.= Dunlap Ck.
C.R.= Cowpasture R.
S.S.= Same River

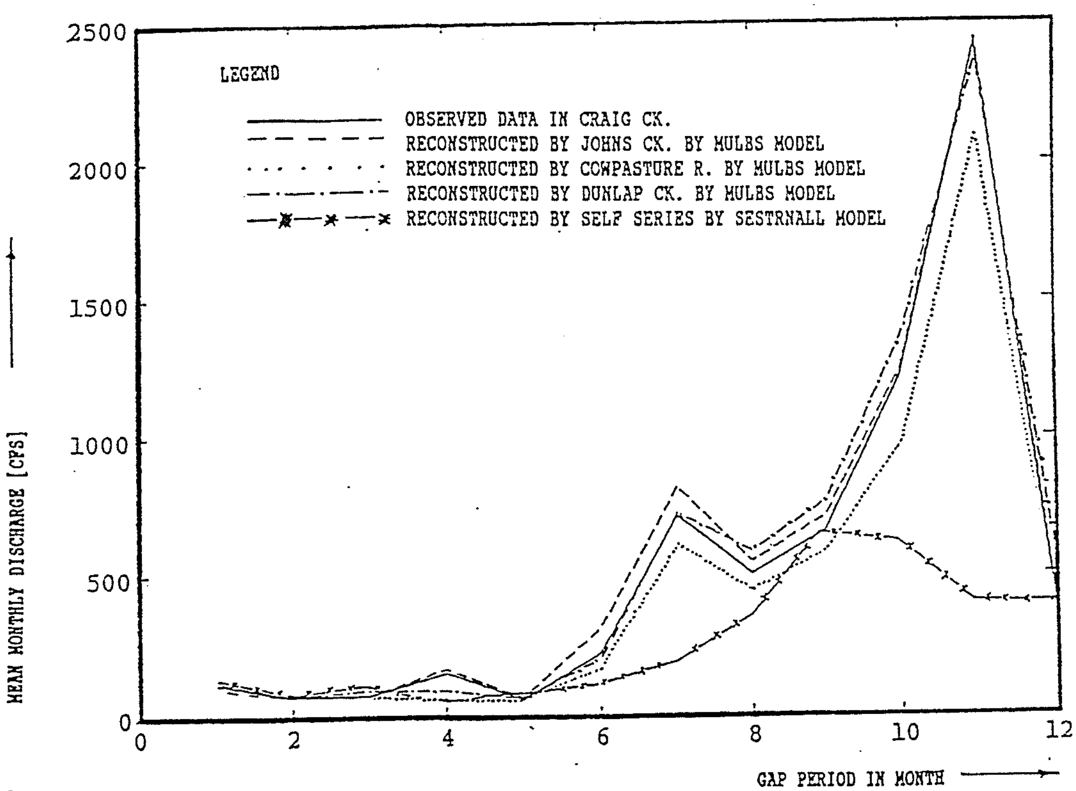


Fig.5.1.2:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by multivariate models using variable information source

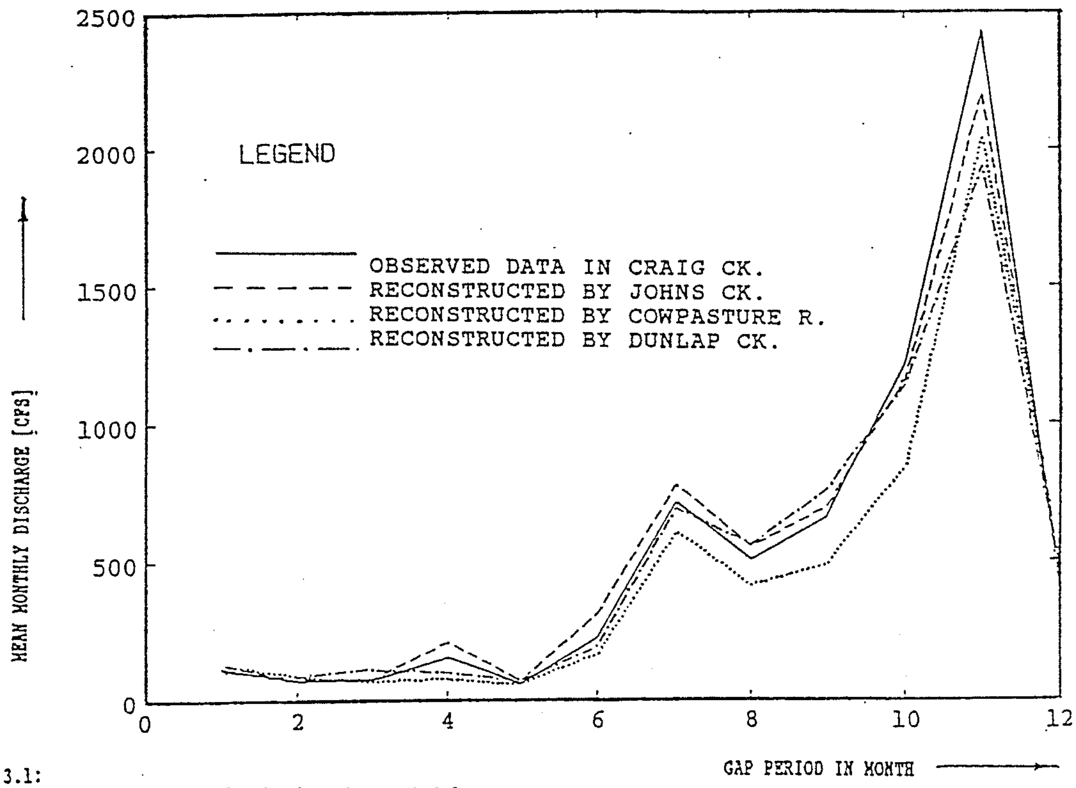


Fig.5.1.3.1:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by REG model using variable information source

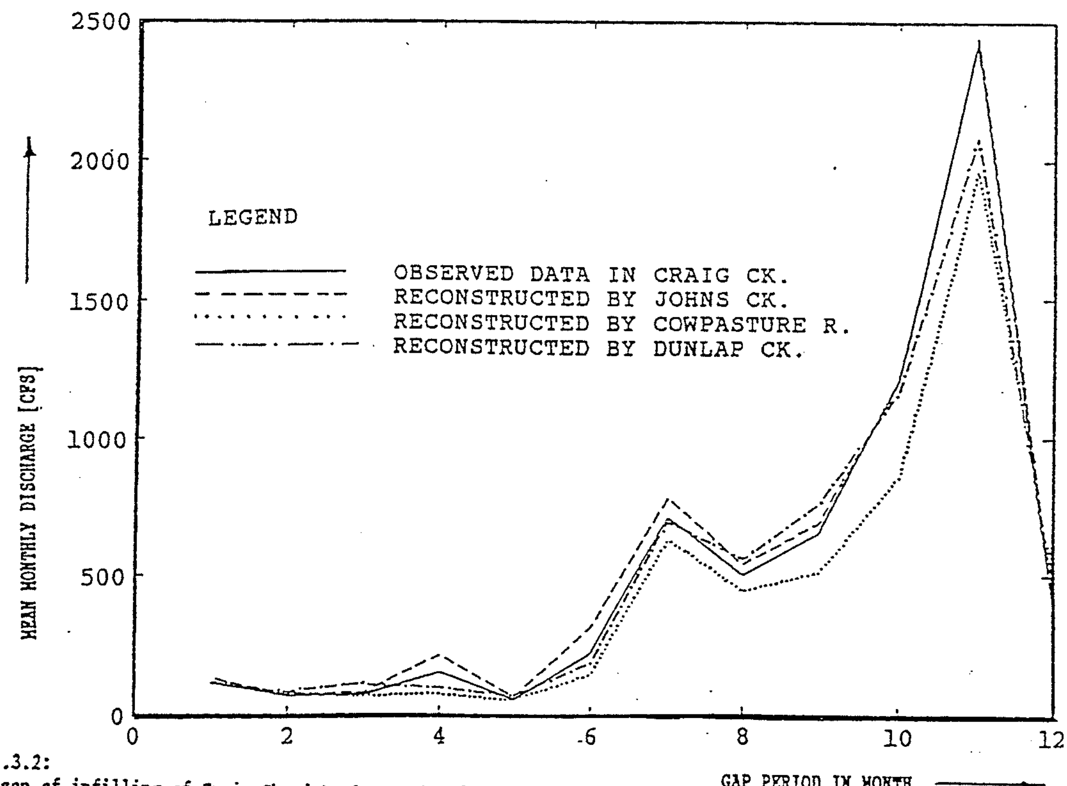


Fig.5.1.3.2:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by SREG model using variable information source

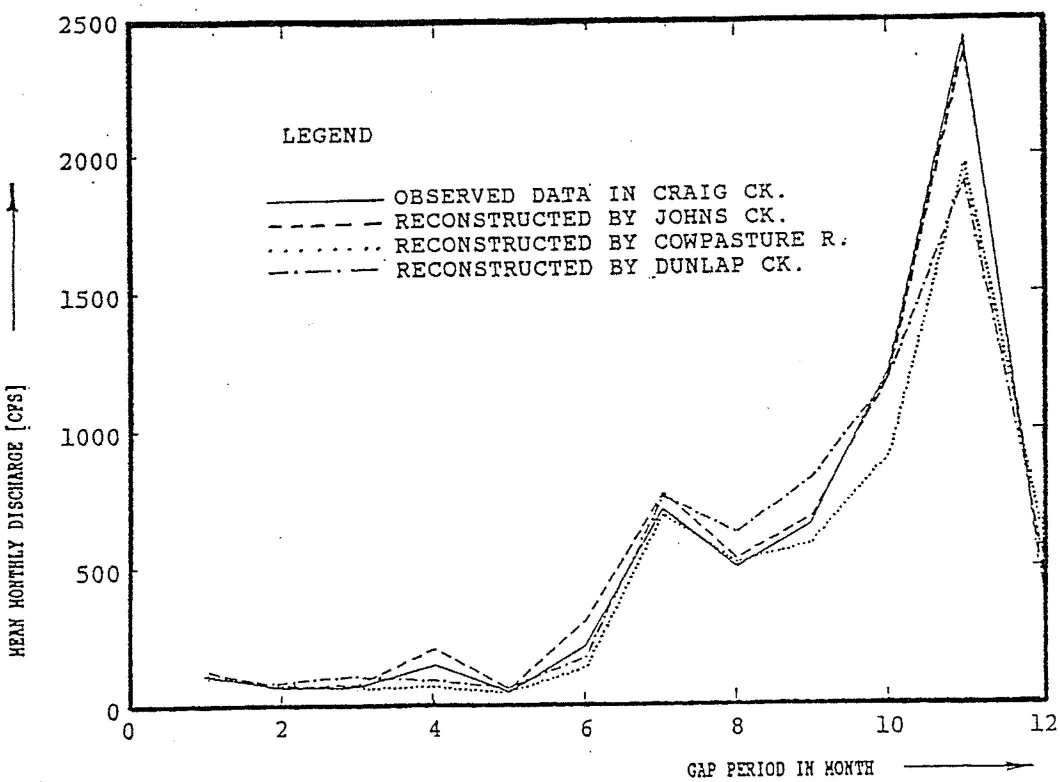


Fig.5.1.3.3:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988]
by SSREG model using variable information source

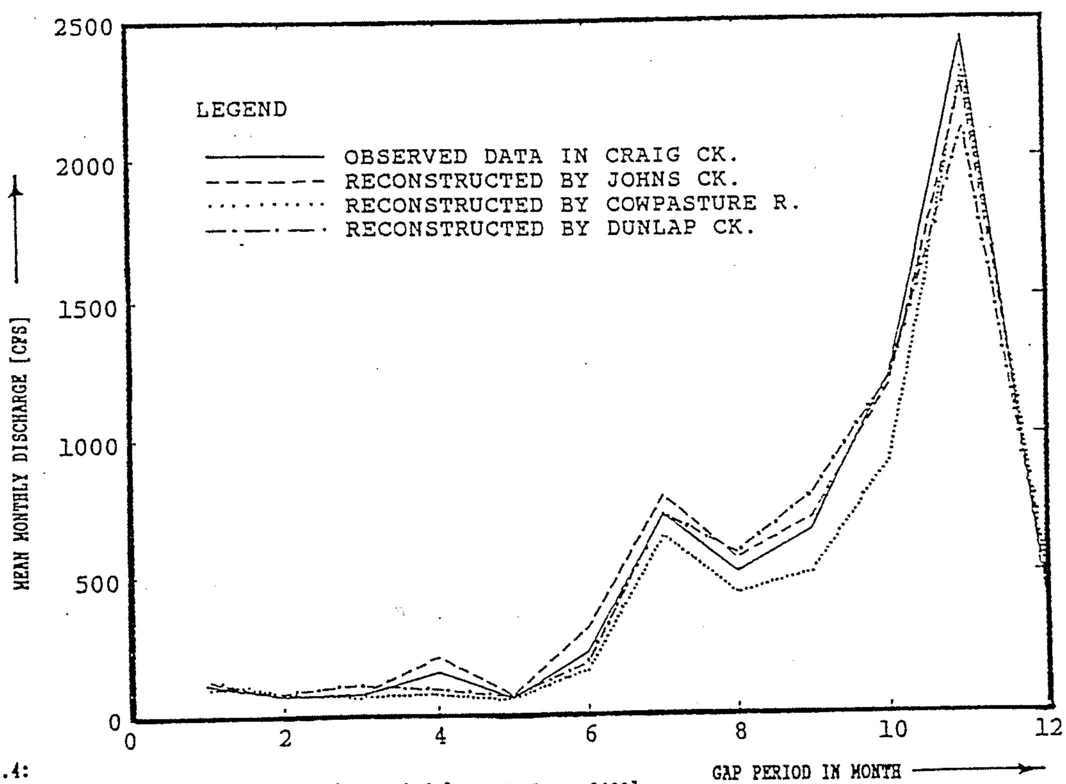


Fig.5.1.3.4:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988]
by AMOVE model using variable information source

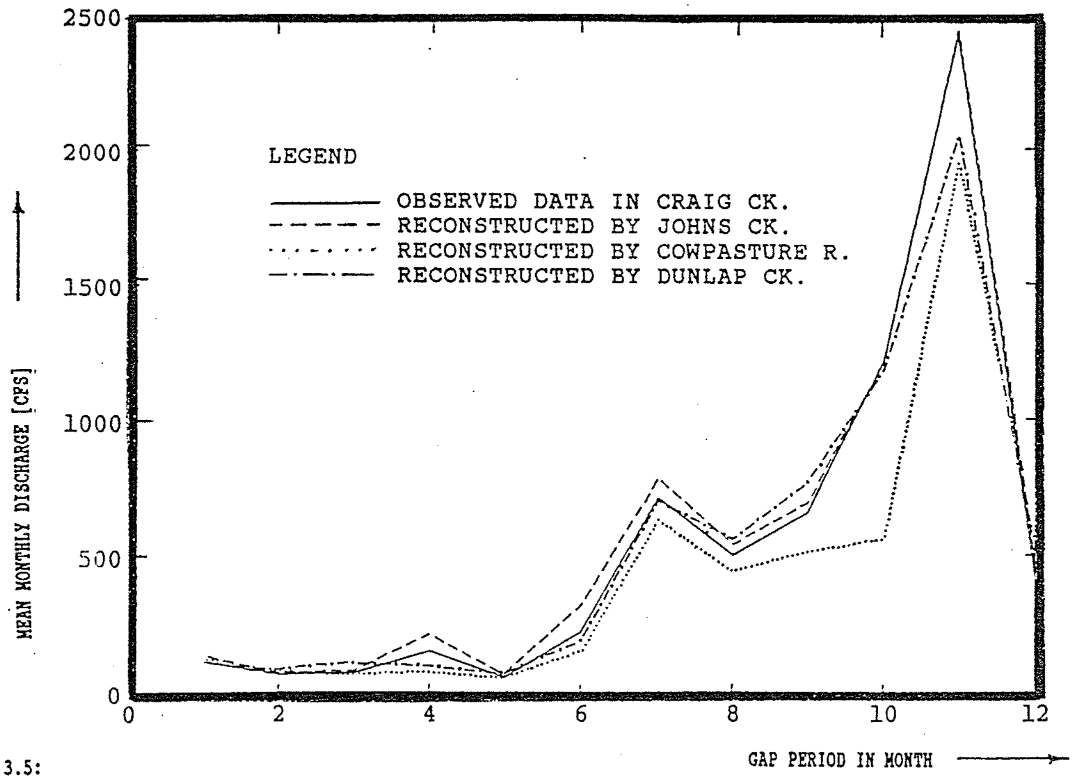


Fig.5.1.3.5:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by SMOVE model using variable information source

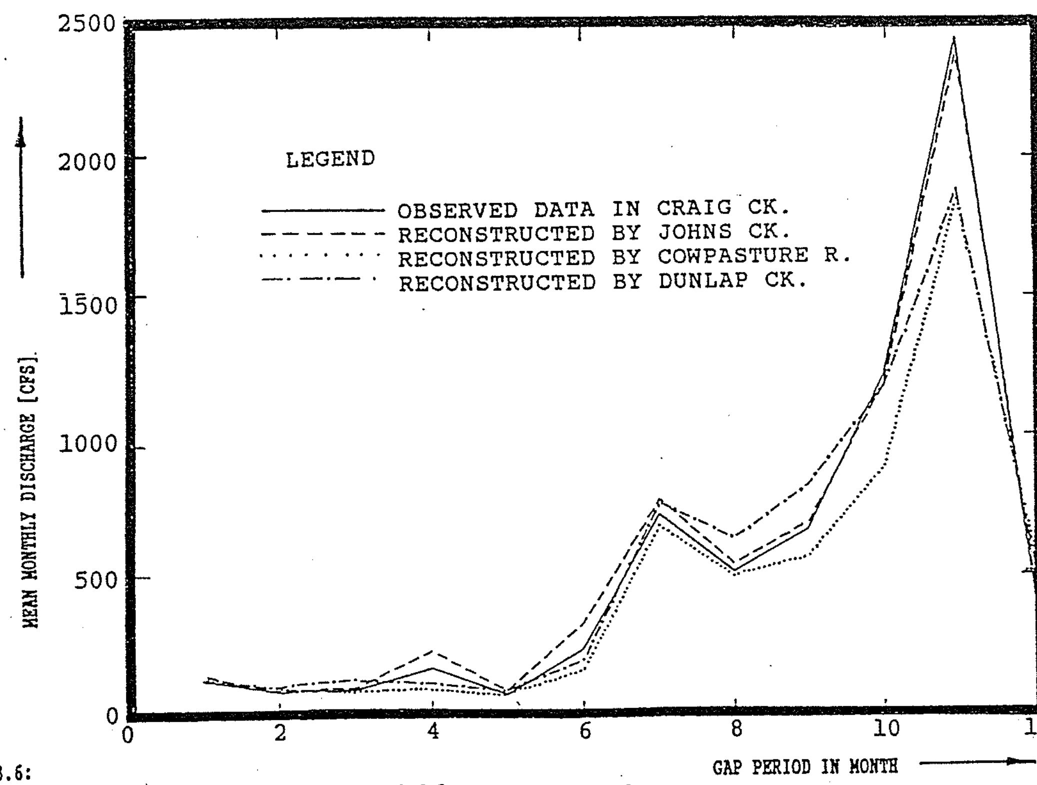


Fig.5.1.3.6:
Comparison of infilling of Craig Ck. data for period [June 1987-May 1988] by SSMOVE model using variable information source

The assessment category two shows that:

. Estimation error in dry season is high in case of seasonal or selected seasonal samples over the one in case of heterogenous sample.

. Although some {base R., river with missing data} pair under seasonal sample show less estimation error in the wet season in contrast to the estimation error in wet season under heterogenous sample, but the overall error is always higher in case of seasonal sample than in case of heterogenous sample.

. Beneficial effect of seasonal consideration in sample selection was noticed in few cases only in wet season.

The assessment category three shows that:

Among the rivers performance of Johns Ck. is best. With respect to some criteria, Cowpasture R. shows better performance over Dunlap Ck. With respect to some other criteria, Dunlap Ck. performs superior to Cowpasture R. The SESTRNALL model performs worst.

Performance of Johns Ck. compared to the other base rivers is best. This performance can be explained by very high cross correlation coefficient of { Craig Ck., Johns Ck.} pair . Johns Ck. In case of MULBS, SMOVE and SREG models, reconstruction by Johns Ck. and Dunlap Ck. are very good in wet season. The close location of these base rivers to Craig Ck., therefore, the resulting high seasonal similarity can be considered as the cause of good performance. Cowpasture R. performed best in case of models incorporating indiscriminately chosen bulk of data [REG, AMOVE] over the models considering seasonal difference. The remoteness of Cowpasture R. from Craig Ck. and therefore the less seasonal similarity can be considered as the reason.

5.2 ASSESSMENT OF QUALITY OF INFILLING OF GAP OF THE MISSING DATA RIVER BELONGING TO THE LB CLUSTER

ASSESSMENT CATEGORY 1

Multivariate level assessment [Table 5.2.1.0] shows that the observed seasonal segment lies within the 95% contour of the predicted configuration in all cases of MULBS model. In case of SESTRNALL model , the observed wet seasonal segment resides beyond the 99.5% predicted contour but the observed dry seasonal segment lies within the 95% contour.

For the MULBS model, in both cases of {Little R., Roanoke R.} and {Little R., Reed Ck.} river pairs, the observed data corresponding to all the elements of both seasonal segment lie within the 95% band of marginal prediction [Table 5.2.1.(1 and 2)]. In case of SESTRNALL model, observed data corresponding to the 2nd element of the dry seasonal segment lies within the 97.5% band, observed data corresponding to 1st element of the wet seasonal segment lies within the 99% band, the observed data corresponding to 5th element of the wet seasonal segment lies beyond the 99.9% band of prediction while all other data corresponding to rest of the elements of both seasonal segment lie within the 95% level of prediction [Table 5.2.1.3].

Entropy reduction by conditioning on simultaneously observed seasonal segment of Raonoke R. and Reed Ck. are significantly better than the one by assuming Markovian nature of seasonal transition of Little R. [Table 5.2.2].

For REG model , for each case of incorporating Roanoke R. and Reed Ck. information, the observed data lie within 95% prediction band [Table 5.2.3.1]. For SREG model [Table 5.2.3.2], incorporating Roanoke R. information, the observed element corresponding to the 6th element of wet seasonal segment lies within 96% prediction band while the rest lie within 95% band of prediction. For the same model, in case of incorporating Reed Ck. information , all the observed data lie within 95% prediction band In case of SSREG model [Table 5.2.3.3], incorporating Roanoke R. information, observed data corresponding to 2nd element of the wet seasonal segment lies within 97% prediction band while the rest lie within 95% band. For SSREG model incorporating Reed Ck. information, all the observed data lie within 95% prediction band of corresponding element in case of both seasonal segments.

Table 5.2.1.0 : Distance of Observed seasonal segment of Little R. with respect to the predicted conditional configuration by the base R. by MULBS Model and by self Series by SESTRNALL Model(multivariate basis appraisal)

		Season 1			Season 2		
Model	Information Source	Observed Distance	DF	Comment	Observed Distance	DF	Comment
MULBS.	Reed Creek	10.41	6	lying within 95% Contour	5.00	6	lying within 95% Contour
MULBS	Roanoke River	9.55	6	lying within 95% Contour	11.80	6	lying within 95% Contour
SESTRNALL	Self Series	9.00	6	lying within 95% Contour	20.00	6	lying outside 99.5% Contour

Table 5.2.1.1: Little R. Data infilling by Reed Ck. by MULBS Model: Predicted vs. Observed data (appraisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Mean Predicted	Element Std. Predicted	Observed Element	Comment
	1	0.066	5.270	0.257	5.080	Contained within 95%
	2	0.047	5.110	0.217	4.720	Contained within 95%
	3	0.048	4.980	0.219	5.110	Contained within 95%
1	4	0.099	5.790	0.315	5.360	Contained within 95%
	5	265.860	55.950	16.305	39.910	Contained within 95%
	6	0.050	5.670	0.224	5.510	Contained within 95%
	1	4800.594	523.827	69.286	470.000	Contained within 95%
	2	6986.777	460.396	83.587	354.000	Contained within 95%
	3	6541.078	565.687	80.877	474.000	Contained within 95%
2	4	9036.719	959.299	95.062	982.000	Contained within 95%
	5	24171.773	1366.951	155.473	1445.000	Contained within 95%
	6	7607.701	654.960	87.222	632.000	Contained within 95%

Table 5.2.1.2: Little R. Data infilling by Roanoke R. by MULBS Model: Predicted vs. Observed data (appraisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Mean Predicted	Element Std. Predicted	Observed Element	Comment
	1	0.052	5.133	0.228	5.075	Contained within 95%
	2	0.035	5.075	0.187	4.718	Contained within 95%
	3	0.023	5.260	0.151	5.112	Contained within 95%
1	4	0.048	5.588	0.220	5.361	Contained within 95%
	5	0.023	4.999	0.151	4.963	Contained within 95%
	6	0.013	5.679	0.115	5.509	Contained within 95%
	1	8.280	38.300	2.877	35.680	Contained within 95%
	2	10.730	38.840	3.276	30.710	Contained within 99%
	3	63.090	52.770	7.943	46.590	Contained within 95%
2	4	17.600	49.670	4.195	52.990	Contained within 95%
	5	3.450	33.680	1.857	31.770	Contained within 95%
	6	14.340	25.680	3.787	25.950	Contained within 95%

Table 5.2.1.3: little River data infilling by SESTRNALL Model
Predicted vs. observed data of Little River
(appraisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Std. Predicted	Element Mean Predicted	Observed Element	Comment
	1	0.0834	0.2888	5.3388	5.0750	Contained within 95%
	2	0.0457	0.2138	5.1521	4.7180	Contained within 97.5%
	3	0.0683	0.2613	5.1508	5.1120	Contained within 95%
1	4	0.1903	0.4362	5.0311	5.3610	Contained within 95%
	5	0.2035	0.4511	5.1207	4.9630	Contained within 95%
	6	0.2233	0.4725	5.4300	5.5090	Contained within 95%
	1	0.0836	0.2891	5.4683	6.1530	Contained within 99%
	2	0.1293	0.3596	5.8231	5.8690	Contained within 95%
	3	0.0996	0.3156	5.9875	6.1610	Contained within 95%
2	4	0.1575	0.3969	6.1467	6.8900	Contained within 95%
	5	0.1397	0.3738	5.8649	7.2360	Not Contained within 99.9
	6	0.1347	0.3670	5.8436	6.4490	Contained within 95%

Table 5.2.2: Entropy reduction in class-membership index sequence of Little R.

Case	Information source	Marginal Entropy of class-membership index sequence of Little R.	Conditional Entropy of class-membership index sequence of Little R.	Entropy Reduction [%]
Conditional Entropy	Reed Ck.	1.2275	0.6582	46.38
Conditional Entropy	Roanoke R.	1.2275	0.3408	72.24
Markovian Entropy	Seasonal transition	1.2275	1.1192	8.82

Table 5.2.3.1: Summary of REG model pertinent to Little R. data infilling

Base River	State	computed model							Residual Analysis					Prediction
		b0	b1	tb0	tb1	DF	Comment	R-sq	Chi-sq(e)	DF	Comment	Const. var.(e)	State of ACF(e)	Observed vs. prediction interval
Roanoke River	lntd*	2.48	0.59	35.07	46.15	358	b0, b1 significant at 5% level	.86	84.0	69	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
Reed Creek	lntd*	2.32	0.63	21.73	32	358	b0, b1 significant at 5% level	.74	65.6	69	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval

Table 5.2.3.2: Summary of SREG model pertinent to Little R. data infilling

Base River	Season	State	computed model							Residual Analysis					Prediction
			b0	b1	tb0	tb1	DF	Comment	R-sq	Chi-sq(e)	DF	Comment	Const. var.(e)	State of ACF(e)	Observed vs. prediction interval
Roanoke River	1	lntd*	2.45	.60	19.68	24.25	178	b0, b1 significant at 5% level	77	38.8	33	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	2.23	.63	20.14	33.78	178	b0, b1 significant at 5% level	.87	21.2	33	normal at 5% level	satisfied	auto-correlated	6th element contained within 96% interval while the rest contained within 95% interval
Reed Creek	1	lntd*	1.73	.76	7.86	16.92	178	b0, b1 significant at 5% level	.62	22.2	33	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	2.12	.66	12.43	22.67	178	b0, b1 significant at 5% level	.74	24.0	33	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval

* lntd : ln transformed data

Table 5.2.3.3: Summary of SSREG model pertinent to Little R. data infilling

Base River	Season	State	computed model							Residual Analysis					Prediction
			b0	b1	tb0	tb1	DF	Comment	R-sq	Chi-sq(e)	DF	Comment	Const. var.(e)	State of ACF(e)	Observed vs. prediction interval
Roanoke River	1	lntd*	2.67	.55	14.78	14.97	136	b0, b1 significant at 5% level	.62	24.78	24	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	2.32	.62	9.74	16.91	52	b0, b1 significant at 5% level	.85	3.78	7	normal at 5% level	satisfied	auto-correlated	2nd element contained within 97% interval while the rest contained within 95% interval
Reed Creek	1	lntd*	2.18	.67	7.55	11.01	124	b0, b1 significant at 5% level	.49	25.98	22	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	2.58	.60	7.16	10.34	70	b0, b1 significant at 5% level	.60	10.06	11	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval

* lntd : ln transformed data

Table 5.2.4: Zero lag Cross correlation Coefficient between Little River and each of the base rivers

Base River	Cross correlation coefficient corresponding to indiscreminately chosen sample	Cross correlation coefficient corresponding to seasonal sample		Cross correlation coefficient corresponding to selected seasonal sample	
		Season 1	Season 2	Season 1	Season 2
Roanoke River	0.93	0.88	0.93	0.79	0.92
Reed Creek	0.86	0.79	0.86	0.70	0.78

ASSESSMENT CATEGORY 2

COMPARISON OF THE REGRESSION MODELS FOR DIFFERENT {RIVER WITH MISSING DATA, BASE R.} PAIRS

The ranking is given in Table 5.2.5.a. The REG model gave the least ed and least eo. In case of {Roanoke R.,Little R.} pair, the SREG and SSREG models respectively performed better over REG in terms of ew and P. In case of { Reed Ck., Little R.} pair, SSREG performed over REG in terms of P. The graphical contrast of infilled versus observed data for various sampling scenarios is given in Figs.5.2.1. (1,2)].

COMPARISON OF MOVE.4 MODELS FOR DIFFERENT {RIVER WITH MISSING DATA, BASE R.} PAIRS

The ranking is given in Table 5.2.5.b. AMOVE gave the least ed in all cases. In case of {Roanoke R., Little R.} pair, SSMOVE performed better over AMOVE with respect to P and eo., while SMOVE performed superior to AMOVE with respect to ew. In case of {Reed Ck., Little R.} pair no beneficial effect of seasonal sampling over the heterogenous sample was noticed. The graphical contrast of infilled versus observed data for various sampling scenarios is given in Figs.5.2.1. (3,4)].

Table 5.2.5.a: Table of ranking of performance of varities of regression model (LB cluster)

River	Model rank in terms of estimation error in dry season [ed]			Model rank in terms of estimation error in wet season [ew]			Model rank in terms of deviation at peak reconstd.- obsvd. [P]			Model rank in terms of overall estimation error [eo]		
	REG	SREG	SSREG	REG	SREG	SSREG	SSREG	REG	SREG	REG	SREG	SSREG
Reed Ck.	0.26	0.27	0.28	0.23	0.23	0.24	-411.1	-456.3	-459.6	0.25	0.25	0.26
	REG	SREG	SSREG	SREG	REG	SSREG	SSREG	SREG	REG	REG	SREG	SSREG
Roanoke R.	0.22	0.24	0.24	0.21	0.22	0.23	-111.4	-180.4	-238.1	0.22	0.23	0.24

Table 5.2.5.b: Table of ranking of performance of varities of MOVE.4 model (LB cluster)

River	Model rank in terms of estimation error in dry season [ed]			Model rank in terms of estimation error in wet season [ew]			Model rank in terms of deviation at peak reconstd.- obsvd. [P]			Model rank in terms of overall estimation error [eo]		
	AMOVE	SMOVE	SSMOVE	AMOVE	SSMOVE	SMOVE	AMOVE	SSMOVE	SMOVE	AMOVE	SMOVE	SSMOVE
Reed Ck.	0.23	0.27	0.27	0.22	0.24	0.24	-341.6	-446.4	-462.5	0.25	0.26	0.29
	AMOVE	SSMOVE	SMOVE	SMOVE	SSMOVE	AMOVE	SSMOVE	AMOVE	SMOVE	SSMOVE	AMOVE	SMOVE
Roanoke R.	0.19	0.25	0.25	0.21	0.22	0.26	-132.5	-159.7	-181.7	0.21	0.23	0.24

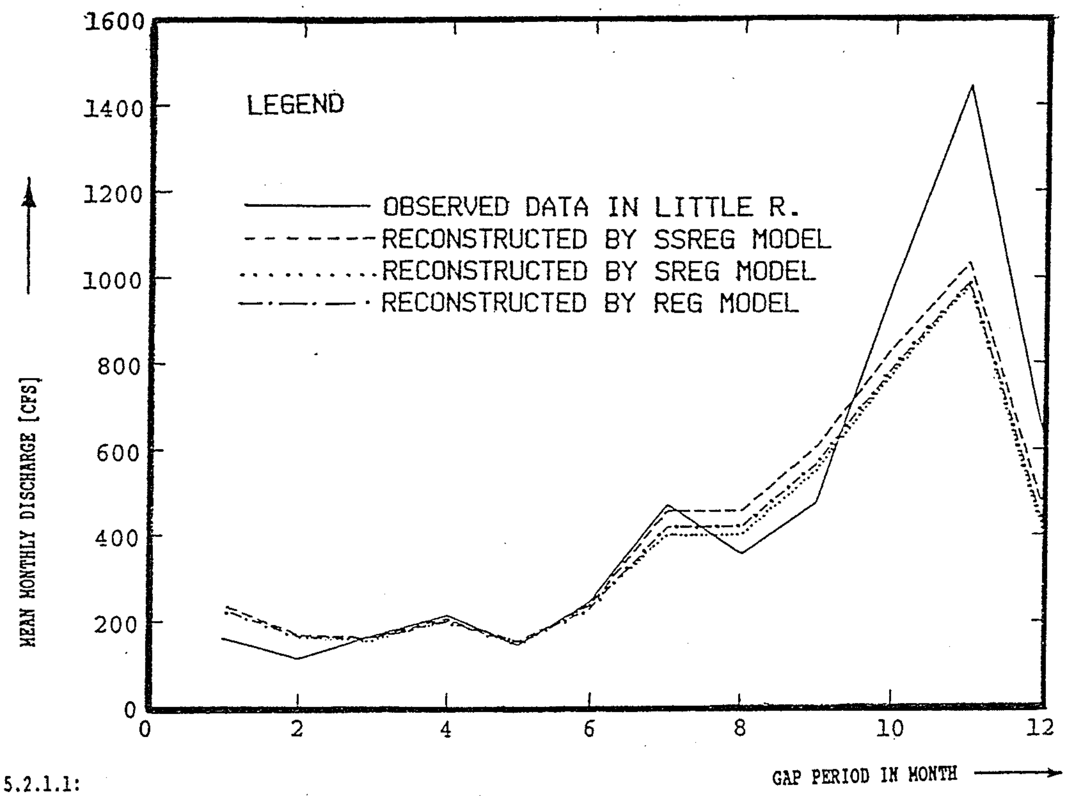


Fig. 5.2.1.1:
Comparison of infilling of Little R. data for period [June 1987-May 1988] by Reed Ck. by REG, SREG, SSREG models (varities of Regression model)

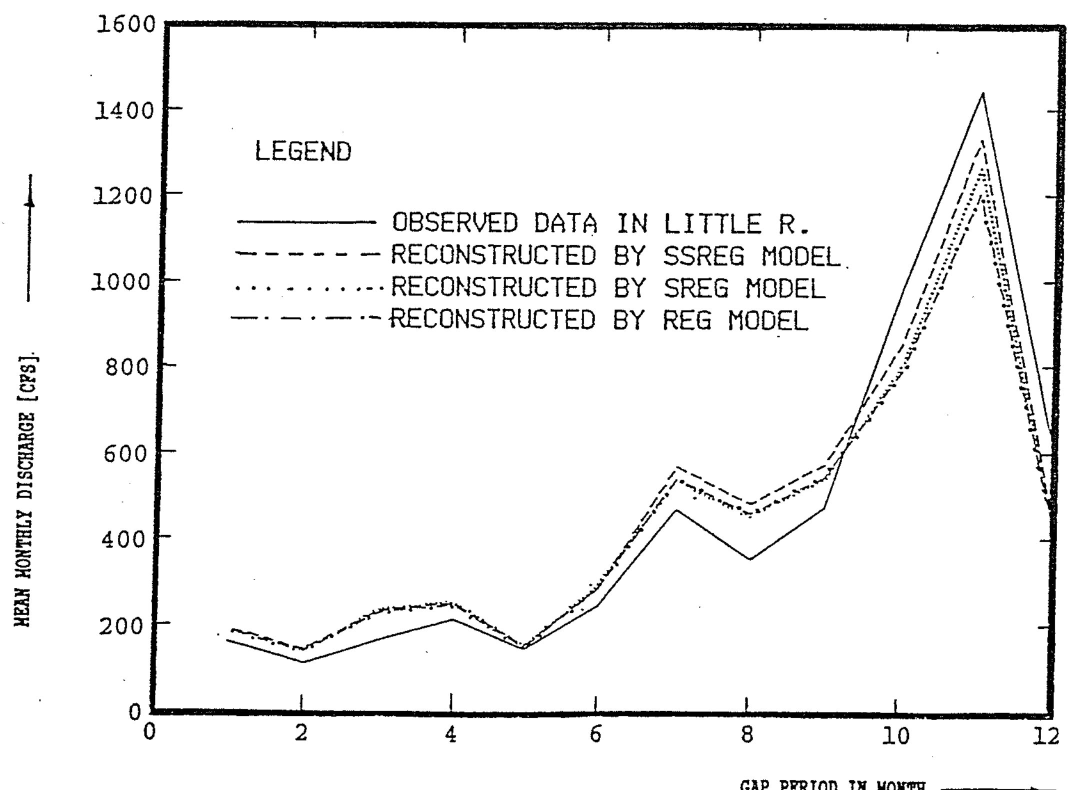


Fig. 5.2.1.2:
Comparison of infilling of Little R. data for period [June 1987-May 1988] by Roanoke R. by REG, SREG, SSREG models (varities of Regression model)

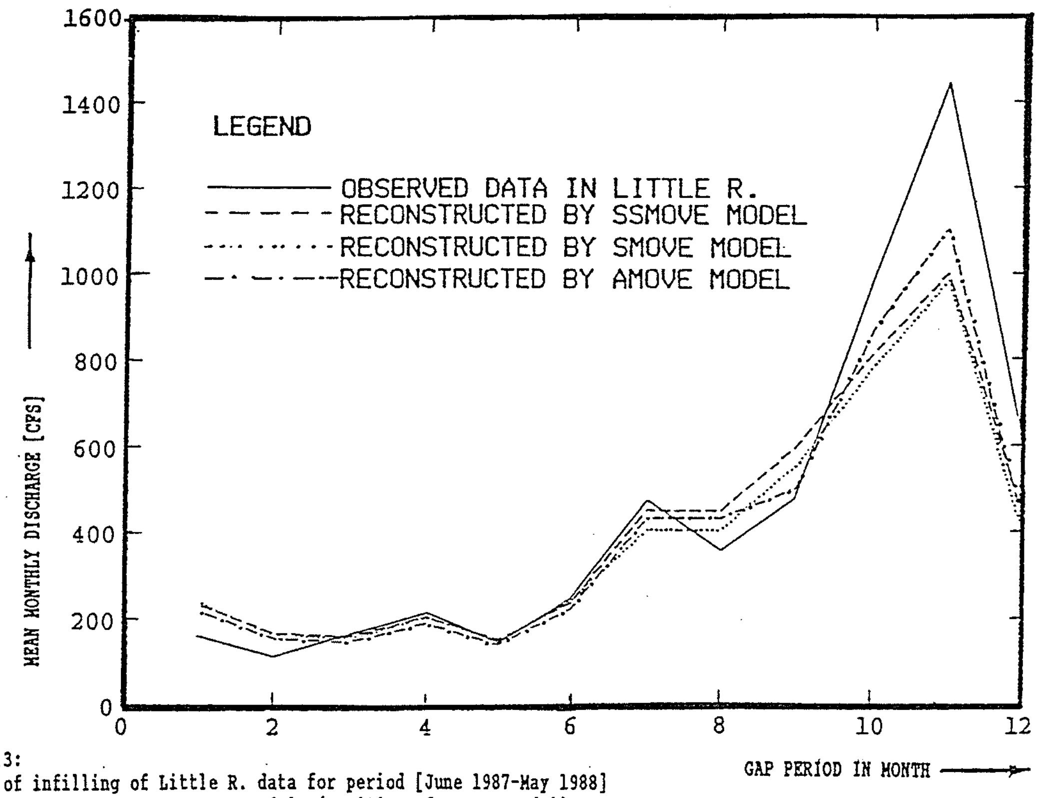


Fig. 5.2.1.3:
Comparison of infilling of Little R. data for period [June 1987-May 1988] by Reed Ck. by AMOVE, SHOVE, SSMOVE models (varities of MOVE.4 model)

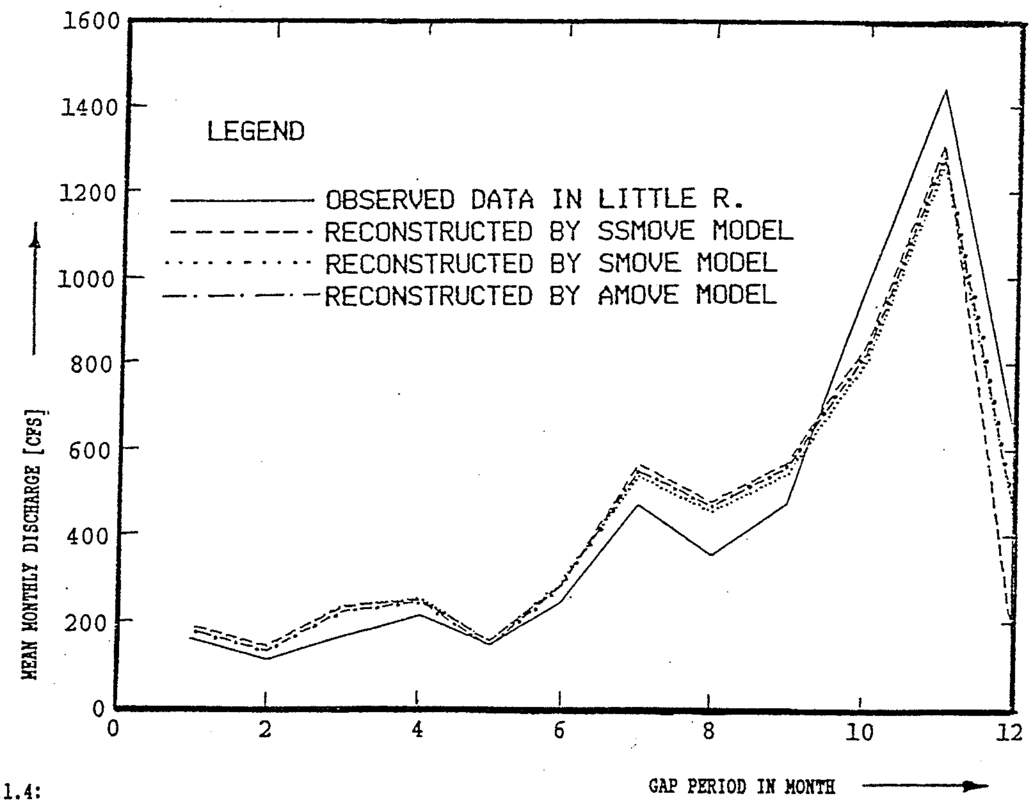


Fig. 5.2.1.4:
Comparison of infilling of Little R. data for period [June 1987-May 1988] by Roanoke R. by AMOVE, SHOVE, SSMOVE models (varities of MOVE.4 model)

ASSESSMENT CATEGORY 3

MULTIVARIATE MODELS (Table 5.2.6.a)

The SESTRNALL model performed very poorly in compared to MULBS model incorporating each of the base rivers information. Roanoke R. performed best with respect to ed, eo. Reed Ck. performed best with respect to ew and P. The contrast of infilled versus observed data by the multivariate models is presented in Fig.5.2.2.

REGRESSION AND MOVE.4 MODEL [Table 5.2.6.(b,c)]

With respect to all the criterion that were chosen for evaluation, in all cases of MOVE.4 and regression models, Roanoke R. performed better over Reed Ck. . The contrast of infilled versus observed data by the multivariate models is presented in Fig.5.2.2.

Table 5.2.6.a: Table of ranking of performance of various sources for the multivariate models (LB cluster)

MODEL	Source rank in terms of estimation error in dry season [ed]			Source rank in terms of estimation error in wet season [ew]			Source rank in terms of deviation at peak reconstd.- obsvd. [P]			Source rank in terms of overall estimation error [eo]		
	Rn.R.	S.S.	Rd.Ck.	Rd.Ck.	Rn.R.	S.S	Rd.Ck.	Rn.R.	S.S.	Rn.R.	Rd.Ck.	S.S.
Multiv. M.	0.23	0.29	0.38	0.11	0.20	0.4	-78.1	207.4	-1092.6	0.21	0.29	0.39

Table 5.2.6.b: Table of ranking of performance of various base R. for each sampling scenario, regression model (LB cluster)

MODEL	River rank in terms of estimation error in dry season [ed]		River rank in terms of estimation error in wet season [ew]		River rank in terms of deviation at peak reconstd.- obsvd. [P]		River rank in terms of overall estimation error [eo]	
	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.
REG	0.22	0.26	0.22	0.23	-238.1	-456.3	0.22	0.25
	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.
SREG	0.24	0.27	0.21	0.23	-180.4	-459.6	0.23	0.25
	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.
SSREG	0.24	0.28	0.28	0.24	-111.4	-411.1	0.24	0.26

Table 5.2.6.c: Table of ranking of performance of various base R. for each sampling scenario, MOVE.4 model (LB cluster)

MODEL	River rank in terms of estimation error in dry season [ed]		River rank in terms of estimation error in wet season [ew]		River rank in terms of deviation at peak reconstd.- obsvd. [P]		River rank in terms of overall estimation error [eo]	
	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.
AMOVE	0.19	0.23	0.22	0.22	-159.7	-341.6	0.20	0.22
	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.
SMOVE	0.25	0.27	0.21	0.24	-181.7	-462.5	0.23	0.25
	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.	Rn.R.	Rd.Ck.
SSMOVE	0.25	0.27	0.22	0.24	-132.5	-411.1	0.24	0.26

Rn.R.= Roanoke R.
Rd.Ck.= Reed Ck.
S.S = self series

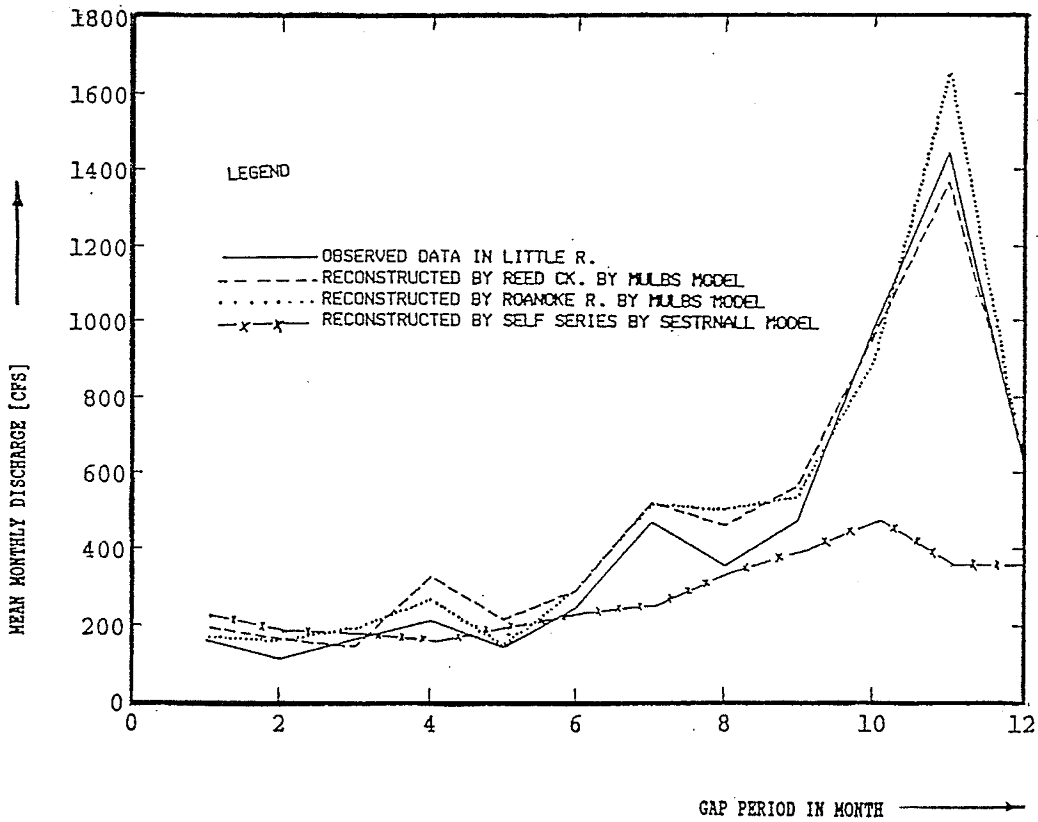


Fig.5.2.2:
Comparison of infilling of Little R. data for period [June 1987-May 1988]
by multivariate models using variable information source

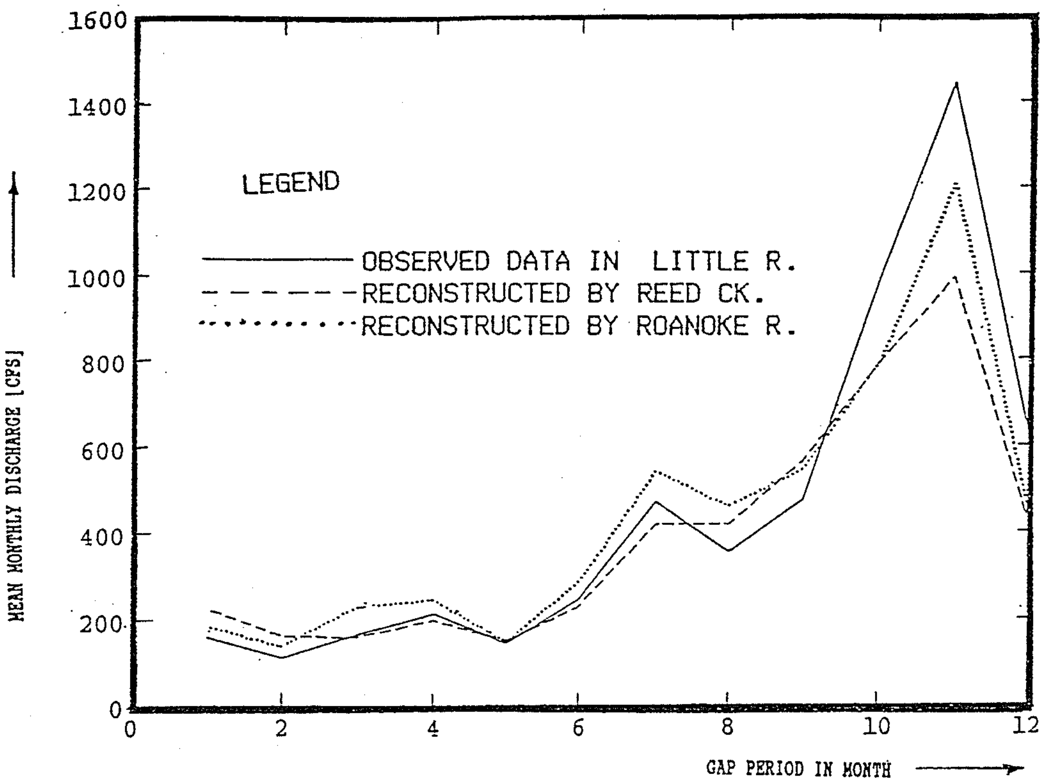


Fig.5.2.3.1:
Comparison of infilling of Little R. data for period [June 1987-May 1988]
by REG model using variable information source

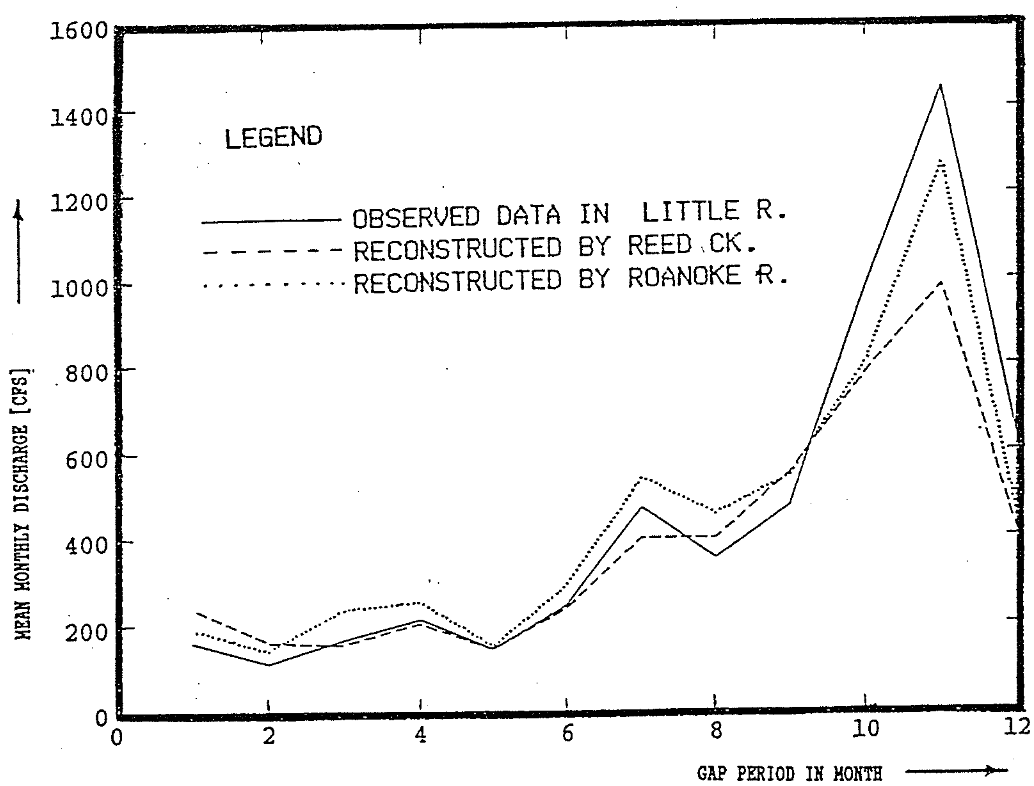


Fig.5.2.3.2:
Comparison of infilling of Little R. data for period [June 1987-May 1988]
by SREG model using variable information source

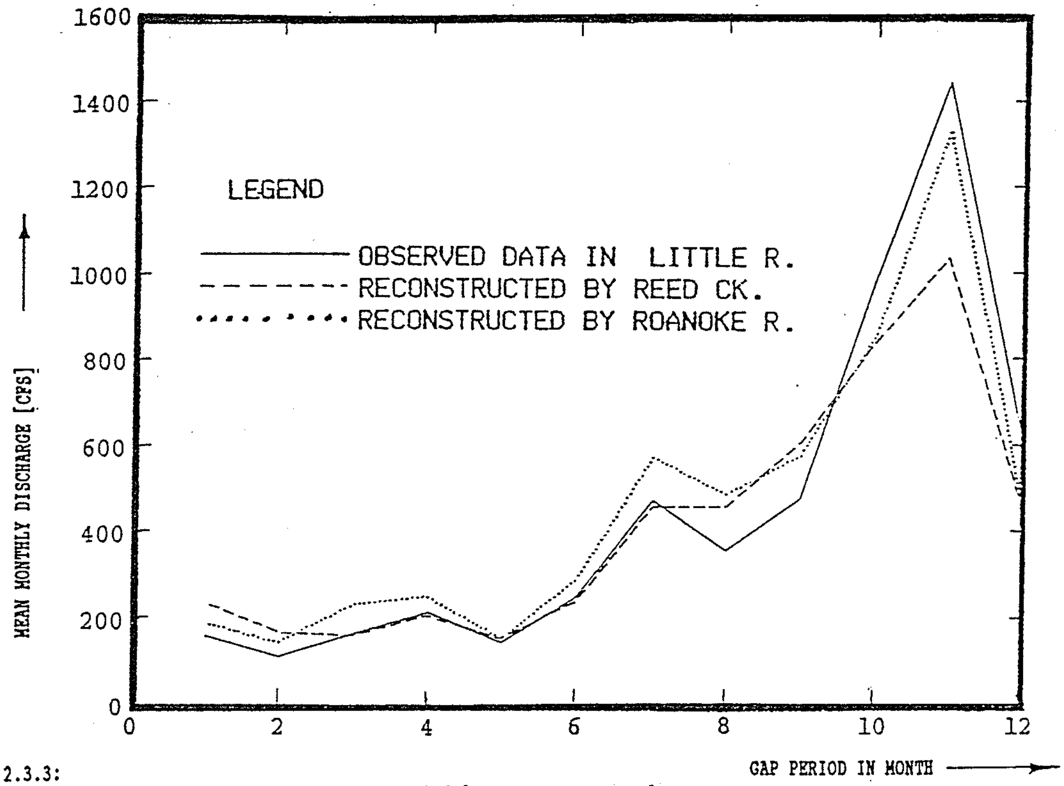


Fig.5.2.3.3:
Comparison of infilling of Little R. data for period [June 1987-May 1988]
by SSREG model using variable information source

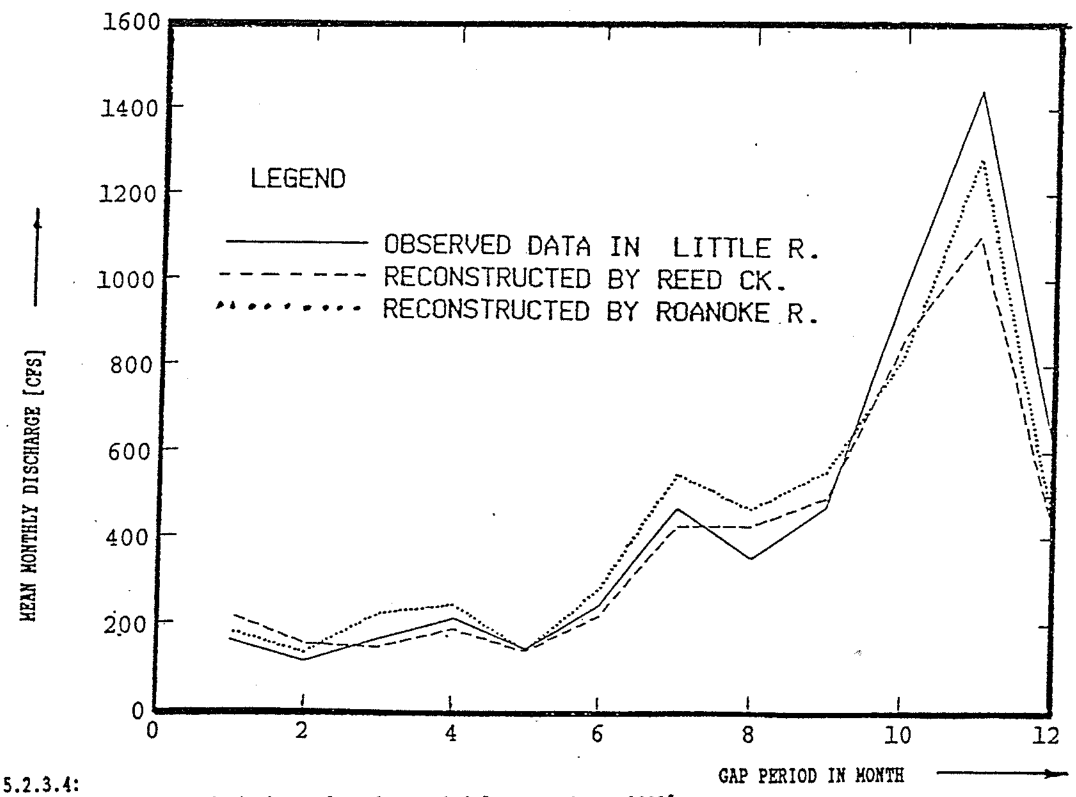


Fig.5.2.3.4:
Comparison of infilling of Little R. data for period [June 1987-May 1988]
by AMOVE model using variable information source

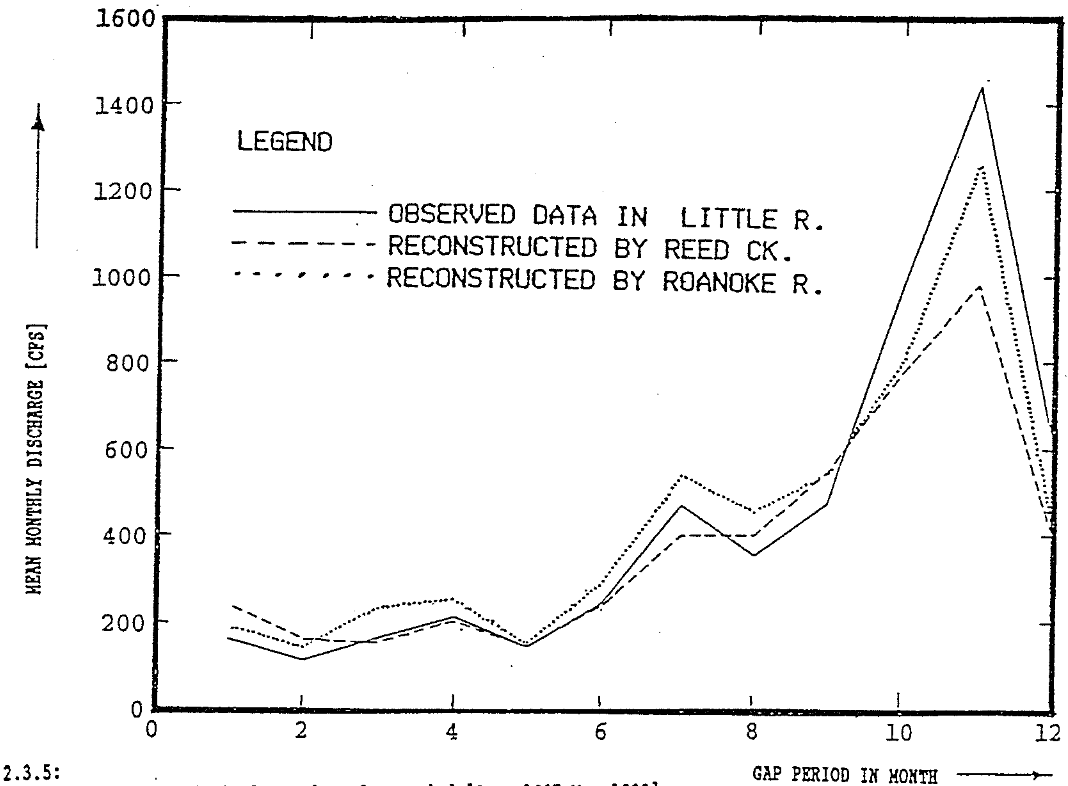


Fig.5.2.3.5:
Comparison of infilling of Little R. data for period [June 1987-May 1988]
by SMOVE model using variable information source

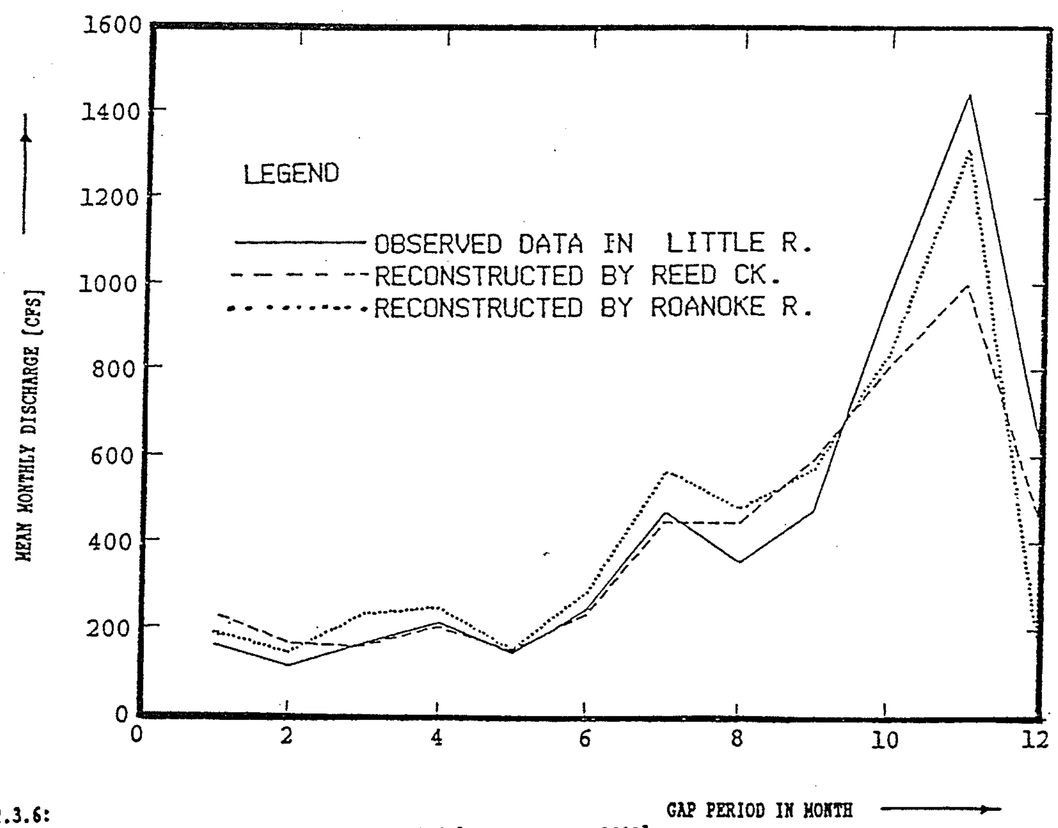


Fig.5.2.3.6:
Comparison of infilling of Little R. data for period [June 1987-May 1988]
by SSMOVE model using variable information source

The assessment category two shows that in all cases the heterogenous sample gave the least ed. In case of {Reed Ck., Little R.} pair, seasonal sampling do not show any improvement over the heterogenous sample.
The assessment category three shows that in all cases of regression and MOVE.4 models Roanoke R. performed superior to the Reed Ck. In case of MULBS model, Roanoke R. performed inferior to Reed Ck. in terms of ew and P. SESTRNALL model performed the worst.

Poor fitting of SESTRNALL model can be explained by the poor information transfer by considering seasonal transition of Markovian nature. In case of MULBS model, fitting of reconstructed data to the observed data is much better in Reed Ck. than in Roanoke R.. Reed Ck. resides on the same drainage basin as that of the Little R. while Roanoke R. is located in different drainage basin. Due to the closer proximity, the seasonal similarity between {Reed Ck., Little R.} pair is more than the seasonal similarity between { Roanoke R., Little .} pair.

In case of all the varieties of regression and MOVE model, better performance of infilling by Roanoke R. over the performance of infilling by Reed Ck. can be explained by the higher zero lag cross-correlation coefficient of {Roanoke R. and Little R} pair than that of {Reed Ck. and Little R} pair

[Table 5.2.4]. In case of incorporating Roanoke R. information, superior performance of SSREG and SSMOVE models can be attributed both to appreciably high zero lag cross correlation coefficient of { Little R. and Roanoke R.} pair as well as to considerable amount of entropy reduction [Table 5.2.2].

5.3 ASSESSMENT OF QUALITY OF INFILLING OF GAP OF THE MISSING DATA RIVER BELONGING TO THE LILL CLUSTER

ASSESSMENT CATEGORY 1

A multivariate level assessment of the MULBS model [Table 5.3.1.0], incorporating Soo R. information transfer, shows that the observed wet seasonal segment lies within the 99% contour of prediction while the observed dry seasonal segment lies within the 95% contour. For the same model incorporating Rutherford Ck. information, all the observed seasonal segment lies within the 95% contour. In case of SESTRNALL model, all the observed seasonal segments lie within the 95% contour of prediction.

For the MULBS model in the case of {Green R., Soo R.} pair[Table 5.3.1.1], observed data corresponding to 5th element of dry seasonal segment lie within 95.7% prediction band while the remainder of the observed data lie within the 95% prediction band of corresponding elements of the associated seasonal segment. In the case of the {Green R., Rutherford Ck.} pair [Table 5.3.1.2], observed data corresponding to the 2nd element of wet seasonal segment lie within the 97.5% prediction band while the remainder of the data lie within the 95% prediction band of corresponding element of the associated seasonal segment. In case of SESTRNALL model [Table 5.3.1.3], observed data corresponding to 5th element of wet seasonal segment lies within the 99% band while the rest of the observed data lie within 95% prediction band of corresponding element of the associated seasonal segment.

Reduction in entropy for the seasonally classified sequence of Green river is appreciable by conditioning on Soo R. than by conditioning on Rutherford Ck. and is very poor under assumption of Markovian nature of seasonal transition [Table 5.3.2].

In case of both the REG and SREG models incorporating the {Green R.,Soo R.} and {Green R., Rutherford Ck.} pairs, all the observed data are contained within 95% band of prediction [Table 5.3.3.(1 and 2)]. In case of the SSREG model incorporating {Green R., Soo R.} pair [Table 5.3.3.3], observed data corresponding to the 5th element of the wet seasonal segment lie within the 98% prediction band while the rest of the data are contained within the 95% prediction level . For the SSREG model incorporating the {Green R., Rutherford Ck.} pair [Table 5.3.3.3], observed data corresponding to the 2nd element of the wet seasonal segment lie within the 99.2% prediction band while the rest of observed data lie within the 95% band.

Table 5.3.1.0 : Distance of Observed seasonal segment of Green R. with respect to the predicted conditional configuration by the base R. by MULBS Model and by self Series by SESTRNALL Model(multivariate basis appraisal)

		Season 1			Season 2		
Model	Information Source	Observed Distance	DF	Comment	Observed Distance	DF	Comment
MULBS	Soo River	3.31	6	lying within 95% Contour	17.02	6	lying within 99% Contour
MULBS	Rutherford Ck.	6.39	6	lying within 95% Contour	10.44	6	lying within 95% Contour
SESTRNALL	Self Series	2.00	6	lying within 95% Contour	8.00	6	lying within 95% Contour

Table 5.3.1.1: Green R. Data infilling by Soo R. by MULBS Model: Predicted vs. Observed data (appraisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Mean Predicted	Element Std. Predicted	Observed Element	Comment
	1	1.180	6.576	1.086	7.140	Contained within 95%
	2	0.840	2.947	0.917	2.660	Contained within 95%
	3	1.750	2.939	1.323	4.070	Contained within 95%
1	4	0.700	3.105	0.837	3.850	Contained within 95%
	5	2.160	2.789	1.470	2.580	Contained within 95.7%
	6	1.050	1.124	1.025	2.090	Contained within 95%
	7	7.780	22.501	2.789	22.280	Contained within 95%
	8	29.240	31.804	5.407	22.040	Contained within 95%
	9	212.430	89.918	14.575	77.840	Contained within 95%
2	10	5065.830	357.645	71.175	408.330	Contained within 95%
	11	2.760	2.912	1.661	-0.452	Contained within 95%
	12	4.680	9.098	2.163	11.160	Contained within 95%

Table 5.3.1.2: Green R. Data infilling by Rutherford Ck. by MULBS Model: Predicted vs. Observed data(appraisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Mean Predicted	Element Std. Predicted	Observed Element	Comment
	1	24.462	34.680	4.946	35.400	Contained within 95%
	2	35.382	20.535	5.948	20.300	Contained within 95%
	3	31.942	18.227	5.652	15.400	Contained within 95%
1	4	16.573	16.331	4.071	13.300	Contained within 95%
	5	18.368	12.775	4.286	9.000	Contained within 95%
	6	29.638	17.457	5.444	12.500	Contained within 95%
	7	72.453	96.766	8.512	80.200	Contained within 95%
	8	159.714	131.911	12.638	103.000	Contained within 97.5%
	9	121.499	104.651	11.023	105.000	Contained within 95%
2	10	48.608	65.121	6.972	72.200	Contained within 95%
	11	11.972	35.748	3.460	31.700	Contained within 95%
	12	89.986	26.587	9.486	22.400	Contained within 95%

Table 5.3.1.3: Green River data infilling by SESTRNALL Model
Predicted vs. observed data of Green River
(apprisal on marginal basis)

Season	Monthly Element	Element Variance Predicted	Element Std. Predicted	Element Mean Predicted	Observed Element	Comment
	1	149.8500	12.2413	31.2400	35.4000	Contained within 95%
	2	106.7100	10.3301	21.1400	20.3000	Contained within 95%
	3	47.1200	6.8644	12.8900	15.4000	Contained within 95%
1	4	50.9700	7.1393	12.4400	13.3000	Contained within 95%
	5	22.0500	4.6957	10.6600	9.0000	Contained within 95%
	6	189.5100	13.7663	29.7800	12.5000	Contained within 95%
	1	497.8900	22.3134	69.4300	80.2000	Contained within 95%
	2	333.9700	18.2748	111.0000	103.0000	Contained within 95%
	3	281.0800	16.7654	99.6400	105.0000	Contained within 95%
2	4	92.9200	9.6395	69.9300	72.2000	Contained within 95%
	5	54.0900	7.3546	53.3800	31.7000	Contained within 99%
	6	225.1400	15.0047	38.0800	22.4000	Contained within 95%

Table 5.3.2: Entropy reduction in class-membership index sequence of Green R.

Case	Information source	Marginal Entropy of class-membership index sequence of Green R.	Conditional Entropy of class-membership index sequence of Green R.	Entropy Reduction [%]
Conditional Entropy	Soo R.	1.1219	0.3937	64.91
Conditional Entropy	Rutherford Ck.	1.1219	0.8955	20.18
Markovian Entropy	Seasonal transition	1.1219	1.0492	6.48

Table 5.3.3.1: Summary of REG model pertinent to Green R. data infilling

Base River	State	computed model							Residual Analysis					Prediction
		b0	b1	tb0	tb1	DF	Comment	R-sq	Chi-sq(e)	DF	Comment	Const. var.(e)	State of ACF(e)	Observed vs. prediction interval
Soo River	lntd*	2.48	0.59	35.07	46.15	358	b0, b1 significant at 5% level	.86	84.0	69	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
Rutherford Creek	lntd*	2.32	0.63	21.73	32	358	b0, b1 significant at 5% level	.74	65.6	69	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval

Table 5.3.3.2: Summary of SREG model pertinent to Green R. data infilling

Base River	Season	State	computed model							Residual Analysis					Prediction
			b0	b1	tb0	tb1	DF	Comment	R-sq	Chi-sq(e)	DF	Comment	Const. var.(e)	State of ACF(e)	Observed vs. prediction interval
Soo River	1	nat**	3.42	2.08	5.27	32.32	130	b0, b1 significant at 5% level	.89	25.02	23	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	.83	1.01	6.87	28.22	130	b0, b1 significant at 5% level	.86	29.35	23	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
Rutherford Creek	1	lntd*	5.15	3.21	4.4	15.8	130	b0, b1 significant at 5% level	.66	36.83	23	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	nat**	18.52	2.92	5.89	18.98	130	b0, b1 significant at 5% level	.73	24.62	23	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval

* lntd : ln transformed data
** nat : raw data

Table 5.3.3.3: Summary of SSREG model pertinent to Green R. data infilling

Base River	Season	State	computed model: b0	b1	tb0	tb1	DF	Comment	R-sq	Residual Analysis: Chi-sq(e)	DF	Comment	Const. var.(e)	State of ACF(e)	Prediction: Observed vs. prediction interval
Soo River	1	lntd*	1.21	.86	11.94	15.28	94	b0, b1 signi-ficant at 5% level	.71	21.17	16	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	lntd*	.58	1.09	3.89	24.49	52	b0, b1 signi-ficant at 5% level	.92	7.81	7	normal at 5% level	satisfied	auto-correlated	5th element contained within 98% interval while the rest contained within 95% interval
Rutherford Creek	1	lntd*	1.67	.88	17.67	13.44	52	b0, b1 signi-ficant at 5% level	.78	24.88	7	normal at 5% level	satisfied	auto-correlated	all elements contained within 95% interval
	2	nat**	10.94	3.52	2.61	16.32	64	b0, b1 signi-ficant at 5% level	.80	9.24	10	normal at 5% level	satisfied	auto-correlated	2nd element contained within 99.2% interval while the rest contained within 95% interval

* lntd : ln transformed data
** nat : raw data

Table 5.3.4: Zero lag Cross correlation Coefficient between Green River and each of the base rivers

Base River	Cross correlation coefficient corresponding to indiscreminately chosen sample	Cross correlation coefficient corresponding to seasonal sample: Season 1	Season 2	Cross correlation coefficient corresponding to selected seasonal sample: Season 1	Season 2
Soo River	0.97	0.94	0.93	0.84	0.96
Rutherford Creek	0.95	0.81	0.86	0.88	0.89

ASSESSMENT CATEGORY 2

COMPARISON OF THE REGRESSION MODELS FOR DIFFERENT {RIVER WITH MISSING DATA, BASE R.} PAIRS

REG performed best in all cases in terms of ew [Table 5.3.5]. In case of the {Green R., Soo R.} pair, SREG performed better over REG in terms of ed, P and eo. In case of the {Green R., Rutherford Ck.} pair, SSREG performed better over REG in terms of ed and eo.and SREG performed better over REG in terms of P. The graphical contrast of infilled versus observed data for various sampling scenarios is given in Figs.5.3.1. (1,2)].

COMPARISON OF MOVE.4 MODELS FOR DIFFERENT {RIVER WITH MISSING DATA, BASE R.} PAIRS

In case of {Soo R., Green R.} pair [Table 5.3.5], AMOVE performed best in terms of ew and P, while SSMOVE performed better over AMOVE in terms of ed. In case of {Rutherford Ck., Green R.} pair, SMOVE performed superior to AMOVE in terms of ew and P, while AMOVE performed best in terms of ed. The graphical contrast of infilled versus observed data for various sampling scenarios is given in Figs.5.3.1. (3,4)].

Table 5.3.5.a: Table of ranking of performance of varities of regression model (LILL cluster)

River	Model rank in terms of estimation error in dry season [ed]			Model rank in terms of estimation error in wet season [ew]			Model rank in terms of deviation at peak reconstd.- obsvd. [P]			Model rank in terms of overall estimation error [eo]		
	SREG	SSREG	REG	REG	SREG	SSREG	SREG	REG	SSREG	SREG	REG	SSREG
Soo R.	0.09	0.11	0.13	0.12	0.15	0.18	4.8	5.4	10.3	0.12	0.13	0.15
	SSREG	SREG	REG	REG	SSREG	SREG	SREG	REG	SSREG	SSREG	SREG	REG
Rutherford Ck.	0.27	0.32	0.48	0.19	0.25	0.26	18.9	24.5	32.9	0.26	0.29	0.37

Table 5.3.5.b: Table of ranking of performance of varities of MOVE.4 model (LILL cluster)

River	Model rank in terms of estimation error in dry season [ed]			Model rank in terms of estimation error in wet season [ew]			Model rank in terms of deviation at peak reconstd.- obsvd. [P]			Model rank in terms of overall estimation error [eo]		
	SSMOVE	SMOVE	AMOVE	AMOVE	SMOVE	SSMOVE	AMOVE	SMOVE	SSMOVE	AMOVE	SMOVE	SSMOVE
Soo R.	0.09	0.10	0.13	0.11	0.15	0.18	4.7	4.8	9.9	0.12	0.13	0.15
	AMOVE	SMOVE	SSMOVE	SMOVE	AMOVE	SSMOVE	SMOVE	AMOVE	SSMOVE	AMOVE	SMOVE	SSMOVE
Rutherford Ck.	0.20	0.26	0.29	0.18	0.25	0.26	24.9	37.2	39.3	0.23	0.23	0.28

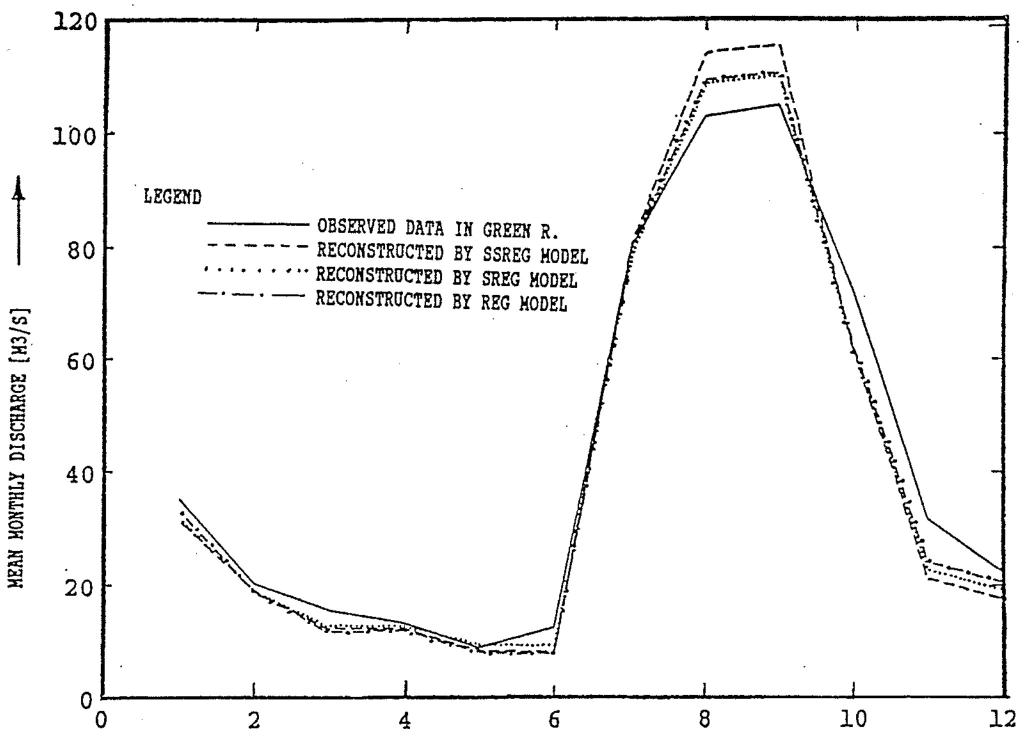


Fig. 5.3.1.1:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945] by Soo R. by REG, SREG, SSREG models (varities of Regression model)

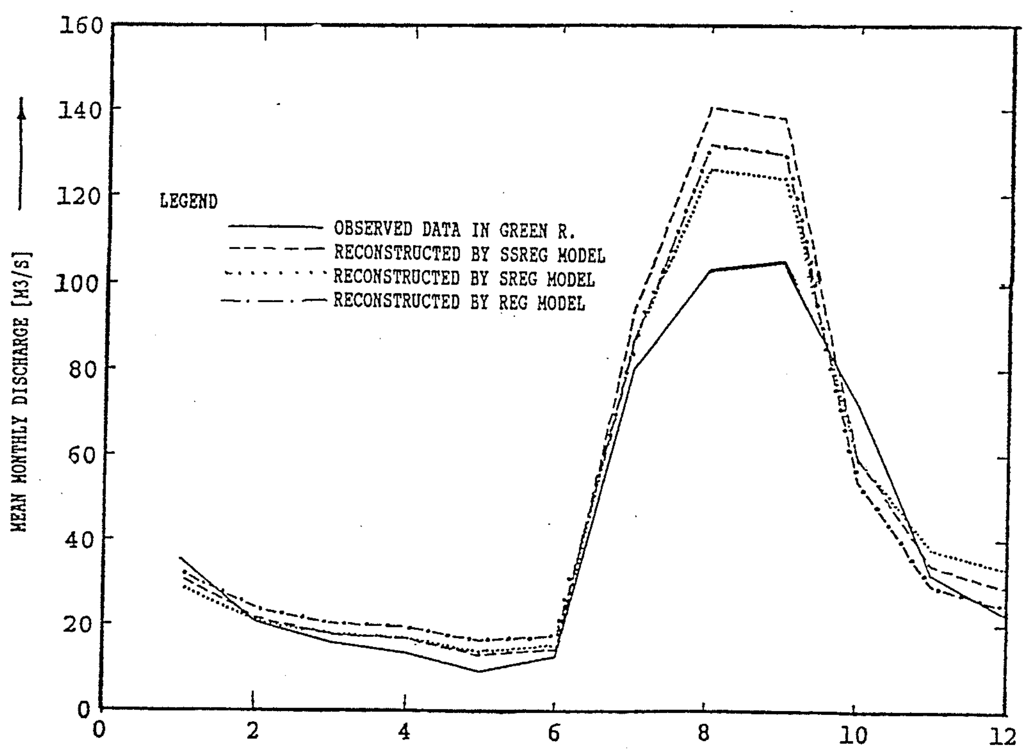


Fig. 5.3.1.2:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945] by Rutherford Ck. by REG, SREG, SSREG models (varities of Regression model)

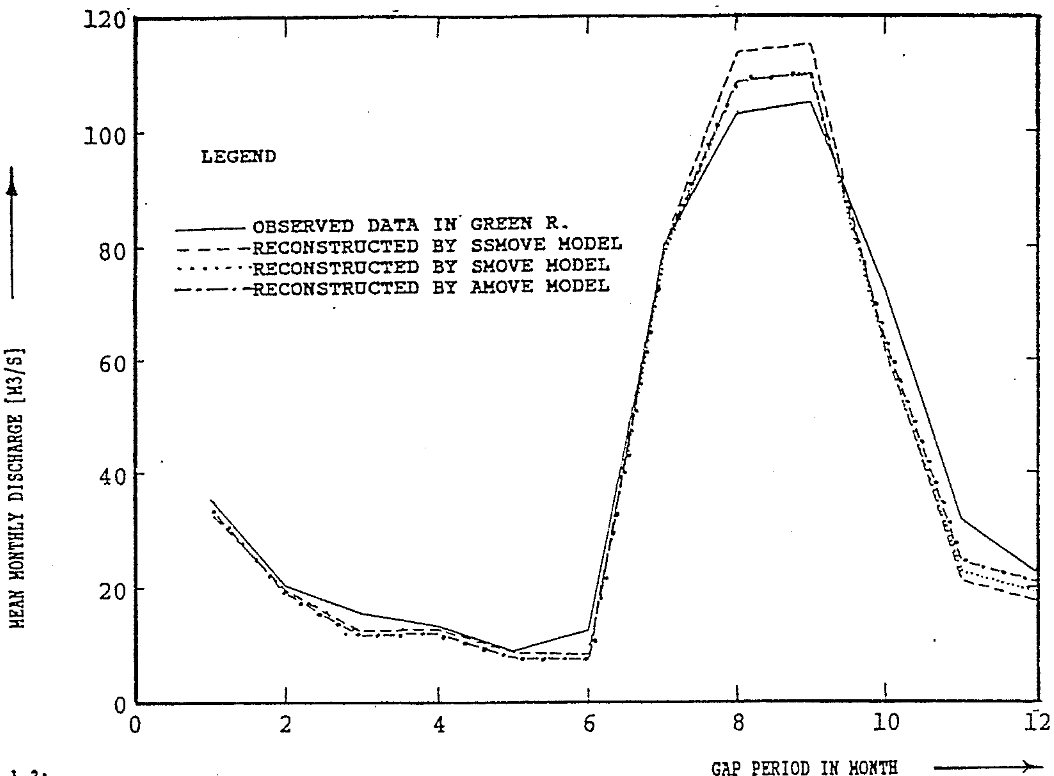


Fig. 5.3.1.3:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945]
by Soo R. by AMOVE, SMOVE, SSMOVE models (varities of MOVE.4 model)

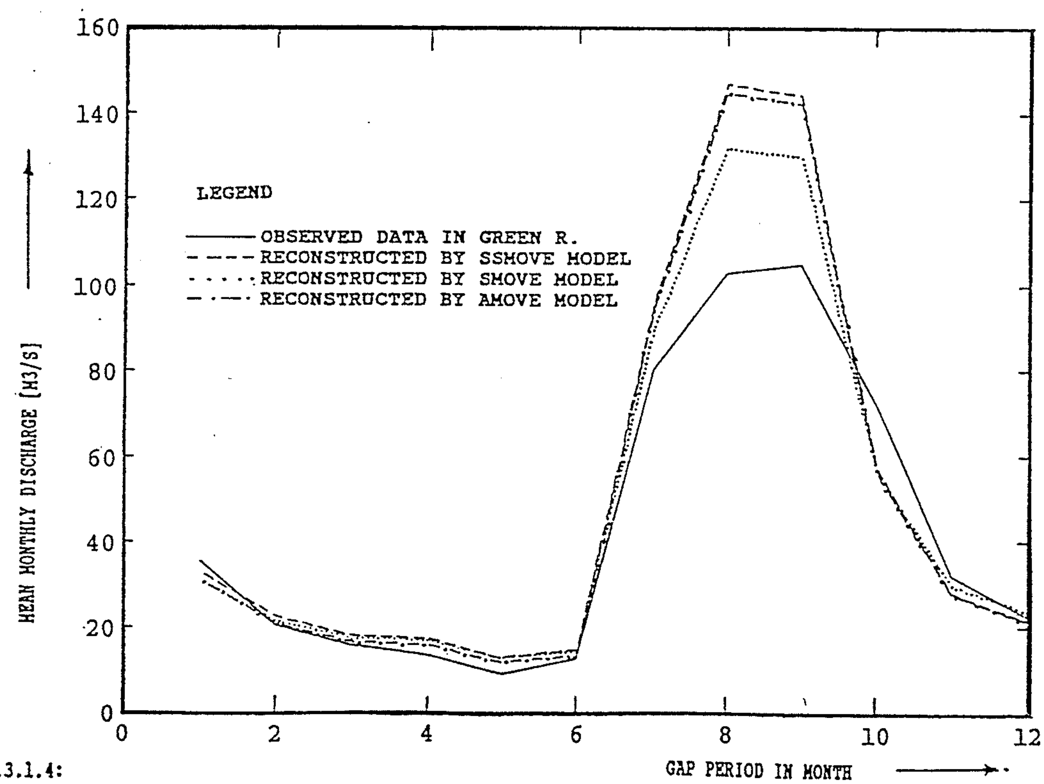


Fig. 5.3.1.4:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945]
by Rutherford Ck. by AMOVE, SMOVE, SSMOVE models (varities of MOVE.4 model)

ASSESSMENT CATEGORY 3

MULTIVARIATE MODELS

Table 5.3.6.a: shows that Soo R. performed best in terms of ed, ew, eo. The Rutherford Ck. performed superior to Soo R. in terms of P. Both SESTRNALL model and MULBS model incorporating Soo river and Rutherford Ck. information [Fig. 5.3.2], showed poor fit of reconstruction shape. This deficiency is specifically prominent at the peak.

REGRESSION AND MOVE.4 MODELS

Both Tables 5.3.6.(b,c) and Figs. 5.3.3.(1, 2, 3 and 4] show that Show better performance of Soo R. over Rutherford Ck.

Table 5.3.6.a: Table of ranking of performance of various sources for the multivariate models (LILL cluster)

MODEL	Source rank in terms of estimation error in dry season [ed]			Source rank in terms of estimation error in wet season [ew]			Source rank in terms of deviation at peak reconstd.- obsvd. [P]			Source rank in terms of overall estimation error [eo]		
-	So.R.	Rd.Ck.	S.S.	So.R.	Rd.Ck.	S.S.	Rd.Ck.	S.S	So.R.	So.R.	Rd.Ck.	S.S.
Multiv. M.	0.16	0.27	0.58	0.15	0.17	0.37	0.3	-5.4	8.0	0.16	0.22	0.49

Table 5.3.6.b: Table of ranking of performance of various base R. for each sampling scenario regression model (LILL cluster)

MODEL	River rank in terms of estimation error in dry season [ed]		River rank in terms of estimation error in wet season [ew]		River rank in terms of deviation at peak reconstd.- obsvd. [P]		River rank in terms of overall estimation error [eo]	
-	So.R.	Rd.Ck.	So.R.	Rd.Ck.	So.R.	Rd.Ck.	So.R.	Rd.Ck.
REG	0.13	0.48	0.12	0.19	5.4	24.5	0.13	0.37
SREG	0.09	0.32	0.15	0.26	4.8	18.9	0.12	0.29
SSREG	0.11	0.27	0.18	0.25	10.3	32.9	0.15	0.26

Table 5.3.6.c: Table of ranking of performance of various base R. for each sampling scenario MOVE.4 model (LILL cluster)

MODEL	River rank in terms of estimation error in dry season [ed]		River rank in terms of estimation error in wet season [ew]		River rank in terms of deviation at peak reconstd.- obsvd. [P]		River rank in terms of overall estimation error [eo]	
-	So.R.	Rd.Ck.	So.R.	Rd.Ck.	So.R.	Rd.Ck.	So.R.	Rd.Ck.
AMOVE	0.13	0.20	0.11	0.25	4.7	37.2	0.12	0.23
SMOVE	0.10	0.26	0.15	0.18	4.8	24.9	0.13	0.23
SSMOVE	0.09	0.29	0.18	0.26	9.9	39.3	0.15	0.28

So. R.= Soo R.
Rd.Ck.= Rutherford Ck.
S.S = self series

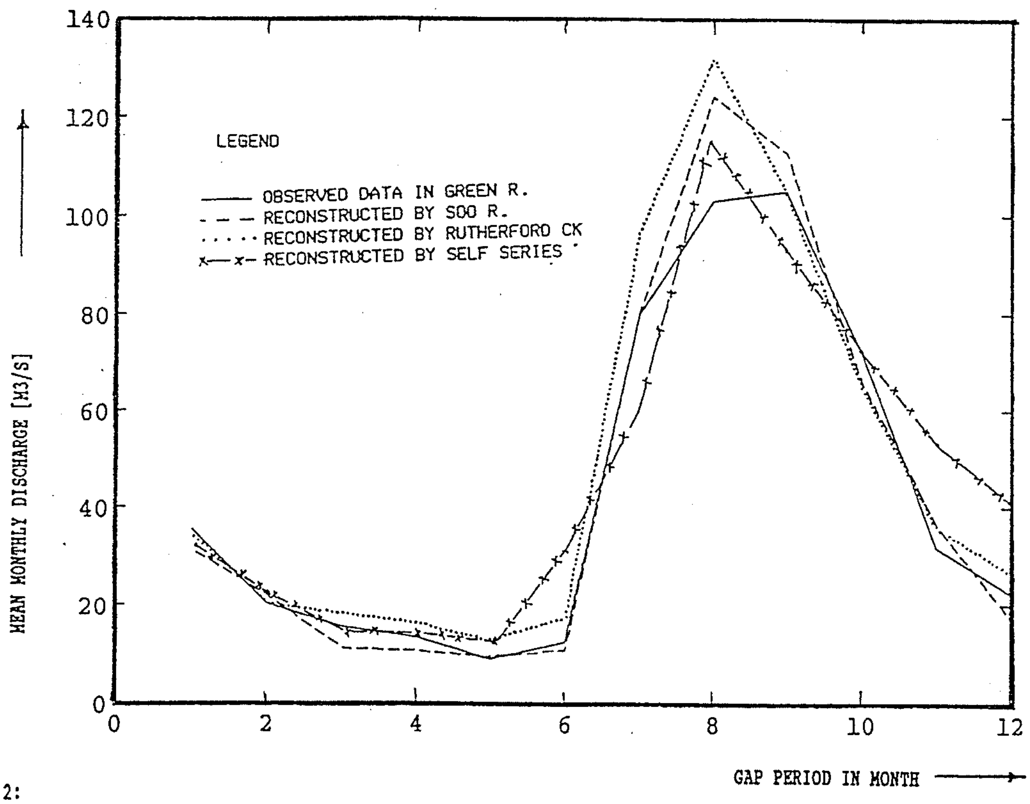


Fig.5.3.2:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945] by multivariate models using variable information source

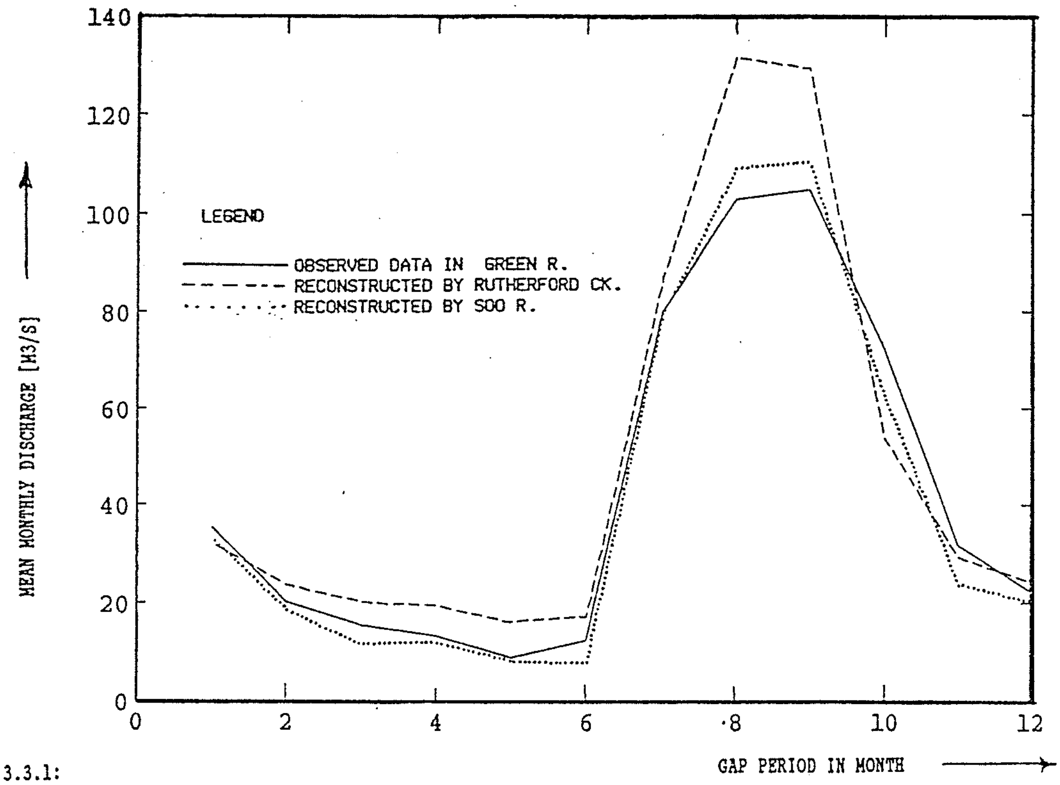


Fig.5.3.3.1:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945]
by REG model using variable information source

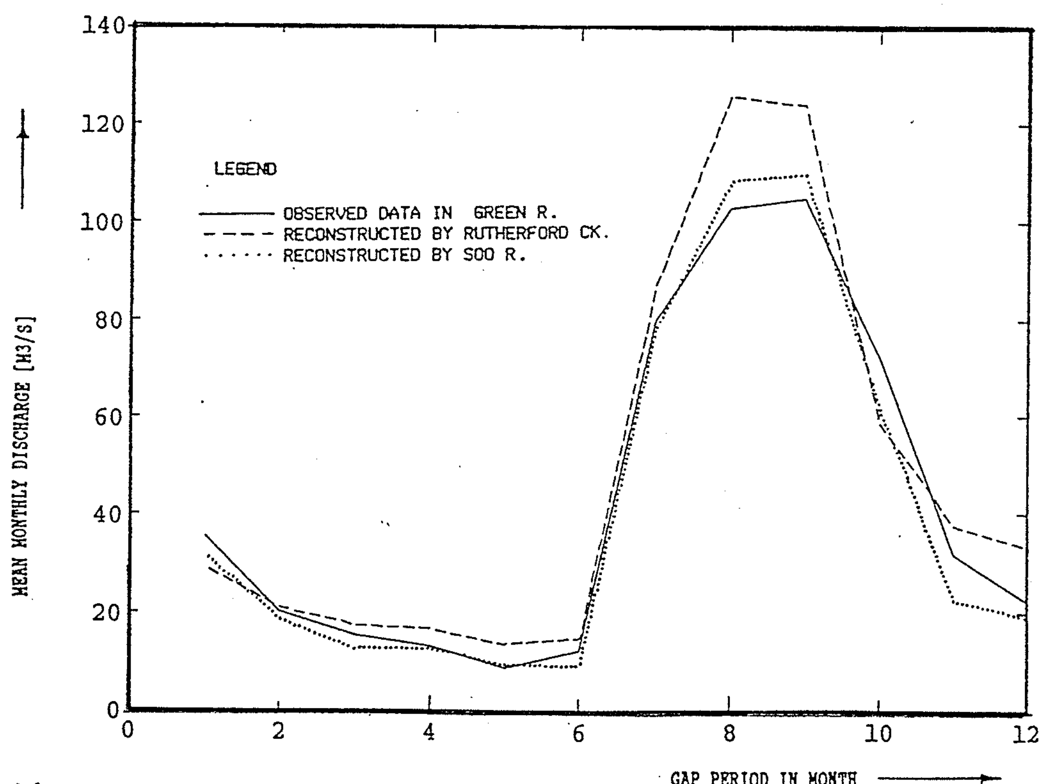


Fig.5.3.3.2:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945]
by SREG model using variable information source

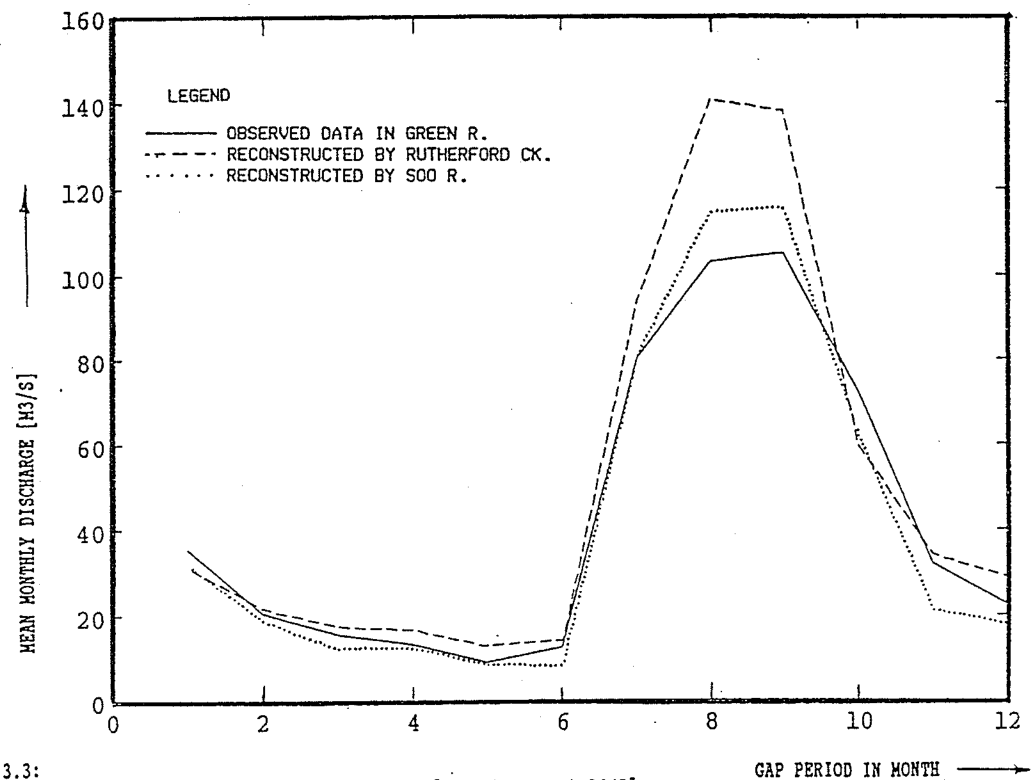


Fig.5.3.3.3:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945] by SSREG model using variable information source

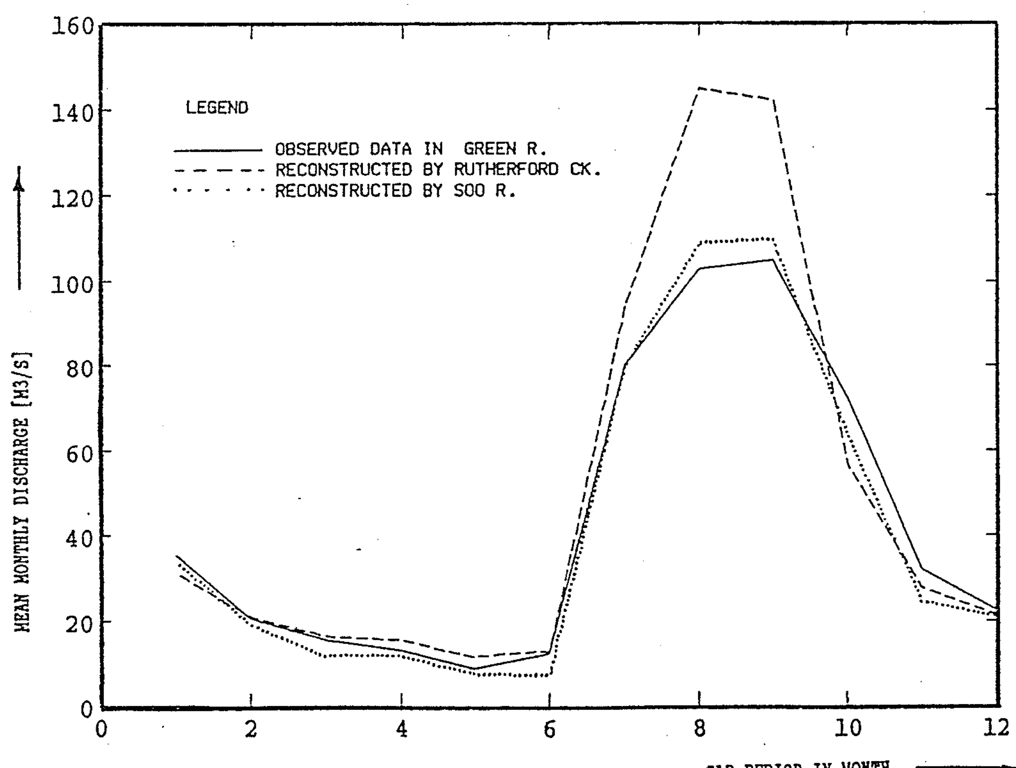


Fig.5.3.3.4:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945] by AMOVE model using variable information source

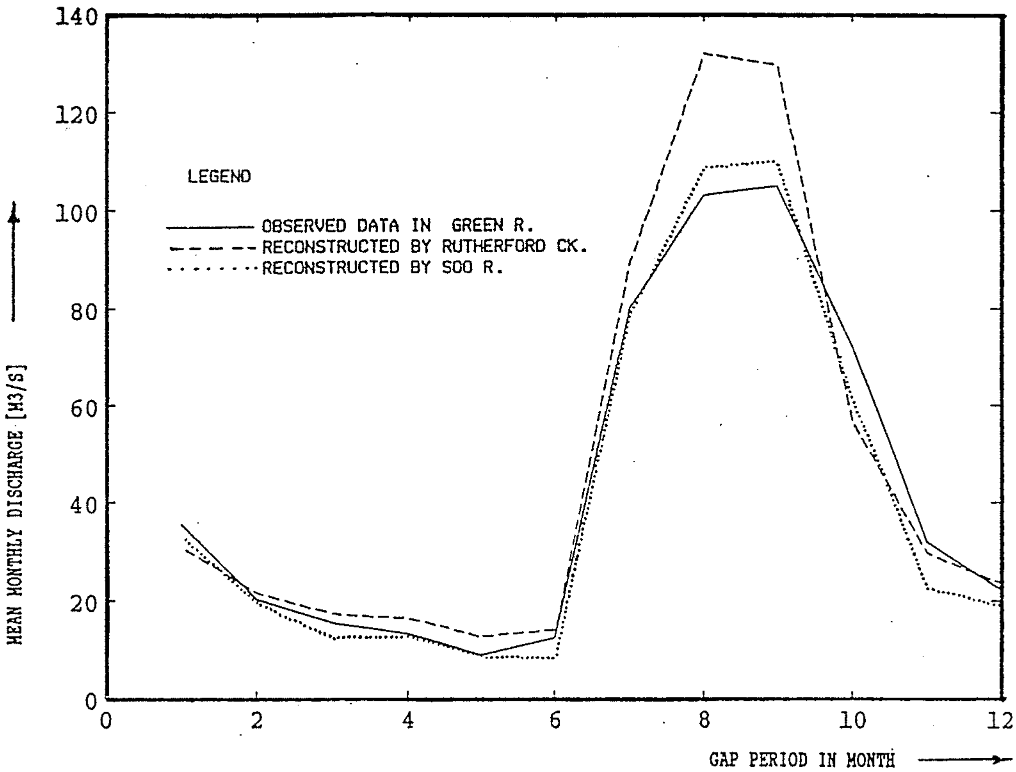


Fig.5.3.3.5:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945]
by SMOVE model using variable information source

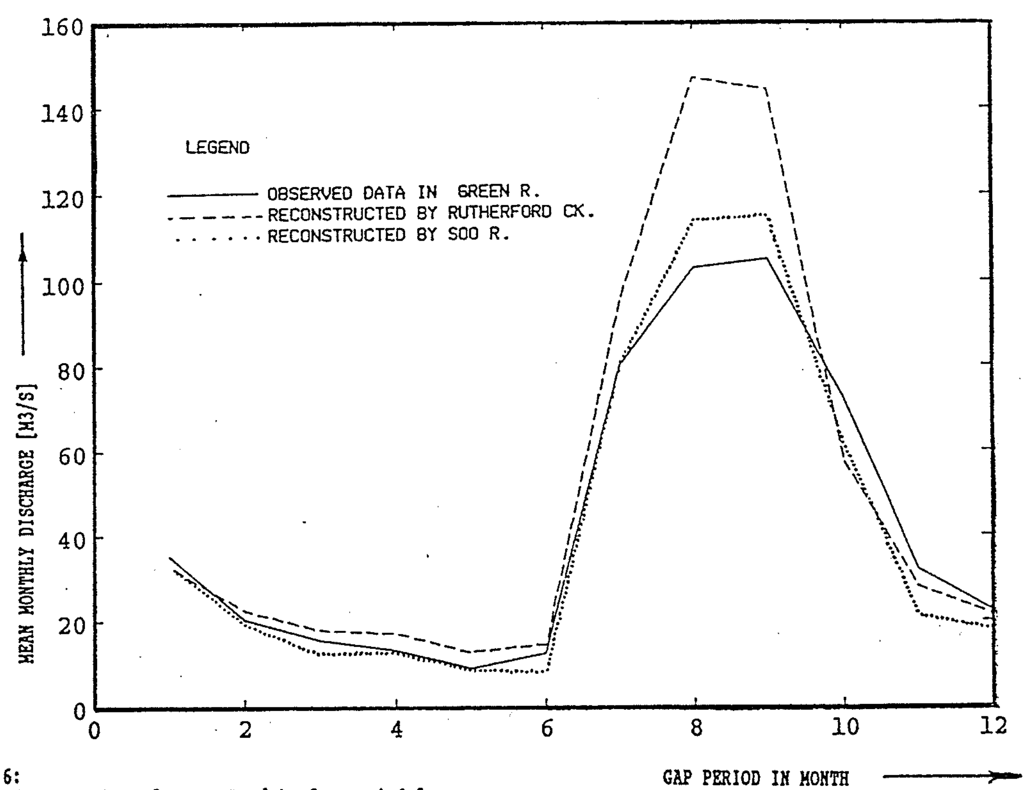


Fig.5.3.3.6:
Comparison of infilling of Green R. data for period [Nov. 1944- Oct.1945]
by SSMOVE model using variable information source

The assessment category two shows that in all cases the regression model, heterogenous sample gave the least ew. This is completely different from the case of UPB and LB cluster. In terms of the remaining criteria seasonal consideration in sampling shows beneficial effects. The positive influence of seasonal sampling was found effective to a lesser degree than the previous case. With respect to the shape reconstruction point of view none of the multivariate model could reconstruct the observed shape.

Poor fit of MULBS models can be explained by the poor fitting of multivariate normal distribution. This lack of fit is most probably due to the inadequate number of multivariate seasonal segments . Poor fit of SESTRNALL model can be attributed to both the poor information transfer by seasonal transition as well as to the inadequate number of multivariate observations. Quality of reconstruction by the REG, SREG and SSREG models are analogous to the AMOVE, SMOVE and SSMOVE models in both cases of {Green R., Soo R.} and {Green R., Rutherford Ck.} pairs. Better infilling performance Soo R. than that of by Rutherford Ck. in all varieties of regression and MOVE models can be explained by the higher zero-lag cross correlation coefficient between {Green R., Soo R.} pair than the one between { Green R., Rutherford Ck.} pair.

DISCUSSION

The proposed multivariate models assume the presence of significant relationship between the predictor and predicted variate and this relationship should be consistent. The dependence is given by Σ_{21} terms of Equation 3.7. The correlation-coefficient matrix in case of MULBS model shows the relationship is of mostly lag-0 nature. But in case of SESTRNALL model the dependence is found to be small. Moreover, the insignificant amount of entropy reduction indicates erratic nature of transition [Table 5.4.1]. Considering these factors, SESTRNALL model should not be expected to perform satisfactorily and in fact it did perform poorly.

The MULBS model performed poorly in case of LILL cluster, it can be conjectured that the available data 22 observations was inadequate to ensure the multivariate normality.

It was expected that the regression and MOVE.4 models would perform very good if the samples are chosen by imposing some homogeneity criteria. It was assumed that the cross-correlation coefficient may increase under the seasonal homogeneity condition. But the Table: 5.4.2 shows decrease in cross correlation coefficient. Only for wet seasonal sample of {Johns Ck.,Craig Ck.} pair, the correlation coefficient increases. In wet season, in very few cases the original cross- correlation coefficient remained constant. In most of

Table 5.4.1: Overall assessment of multivariate models

Cluster	Information Source	Entropy Reduction [%]	Estimation Error in Dry S.	Estimation Error in Wet S.	Overall Error
	Johns Ck.	78.88	0.18	0.08	0.14
	Dunlap Ck.	70.42	0.23	0.21	0.22
UPB	Cowpasture R.	59.02	0.27	0.14	0.22
	Self Series	5.00	0.39	0.52	0.46
	Reed Ck.	46.38	0.38	0.16	0.29
LB	Roanoke R.	72.24	0.23	0.20	0.21
	Self Series	8.82	0.29	0.47	0.39
	Soo R.	64.91	0.16	0.15	0.16
LILL	Rutherford Ck.	20.18	0.27	0.17	0.22
	Self Series	6.48	0.58	0.39	0.49

Table 5.4.2: Cross correlation coefficient in various samples

Cluster	River	Heterogenous sample	Seasonal sample dry	Seasonal sample Wet	Selected seasonal sample dry	Selected seasonal sample Wet
	Johns Ck.	0.98	0.97	0.99	0.95	0.98
UPB	Dunlap Ck.	0.95	0.89	0.93	0.86	0.93
	Cowpasture R.	0.93	0.87	0.90	0.83	0.88
	Roanoke R.	0.93	0.88	0.93	0.79	0.92
LB	Reed Ck.	0.86	0.79	0.86	0.70	0.78
	Soo R.	0.97	0.94	0.93	0.84	0.96
LILL	Rutherford Ck.	0.95	0.81	0.86	0.88	0.89

the cases it dropped. This reduction in cross correlation coefficient in dry seasonal sample was more than the cross correlation coefficient in wet seasonal sample. In the dry season, the estimation errors were on an average more than the errors in wet season.

The procedure of seasonal segmentation or assumption of six months long season over a fixed period of time could be considered to be a factor behind such a drop in correlation. For example in table A.2.3, in the Craig Ck. data matrix, there are very high flow in the month of June of the 15th year and of the 19th year. There were also high flows in the month of November of the 20th year and of the 28th year. The months of November and June have been assigned to dry season. It can be conjectured that these high flows may affect the homogeneity or in extreme cases behave like an outlier. In a dense sample such as the original data set it is balanced to some extent, but when the data is selected, the size of the sample reduces and if the reduced sample contains such abnormal flow, it may degrade the fitted line. The efficacy of seasonal group characteristics for sample selection is thus found to be beneficial for limited cases.

Chapter 6

CONCLUSION

In this thesis an effort was made to investigate the efficacy of seasonal group characteristics for monthly flow data infilling purpose. Two multivariate models were developed: MULBS and SESTRNALL models. The MULBS model reconstruct a seasonal flow group of a river with missing data conditioned on the simultaneously observed flow group in the nearby located river. The MULBS model was based on the assumption of consistent nature of simultaneously observed inter dependent flow groups {missing data river, base river}.

The SESTRNALL model on the other hand reconstruct a flow group of a river conditioned on the preceding flow group in the same river. This model was based on the assumption of consistent inter seasonal dependence.

The models were applied to the real world data. The model assumptions were also tested in the by using conditional entropy principle. The SESTRNALL model performed very poorly. The MULBS model showed satisfactory performance only under the constraints of i) longer period of concurrent data, ii) close proximity of and seasonal similarity of missing data river and a base river. Moreover, it involves very complex multidimensional computational procedures which assumes the data to be multivariate normal.

Further in this thesis, the use of group characteristics of monthly flow data in the extraction of relative homogeneous samples for regression models were investigated. The occurrence of groups of low and high flow vary over the length of a year. Thus a segmentation scheme (such as the one used in this thesis) based on the assumption that the flow groups occur over some fixed calendar months runs the risk of assigning the flows to an incorrect season. Caution should be exercised in using such a sample. Under such a situation, the use of 12 regression equations for 12 months may run into even greater risk of incorrect estimate of missing data because the sample size would be reduced by a factor of six.

SCOPE FOR FURTHER RESEARCH

Following issues were identified during the course of this thesis, which are open to future study.

1. Multivariate models developed here are capable of reconstruction of completely missing seasonal segments and these models do not deal with the situation of partially missing seasonal segments. Considering $\mathbf{X}=[\mathbf{X}_1,\mathbf{X}_2]^T$ as the simultaneously observed seasonal segment, **EM** algorithm [Johnson and Wichern 1988] could be considered as a candidate solution for the problem.

2. In some cases, poor performance of SSREG and SSMOVE over REG and AMOVE is suspected to be due to the specific clustering algorithm used in this thesis. The selected seasonal segments as well as well as the entropy reduction is expected to vary with the variation of clustering algorithm and the metric under consideration. One can study the entropy reduction and the corresponding infilling performance under variation of the clustering algorithm as well as the clustering metric.

3. The study of the data matrix shows that groups of flow do occur, but the occurrences are neither fixed by some calendar months nor do they stretch over six months time (also supported by the cross-correlation matrix). The entropy analysis shows consistency in the simultaneous occurrence of flows in the neighbouring rivers. Therefore a multivariate model of lesser dimension can be considered for the future research with the parameters calibrated from all available data. This model can be compared to the existing multivariate models that uses standardized data [Kottegoda and Elgy (1977)].

APPENDIX A.1

A.1.1 K- MEANS ALGORITHM

In this algorithm, an initial number of clusters **k** are assumed and **k** cluster centres are arbitrarily chosen from the data. The pattern vectors are assigned to the cluster with which it has the minimum **Euclidian distance** . The steps of this algorithm are [Tou and Gonzalez (1974)] :

1. At the 1st iteration step, as initial seed, choose **k** cluster means $Z_1(1)$ .... $Z_{k(1)}$.
2. At any **k**th iteration step, assign a pattern vector **X** to any of the **k** clusters satisfying following inequality :

   $X \in S_j$ if $\| X - Z_j(k) \| < \| X - Z_i \|$ ......[A.1.1]

   with :

   $i=1.....k$ and $i \neq j$

   . $\| X - Z_i \|$ = Euclidean distance between **X** and Z_i

   $= \sqrt{(X - Z_i)^T (X - Z_i)}$

   . $S_j(k)$ = Set of samples belonging to the cluster represented by $Z_j(k)$ at **k**th iteration step
3. Compute new cluster centres from the clusters formed in step 2.[Consider the estimate of the sample mean of the set S_j **(k)** to be the new cluster centre $Z_j(k+1)$].
4. For all the clusters , repeat step-2 and step-3 till the cluster mean at **k**th iteration is not significantly different from the cluster mean at **(k+1)**th iteration step.

A.1.2 PANU PROPOSED INFILLING PROCEDURE AND THE RATIONALE BEHIND THE MODIFICATION

Panu (1978) proposed a data infilling algorithm consisting of following steps:

- Finding of the most probable sub-cluster representing the missing seasonal segment by utilizing transitional probability of the seasonal segments.
- Finding of the most probable Mahalanobis distance of the missing seasonal segment from the relationship between the transitional seasonal segments.
- Generating seasonal segment with a configuration specified by the configuration of the most probable sub-cluster and constraning it to have the most probable Mahalanobis distance .

This procedure assumes that the Markovian transition pattern is adequate to recognize the sub-cluster to which the missing segment is expected to belong to and subsequent generation of a seasonal segment by specifying the marginal configuration of the sub-cluster obtained from the historical data set is adequate to mimic the missing seasonal segment provided it is constrained within the Mahalanobis distance.

Even if the seasonal transition were of perfectly Markovian nature, thus, rendering the correct sub-cluster, the Mahalanobis distance alone is no binding constraint for obtaining the unique estimate of missing seasonal segment. Let μ, Σ denote the configuration of the most probable sub-cluster , then for a known Mahalanobis distance c^2 , any seasonal segment X will satisfy the equation of equiprobability contour given by :

$$(X-\mu)^T \Sigma^{-1} (X-\mu) = c^2 \text{ .... [A.1.2]}$$

Since μ and Σ are estimated from the historical data conforming to the configuration of the sub-cluster, the Mahalanobis distance is the only constraint considered here. But there are infinite combinations of components of X satisfying equation A.1.2. This is shown in [Fig. A.1.1] where two pattern vector $X_1 = [x_{11}, x_{12}]^T$ and $X_2 = [x_{21}, x_{22}]^T$ represented by the points P_1 and P_2 both satisfy [Eq. A.1.2] but each consisting of combinations of elements of totally different magnitude. Unless, the configuration is adapted according to preceding observed segment, the model proposed by Panu (1978) would generate estimates of missing segment within a certain probability level (given by the Mahalanobis distance) from a cluster with a rigid configuration computed from the historical data. An empirical data reconstruction model is developed which uses the inter-seasonal dependence (Panu 1978) and furthermore incorporate the flexibility in distributional configuration of the prospective candidate

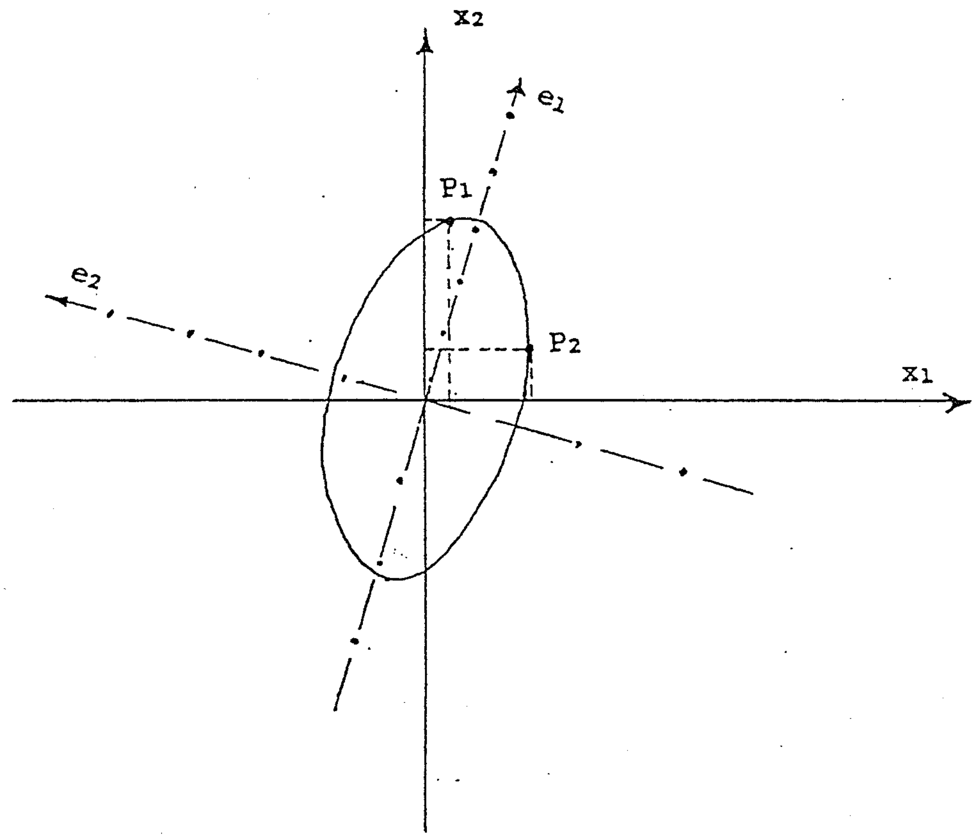


Fig.A.1.1: Equiprobabilty contour of a bivariate normal distribution with

$[\sigma^2_{22} > \sigma^2_{11}]$ and $[\rho_{12} > 0]$

seasonal segment corresponding to a missing seasonal segment.

A.1.3 DESCRIPTION OF THE INFILLING EQUATION CORRESPONDING TO MOVE.4 MODEL

In this thesis, MOVE.4 [Vogel and Stedinger (1985)] is used under three different sampling scenarios. In each case, n_2 period of missing record of the series **Y** are estimated by the corresponding observations existing in the series **X**. Both series are assumed to have a common record of n_1 period. The general infilling estimate of the missing data by MOVE.4 is :

$$\hat{y}_i = a' + b\,(x_i - \bar{x}_2) \quad \text{.............} \quad [A.1.3]$$

where,

$$a' = \frac{(n_1 + n_2)\hat{\mu}_y^{*} - n_1 \bar{y}_1}{n_2}$$

$$b^2 = \frac{[(n_1+n_2-1)\,\hat{\sigma}_y^{*2} - (n_1-1)s_{y1}^2 - n_1(\bar{y}_1 - \hat{\mu}_y^{*})^2 - n_2(a' - \hat{\mu}_y^{*})^2]}{[\,(n_2-1)s_{x2}^2\,]}$$

$$\hat{\mu}^{*} = \bar{y}_1 + \frac{(n_1-3)r^2}{(n_1-4)r^2+1} \cdot \frac{n_2}{(n_1+n_2)} \cdot \hat{\beta} \cdot (\bar{x}_2 - \bar{x}_1)$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n1}(x_i - \bar{x}_1)\cdot(y_i - \bar{y}_1)}{\sum_{i=1}^{n1}(x_i - \bar{x}_1)^2}$$

$$r = \hat{\beta} \cdot \frac{s_{x1}}{s_{y1}}$$

$$\overline{y_1} = \frac{1}{n_1} \sum_{i=1}^{n1} y_i$$

$$\overline{x_1} = \frac{1}{n_1} \sum_{i=1}^{n1} x_i$$

$$s_{y1}^2 = \frac{1}{(n_1 - 1)} \sum_{i=1}^{n1} (y_i - \overline{y_1})^2$$

$$s_{x1}^2 = \frac{1}{(n_1 - 1)} \sum_{i=1}^{n1} (x_i - \overline{x_1})^2$$

$$s_{x2}^2 = \frac{1}{(n_2 - 1)} \sum_{i=n1+1}^{n1+n2} (x_i - \overline{x_2})^2$$

$$\overline{x_2} = \frac{1}{n_2} \sum_{i=n1+1}^{n1+n2} x_i$$

A.1.4 DESCRIPTION OF THE INFILLING EQUATION CORRESPONDING TO REGRESSION MODEL

Least square regression model is used to infill $\mathbf{n_2}$ period of missing record of the series **Y** by the corresponding observations existing in the series **X** . For both of these series, common record of $\mathbf{n_1}$ period is considered. The general infilling equation
is :

$$\hat{y}_i = b_0 + b_1 x_i + e_i \quad \text{..............[A.1.4]}$$

where,

$$b_1 = \frac{\sum_{i=1}^{n1} (x_i - \overline{x}_1)(y_i - \overline{y}_1)}{\sum_{i=1}^{n1} (x_i - \overline{x}_1)^2}$$

$$b_0 = \overline{y} - b_1 \overline{x_1}$$

e_i = normally distributed random error term

APPENDIX A.2

A.2.1 SEASONAL SEGMENTATION AND SLIDING OF DATA MATRIX

These procedures are explained in case of Dunlap Creek (a member of UPB cluster). From the reported data matrix, 31 years of data for period of (Oct. 1958- Sept. 1988) are taken as input data matrix [Table A.2.1]. The correlogram [Fig. 1.3] indicates two seasons per year. For each row of the data matrix, the monthly flows are ranked in ascending order of magnitude, the six months having the 1st six lowest flows are assigned to the group-1 and the six months corresponding to the six next lowest flows are assigned to the group-2 [Table A.2.1]. This procedure is repeated for each year. The number of assignment of each month to each of the two groups are counted. The chain of six months [June- Nov.] and [Dec.- May] respectively shows maximum number of assignment to the group-1 (dry season) and group-2 (Wet season). This procedure adopted for finding of the association of the months to the season is referred as **Seasonal segmentation** .

In order to facilitate the beginning of the data matrix to coincide with the beginning of the season, the data matrix is slided forward by deleting eight data [Oct.1958-May 1958] of the 1st row and appending eight more data [Oct.1989 -May 1989] at the end of the data matrix . Thus, the initially entered data matrix [Oct.1958 - Sep.1988] is **slided** to the data matrix [June 1958 - May 1989]. The segmentation done on the slided

Table A.2.1: Segmentation of the original data matrix of Dunlap Creek (Oct.1958-Sep.1988)

Input Data matrix

Year	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sep.
1	61.9	130.0	278.0	235.0	357.0	582.0	529.0	395.0	46.7	93.4	145.0	27.4
2	24.1	22.7	41.9	115.0	96.0	237.0	420.0	103.0	79.9	24.3	29.2	21.9
3	59.1	98.7	176.0	130.0	299.0	476.0	393.0	271.0	65.9	26.5	22.6	37.8
4	21.2	25.8	28.8	60.6	396.0	292.0	317.0	205.0	167.0	33.2	23.3	20.0
5	157.0	114.0	376.0	259.0	333.0	557.0	182.0	64.7	47.1	30.1	27.9	16.8
6	31.5	98.2	129.0	314.0	120.0	845.0	72.5	48.4	47.0	23.8	16.2	15.8
7	18.8	26.4	31.4	247.0	208.0	440.0	193.0	62.8	26.9	16.6	15.0	19.7
8	33.5	48.4	123.0	344.0	294.0	466.0	288.0	119.0	45.6	33.9	23.5	16.9
9	30.4	20.2	22.7	33.7	321.0	175.0	121.0	310.0	49.0	14.3	20.4	71.5
10	160.0	113.0	168.0	330.0	220.0	689.0	142.0	339.0	89.0	30.3	22.8	20.3
11	53.0	38.0	202.0	294.0	207.0	235.0	159.0	176.0	71.0	22.6	23.3	13.3
12	60.5	107.0	75.4	146.0	195.0	189.0	141.0	81.3	72.5	31.9	380.0	30.8
13	23.5	29.2	345.0	228.0	304.0	129.0	263.0	92.9	39.6	17.3	20.4	11.0
14	28.1	120.0	113.0	208.0	580.0	267.0	247.0	453.0	124.0	35.3	46.4	42.5
15	111.0	79.4	201.0	277.0	658.0	242.0	394.0	287.0	584.0	358.0	79.4	24.4
16	75.0	487.0	603.0	189.0	431.0	622.0	625.0	506.0	157.0	84.2	47.0	20.4
17	68.2	159.0	694.0	515.0	200.0	405.0	220.0	276.0	174.0	69.0	41.3	49.5
18	38.5	41.4	220.0	361.0	506.0	608.0	268.0	453.0	126.0	65.5	38.8	39.3
19	153.0	83.9	74.7	296.0	213.0	181.0	105.0	117.0	171.0	32.8	20.7	17.4
20	327.0	102.0	206.0	47.6	162.0	321.0	524.0	49.5	31.9	22.7	22.5	27.6
21	94.5	353.0	229.0	462.0	109.0	827.0	281.0	336.0	50.4	35.4	31.8	18.2
22	17.2	23.1	108.0	468.0	542.0	411.0	200.0	289.0	237.0	61.4	46.8	164.0
23	243.0	347.0	173.0	366.0	140.0	657.0	547.0	155.0	49.7	53.0	31.3	27.6
24	20.5	34.7	29.5	24.2	119.0	94.4	176.0	355.0	246.0	36.3	16.4	17.2
25	30.5	21.0	113.0	249.0	461.0	422.0	225.0	105.0	330.0	32.9	26.5	17.7
26	33.3	89.7	206.0	61.0	291.0	401.0	603.0	211.0	74.7	31.0	17.9	11.2
27	45.8	70.8	273.0	125.0	619.0	515.0	436.0	340.0	40.0	60.7	514.0	72.0
28	57.4	141.0	180.0	218.0	446.0	199.0	136.0	105.0	40.5	25.8	98.3	21.2
29	19.4	659.0	153.0	58.5	276.0	305.0	54.7	265.0	39.1	29.1	39.4	34.4
30	23.2	73.8	329.0	257.0	365.0	581.0	1071.0	202.0	63.0	38.0	17.3	61.3
31	23.9	58.6	108.0	171.0	121.0	59.1	91.5	170.0	28.8	18.7	16.0	25.6

Yearly Ranking of monthly data

Year	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sep.
1	3	5	8	7	9	12	11	10	2	4	6	1
2	3	2	6	10	8	11	12	9	7	4	5	1
3	4	6	8	7	10	12	11	9	5	2	1	3
4	2	4	5	7	12	10	11	9	8	6	3	1
5	7	6	11	9	10	12	8	5	4	3	2	1
6	4	8	10	11	9	12	7	6	5	3	2	1
7	3	5	7	11	10	12	9	8	6	2	1	4
8	3	6	8	11	10	12	9	7	5	4	2	1
9	5	2	4	6	12	10	9	11	7	1	3	8
10	7	5	8	10	9	12	6	11	4	3	2	1
11	5	4	9	12	10	11	7	8	6	2	3	1
12	3	7	5	9	11	10	8	6	4	2	12	1
13	4	5	12	9	11	8	10	7	6	2	3	1
14	1	6	5	8	12	10	9	11	7	2	4	3
15	4	3	5	7	12	6	10	8	11	9	2	1
16	3	8	10	6	7	11	12	9	5	4	2	1
17	3	5	12	11	7	10	8	9	6	4	1	2
18	1	4	7	9	11	12	8	10	6	5	2	3
19	8	5	4	12	11	10	6	7	9	3	2	1
20	11	7	9	5	8	10	12	6	4	2	1	3
21	5	10	7	11	6	12	8	9	4	3	2	1
22	1	2	5	11	12	10	7	9	8	4	3	6
23	8	9	7	10	5	12	11	6	3	4	2	1
24	3	6	5	4	9	8	10	12	11	7	1	2
25	4	2	7	9	12	11	8	6	10	5	3	1
26	4	7	8	5	10	11	12	9	6	3	2	1
27	2	4	7	6	12	11	9	8	1	3	10	5
28	4	8	9	11	12	10	7	6	3	2	5	1
29	1	12	8	7	10	11	6	9	4	2	5	3
30	2	6	9	8	10	11	12	7	5	3	1	4
31	3	6	9	12	10	7	8	11	5	2	1	4

(Cotd.)

(Cotd.)

Classified data matrix

Year	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sep.
1	1	1	2	2	2	2	2	2	1	1	1	1
2	1	1	1	2	2	2	2	2	2	1	1	1
3	1	1	2	2	2	2	2	2	1	1	1	1
4	1	1	1	2	2	2	2	2	2	1	1	1
5	2	1	2	2	2	2	2	1	1	1	1	1
6	1	2	2	2	2	2	2	1	1	1	1	1
7	1	1	2	2	2	2	2	2	1	1	1	1
8	1	1	2	2	2	2	2	2	1	1	1	1
9	1	1	1	1	2	2	2	2	2	1	1	2
10	2	1	2	2	2	2	1	2	1	1	1	1
11	1	1	2	2	2	2	2	2	1	1	1	1
12	1	2	1	2	2	2	2	1	1	1	2	1
13	1	1	2	2	2	2	2	2	1	1	1	1
14	1	1	1	2	2	2	2	2	2	1	1	1
15	1	1	1	2	2	1	2	2	2	2	1	1
16	1	2	2	1	2	2	2	2	1	1	1	1
17	1	1	2	2	2	2	2	2	1	1	1	1
18	1	1	2	2	2	1	2	2	1	1	1	1
19	2	1	1	2	2	2	2	2	2	1	1	1
20	2	2	2	1	2	2	2	1	1	1	1	1
21	1	2	2	2	1	2	2	2	1	1	1	1
22	1	1	1	2	2	2	2	2	2	1	1	1
23	2	2	2	2	1	2	2	1	1	1	1	1
24	1	1	1	1	2	2	2	2	2	2	1	1
25	1	1	2	2	2	2	2	1	2	1	1	1
26	1	2	2	1	2	2	2	2	1	1	1	1
27	1	1	2	1	2	2	2	2	1	1	2	1
28	1	2	2	2	2	2	2	1	1	1	1	1
29	1	2	2	2	2	2	1	2	1	1	1	1
30	1	1	2	2	2	2	2	2	1	1	1	1
31	1	1	2	2	2	2	2	2	1	1	1	1

Total number of assignment of the months to the yearly seasonal groups

Group	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May	June	July	Aug.	Sep.
1	26	22	9	6	2	1	3	7	22	29	29	30
2	5	9	22	25	29	30	29	24	9	2	2	1

Table A.2.2: Segmentation of the slided data matrix of Dunlap Creek (June 1958-May 1989)

Slided Data matrix

Year	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
1	46.7	93.4	145.0	27.4	24.1	22.7	41.9	115.0	96.0	237.0	420.0	103.0
2	79.9	24.3	29.2	21.9	59.1	98.7	176.0	130.0	299.0	476.0	393.0	271.0
3	65.9	26.5	22.6	37.8	21.2	25.8	28.8	60.6	396.0	292.0	317.0	205.0
4	167.0	33.2	23.3	20.0	157.0	114.0	376.0	259.0	333.0	557.0	182.0	64.7
5	47.1	30.1	27.9	16.8	31.5	98.2	129.0	314.0	120.0	845.0	72.5	48.4
6	47.0	23.8	16.2	15.8	18.8	26.4	31.4	247.0	208.0	440.0	193.0	62.8
7	26.9	16.6	15.0	19.7	33.5	48.4	123.0	344.0	294.0	466.0	288.0	119.0
8	45.6	33.9	23.5	16.9	30.4	20.2	22.7	33.7	321.0	175.0	121.0	310.0
9	49.0	14.3	20.4	71.5	160.0	113.0	168.0	330.0	220.0	689.0	142.0	339.0
10	89.0	30.3	22.8	20.3	53.0	38.0	202.0	294.0	207.0	235.0	159.0	176.0
11	71.0	22.6	23.3	13.3	60.5	107.0	75.4	146.0	195.0	189.0	141.0	81.3
12	72.5	31.9	380.0	30.3	23.3	29.2	345.0	228.0	304.0	129.0	263.0	92.9
13	39.6	17.3	20.4	11.0	28.1	120.0	113.0	208.0	580.0	267.0	247.0	453.0
14	124.0	35.3	46.4	42.5	111.0	79.4	201.0	277.0	658.0	242.0	394.0	287.0
15	584.0	358.0	79.4	24.4	75.0	487.0	603.0	189.0	431.0	622.0	625.0	506.0
16	157.0	84.2	47.0	20.4	68.2	159.0	694.0	515.0	200.0	405.0	220.0	276.0
17	174.0	69.0	41.3	49.5	38.5	41.4	220.0	361.0	506.0	608.0	258.0	453.0
18	126.0	65.5	38.8	39.3	153.0	83.9	74.7	296.0	213.0	181.0	105.0	117.0
19	171.0	32.8	20.7	17.4	327.0	102.0	206.0	47.6	162.0	321.0	524.0	49.5
20	31.9	22.7	22.5	27.6	94.5	353.0	229.0	462.0	109.0	827.0	281.0	336.0
21	50.4	35.4	31.8	18.2	17.2	23.1	108.0	468.0	542.0	411.0	200.0	289.0
22	237.0	61.4	46.8	164.0	243.0	347.0	173.0	366.0	140.0	657.0	547.0	155.0
23	49.7	53.0	31.3	27.6	20.5	34.7	29.5	24.2	119.0	94.4	176.0	355.0
24	246.0	36.3	16.4	17.2	30.5	21.0	113.0	249.0	461.0	422.0	225.0	105.0
25	330.0	32.9	26.5	17.7	33.3	89.7	206.0	61.0	291.0	401.0	603.0	211.0
26	74.7	31.0	17.9	11.2	45.8	70.8	273.0	125.0	619.0	515.0	436.0	340.0
27	40.0	60.7	514.0	72.0	57.4	141.0	180.0	218.0	446.0	199.0	136.0	105.0
28	40.5	25.8	98.3	21.2	19.4	659.0	153.0	58.5	276.0	305.0	54.7	265.0
29	39.1	29.1	39.4	34.4	23.2	73.8	329.0	257.0	365.0	581.0	1071.0	202.0
30	63.0	38.0	17.3	61.3	23.9	58.6	108.0	171.0	121.0	59.1	91.5	170.0
31	28.8	18.7	16.0	25.6	22.0	74.6	96.4	206.0	172.0	247.0	308.0	536.0

Yearly Ranking of monthly data

Year	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
1	5	6	10	3	2	1	4	9	7	11	12	8
2	5	2	3	1	4	6	8	7	10	12	11	9
3	8	4	2	6	1	3	5	7	12	10	11	9
4	7	3	2	1	6	5	11	9	10	12	8	4
5	5	3	2	1	4	8	10	11	9	12	7	6
6	7	4	2	1	3	5	6	11	10	12	9	8
7	4	2	1	3	5	6	8	11	10	12	9	7
8	8	7	4	1	5	2	3	6	12	10	9	11
9	3	1	2	4	7	5	8	10	9	12	6	11
10	6	3	2	1	5	4	9	12	10	11	7	8
11	5	2	3	1	4	8	6	10	12	11	9	7
12	5	4	12	3	1	2	11	8	10	7	9	6
13	5	2	3	1	4	7	6	8	12	10	9	11
14	6	1	3	2	5	4	7	9	12	8	11	10
15	9	5	3	1	2	7	10	4	6	11	12	8
16	5	4	2	1	3	6	12	11	7	10	8	9
17	6	5	2	4	1	3	7	9	11	12	8	10
18	8	3	1	2	9	5	4	12	11	10	6	7
19	8	3	2	1	11	6	9	4	7	10	12	5
20	4	2	1	3	5	10	7	11	6	12	8	9
21	6	5	4	2	1	3	7	11	12	10	8	9
22	7	2	1	5	8	9	6	10	3	12	11	4
23	7	8	5	3	1	6	4	2	10	9	11	12
24	9	5	1	2	4	3	7	10	12	11	8	6
25	10	3	2	1	4	6	7	5	9	11	12	8
26	6	3	2	1	4	5	8	7	12	11	10	9
27	1	3	12	4	2	7	8	10	11	9	6	5
28	4	3	7	2	1	12	8	6	10	11	5	9
29	4	2	5	3	1	6	9	8	10	11	12	7
30	7	3	1	6	2	4	9	12	10	5	8	11
31	5	2	1	4	3	6	7	9	8	10	11	12

(Cotd.)

Classified data matrix (Cotd.)

Year	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
1	1	1	1	1	1	1	2	2	2	2	2	2
2	1	1	2	1	1	1	1	2	2	2	2	2
3	2	1	1	1	1	1	1	2	2	2	2	2
4	2	1	1	1	1	1	2	2	2	2	2	1
5	1	1	1	1	1	2	2	2	2	2	2	1
6	2	1	1	1	1	1	1	2	2	2	2	2
7	1	1	1	1	1	1	2	2	2	2	2	2
8	2	2	1	1	1	1	1	1	2	2	2	2
9	1	1	1	1	2	1	2	2	2	2	1	2
10	1	1	1	1	1	1	2	2	2	2	2	2
11	1	1	1	1	1	2	1	2	2	2	2	2
12	1	1	2	1	1	1	2	2	2	2	2	1
13	1	1	1	1	1	2	1	2	2	2	2	2
14	1	1	1	1	1	1	2	2	2	2	2	2
15	2	1	1	1	1	2	2	1	1	2	2	2
16	1	1	1	1	1	1	2	2	2	2	2	2
17	1	1	1	1	1	1	2	2	2	2	2	2
18	2	1	1	1	2	1	1	2	2	2	1	2
19	2	1	1	1	2	1	2	1	2	2	2	1
20	1	1	1	1	1	2	2	2	1	2	2	2
21	1	1	1	1	1	1	2	2	2	2	2	2
22	2	1	1	1	2	2	1	2	1	2	2	1
23	2	2	1	1	1	1	1	1	2	2	2	2
24	2	1	1	1	1	1	2	2	2	2	2	1
25	2	1	1	1	1	1	2	1	2	2	2	2
26	1	1	1	1	1	1	2	2	2	2	2	2
27	1	1	2	1	1	2	2	2	2	2	1	1
28	1	1	2	1	1	2	2	1	2	2	1	2
29	1	1	1	1	1	1	2	2	2	2	2	2
30	2	1	1	1	1	1	2	2	2	1	2	2
31	1	1	1	1	1	1	2	2	2	2	2	2

Total number of assignment of the months to the yearly seasonal groups

Group	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
1	19	29	27	31	27	23	9	6	3	1	4	7
2	12	2	4		4	8	22	25	28	30	27	24

data matrix is shown in Table A.2.2. In case of this slided data matrix , each row is divided into two each six month long links corresponding to the two different seasonal segments namely [June - Nov.] and [Dec.- May]. Each row corresponding to the period [June- May] is considered as **one year.** The sliding process requires the coincidence of the beginning of the data matrix with the beginning of any season. In this thesis, dry season is arbitrarily considered. In the real world situation, in case of meager data, the sliding process may necessitates deleting of a number of months (< 12) of data.

A.2.2 SEASONAL SEGMENT

For any river, for any season, the set of monthly flows belonging to a common season is called a **seasonal segment.** In case of Dunlap Ck. slided data matrix, set of monthly flows corresponding to [June - Nov.] and [Dec.- May] period are considered as **Dry and Wet seasonal segments.**

A.2.3 SIMULTANEOUSLY OBSERVED SEASONAL SEGMENT

For a particular year, for a particular season, concurrently observed seasonal segments in the river with missing data and a base river is referred as **simultaneously observed seasonal segment.** In case of both the rivers having same number of seasons per year and same association of months to season, the simultaneously observed seasonal segment in any year is represented by a **2m**- dimensional vector (m= length of season ZAin month). In case of { Craig Ck., Dunlap Ck.} pair, considering Craig Ck. and Dunlap Ck. respectively as river with missing data and base river, the simultaneously observed dry seasonal segment on 11th year is illustrated in Table A.2.3.

A.2.4 TRANSITIONAL SEASONAL SEGMENT

For the river with missing data , the transitional seasonal segment is the integrated representation of the source and destiny seasonal segment corresponding to a seasonal transition. Mathematically, it is represented by a **2m**-dimensional vector comprising of the monthly flows of the source and the destiny seasonal segment pertinent to the seasonal transition. Considering Craig Ck. as the river with missing data , the transitional seasonal segment corresponding to the **dry - wet** seasonal transition of 11th year is featured in Table A.2.4.

A.2.5. SELECTED SEASONAL SEGMENTS

The concept of selected seasonal segment is developed in chapter 3, it is explained in case of {Craig Ck., Dunlap Ck.} pair considering Craig Ck. and Dunlap Ck. to be the river with missing data and a base river. For each of these rivers , the correlogram analysis shows two seasons per year and segmentation assigns the periods [June- Nov.] and [Dec.- May]

Table A.2.3: Demonstration of simultaneously observed seasonal segment of Craig Ck. and Dunlap Ck. (June1958-May 1989)

Craig Ck. data matrix

Year	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
1	137.0	155.0	146.0	56.3	52.5	59.5	147.0	244.0	186.0	428.0	998.0	192.0
2	112.0	44.6	63.0	81.7	476.0	442.0	611.0	402.0	1028.0	1073.0	915.0	554.0
3	186.0	54.5	44.9	54.9	44.7	57.3	54.3	101.0	663.0	691.0	738.0	488.0
4	238.0	77.0	216.0	77.6	201.0	342.0	903.0	593.0	691.0	1109.0	663.0	212.0
5	246.0	93.5	163.0	56.6	64.6	520.0	375.0	745.0	232.0	1412.0	180.0	138.0
6	87.2	50.3	41.3	39.6	40.9	66.2	85.4	641.0	593.0	912.0	680.0	160.0
7	87.3	43.8	35.6	46.4	91.2	151.0	256.0	680.0	809.0	907.0	464.0	307.0
8	97.8	63.5	42.6	39.7	85.6	54.9	48.9	56.9	904.0	408.0	197.0	602.0
9	87.0	33.5	58.0	159.0	397.0	224.0	367.0	747.0	482.0	1094.0	257.0	419.0
10	176.0	101.0	147.0	78.8	161.0	104.0	542.0	711.0	530.0	656.0	351.0	336.0
11	256.0	69.3	77.4	34.1	288.0	308.0	210.0	370.0	580.0	584.0	310.0	145.0
12	221.0	132.0	240.0	114.0	78.7	142.0	551.0	637.0	663.0	291.0	490.0	256.0
13	77.3	52.4	150.0	48.1	107.0	457.0	231.0	354.0	1208.0	520.0	534.0	1091.0
14	542.0	126.0	114.0	131.0	432.0	288.0	394.0	537.0	1051.0	593.0	746.0	758.0
15	1134.0	307.0	175.0	96.9	259.0	872.0	1061.0	440.0	933.0	1205.0	1331.0	892.0
16	415.0	234.0	157.0	78.0	128.0	230.0	1105.0	1080.0	496.0	655.0	500.0	441.0
17	263.0	138.0	146.0	108.0	95.3	91.7	467.0	737.0	938.0	1457.0	571.0	940.0
18	228.0	111.0	86.3	294.0	340.0	175.0	156.0	724.0	406.0	329.0	394.0	333.0
19	1023.0	141.0	80.4	55.6	678.0	289.0	470.0	164.0	282.0	611.0	906.0	131.0
20	85.5	62.9	43.0	55.0	109.0	1009.0	529.0	994.0	310.0	1575.0	1085.0	1041.0
21	219.0	96.0	139.0	69.8	49.4	70.6	193.0	1030.0	1096.0	1193.0	661.0	601.0
22	566.0	162.0	138.0	875.0	659.0	968.0	453.0	805.0	301.0	1148.0	1316.0	357.0
23	156.0	106.0	81.2	68.9	64.6	84.9	84.4	63.1	202.0	202.0	277.0	543.0
24	511.0	102.0	44.1	44.5	83.9	58.9	173.0	607.0	1010.0	868.0	377.0	285.0
25	793.0	109.0	85.3	46.1	85.2	192.0	585.0	212.0	809.0	1238.0	1503.0	397.0
26	205.0	77.4	45.0	39.3	138.0	191.0	312.0	345.0	1084.0	1027.0	1085.0	731.0
27	92.8	82.1	306.0	126.0	85.9	164.0	246.0	445.0	805.0	355.0	295.0	347.0
28	109.0	61.7	407.0	65.7	64.7	2112.0	458.0	165.0	544.0	633.0	157.0	403.0
29	114.0	65.3	68.4	150.0	55.5	222.0	712.0	503.0	655.0	1211.0	2427.0	402.0
30	115.0	74.7	42.6	197.0	54.6	170.0	248.0	373.0	292.0	141.0	372.0	309.0
31	86.1	69.5	57.1	80.5	66.5	163.0	136.0	326.0	285.0	572.0	324.0	1183.0

Dry Season Wet Season

Dunlap Ck. data matrix

Year	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
1	46.7	93.4	145.0	27.4	24.1	22.7	41.9	115.0	96.0	237.0	420.0	103.0
2	79.9	24.3	29.2	21.9	59.1	98.7	176.0	130.0	299.0	476.0	393.0	271.0
3	65.9	26.5	22.6	37.8	21.2	25.8	28.3	60.6	396.0	292.0	317.0	205.0
4	167.0	33.2	23.3	20.0	157.0	114.0	376.0	259.0	333.0	557.0	182.0	64.7
5	47.1	30.1	27.9	16.8	31.5	98.2	129.0	314.0	120.0	845.0	72.5	48.4
6	47.0	23.8	16.2	15.8	18.3	26.4	31.4	247.0	208.0	440.0	193.0	62.8
7	26.9	16.6	15.0	19.7	33.5	48.4	123.0	344.0	294.0	466.0	288.0	119.0
8	45.6	33.9	23.5	16.9	30.4	20.2	22.7	33.7	321.0	175.0	121.0	310.0
9	49.0	14.3	20.4	71.5	160.0	113.0	168.0	330.0	220.0	689.0	142.0	339.0
10	89.0	30.3	22.9	20.3	53.0	38.0	202.0	294.0	207.0	235.0	159.0	176.0
11	71.0	22.6	23.3	13.3	60.5	107.0	75.4	146.0	195.0	189.0	141.0	81.3
12	72.5	31.9	380.0	30.8	23.8	29.2	345.0	228.0	304.0	129.0	263.0	92.9
13	39.6	17.3	20.4	11.0	28.1	120.0	113.0	208.0	580.0	267.0	247.0	453.0
14	124.0	35.3	46.4	42.5	111.0	79.4	201.0	277.0	658.0	242.0	394.0	287.0
15	584.0	358.0	79.4	24.4	75.0	487.0	603.0	189.0	431.0	622.0	625.0	506.0
16	157.0	84.2	47.0	20.4	68.2	159.0	694.0	515.0	200.0	405.0	220.0	276.0
17	174.0	69.0	41.3	49.5	38.5	41.4	220.0	361.0	506.0	608.0	268.0	453.0
18	126.0	65.5	38.8	39.3	153.0	83.9	74.7	296.0	213.0	181.0	105.0	117.0
19	171.0	32.8	20.7	17.4	327.0	102.0	206.0	47.5	162.0	321.0	524.0	49.5
20	31.9	22.7	22.5	27.6	94.5	353.0	229.0	462.0	109.0	827.0	281.0	336.0
21	50.4	35.4	31.8	18.2	17.2	23.1	108.0	468.0	542.0	411.0	200.0	289.0
22	237.0	61.4	46.8	164.0	243.0	347.0	173.0	366.0	140.0	657.0	547.0	155.0
23	49.7	53.0	31.3	27.6	20.5	34.7	29.5	24.2	119.0	94.4	176.0	355.0
24	246.0	36.3	16.4	17.2	30.5	21.0	113.0	249.0	461.0	422.0	225.0	105.0
25	330.0	32.9	26.5	17.7	33.3	89.7	206.0	61.0	291.0	401.0	603.0	211.0
26	74.7	31.0	17.9	11.2	45.8	70.3	273.0	125.0	619.0	515.0	436.0	340.0
27	40.0	60.7	514.0	72.0	57.4	141.0	180.0	218.0	446.0	199.0	136.0	105.0
28	40.5	25.8	98.3	21.2	19.4	659.0	153.0	58.5	276.0	305.0	54.7	265.0
29	39.1	29.1	39.4	34.4	23.2	73.8	329.0	257.0	365.0	581.0	1071.0	202.0
30	63.0	38.0	17.3	61.3	23.9	58.6	108.0	171.0	121.0	59.1	91.5	170.0
31	28.8	18.7	16.0	25.6	22.0	74.6	96.4	206.0	172.0	247.0	308.0	536.0

Dry Season Wet Season

Simultaneously observed dry seasonal segment on 11 th year $= [256.0 \quad 69.3 \quad 77.4 \quad 34.1 \quad 288.0 \quad 308.0 \quad 71.0 \quad 22.6 \quad 23.3 \quad 13.3 \quad 60.5 \quad 107.0]^T$

Table A.2.4: Demonstration of transitional seasonal segment of Craig Ck. (June 1958- May 1989)

Craig Ck. data matrix

Year	June	July	Aug.	Sep.	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	April	May
1	137.0	155.0	146.0	56.3	52.5	59.5	147.0	244.0	186.0	428.0	998.0	192.0
2	112.0	44.6	63.0	81.7	476.0	442.0	611.0	402.0	1028.0	1073.0	915.0	554.0
3	186.0	54.5	44.9	54.9	44.7	57.3	54.8	101.0	663.0	691.0	738.0	488.0
4	238.0	77.0	216.0	77.6	201.0	342.0	903.0	593.0	691.0	1109.0	663.0	212.0
5	246.0	93.5	163.0	56.5	64.5	520.0	375.0	745.0	232.0	1412.0	180.0	138.0
6	87.2	50.3	41.3	39.5	40.9	66.7	85.4	641.0	593.0	912.0	680.0	160.0
7	87.3	43.8	35.6	46.4	91.2	151.0	256.0	680.0	809.0	907.0	464.0	307.0
8	97.3	63.5	42.6	39.7	85.5	54.9	48.9	56.9	904.0	408.0	197.0	602.0
9	87.0	33.5	58.0	159.0	397.0	224.0	367.0	747.0	482.0	1094.0	257.0	419.0
10	176.0	101.0	147.0	78.3	161.0	104.0	542.0	711.0	530.0	656.0	351.0	336.0
11	256.0	69.3	77.4	34.1	288.0	308.0	210.0	370.0	580.0	584.0	310.0	145.0
12	221.0	132.0	240.0	114.0	78.7	142.0	551.0	637.0	663.0	291.0	490.0	256.0
13	77.3	52.4	150.0	48.1	107.0	457.0	231.0	354.0	1208.0	520.0	534.0	1091.0
14	542.0	126.0	114.0	131.0	432.0	288.0	394.0	537.0	1051.0	593.0	746.0	758.0
15	1134.0	307.0	175.0	96.9	269.0	872.0	1061.0	440.0	933.0	1205.0	1331.0	892.0
16	415.0	234.0	157.0	78.0	128.0	230.0	1105.0	1080.0	496.0	655.0	500.0	441.0
17	263.0	138.0	146.0	108.0	95.3	91.7	467.0	737.0	938.0	1457.0	571.0	940.0
18	228.0	111.0	86.3	294.0	340.0	175.0	156.0	724.0	406.0	329.0	394.0	333.0
19	1023.0	141.0	80.4	55.6	678.0	289.0	470.0	164.0	282.0	611.0	906.0	131.0
20	85.5	62.9	43.0	55.0	109.0	1009.0	529.0	994.0	310.0	1575.0	1085.0	1041.0
21	219.0	96.0	139.0	69.3	49.4	70.5	193.0	1030.0	1096.0	1193.0	661.0	601.0
22	566.0	162.0	138.0	875.0	659.0	968.0	453.0	805.0	301.0	1148.0	1316.0	357.0
23	156.0	106.0	81.2	68.9	64.6	84.9	84.4	63.1	202.0	202.0	277.0	543.0
24	511.0	102.0	44.1	44.5	83.9	58.9	173.0	607.0	1010.0	868.0	377.0	285.0
25	793.0	109.0	85.3	46.1	85.2	192.0	585.0	212.0	809.0	1238.0	1503.0	397.0
26	205.0	77.4	45.0	39.3	138.0	191.0	812.0	345.0	1084.0	1027.0	1085.0	731.0
27	92.3	82.1	306.0	126.0	85.9	164.0	246.0	445.0	805.0	355.0	295.0	347.0
28	109.0	61.7	407.0	65.7	64.7	2112.0	458.0	165.0	544.0	633.0	157.0	403.0
29	114.0	65.3	68.4	150.0	55.5	222.0	712.0	503.0	655.0	1211.0	2427.0	402.0
30	115.0	74.7	42.6	197.0	54.6	170.0	248.0	373.0	292.0	141.0	372.0	309.0
31	86.1	69.5	57.1	80.5	66.5	163.0	136.0	326.0	285.0	572.0	324.0	1183.0

Dry Season Wet Season

Transitional seasonal Segment of dry- wet transition on 11th year $= [\ 256.0 \quad 69.3 \quad 77.4 \quad 34.1 \quad 288.0 \quad 308.0 \quad 210.0 \quad 370.0 \quad 580.0 \quad 584.0 \quad 310.0 \quad 145.0\ ]^T$

to dry and wet seasons. For each of these rivers, the data matrix is assumed to coincide with the beginning month of dry season. For each of these rivers, for each season, the seasonal segments represented by pattern vectors are sub-clustered into two sub-clusters by **k - means algorithm** as explained in A.1.1. For each river, the two sub-clusters corresponding to dry seasonal clusters are denoted by **1** and **2**. The two wet seasonal sub-clusters are denoted by **3** and **4**. For each river, the seasonally segmented time wave-form is replaced by the sequence of corresponding class-membership indices. These sequences of class-membership indices in case of both rivers are juxtaposed in Table A.2.5. For Craig Ck., each of two seasonal period long gaps on 29th year is denoted by zero. The probability of occurrence of any class-membership index $\mathbf{R_{ai}}$ in Craig Ck. is computed by conditioning on the simultaneously observed class-membership index $\mathbf{R_{bj}}$ in Dunlap Ck. for (i=1..4, j=1..4). This conditional probability $\mathbf{P(R_{ai} \mid R_{bj})}$ is computed and summarized in Table A.2.5. During the dry seasonal gap on 29th year, the observed dry seasonal segment (SYN) in Dunlap Creek is of type 2 . The conditional probability matrix shows that the most probable dry seasonal segment (MPR) of Craig Ck. to be of type 1 because, $P(R_a=1 \mid R_b=2)=1.0$ and greater than $P(R_a=2 \mid R_b=2)=0.0$. The juxtaposition of the sequences of class-membership indices show that the combinations (MPR=1, SYN=2) corresponding to dry season are observed 26 times in years (1, 2, 3, 4, 5, 6, 7, 8,

Table A.2.5 Juxtaposition of seasonal class-membership index sequences of Craig Ck. and Dunlap Ck. and the conditional probability matrix

River	Year														
-	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Craig Ck.	1 4	1 3	1 4	1 3	1 4	1 4	1 4	1 4	1 4	1 4	1 4	1 4	1 4	1 4	2 3
Dunlap Ck.	2 4	2 3	2 4	2 3	2 3	2 4	2 4	2 4	2 3	2 3	2 4	2 4	2 4	2 4	1 3

River	Year															
-	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Craig Ck.	1 3	1 3	1 4	1 4	2 3	1 3	2 3	1 4	1 4	1 3	1 3	1 4	2 4	0 0	1 4	1 4
Dunlap Ck.	2 3	2 3	2 4	2 4	1 3	2 4	1 3	2 4	2 4	2 3	2 3	2 4	1 4	2 3	2 4	2 4

Matrix of probability of occurrence of Craig Ck. sub-cluster index conditioned on the simultaneous occurrence of Dunlap Ck. sub-cluster index $[P(R_a|R_b)]$

Dunlap Ck. Sub-cluster index

Craig Ck. Sub-cluster index	1	2	3	4
1	0.00	1.00	0.00	0.00
2	1.00	0.00	0.00	0.00
3	0.00	0.00	0.82	0.05
4	0.00	0.00	0.18	0.95

9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 21, 23, 24, 25, 26, 27, 30, 31) . The set of simultaneously observed dry seasonal segments in these years is defined as **selected seasonal segments corresponding to the dry seasonal gap** of Craig Ck. on 29^{th} year . The MPR corresponding to the wet seasonal gap of Craig Ck. on 29^{th} year is of type 3 because, $P(R_a=3|R_b=3)=0.82$ and greater than $P(R_a=4|R_b=3)=0.18$. The combination (MPR=3,SYN=3) are observed nine times in years (2, 4, 15, 16, 17, 20, 22, 25, 26). The set of simultaneously observed wet seasonal segments in these years is defined as selected seasonal segments corresponding to the wet seasonal gap of Craig Ck. on 29^{th} year .

A.2.6 SAMPLING SCENARIOS FOR MOVE.4 AND REGRESSION MODELS

This is explained in case of { Craig Ck., Dunlap Ck.} pair considering one year long gap on 29^{th} year of Craig Ck.

- In case of AMOVE and REG models, one infilling equation is considered. The sample consists of all the concurrently observed monthly data of Dunlap Ck. and Craig Ck. by by considering the former as predictor variable and the later as predicted variable. Thus, the sample consists of [12*30 = 360] concurrent observations of the predictor and predicted variables. Each of the missing data of Craig Ck. on 29^{th} year is estimated by equation A.1.3 and equation

A.1.4 by AMOVE and REG model corresponding to the during gap observed monthly flow data in Dunlap Ck.

. In case of SMOVE, SREG models, corresponding to the missing data belonging to a particular season , say for example dry season, one infilling equation is calibrated from the sample consisting of all the concurrently observed data belonging to the dry season of Dunlap Ck. and Craig Ck. considering the former as predictor river and the later as the predicted river. Thus, for this case, the sample consists of [6*30=180] values of simultaneously observed predictor and predicted variables. The values of the missing data belonging to the dry season of Craig Ck. on 29th year are estimated corresponding to the existing data of Dunlap Ck. by SMOVE, SREG models using equation A.1.3 and equation A.1.4, calibrated on the dry seasonal sample . Similarly, the values of missing data belonging to the wet season of Craig Ck. on 29th year are estimated corresponding to the existing data of Dunlap Ck. by SMOVE, SREG models using equation A.1.3 and equation A.1.4 calibrated on the wet seasonal sample.

. In case of SSMOVE and SSREG models, corresponding to the missing data belonging to the dry season of Craig Ck. , the infilling equation is calibrated from the selected seasonal sample corresponding to the dry seasonal gap [Appendix

A2.5]. Thus,the sample consists of [6* 26 = 156] concurrent observations of predictor and predicted variable corresponding to the elements of 26 selected seasonal segments of the dry seasonal gap. The estimates of the missing data are computed corresponding to the existing monthly values of Dunlap Ck. by SSMOVE and SSREG models using equation A.1.3 and equation A.1.4 calibrated on this sample. Similarly, the values of missing data belonging to the wet season of Craig on 29th year are estimated corresponding to the existing data of Dunlap Ck. by SSMOVE and SSREG model using equation A.1.3 and equation A.1.4 calibrated on the [9*6=54] elements of nine selected seasonal segments corresponding to the wet seasonal gap.

A.2.7 SAMPLE FOR MULBS MODEL

This is explained in case of {Craig Ck.,Dunlap Ck.} pair considering two seasons long gap on 29^{th} year in Craig Ck. For dry seasonal gap segment, the conditional distributional parameters are estimated from equation 3.7 corresponding to the during gap observed dry seasonal segment in Dunlap Ck.. The sample consists of all the simultaneously observed dry seasonal segments of {Craig Ck., Dunlap Ck.} pair . Here, dry seasonal segments of Dunlap Ck. and Craig Ck. are considered as predictor and predicted vectors. For this particular case, the sample consists of 30 multivariate observations [Table

A.2.3]. Similarly, corresponding to the wet seasonal gap segment, the sample consists of 30 simultaneously observed wet seasonal segment of { Dunlap Ck. and Craig Ck.} pair. Configuration of missing wet seasonal segment of Craig Ck. is estimated from this sample by conditioning on the observed wet seasonal segment in Dunlap Ck . [Eq.3.7].

A.2.8 SAMPLE FOR SESTRNALL MODEL

This is explained in case of Craig Ck. assuming two seasons long gap on 29th year. For infilling of the dry seasonal gap segment, the sample is considered to be consisting of all the wet and dry seasonal segments corresponding to complete [wet-dry] seasonal transition. All the successively observed wet and dry seasonal segments are considered as predictor and predicted vectors with respect to the infilling of the dry seasonal gap. The configuration of the dry seasonal segment is estimated from equation 3.7 with the parameters calibrated from this sample conditioned on the observed wet seasonal segment on 28th year. There are 29 such transitions [Table A.2.6]. Similarly, for infilling of the wet seasonal gap segment, the sample is considered to be consisting of all the wet and dry seasonal segments corresponding to complete [dry-wet] seasonal transition. All the successively observed dry and dry and wet seasonal segments are considered as predictor and predicted vectors with respect to the infilling of the wet seasonal gap. There are 30 such transitions [Table A.2.6].

Table A.2.6 Demontration of extraction of samples for SESTRNALL model
for Craig Ck. data infilling

Year	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
Season	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W	D W

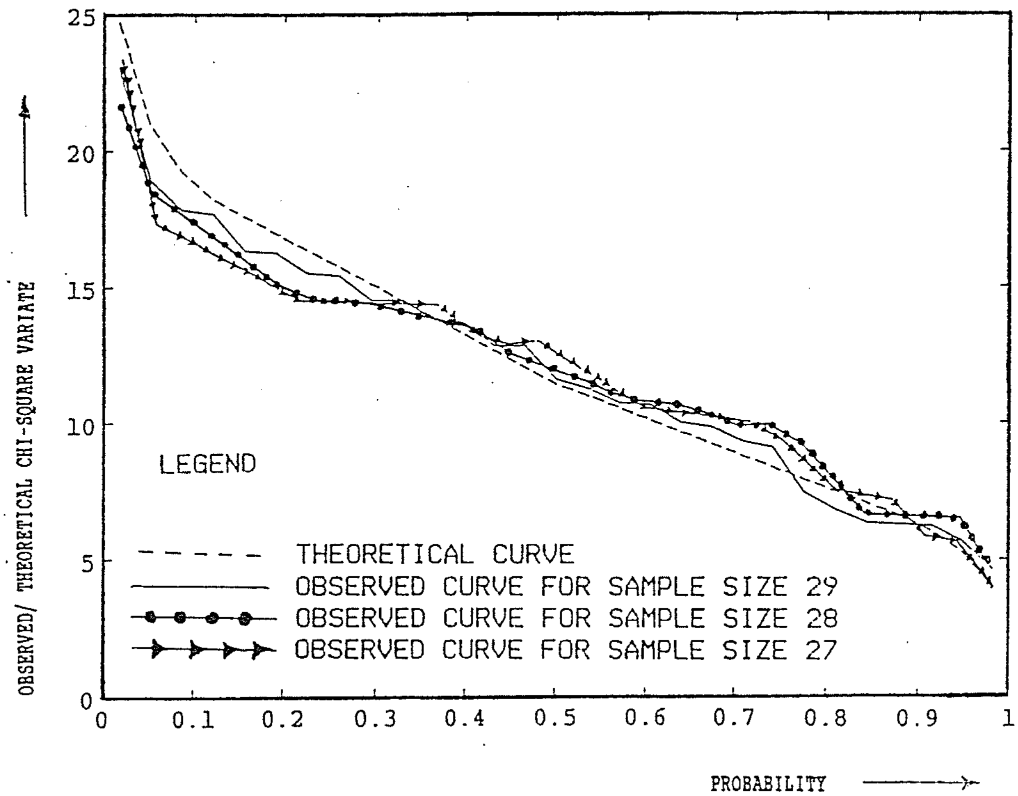
gap

Years corresponding to complete seasonal transition

[Wet-Dry] s. transition		[Dry- Wet] s. transition	
Wet S. →	Dry S.	Dry S. →	Wet S.
1	2	1	1
2	3	2	2
3	4	3	3
4	5	4	4
5	6	5	5
6	7	6	6
7	8	7	7
8	9	8	8
9	10	9	9
10	11	10	10
11	12	11	11
12	13	12	12
13	14	13	13
14	15	14	14
15	16	15	15
16	17	16	16
17	18	17	17
18	19	18	18
19	20	19	19
20	21	20	20
21	22	21	21
22	23	22	22
23	24	23	23
24	25	24	24
25	26	25	25
26	27	26	26
27	28	27	27
30	31	28	28
31	1	30	30
		31	31
	Σ 29		Σ 30

LEGEND
D = Dry seasonal segment
W = Wet seasonal segment

The configuration of the wet seasonal segment is estimated from equation 3.7 with the parameters calibrated from this sample and by conditioning on the estimate of the reconstructed dry seasonal segment on 29th year.



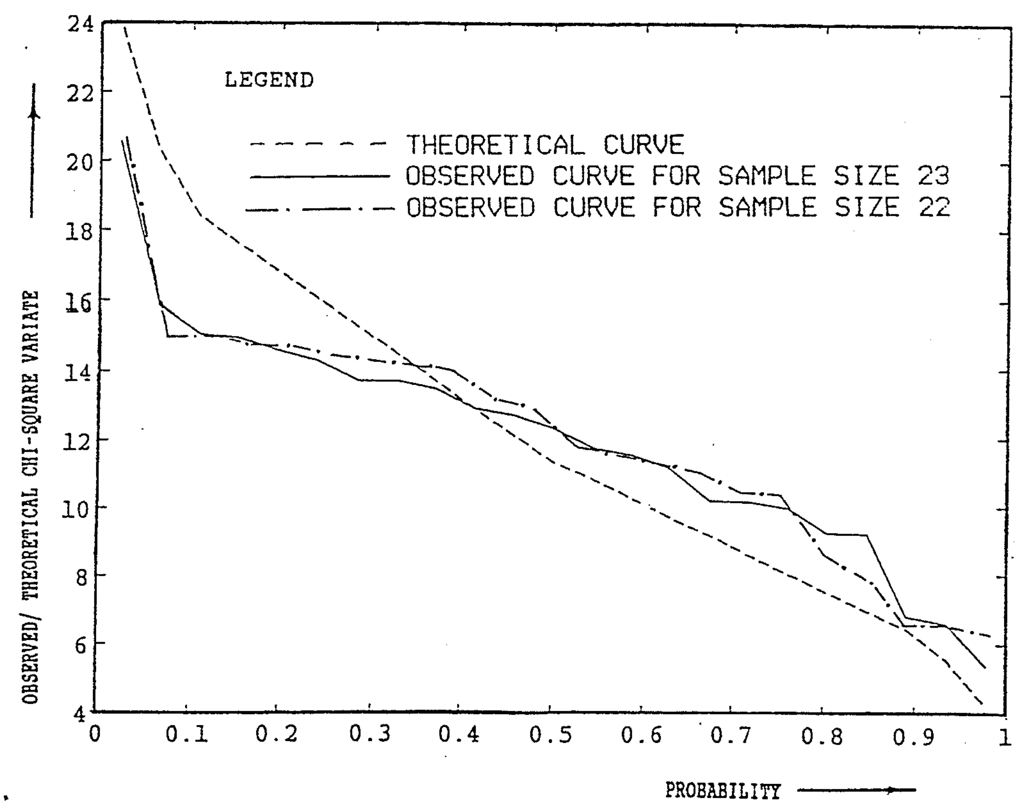


Fig. A.2:
Chi-square plot of Mahalanobis distance of the wet seasonal segment of { Craig Ck., Johns k.} pair for variable sample size

REFERENCES

1. Anderson,T.W., " *Maximum likelihood estimate for a multivariate normal distribution* " , J. of Am. Stat. Assoc. 52, 200-203, 1957.

2. Bras, R.L. and Iturbe, I.R., " *Random functions and Hydrology*" , Addison - Wesley publishing Company, 1985.

3. Chatfield, C., " *The analysis of time series theory and practice* " , Chapman and Hall London , 1975.

4. Denny, J.L. , Kisiel , C.C. and Yakowitz, S. J., " *Procedure for determining the order of dependence in streamflow record* ", Water Resour. Res. 10(5), 1974.

5. Edgett, G.L., " *Multiple regression with missing observations among the independent variables* ", J. of Am. Stat. Assoc., 51, 122-132, 1956.

6. Fiering, M. B., " *Use of correlation to improve estimate of the mean and variance* " , U.S. Geol. Surv. Prof. Pap., 434-C, C1-C9, 1963.

7. Fiering, M.B. and Jackson, B.B., " *Synthetic Streamflows* ", Water Resources monograph, American geophysical union, Washington D.C. 1971.

8. Freund, J.E. and Walpole , R.E., " *Mathematical Statistics* " , Prentice - Hall, N.J. 1980.

9. Gilory, E.J., " *Reliability of a variance estimate obtained from a sample augmented by multivariate regression* ", Water Resour. Res., 6(6), 1595-1600, 1970.

10. Haan, C.T., " *Statistical methods in hydrology* " , the Iowa state university Press/Ames, 1977.

11. Harmancioglu, N. and Yevjevich, V., " *Transfer of hydrologic information among river points* ", J. Hydrol., 91, 103-118, 1987.

12. Hipel, K.W. , Mcleod, A.I. and Lennox, W.C., " *Advances in Box- Jenkin Modelling [1]* ", Water Resour. Res. 13(3), 567-575, 1977.

13. Hipel, R.W. , Mcleod , A.I. and Lennox, W.C., " *Advances in Box-Jenkins Modelling [2]* ", Water Resour. Res. 13(3), 577-586, 1977.

14. Hirsh, R.M., " *An evaluation of some record reconstruction techniques* ", Water Resour. Res. , 15(6), 1781-1790, 1979.

15. Hirsh, R.M., " *A comparison of four record extension techniques* ", Water Resour. Res. 18(4), 1081-1088, 1982.

16. Hogben, L., " *Elementary Linear Algebra* ", West Publishing Company , 1987.

17. Johnson, R.A. and Wichern , D.W., " *Applied multivariate statistical analysis* ", Prentice Hall N.Y., 1988.

19. Khinchin, A.I., " *Mathematical foundation of information theory* ", Dover Publications, Inc. N.Y. 1957.

20. Kottegoda, N.T. and Elgy, J., " *Infilling missing flow data* ", 3rd. international hydrologic symposium, 27-29, (60-73), July 1977.

21. Kuczera, G., " *On Maximum likelihood estimators for the multisite lag- one stream-flow model: Complete and incomplete data cases* ", Water Resour. Res. 23(4), 641-645, 1987.

22. Matalas, N.C. and Jacobs, B., " *A correlation procedure for augmenting hydrological data* ", U.S. Geol. Survey. Prof. Pap., 434- E, E1-E7, 1964.

23. Moran, M.A., " *On estimators obtained from a sample augmented by multiple regression* ", Water Resour. Res. 10(1), 81-85, 1974.

24. Neter, J., Wasserman, W. and Kutner, M.H., " *Applied linear statistical models* ", Richard D. Irwin, Inc. Illinois, 1985.

25. Nicholson, G.E. Jr., " *Estimation of parameters from incomplete multivariate samples*", Journal of the American Statistical Association, 52, 523-526, 1952.

26. Panu, U.S. and Unny, T.E., " *Stochastic synthesis of hydrologic data based on concepts of pattern recognition : I. General Methodology of the Approach* ", J. of Hydrol. , 46, 5-34, 1980.

27. Panu, U.S. and Unny, T.E., " *Stochastic synthesis of hydrologic data based on concepts of pattern recognition : II. Application of Natural Watersheds* " , J. of Hydrol., 46, 197-217, 1980.

28. Panu, U.S. and Unny, T.E. , " *Stochastic synthesis of hydrologic data based on concepts of pattern recognition : III. Performance Evaluation of the Methodology* " , J. of Hydrol., 46, 219-237, 1980.

29. Panu, U.S., " *Stochastic synthesis of monthly stream-flow data based on pattern recognition* ", Ph. D. Dissertation, University of Waterloo, Waterloo, Ont. 1978.

30. Romesburg, H.C., " *Cluster analysis for researchers* ", Lifetime learning Publications, California, 1974.

31. Tou, J.T. and Gonzalez, R.C., " *Pattern recognition principles* ", Addison - Wesley , Reading, Mass. 1974.

32. Tuckwell, H.C. " *Elementary application of probability theory* ", Chapman and Hall , N.Y., 1988.

33. Vogel, R.M. and Stedinger, J.R., " *Minimum variance streamflow record augmentation procedure* ", Water Resour. Res. 21(5), 715-723, 1985.

34. Young, G.K., Orlob, G.T. and Rosener, L.A., " *Design criteria for using stochastic hydrology* ", J. Hydraulics. Division. ASCE 96, (HY4), 911-926.

35. Jones, D.S., " *Elementary information theory* ", Clarendon press, Oxford, 1979.