



City Research Online

City St George's, University of London

Citation: Cavadias, G. S. (1975). Stochastic models in hydrology. (Unpublished Doctoral thesis, The City University)

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: <https://openaccess.city.ac.uk/id/eprint/38330/>

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

STOCHASTIC MODELS IN HYDROLOGY

BY

G. S. CAVADIAS

A THESIS

PRESENTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF CIVIL ENGINEERING

THE CITY UNIVERSITY, LONDON, ENGLAND

1975

STOCHASTIC MODELS IN HYDROLOGY

G. CAVADIAS

ACKNOWLEDGEMENT

I would like to express my thanks to Professor Peter O. Wolf for his comments and encouragement during the course of this study.

STOCHASTIC MODELS IN HYDROLOGY

G. Cavadias

PAGES

<u>INTRODUCTION AND SUMMARY AND LIST OF SYMBOLS</u>	i - x
1. <u>A CLASSIFICATION OF HYDROLOGICAL MODELS</u>	1.1 - 1.14
2. <u>OPERATIONAL MODELS WITH STOCHASTIC INPUT</u>	
2.1 <u>Univariate models</u>	2.1 - 2.20
2.2 <u>Transfer function models with one input and one output</u>	2.21 - 2.31
2.2.1 The method of Box and Jenkins	
2.2.2 The method of L. Haugh	
2.3 <u>Transfer function models with multiple input and output</u>	2.32 - 2.42
2.3.1 Multiple input and single output	
2.3.2 Multiple input and multiple output	
2.4 <u>The choice of a stochastic models</u>	2.43 - 2.54
2.4.1 Choice of estimation method on the basis of the partition of the sum of squares.	
2.4.2 General consideration on the properties of stochastic models	
2.5 <u>Two hydrological examples</u>	2.55 - 2.79
2.5.1 Hydrological system with one input and one output (Megiscane and Bell rivers)	
2.5.2 Hydrological system with multiple input and output (Groundwater levels)	
3. Appendix 2.1 Estimation of ARIMA models	2.80 - 2.81
Appendix 2.2 An outline of canonical correlation	2.82 - 2.91
3. <u>REGRESSION WITH STOCHASTIC CONSTRAINTS (MIXED REGRESSION)</u>	3.1 - 3.8
4. <u>APPLICATION OF MIXED REGRESSION TO THE ESTIMATION OF PRECIPITATION AND RUNOFF FROM A MULTI-BASIN REGION</u>	4.1 - 4.19
Appendix 4.1 Data	4.20 - 4.25
Appendix 4.2 Detailed computation results	4.26 - 4.37

PAGES

5. AGGREGATION PROPERTIES OF STOCHASTIC MODELS

5.1	Aggregation properties of univariate models	5.3 - 5.4
5.2	Aggregation properties of transfer function models	5.4 - 5.5
5.3	The application of mixed regression to the aggregation problem	5.5 - 5.6
5.4	A hydrological example	5.6 - 5.7
	Appendix 5.1 Derivation of aggregation properties	5.9 - 5.11
	Appendix 5.2 Mixed regression applied to the aggregation problem	5.12 - 5.19
	Appendix 5.3 A computer program for mixed regression	5.20 - 5.37

6. REFERENCES

5.38 - 5.45

INTRODUCTION AND SUMMARY

A hydrological model is a set of mathematical relations between the variables of a river basin, such that the analytically derived behavior of the basin is similar to the actual behavior. The criterion of similarity depends in principle on the purpose of the study but in practice it is assumed to have a simple mathematical form such as the sum of squares of the differences between estimated and observed values of some criterion variable.

There is a great variety of models in the hydrological literature, which in most cases are adapted to the nature of the problem to be solved. There have also been several attempts to classify these models and to define their differences and particularly their similarities, prominent among which is the paper by Dooge (1972).

The main purpose of this study is to examine the uses and limitations of certain types of operational models with single or multiple stochastic input.

In addition, the study aims to explore the boundary between stochastic and deterministic models and to propose techniques by which prior hydrological information can be incorporated in the estimation of the parameters of hydrological models.

Before dealing with the main topic of the study, it was considered necessary to present another classification of existing hydrological models, mainly in order to clarify some overlapping areas. More specifically, certain types of models with deterministic input appear to be similar to stochastic transfer function models but the presence of stochastic inputs makes it necessary to use different methods of estimation for the parameters of the system. This classification is the subject of the first section of the study.

The second section deals with operational models with stochastic input, and after a brief survey of the well-known methodology concerning the univariate case a detailed analysis of transfer function models is presented. In the case of transfer function models with one input and one output, two methods of estimation by Box and Jenkins and L. Haugh are described and their areas of applicability delineated. The next sub-section is devoted to models with multiple input and either single or multiple output. In this latter case the method of canonical correlation is proposed as a way to simplify the relationships between pairs of input and output variables. Other methods of canonical decomposition of the input of multivariate systems have been proposed recently (Young et al. 1975). The method suggested in this study has the desirable characteristics of conceptual and computational simplicity.

Considering the variety of models and methods, the question is raised concerning the appropriate choice of estimation method. In sub-section 2.4 a graphical method of classification of transfer function models is proposed which assists the choice among deterministic, univariate stochastic or transfer function stochastic models. The next part of this section deals with certain general questions on hydrological modelling and particularly with the properties of hydrological variables reproduced by various models and the relevance of these properties for the design of various projects. The last part (2.5) of this section gives two hydrological applications illustrating the use of the methods of transfer function modelling discussed previously.

Section 3 of the study describes the method of mixed regression which permits to combine information from two different sources in the estimation of multiple regression coefficients. This method is a very versatile tool in hydrologic research because it does not require the assumptions on the prior distributions of the parameters which are necessary for the application of Bayesian methods. An important feature of the mixed regression method is that it includes a test of compatibility of the two sources of information.

Section 4 of the study describes the application of mixed regression to the estimation of precipitation and runoff of a multi-basin region.

Given that in some areas the coverage with precipitation measurements is not representative of the overall precipitation pattern whereas the measured river flow includes the contributions from areas without precipitation measurements, the question may be asked whether runoff information may be used to improve the regional estimates of precipitation. This problem has been discussed by Solomon et al (1968) who used an iterative method to incorporate the information on regional runoff in the estimation of the precipitation. The example used in this section is the precipitation and runoff distribution in the Gaspé peninsula, a region in eastern Canada.

The study of this example starts with the application of the iterative method and then continues with the comparison of the results with the corresponding results of the mixed regression method. It appears from this comparison that the mixed regression method permits to test the compatibility of the two sources of data; it gives their relative contributions to the final result and therefore is an improvement over the iterative method.

Section 5 of the study is devoted to the aggregation properties of river flows and their modelling. There are two basic problems in hydrology for which the properties of aggregated variables are important. The first problem is the estimation of variables for which only aggregated values are

observed (e.g. daily values of precipitation). The second problem, which has attracted more attention in hydrology, is whether the aggregated values (e.g. annual) of simulated flows have the appropriate long-term properties. After a review of some basic results concerning the mathematically derived properties of aggregated observations from a given stochastic process, a procedure for testing the compatibility of the short- and long-term models is proposed. In cases where the two models are not compatible, it is proposed to use the method of mixed regression for estimating the parameters of the short-term process in order to take into account the properties of the long-term aggregated flows.

In conclusion, the purpose of this study is to classify certain basic concepts in hydrological modelling and to contribute to bridging the gap between deterministic and stochastic modelling methods. It is well-known that this gap is still substantial and it will take many efforts such as this study to reduce it to a minimum realistic size.

LIST OF SYMBOLS

1. - A CLASSIFICATION OF HYDROLOGICAL SYSTEMS

Page

1.7	I	= Inflow
1.7	O	= Outflow
1.7	S	= Storage
1.7	k, x	= Coefficients in the Muskingum equation
1.12	e_i	= Residuals
1.12	θ_i	= Model parameter

2. - OPERATIONAL MODELS WITH STOCHASTIC INPUT

2.4	<u>2.1</u>	Univariate models
2.4	Q_i	= River flow of the i^{th} month (Thomas and Fiering Model)
2.4	$\bar{Q}_j$	= Mean flow for the j^{th} month where $j=1,2,\dots,12$.
2.4	σ_j	= Standard deviation of the river flow of the i^{th} month.
2.4	p_j	= Correlation coefficient between the flows of the j^{th} and $(j+1)^{\text{st}}$ months.
2.4	ϵ_i	= Standard normal variable
2.5	Q_i	= River flow of the t^{th} month (Roesner and Yevjevich model)
2.5	$\bar{Q}_j$	= Mean flow of the t^{th} month where $t=1,2,\dots, 12$.
2.5	e_i	= Independently distributed random variable
2.5	ϕ_i	= Coefficient of autoregressive model
2.5	x_t	= Deviation from the mean of the process at time t .
2.9	R_n^*	= Range
2.9	R_n	= Rescaled range
2.11	k	= Coefficient in Hurst's equation
2.11	H	= Exponent in Hurst's equation
2.13	a_t	= Independently distributed random variable.
2.14	θ_i	= Coefficient of moving average process

Page

- 2.15 ζ }
 i } Symbols of the equation of the single
 t' } broken line process defined in
 k_i } figure 2.2
 a_i }
 $\eta_{i,n}$ }
- 2.17 $Q(t)$ = Sum of single broke line processes
- 2.21 s = Number of terms in the transfer function
- 2.21 ω_i = Coefficient of the transfer function
- 2.21 μ_t = Noise of the transfer function model
- 2.22 β_t = Residuals of the input model computed from the output observations (Box and Jenkins method)
- 2.23 $r(k)$ = Correlation coefficient of lag k
- 2.23 N_t = Noise of transfer function model
- 2.23 ϕ_i = Coefficient of the i^{th} , autoregressive term of the ARIMA model for N_t .
- 2.23 Ω_i = Coefficient of the i^{th} moving average term of the ARIMA model for N_t .
- 2.24 μ_t = Residual of the univariate model of the input x_t .
- 2.24 v_t = Residual of univariate model of output y_t .
- 2.24 ϕ_i = Coefficient of the i^{th} autoregressive term of the ARIMA model of the input x_t .
- 2.24 θ_i = Coefficient of the i^{th} moving average term of the ARIMA model of the input x_t .
- 2.24 ϕ'_i = Coefficient of the i^{th} autoregressive term of the ARIMA model of the output y_t .
- 2.25 θ'_i = Coefficient of the i^{th} moving average term of the ARIMA model of the output y_t .
- 2.28 ϕ'_i = Coefficient of the i^{th} autoregressive term of the ARIMA model of the residual v_t of the output process.
- 2.28 Ω'_i = Coefficient of the i^{th} moving average term of the ARIMA model of the residual v_t of the output process.

Page

- 2.30 U = Statistic for testing the independence between the residuals u and v .
- 2.31 U' = Statistic for testing the independence between the residuals u and a .
- 2.37 $M(x)$ = A model containing terms in x .
- 2.40 ζ_i = The i^{th} canonical variable.
- 2.43 SSy = Total sum of squares of y_t
- 2.43 SSy^* = Sum of squares attributable to autogressive terms.
- 2.43 SSx = Sum of squares attributable to the input variables.
- 2.43 SSu = Sum of squares attributable to the residuals.
- 2.80 ϕ_0, θ_0 = Initial values of the parameters of an ARIMA process.
- 2.84 R = Correlation matrix of original variables.
- 2.85 λ_i = The i^{th} characteristic root of the matrix used for estimating the canonical variables.

3.- REGRESSION WITH STOCHASTIC CONSTRAINTS (MIXED REGRESSION)

- 3.1 y = Dependent variable of the main regression.
- 3.1 e = Residuals of main regression
- 3.1 w = Dependent variable of the constraints
- 3.1 v = Residuals of the constraints.
- 3.2 X = Matrix of observations of the independent variables in the main regression
- 3.3 V = Symmetric positive definite matrix
- 3.3 σ^2 = Positive parameter multiplying V to give the covariance matrix of the disturbance term of a generalised regression.
- 3.3 V_m = Symmetrix positive definite matrix
- 3.3 σ_m^2 = Positive parameter multiplying V_m to give the covariance matrix of the disturbance term of the main regression.
- 3.4 Z = Matrix of observations of the independent variables in the constraints.
- 3.4 V_c = Symmetric positive definite matrix
- 3.4 σ_c^2 = Positive parameter multiplying V_c to give the covariance matrix of the disturbance term of the constraints.

Page

- 3.4 b_M = Vector of regression coefficients of the mixed regression equation.
- 3.4 $V(b_M)$ = Covariance matrix of b_M
- 3.4 b_m = Vector of regression coefficients of the main regression equation
- 3.4 b_c = Vector of regression coefficients of the constraints
- 3.6 Q = Theil's compatibility statistic
- 3.6 θ_m = Contribution of the main regression to the mixed regression
- 3.6 θ_c = Contribution of the constraints to the mixed regression
- 3.8 $\sigma_o^2(y)$ = Variance of errors of observation of the dependent variable of the main regression
- 3.8 $\sigma_o^2(w)$ = Variance of errors of observation of the dependent variable of the constraints

4.- APPLICATION OF MIXED REGRESSION TO THE ESTIMATION OF PRECIPITATION AND RUNOFF FROM A MULTI-BASIN REGION

- 4.5 E = Mean annual evaporation
- 4.5 P = Mean annual precipitation
- 4.5 T = Mean annual temperature
- 4.5 k = Ratio of recorded to computed runoff
- 4.14 G_i = Physiographic variable of a square
- 4.15 $\bar{G}_i$ = Mean value of the i^{th} physiographic variable of a gauged basin
- 4.16 σ_{PT}^2 = Total variance of precipitation
- 4.16 σ_{PE}^2 = Variance of observation errors of precipitation
- 4.16 σ_{QT}^2 = Total variance of flow
- 4.16 σ_{QE}^2 = Variance of observation errors of flow

Page5.- AGGREGATION PROPERTIES OF STOCHASTIC MODELS

- 5.6 w_t = Monthly standardised flows
- 5.6 $\bar{w}_t$ = Annual standardised flows
- 5.12 MN = Index for short-period flows numbered sequentially
- 5.12 M = Index of short-period flows
- 5.12 N = Index for long-period flows
- 5.12 w_{mm} = Standardised short-period flow
- 5.13 $\bar{y}_n$ = Mean flow of the n^{th} long period
- 5.13 $\bar{y}$ = Long-term mean flow
- 5.13 $\bar{s}$ = Mean of the short-term standard deviations
- 5.13 $\bar{w}_n$ = Weighted average of standardised short-period flows
- 5.13 ϕ, A = Coefficients of the ARIMA model fitted to the weighted average of the standardised short-period flows
- 5.13 ϕ, θ = Coefficients of the ARIMA model fitted to the standardised short-period flows.

1. A CLASSIFICATION OF HYDROLOGICAL MODELS

1. A CLASSIFICATION OF HYDROLOGICAL MODELS

Hydrological models are mathematical representations of the interrelated phenomena occurring in a river basin. Given the degree of complexity of these phenomena and the usual scarcity of relevant data, it is almost impossible to develop perfectly satisfactory models.

Fortunately, the task of formulating hydrological models is made more tractable because it is not necessary to find a mathematical representation for all possible characteristics of a basin, but only for those relevant for a given project. This is one of the basic reasons for the large number of models proposed by hydrologists, ranging from direct applications of the equations of hydrodynamics to deterministic relationships between storm precipitation and flood flow and finally to statistically estimated relationships between precipitation and runoff of large basins.

The models proposed in the hydrological literature such as in Cavadias (1966), Dawdy (1968), Dooge (1972), Clarke (1973), are based on various assumptions concerning the type of input and the properties of the system. General classifications of the type shown in Fig. 1.1 (Cavadias 1966) take into account the possibility of non-linearity, time variability, space-variability and feedback in the hydrological system but do not generally include any details of the various types of equations representing the hydrological system and of the methods of estimating the parameters.

INPUT		SYSTEM			
DEGREE OF COMPLEXITY		DEGREE OF NON-LINEARITY	DEGREE OF TIME-VARIABILITY	DEGREE OF SPACE VARIABILITY	DEGREE OF FEEDBACK
DETERMINISTIC		LINEAR	TIME-INVARIANT	SPACE-INVARIANT	WITHOUT FEEDBACK
		QUADRATIC			
STOCHASTIC (NO AUTO-CORRELATION)		CUBIC	TIME-VARIABLE	SPACE-VARIABLE	WITH FEEDBACK
STOCHASTIC (FIRST ORDER AUTO-CORRELATION)					
STOCHASTIC (SECOND ORDER AUTO-CORRELATION)					

FIG. 1.1

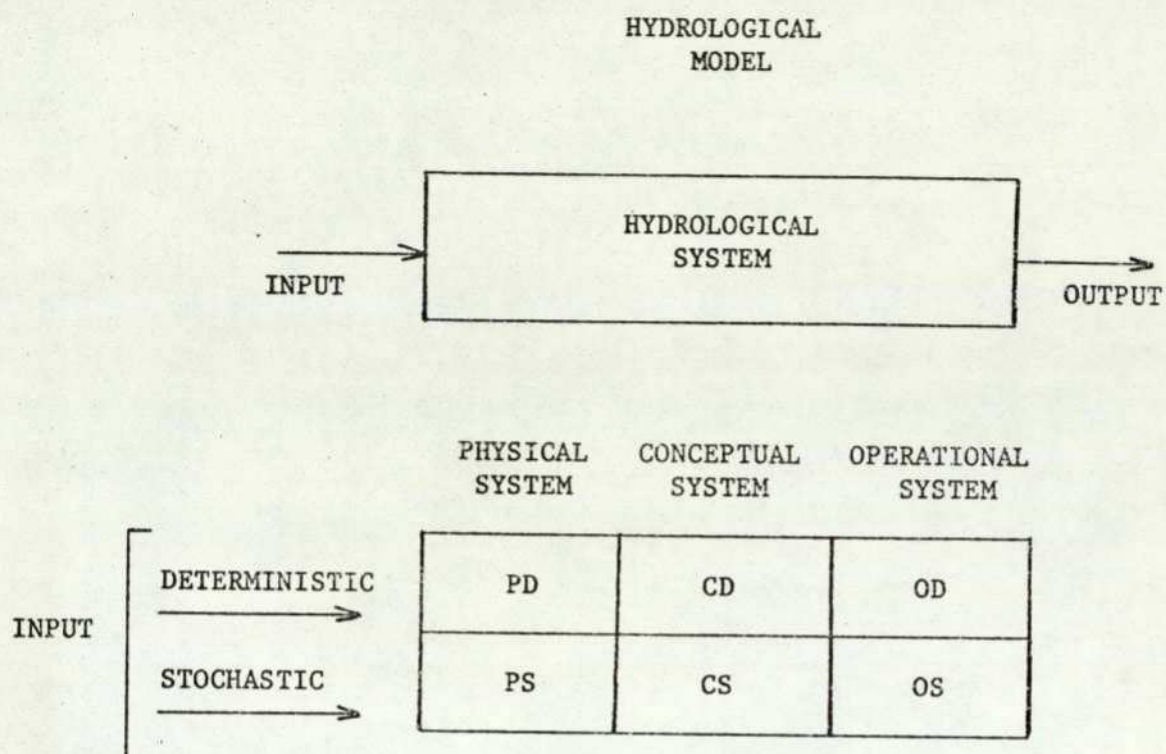
In this study we are concerned only with lumped, time-invariant systems without feedback. Therefore, in order to define the systems that form the subject of our analysis, we propose a more detailed classification of the above category of systems (Fig. 1.2).

The basis for this classification is a two-way distinction between hydrological models according to the category of system (physical, conceptual or operational), i.e. the type of equation representing the phenomena occurring in the basin, and to the category of input (deterministic or stochastic).

Within each combination of system and input categories, it is possible to use one of the three following methods for estimating the parameters of the equations:

- (1) Deterministic method.
- (2) Deterministic method taking into account noise in the input.
- (3) Stochastic method.

We shall now give a more detailed description of the individual models shown in Fig.1.2 and of the methods of estimation of the parameters.



CLASSIFICATION OF HYDROLOGICAL SYSTEMS

FIGURE 1.2

1. Physical Systems With Deterministic Input (PD)

This is the basic type of engineering model in which the physical laws governing the phenomenon are known. Examples of such models are the deflection of an elastic spring and the flow in a pipe. A more complex example of particular interest to the hydrologist is the method of hydraulic routing of a flood wave through a channel, which uses the equations of continuity and momentum and assumes that the appropriate initial and boundary conditions are known (e.g. Harbaugh 1967, Garrison et al. 1969).

Another example of a physical system with deterministic input is given in Freeze and Harlan (1969), Stevenson and Freeze (1974) where a physically based general equation is proposed, representing infiltration and soil moisture flow. This equation has been applied to the study of the runoff from small plots under various boundary conditions and soil properties (Stevenson and Freeze 1974).

The parameters of the model are estimated by some form of the steepest descent method. A detailed description of the calibration of a physical model is given in the above mentioned paper by Stevenson and Freeze and in Cormary and Guilbot (1971). This method of estimation of the parameters can be called deterministic because it is based on the following assumptions:

- (1) There exists a unique transformation from the input to the output.
- (2) The input and the output are observed without error.

Under the additional assumption that the error surface $S(\theta_1, \dots, \theta_k)$ is reasonably smooth, the steepest descent procedures converge to the minimum value of the error surface and the parameters $\theta_1, \dots, \theta_k$ are thus estimated.

The above deterministic method for estimating the optimal values of the parameters does not generally converge when there is appreciable noise in the measurements of the input. In this case it is possible to use the algorithm of Kiefer and Wolfowitz (1952) and for the multidimensional case the algorithm of Blum (1954). These methods are also described in Cormary and Guilbot (1971) and in Lavi and Vogl (1966). In certain cases, the parameters of physically based models with deterministic input are estimated by statistical methods as for example in Stevenson and Freeze (1974).

2. Conceptual Systems With Deterministic Input (CD)

In the analysis of hydrological systems we often do not have sufficient information to formulate the continuity and momentum equations and to establish the appropriate initial and boundary conditions. In such cases we may still use the equation of continuity and other equations based on physical concepts.

Consider for example the system shown in Fig. 1.3, consisting of a river reach where the input and output are the

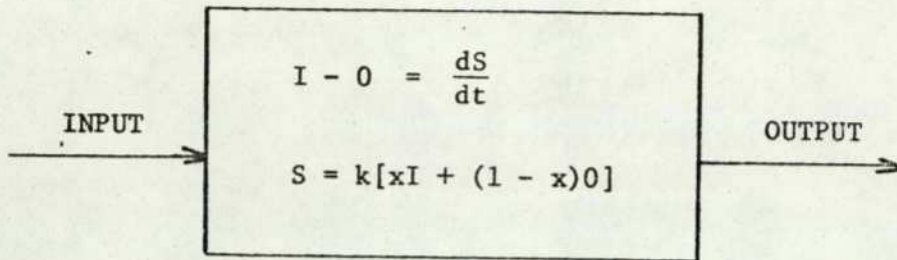


FIG. 1.3

upstream and downstream flood hydrographs respectively. The equation of continuity

$$(1.2) \quad I - O = \frac{dS}{dt}$$

where

I = Inflow

O = Outflow

$\frac{dS}{dt}$ = Rate of change of storage

forms the basis for all flood routing methods. In order to complete the mathematical description of the system, we may assume that the storage is expressed by the Muskingum equation

$$(1.3) \quad S = k[xI + (1 - x)O]$$

It is well known that this equation is based on physical concepts (Ven Te Chow 1959) such as the discharge characteristics of control sections at each end of the reach and the slope of the water

surface. However, the correspondence between the physical laws governing the phenomenon and equation (1.3) is only approximate and therefore we call the model shown in Fig. 1.3 a conceptual model.

In contrast to physical-deterministic models which are mainly research oriented, conceptual-deterministic models have been used extensively for studying large basins and for the design of water resources projects. Examples of such models are the Stanford model described in Crawford and Linsley (1964) and the models described in the papers by Sittner et al (1969), Dawdy and O'Donnell (1965), Roche and Slivitzky (1967) and others.

These models consist of a series of equations based on physical concepts and containing many (10-20) parameters. The values of some of the parameters are estimated on the basis of available information taken from maps and surveys and the remaining parameters are estimated by an optimization algorithm which minimizes a measure of the discrepancy between the estimated and the actual output. The most common measure of this type is the quantity $S = \sum e_i^2$ where e_i are the differences between observed and estimated output. The mathematical form of the response surface is unknown and therefore the minimum value of S and the corresponding values of the parameters are determined by the methods of steepest descent mentioned in the previous section.

3. Operational Systems With Deterministic Input (OD)

As in the types of models discussed in the previous sections, the output of an operational-deterministic (OD) model is a single-valued function of the input. The main difference is that the form of the transformation from input to output is not based on physical laws (e.g. the momentum and continuity equations) but it is selected on the basis of simplicity, convenience and previous experience with similar systems.

Given the input and the output of the system, the constants of the equation are determined by the operation of "deconvolution" either graphically or by the solution of a system of simultaneous equations. One of the basic examples of an OD-model is the linear transformation from precipitation to runoff described by the unit hydrograph (e.g. Nash 1959 and 1969).

As in previous cases, the parameters of OD-models can also be estimated by statistical methods such as multiple regression (e.g. Linsley, Kohler, Paulhus 1975). This approach uses all available input and output data and results in more stable estimates of the parameters of the transformation equation (e.g. the unit hydrograph ordinates).

4. Operational Systems With Stochastic Input (OS)

In the previous sub-sections we have defined physical, conceptual and operational hydrological systems in which the input is assumed to be deterministic and observable with or without error. This assumption is not valid for large basins in which the input is not even defined. For example, the precipitation on a large basin is distributed both in time and space and therefore the available point measurements of precipitation can be considered only as statistical indices of the real input. Given that we cannot establish a one-to-one relation between input and output, the only possibility is to establish a statistical relation i.e. to estimate the conditional probability distribution of the output given the observed values of the input index.

For very large basins where the lags between the multiple inputs and the output become very long, it is difficult to establish even a statistical relationship. In this case the hydrologist may model only the output by fitting a stochastic process to the observed outflow. (Thomas and Fiering 1962).

In some cases, however, it might be possible to have a reliable index of the input and therefore an attempt can be made to relate the input indices to the measured output. In such cases the only approach available to the hydrologist is to use statistical estimation methods, the simplest form of which is linear regression analysis. The regression methods used in this case do not use any knowledge of the physical aspects of the hydrologic cycle and are

limited to reproducing the statistical properties of the input and the output. Due to the fact that the input is an index, the continuity equation and even dimensional homogeneity cannot and need not be satisfied.

The result of the statistical estimation is the expected value of the output given the input and the variance of the residuals. The usual regression methods are based on certain assumptions concerning the residuals from the model, namely that they have zero expected value, equal variance and no autocorrelation.

These assumptions are usually not satisfied by hydrological time series with seasonal characteristics and therefore several attempts have been made by hydrologists to use other statistical methods (ARIMA models, Spectral analysis, Fractional Gaussian noise models, broken line processes) which reproduce some of the most important properties of hydrologic series.

From the point of view of our classification, all these models are still operational-stochastic (OS), i.e. they do not take into account the physical laws determining the transformations of the hydrologic variables within the basin. Nevertheless, due to the great difficulty in formulating even approximate conceptual-stochastic models for large basins, OS-models with carefully formulated and verified statistical assumptions will remain a very useful tool for the hydrologist for a long time.

5. Conceptual Stochastic (CS) and
6. Physical Stochastic (PS) Models

The ideal hydrological model should take into account all the important characteristics of hydrological phenomena:

- (1) The physical laws governing precipitation, evaporation, interception, infiltration, soil moisture, flow in surface channels and groundwater flow.
- (2) The nature and magnitude of the errors of measurement of all variables.
- (3) The probability distributions, intercorrelations and other statistical properties of the inputs and outputs.

It is obviously impossible to satisfy all these requirements at the present time. The only way to approach this ideal would be to start from the simpler conceptual or operational models and attempt to satisfy the following conditions: (See also Clarke op. cit.).

- (1) The residuals e_i from the overall model and whenever possible from the partial models should satisfy the usual regression assumptions.
- (2) The model should include the estimation of the variances of all parameters and their correlation matrix.
- (3) The model should provide the means for a sensitivity analysis of the surface $S = \sum(e_i^2) = f(\theta_1, \dots, \theta_k)$.

It is well known that the present estimation methods of conceptual models do not satisfy any of the above conditions. The main difficulty is that most of the intermediate inputs and outputs are not measurable and the corresponding constants (e.g. "reservoir capacities") have only an approximate physical meaning.

The road to satisfactory conceptual stochastic and particularly stochastic physical models is long. Given that the only way to improve the operational stochastic models now available is to replace the simple equations of current OS-models with conceptual or physical equations, it can be predicted that this will be the future direction of hydrologic modelling. Perhaps a start could be made in this direction by estimating some of the simpler conceptual models under the condition that the residuals of the computed output have mean zero, equal variance and no autocorrelation.

It is however interesting to note that, once the characteristics of a physical or conceptual hydrological system are determined, we may also examine its behavior under a stochastic input. If for example the input to the system is a random variable x with density function $g_1(x)$ and the output y can be expressed as a deterministic function $y = f(x)$ of x , then we may derive the density function $g_2(y)$ of y from the formula

$$(1.4) \quad g_2(y) = g_1(x) \cdot \left| \frac{dx}{dy} \right|, \quad \frac{dy}{dx} \neq 0$$

In such a system, the methods of physics (e.g. hydraulics) are used to derive the output from the input, and the theory of probability and not statistics is used to derive the probability distribution of the output from that of the input. For example in the hydraulic routing discussed previously, we can make probability statements about the discharge at the end of the channel, given the probability distribution of the inflow to the channel.

2. OPERATIONAL MODELS WITH STOCHASTIC INPUT

2. OPERATIONAL MODELS WITH STOCHASTIC INPUT

2.1 Univariate models

In section 1 a classification of hydrological models was proposed and the conclusion drawn that conceptual stochastic (CS) and particularly physical stochastic (PS) models are preferable to other types of models but very difficult to formulate. These two types of models can be partitioned into a number of sub-types representing specific transformations within the runoff phase of the hydrologic cycle. Each sub-type consists of a number of equations and constraints based as far as possible on the physical laws governing the transformation represented by the sub-type.

Both the equations and the constraints contain a number of parameters which are usually estimated by steepest descent methods that may take into account the errors in the observations of the input. The result of this "calibration" is a set of equations which transforms the observed input into an estimated output, similar to the observed output according to some criterion, which usually involves a function of the residuals which has to be minimised.

One of the main problems in fitting hydrological models is that even when the residuals are sufficiently small, i.e. the observed and estimated outputs are sufficiently close, the residuals are usually not random and therefore the modelling process cannot be considered as complete. The question may therefore be raised as to

how to calibrate the model in order to make the residuals independently distributed random variables. In the case of physical or conceptual systems consisting of a set of subsystems represented by complex non-linear equations, it is difficult to see at the present time how one could estimate ten or twenty constants under the constraint that residuals be random. The situation however is different when we come to the operational models where the forms of the equations are simple convolutions or multiple regressions on lagged values of the input. In this case, given the inherent simplicity of the model, it is possible by using appropriate statistical estimation procedures to ensure that the residuals be independently distributed random variables.

An important part of the formulation of hydrological models is the verification of the degree of similarity between the model and the basin it represents. This verification is a complex process which, in the case of physical and conceptual systems, consists in verifying that the equations of the model reflect as much as possible the physical laws governing the phenomena. This "theoretical" part of the verification must be supplemented by the "empirical" verification that the input-output transformation incorporated in the model reproduces satisfactorily the available input-output observations. In the case of operational systems, the equations are not related to any physical laws and therefore only the empirical part of the verification is possible.

For many of the practically important basins, the only knowledge we have of the actual input is in the form of statistical indices (e.g. precipitation and snow survey observations). Thus we must verify that the properties of the stochastic processes fitted to the inputs are similar to the corresponding properties of the observed inputs.

It is important to realize that for the purpose of applications it is never necessary for a hydrological model to preserve all the properties of the observed flow sequence. For example, the Hurst property is not relevant for sequences that will be used for the design of a cofferdam, and conversely, it is not necessary to reproduce the fine structure of the daily inflow if the purpose of the study is to design a very large storage reservoir.

The purpose of stochastic modelling is to apply statistical methods to hydrological problems in such a way as to include the verification of the basic assumptions in the estimating procedure. These statistical methods are related to the estimation of stochastic-operational models and in order to put them in the proper frame of reference, we shall begin by a brief look at the development of (SO)-models.

Present-day development of (SO)-models dates from the work of the Harvard Water Program (Thomas and Fiering 1962) which was made possible by the advent of electronic computers. This development was however preceded by the attempts of early pre-computer age hydrologists like Hazen (1914) and Sudler (1927) who felt that the

mass-curve, based only on the record, should be supplemented by some other combinations of flows. In order to achieve this Hazen used standardized flows from many rivers in a region and Sudler wrote historical records on cards and created synthetic sequences by drawing a set of cards at random. These primitive attempts had obvious limitations and were not generally used by practising hydrologists.

It must be noted at this point that the traditional mass curve method, the approaches of Hurst and Sudler mentioned above and the majority of stochastic models to be discussed briefly in this section concentrate on fitting a stochastic process to the output of large basins, mainly for the purpose of simulating various flow sequences and using them in the design of water resources projects.

2.1 Modelling the Output

The first attempt to formulate a mathematical model for monthly flows was made by Thomas and Fiering (1962) who proposed the model

$$(2.1) \quad Q_{i+1} = \bar{Q}_{j+1} + b_j(Q_i - \bar{Q}_j) + \epsilon_i \sigma_{j+1}(1 - \rho_j^2)^{\frac{1}{2}}$$

where Q_i and Q_{i+1} are the flows of the i^{th} and the $(i + 1)^{\text{th}}$ months respectively; $\bar{Q}_j$ and $\bar{Q}_{j+1}$ are the mean monthly flows for the j^{th} and $(j + 1)^{\text{st}}$ months; b_j is the regression coefficient; ϵ_i is a standard normal random variable; σ_{j+1} is the standard deviation of flows of the $(j + 1)^{\text{st}}$ month and ρ_j is the correlation coefficient between the flows of the j^{th} and $(j + 1)^{\text{st}}$ months. In the above equation the index i takes the values $1, 2, \dots, n$ where n is the total number of observations and the index j takes the values $1, 2, \dots, 12$.

A related type of model was proposed by Roesner and Yevjevich (1966). The equation of the model is:

$$(2.2) \quad \frac{Q_i - \bar{Q}_j}{\sigma_j} = \rho_1 \frac{Q_{i-1} - \bar{Q}_{j-1}}{\sigma_{j-1}} + \epsilon_i$$

where Q_i is the value of the process at time i , $\bar{Q}_j$ is the mean value for the month j , σ_j is the standard deviation for the month j , ρ_1 is the serial correlation coefficient of order one and ϵ_i is an independently distributed random variable.

The main difference with the Thomas and Fiering model is the assumption that the correlation coefficients between consecutive months do not have seasonal fluctuations. Quimpo (1967) attempted to extend this approach to daily flows and Beard (1965) used this model assuming that the independently distributed random component is given by a log-Pearson III distribution.

This type of model can be extended to more general autoregressive models as proposed by Yevjevich (1963). This model is given by the equation:

$$(2.3) \quad X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_n X_{t-n} + a_t$$

where X_t is the value of the deviation from the mean of the process at time t and ϵ_t is an independently distributed random variable, and where it is again assumed that there is no seasonal variation in the parameters.

The models described above were applied to several actual cases with varying success. Given the complexity of the behaviour of observed flows and the variety of purposes for which hydrological modelling is undertaken, it follows that the adequacy of the model can be assessed from several points of view. For example, attention may be focussed on the reproduction of the short-term or the long-term properties of the observed flows according to which category of properties is more relevant for the project under study.

The question as to how to test the adequacy of a hydrological model is difficult to answer. If the criterion of similarity is simple as for example minimizing the sum of the squares of the residuals from the model, it is possible to apply it to the recorded flows. If on the other hand we need to consider more complex aspects of this similarity such as autocorrelation coefficients at longer lags, cross-correlations, distribution-free properties such as runs, number of crossings etc. and spectral properties, it is not possible to evaluate the performance of the model in a simple way.

A practical way to test the similarity between a set of observed flows and a stochastic model is to use it to generate simulated flow sequences by the Monte-Carlo method and then compare the actual and artificial flow sequences from all relevant points of view. This approach was used e.g. in the simulation study of the inflow to the Great Lakes (Clark and Cavadias 1967, Megerian and Pentland 1968), where the following properties of the simulated flow sequences were compared to those of the observed flows:

- (1) Basic statistical parameters (Mean, standard deviation etc.)
- (2) Probability distribution
- (3) Autocorrelation function
- (4) Cross-correlation function
- (5) Spectral properties (Location of peaks of spectrum)
- (6) Distribution-free properties (runs, number of crossings etc.)
- (7) Long-term persistence (Hurst's H)

Thus, sequences of simulated flows or "simflows"* are one way of understanding the properties of the fitted process and comparing them to the properties of the observed flows.

The proper use of simulated flows has been the subject of long discussions among hydrologists. In particular, engineers dealing with the design of water resources projects have an understandable reluctance to consider simulated flows on the same footing with observed flows for making design decisions. In order to understand the rôle of simulated flows in the design of projects we must look briefly at the more elementary problem of fitting a probability distribution to a given set of data. In this case, the well-known procedure is

- (1) Construct a histogram.
- (2) Select the type of probability distribution to be fitted.

* This term is introduced to denote "simulated sequences of numbers generated by the application of the Monte-Carlo method to the fitted stochastic process".

- (3) Using the method of moments or of maximum likelihood, estimate the parameters of the postulated probability distribution and establish confidence intervals.
- (4) Use published tables of the fitted distribution to make probability statements concerning the random variable represented by the original data.

Looking now at the problem of fitting a stochastic process to a set of observed flows, we see that, due to their complexity, it is not possible to tabulate the properties of these processes. Consequently, the only way to understand them is to generate sequences of "simflows", study their various properties and compare them to the corresponding properties of real flows. We come therefore to the conclusion that simflows are analogous to the tabulations of theoretical probability distributions and they are in reality no more controversial. If we are satisfied that the stochastic process has been fitted correctly, we can use simflows as the input to a water resource system and study its behaviour. In following this procedure, however, we should keep in mind that the stochastic processes in question are much more complex than probability distributions and therefore we should not expect that the generated sequences will reproduce all the properties of the observations. We must usually be satisfied with a compromise between the degree of simplicity of the fitted model and the nature of the properties which are reproduced in a satisfactory way.

The parameters of the models discussed in this section are estimated on the basis of short-term properties of the observed flows (means, variances, autocorrelations, etc.) and therefore these models attempt mainly to preserve short-term properties. When however these models are applied to the design of projects, it very often happens that the dimensions of the structures depend on long-term properties of the record such as persistent sequences of droughts or rare occurrences of large floods.

The preservation of these properties is not incorporated in the structure of the models already described and when simulated and observed flows are compared empirically, their long-term properties do not agree.

The first study of long-term properties of river flows goes back to the classical paper by Hurst (1951) in which he used the long record of Nile river to investigate the reservoir capacity required for complete regulation of the river. In his paper, Hurst studied a statistic which can be defined as follows: (Fig. 2.1): Given the sequence of flows $X_1, \dots, X_k, \dots, X_n$ we compute the quantity

$$(2.4) \quad S_k^* = \frac{k}{n} \sum_{i=1}^k x_i - \frac{k}{n} \sum_{i=1}^n x_i$$

and then form the statistic

$$(2.5) \quad R_n = \frac{R_n^*}{\hat{\sigma}_n}$$

where

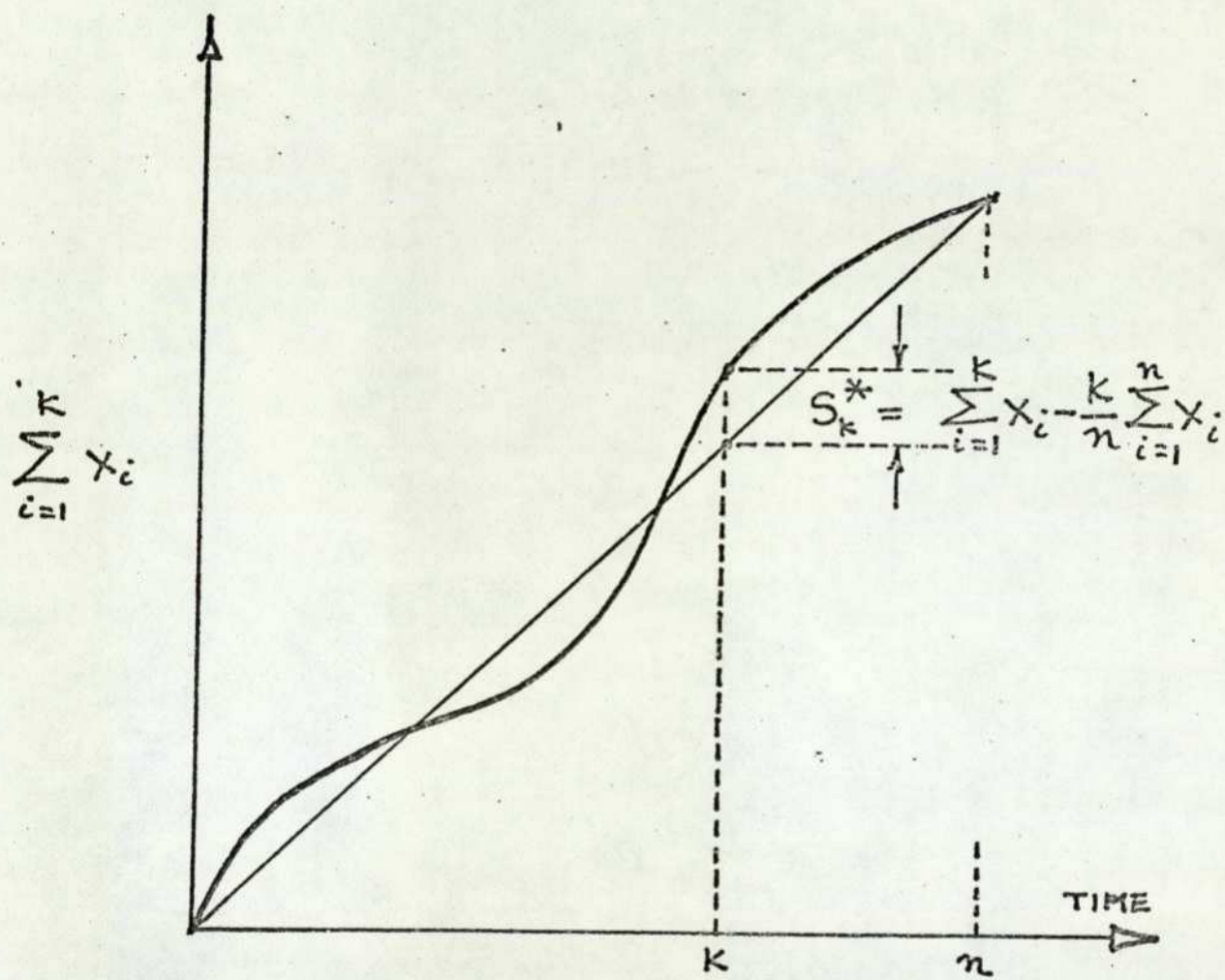


FIGURE 2.1

$$(2.6) \quad R_n^* = M_n^* - m_n^*$$

and

$$(2.7) \quad M_n^* = \max [S_1^*, S_2^*, \dots, S_k^*, \dots, S_n^*]$$

$$(2.8) \quad m_n^* = \min [S_1^*, S_2^*, \dots, S_k^*, \dots, S_n^*]$$

In intuitive terms, the statistic R_n^* is the capacity that a reservoir must have in order to maintain a regulated flow equal to the mean flow of the river. Finally, the statistic R_n is defined as the ratio of R_n^* to the standard deviation of the flows and is called "rescaled range". Hurst investigated the behaviour of the rescaled range not only for the record of the Nile river flow but also for other long geophysical series such as precipitation, tree rings and varves and established the following empirical equation:

$$(2.9) \quad R(n) = k \cdot n^H$$

where the mean value of H was found to be 0.73 with a standard deviation of 0.08. In contrast to this result, Hurst derived the following for the expected value of $R(n)$ for a sequence of independent standard normal variables:

$$(2.10) \quad E[R(n)] = 1.25n^{0.5}$$

The discrepancy in the exponents of the two equations was attributed by Feller (1951) to the autocorrelation of the geophysical series. It was shown however by Barnard (1956) and Moran (1959) that

the asymptotic value of H for autoregressive (Markovian) models is also 0.5. In addition, it was found empirically by Yevjevich (1965) that the range R_n^* and the rescaled range R_n of Markov processes have large sampling variations. Using also empirical approaches Fiering (1967) and Mandelbrot and Van Ness (1968) showed that the Hurst exponent of 0.70 can be simulated as a transient phenomenon of Markovian processes with a large number of autoregressive lags.*

A direct approach to the problem was made by Mandelbrot and Van Ness (1968) who proposed a special process called "fractional Gaussian noise" (FGN). The main characteristic of this process is that the amplitude of its periodic fluctuations increases with the length of the period. This means that if we decompose a flow time series in a sum of sine and cosine curves, the curves with lowest frequencies (longest periods) will have larger amplitudes.

The Hurst exponent H is a parameter of the FGN-process and therefore a fitted FGN process reflects the value of H determined from the observations.

It must be noted however that by concentrating on the reproduction of Hurst's H , FGN-processes do not reflect, in their structure, and therefore do not reproduce the short-term properties of the observed flows. It is also worth noting at this point that there are many ways in which geophysical series can show long periodic or cyclical fluctuations and there is no reason to believe that these very elusive phenomena on which very little information

*It must be noted, however, that Hurst (1951) obtained very high values of H for very long series of natural data and this leaves open the question of the non-markovian nature of such series.

is available, can be best represented by the FGN process. The merit of Mandelbrot and his collaborators in this connection, consists mainly in pointing out the need for reproducing long-term properties such as Hurst's H and proposing one possible solution to the problem.

It appears therefore that in trying to formulate models for river flows, hydrologists have arrived at two different classes, autoregressive and fractional Gaussian noise, which reproduce the short- and long-term properties of the record respectively. Given the inability of FGN processes to reproduce the short-term properties of flows, attempts have been made to use the class of linear stationary processes of which the autoregressive process is a special case to simulate long-term properties of river flows and in particular the Hurst property (O'Connell 1971). This class of "autoregressive moving average integrated (ARIMA) processes" has been studied extensively in Box and Jenkins (1970).

ARIMA stochastic processes consist of the following cases:

- (1) Autoregressive (AR) processes of order p given by the equation

$$(2.11) \quad X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + a_t$$

where X_t are the original observations or suitable transformations, ϕ_i are parameters and a_t are identically and independently distributed random variables.

- (2) Moving average (MA) process of order q given by the equation

$$(2.12) \quad X_t = a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}$$

where θ_i are parameters and a_t is defined as above.

- (3) Mixed (ARIMA) process of order (p, q) given by the equation

$$(2.13) \quad X_t = \phi_1 X_{t-1} + \dots + \phi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q}$$

where ϕ_i , θ_i and a_t are defined as above.

Box and Jenkins (1970) describe a systematic approach for fitting ARIMA processes, which consists of the following steps:

(a) Identification

The correlogram of the data is compared with the autocorrelation functions of various orders and types of ARIMA processes and the appropriate process identified.

(b) Estimation

The parameters of the process are estimated using iterative least squares from the condition that the sum of the squares of the residuals is minimized. (See Appendix 2.1).

(c) Diagnostic testing

Diagnostic tests of randomness are carried out on the residuals, to verify the adequacy of the model.

(4) Forecasting

If the model has proved satisfactory, it is used for forecasting future values of the observed series.

Given the extensive literature and the availability of computer programs on ARIMA processes and particularly the possibility of using them for reproducing both the short- and long-term properties of hydrological series, it is apparent that ARIMA processes are a very useful tool for hydrological modelling.

Another type of process which reproduces the long-term properties of the record and is intuitively more attractive than the fractional Gaussian noise process was proposed by Mejia, Rodriguez-Iturbe and Dawdy (1972). This broken-line process, is defined in the following way:

We consider first the single broken line process (Fig. 2.2) given by the equation

$$(2.14) \quad \xi_i(t' - k_i a_i) = \eta_{i,n} + \frac{(t' - na_i)}{a_i} (\eta_{i,n+1} - \eta_{i,n})$$

where $na_i \leq t' \leq (n+1)a_i$

where η_i are mutually independent, identically distributed random variables with mean zero and variance σ_i^2 and k_i are random variables independent of the η_i , which are uniformly distributed in the interval

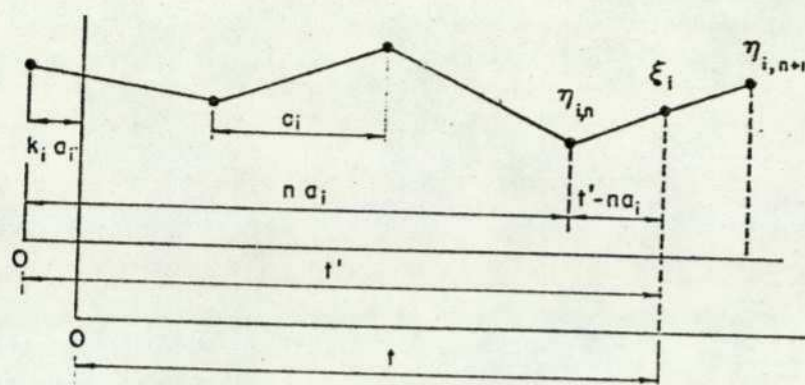


FIG. 2.2

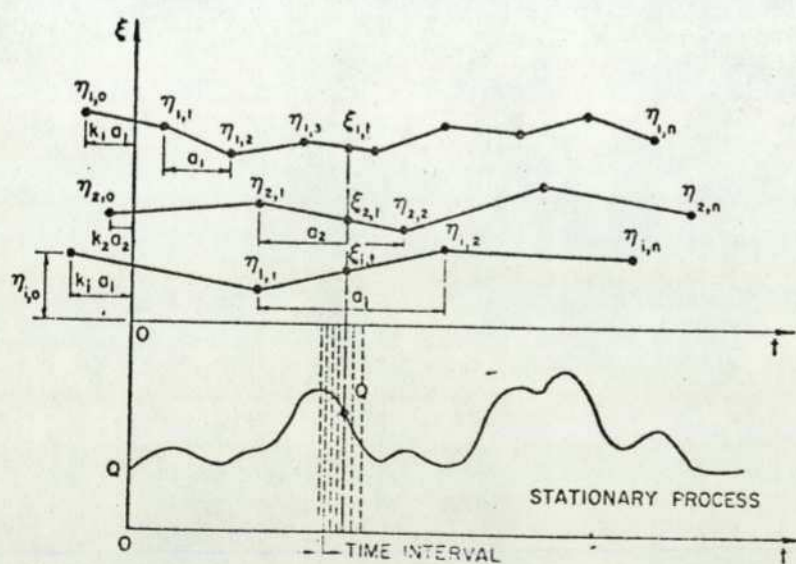


FIG. 2.3

$[0, 1]$. The two parameters of the single broken line process are σ_1^2 and a_1 , the time interval between the occurrences of two consecutive values of η_1 .

The broken line process is now defined by the following equation

$$(2.15) \quad Q(t) = \sum_{i=1}^N \xi_i(t)$$

as the sum of N single broken line processes. (Figure 2.3).

A practical application of this process is given in Garcia et al (1972).

The basis of the broken line process is that the observed record is the result of the superposition of phenomena occurring at regular intervals of various lengths. This idea is intuitively appealing, but the structure of the individual components (single broken line processes) is very schematic. Hydrological processes are only a part of the water cycle which in turn is influenced by other geophysical phenomena and therefore the long-term properties of the variables of the hydrologic cycle have physical causes.

One of the best known examples of such a causal relationship is the theory by Milankovic concerning the relationship between the motion of the earth, the resulting fluctuations of incoming radiation and the observed occurrence of ice ages on the earth (Fig. 2.4). Assuming that this theory is correct, it appears that the very long-term trend of the solar radiation and therefore of the mean temperature of the Earth's surface is deterministic and can be computed using the methods of celestial mechanics.

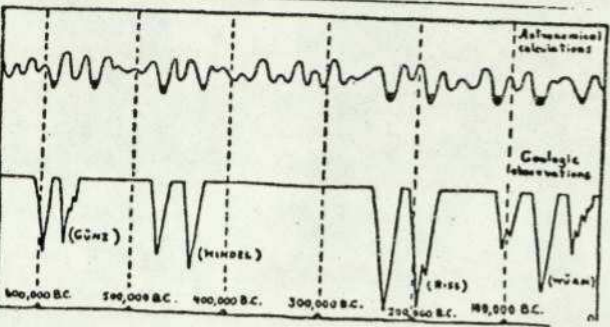


FIG. 2.4

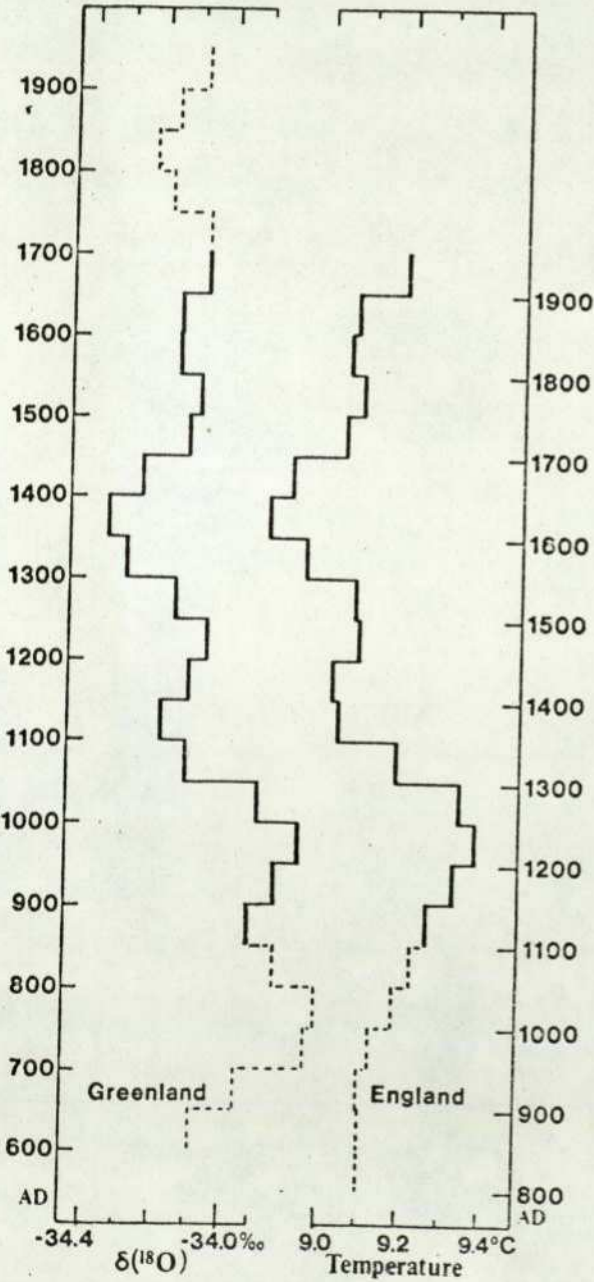


FIG. 2.5(b)

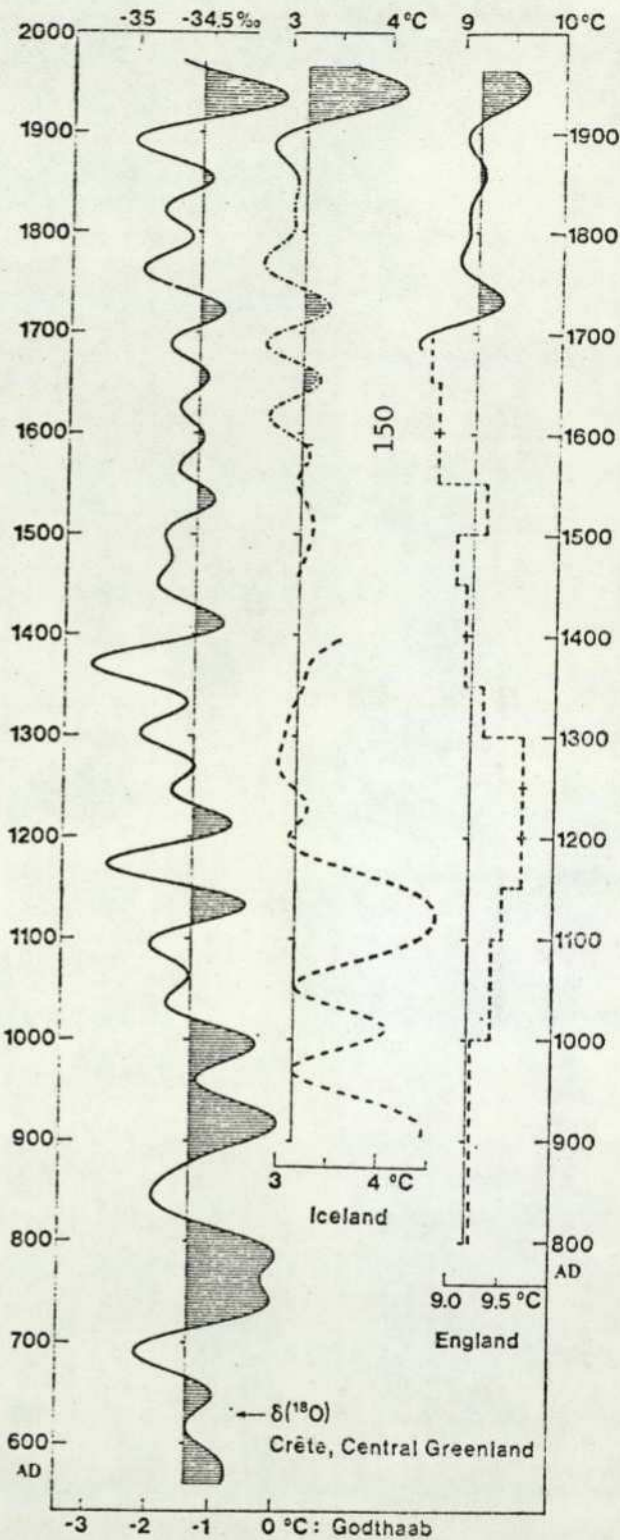


FIG. 2.5(a)

Turning now to a shorter time scale which is of more interest to the engineer, we compare the temperature fluctuations in Greenland, Iceland and England shown in Fig. 2.5(a) (Dansgaard et al. 1975). An examination of this figure indicates that, when smoothed by a 60-year low pass digital filter, the temperatures in Iceland and England for the last 350 years show good correlation with the δ -index of temperature in Greenland defined as the concentration of the heavy oxygen isotope ^{18}O on the ice sheet.

If on the other hand we use a 250-year low-pass digital filter for all variables, we find that the long-term fluctuations of temperatures in Iceland and England by 150 and 240 years respectively (Fig. 2.5(b)).

The above comparison shows that the long-term temperature fluctuations in Iceland and England and probably in Western Europe lag behind the corresponding fluctuations in Greenland and Dansgaard et al. (1975) suggest a meteorological explanation for this phenomenon. Regardless of whether their explanation is correct, it is apparent that the long-term properties of meteorological phenomena in various parts of the earth are related in complex ways and they can be understood only by elucidating their physical causes.

It appears from the above considerations that, if we want to take into account the long-term properties of the variables influencing the hydrologic cycle, we must investigate their relations with causative factors such as the solar radiation and, for shorter period phenomena, the disturbances in the atmospheric circulation. Attempts to postulate processes such as fractional gaussian noise or broken line, which conserve a number of statistical properties of the complex time behavior of hydrological variables, do not use the available and very valuable information on the causative factors affecting the hydrologic cycle.

Before leaving the subject of univariable modelling of river flows we must note that, in order to find processes that reproduce both short- and long-term properties of the flow record, it may be necessary to consider the observations as the result of a superposition of the effects of two or more stochastic processes with different characteristics. This approach would be similar to the time series analysis methods used in economics, which attempt to estimate the components of the series i.e. the trend, the cyclical and the seasonal component. The difficulties and limitations of this approach are well known and hydrologists must proceed with care if they want to explore this possibility.

2.2 Transfer function models with one input and one output

In the general discussion on hydrological modelling we mentioned the limitations of the regression approach for the estimation of operational models with stochastic input, and particularly the fact that the residuals of an ordinary regression are not independently distributed. In this section we shall deal with the subject of stochastic modelling of relationships (Transfer functions) between hydrological variables in a more rigorous way.

Let us consider a linear equation representing the relationship between the input x_t and the output y_t of a hydrological system

$$(2.16) \quad y_t = \sum_{i=1}^s \omega_i x_{t-i} + u_t$$

where u_t is a stochastic process the autocorrelation properties of which can be estimated together with the coefficients ω_i of the transfer or impulse response function. We shall discuss briefly two methods for identifying and estimating models of the above type.

- (1) The Transfer function method of Box and Jenkins.
- (2) The Transfer function method of L. Haugh.

The method of Box and Jenkins

Suppose that the relation between input and output is given by equation (2.13) and that the input x_t is generated by a stochastic process. The Box and Jenkins approach consists of the following steps:

Step 1

Estimate the parameters of the stochastic process generating x_t and compute the residuals from the fitted model. For example, if the input can be represented by the ARIMA model

$$(2.17) \quad x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + \alpha_t + \omega_1 \alpha_{t-1} + \dots + \omega_q \alpha_{t-q}$$

We estimate the coefficients ϕ_1 and ω_1 of the model and compute the residuals α_t recursively from the equation

$$(2.18) \quad \alpha_t = x_t - \phi_1 x_{t-1} - \dots - \phi_p x_{t-p} - \omega_1 \alpha_{t-1} - \dots - \omega_q \alpha_{t-q}$$

Step 2

Using the observations of the output y_t we compute the residuals β_t from the same model, using the equation

$$(2.19) \quad \beta_t = y_t - \phi_1 y_{t-1} - \dots - \phi_p y_{t-p} - \omega_1 \beta_{t-1} - \dots - \omega_q \beta_{t-q}$$

Step 3

We estimate the cross-correlation coefficient $r(\alpha_t, \beta_{t-k})$ between β_t and α_{t-k} for all positive and negative lags k .

Step 4

The coefficients H_k of the impulse response function are given by the formula:

$$(2.18) \quad H_k = \frac{r_{\alpha\beta}(k) \cdot s_\beta}{s_\alpha}$$

We note that equation (2.18) is identical to the well known equation for the regression coefficient of simple regression:

$$(2.19) \quad b = \frac{r_{xy} \cdot s_y}{s_x}$$

Equation (2.18) does not give efficient estimates of the impulse response (or Transfer function) coefficients and is used only to provide initial values for the estimation procedure.

Step 5

Using the above initial values, the next step is the simultaneous and efficient estimation of the model (2.13) including the parameters (ϕ_i, Ω_i) of the stochastic process generating the noise N_t :

$$(2.20) \quad N_t = \phi_1 N_{t-1} + \dots + \phi_p N_{t-p} + a_t + \Omega_1 a_{t-1} + \dots + \Omega_q a_{t-q}$$

Step 6

In the final step, the estimations of the transfer function and the noise models are combined in the initial model

$$(2.14) \quad Y_t = \sum_{j=0}^M H_j x_{t-j} + N_t$$

The method of L. Haugh

The basic idea of this method is to fit univariate stochastic models to both the input and the output of the system. The method is described in Haugh (1972).

Consider a hydrological system with input x_t and output y_t and assume that we can fit the two ARIMA models

$$(2.23) \quad x_t = \phi_1 x_{t-1} + \dots + \phi_p x_{t-p} + u_t + \theta_1 u_{t-1} + \dots + \theta_q u_{t-q}$$

$$(2.24) \quad y_t = \phi'_1 y_{t-1} + \dots + \phi'_p y_{t-p} + v_t + \theta'_1 v_{t-1} + \dots + \theta'_q v_{t-q}$$

to the input and the output. The method can be described in the following steps, shown in figure (2.2) adapted from Korfage (1974).

Step 1

Compute the residuals u_t and v_t of the above models.

Step 2

Compute the cross-correlation function $r(u_{t-k}, v_t)$ between u_{t-k} and v_t for all positive and negative values of the lag k .

Step 3

Examine the cross-correlation function in order to distinguish between the following four types of relationship between u_{t-k} and v_t .

(See Figure 2.6 and Table 2.1).

TABLE 2.1

	TYPE OF STATISTICAL ASSOCIATION BETWEEN x AND y	Values of $r(u_{t-k}, v_t)$		
		$k < 0$	$k=0$	$k > 0$
(1)	x, y INDEPENDENT	0	0	0
(2)	FEEDBACK BETWEEN x AND y	$\neq 0$ AND		$\neq 0$
(3)	SIMULTANEOUS ASSOCIATION		$\neq 0$	
(4)	SIMULTANEOUS ASSOCIATION WITHOUT PREDICTABILITY	0	$\neq 0$	0
(5)	ONLY SIMULTANEOUS ASSOCIATION IF AT ALL	0		0
(6)	x PREDICTS y			$\neq 0$
(7)	SIMULTANEOUS AND LAGGED ASSOCIATION		$\neq 0$	$\neq 0$
(8)	x PREDICTS y BUT NOT INSTANTANEOUSLY	0	0	$\neq 0$
(9)	UNIDIRECTIONAL PREDICTABILITY] EITHER FROM x TO y OR	0	0	AND $\neq 0$ AND $\neq 0$
(10)	y DOES NOT PREDICT x	0		
(11)	y DOES NOT PREDICT x AT ALL	0	0	

THE HAUGH METHOD OF ESTIMATION OF TRANSFER FUNCTION MODELS

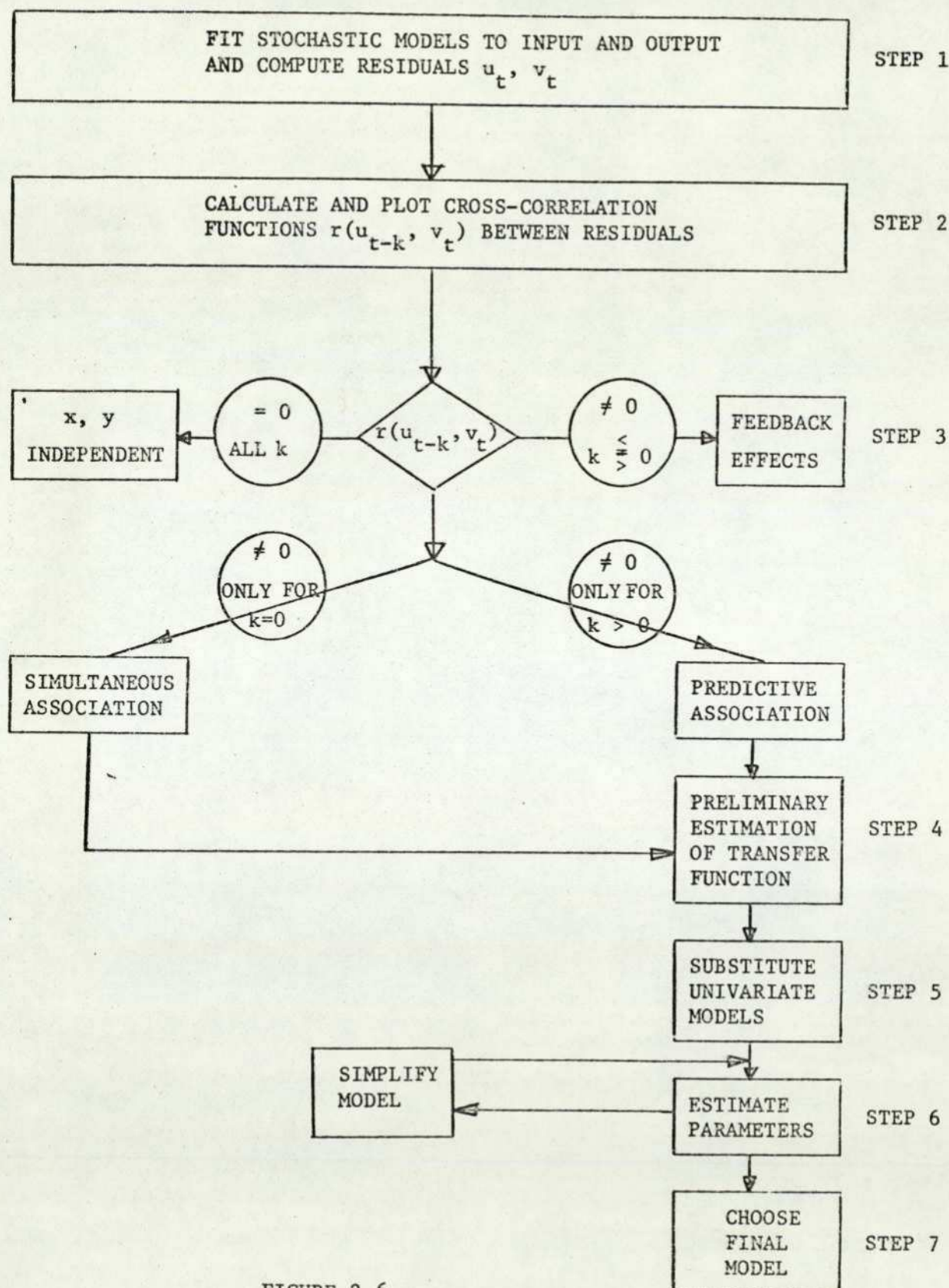


FIGURE 2.6

$$(a) \quad r(u_{t-k}, v_t) = 0 \text{ for all } k$$

The residuals u_t and v_{t-k} are uncorrelated for all k , i.e. there is no relationship between the input x_t and the output y_t of the system.

$$(b) \quad r(u_{t-k}, v_t) \neq 0 \text{ for } k = 0$$

In this case, called by Granger (1969) "instantaneous causality", we are dealing with a simultaneous statistical association and a knowledge of the input x_t cannot help to forecast the output y_t . However, the hydrological situation may suggest that there is a causal relationship which we can study by examining the form of the transfer function.

$$(c) \quad r(u_{t-k}, v_t) \neq 0 \text{ for } k > 0$$

This is a case of predictive association between x and y . Cases (b) and (c) together can be called "simultaneous and lagged statistical association between x and y ". Haugh uses the term "dynamic regression" for this combined case. This case is appropriate when the hydrological situation suggests a direction of causality from x_t to y_t (e.g. x_t = precipitation, y_t = runoff), and is the main subject of the present section.

$$(d) \quad r(u_{t-k}, v_t) \neq 0 \quad \text{for } k \begin{matrix} \geq \\ < \end{matrix} 0$$

This is the most complex case, where feedback is present in the hydrological system, as for example when x_t = ground water level and y_t = river flow. Haugh gives a brief description of this case which is undoubtedly of great importance in hydrology and a promising area for future research.

Step 4

Returning now to the case (c), we estimate the impulse response function of the residual series by considering the following model:

$$(2.25) \quad v_t = \phi_1' v_{t-1} + \dots + \phi_p' v_{t-p} + u_t + \psi_1' u_{t-1} + \dots + \psi_q' u_{t-q} \\ + a_t + \Omega_1' a_{t-1} + \dots + \Omega_q' a_{t-q}$$

Given that u_t is not autocorrelated, model (2.25) can be fitted in the same way as the model (2.16) described in the previous section on the Box and Jenkins method.

Step 5

The univariate models (2.23) and (2.24) fitted to x_t and y_t are now combined with equation (2.25) to give:

$$\begin{aligned}
 (2.26) \quad y_t = & \phi'_1 y_{t-1} + \dots + \phi'_p y_{t-p} \\
 & + \phi'_1 x_{t-1} + \dots + \phi'_p x_{t-p} + \psi'_1 u_{t-1} + \dots + \psi'_q u_{t-q} \\
 & + a_t + \Omega'_1 a_{t-1} + \dots + \Omega'_q a_{t-q} \\
 & + \theta'_1 v_{t-1} + \dots + \theta'_q v_{t-q}
 \end{aligned}$$

In many cases it is possible to simplify in equation (2.26) by observing the values of the coefficients ϕ'_1 , ϕ'_i , ψ'_1 , Ω'_1 and θ'_1 . Haugh (1972 p. 119) discusses the conditions and gives examples of this simplification of the model.

Step 6

The parameters of the transfer function model are now estimated efficiently by using a non-linear least squares algorithm described in Box and Jenkins (1970).

Step 7

The final step is to choose the best model according to a criterion dictated by the nature of the problem. For example it may be possible to reduce the variance of the residuals a_t by using a more complicated model but a simpler model may be finally selected because it contains a smaller number of parameters.

The model fitted to the data must now be used to carry out tests of hypotheses concerning the type of relationship between x and y . Given the large variety of possible hypotheses shown in Table 2.1 this is a difficult and as yet unsolved statistical problem. (Pierce and Haugh 1975). Fortunately, it is possible to test two important hypotheses concerning the nature of the statistical association between x and y :

Case 1: x and y are independent.

Case 2: x is related to y by dynamic regression (Unidirectional predictability from x to y).

Case 1

We consider the cross-correlation function $r(u_{t-k}, v_t)$ between the residuals u_t and v_t of the univariate models fitted to the variables x_t and y_t respectively and select two constants N and M large enough to include any expected non-negligible non-zero coefficient of the transfer from u_t to v_t . We then compute the statistic

$$(2.27) \quad U = m \sum_{k=-N}^M \hat{r}^2(u_{t-k}, v_t)$$

where m = number of pairs of residuals available for the computation (Box and Jenkins 1970, p. 394). Under the null hypothesis of independence between u and v and therefore between x and y , the statistic U has a chi square distribution with $N + u + 1$ degrees of freedom.

2.3 Transfer function models with multiple input and output

For most hydrological systems it is necessary to consider multiple inputs and outputs. We shall now extend the methodology of the previous section to such systems and conclude with a hydrological example.

2.3.1 Systems with multiple input and single output

In the previous section we have considered the transfer function model

$$(2.29) \quad y_t = \sum_{i=1}^p \phi_i y_{t-i} + \sum_{i=0}^s \omega_i x_{t-i} + u_t$$

where the noise u_t is generated by the univariate ARIMA model

$$(2.30) \quad u_t = \sum_{i=1}^r \delta_i u_{t-i} + \sum_{i=1}^q \theta_i a_{t-i}$$

In the case of multiple inputs $x_1, x_2, \dots, x_M$, equation (1) can be extended in the following way:

$$(2.31) \quad y_t = \sum_{i=1}^p \phi_i y_{t-i} + \sum_{i=0}^{S1} \omega_{1i} x_{1(t-i)} + \dots + \sum_{i=0}^{SM} \omega_{Mi} x_{M(t-i)} + u_t$$

where again u_t is given by the ARIMA process of equation 2.30.

This model is fitted to the data according to the stage-wise procedure described in (Draper and Smith 1966). The procedure starts with the most important variable x_1 and then other input

variables are added one at a time after testing the statistical significance of their contributions to the variance of the output variable. Figure (2.7) shows the flow diagram of the computational procedure which can be described as follows:

STEP 1 Preliminary investigation

The initial input variable x_1 is selected on the basis of a preliminary examination of the correlation matrix of all variables and of the cross-correlograms of the output variable with each of the input variables.

STEPS 2, 3 AND 4 Bivariate modelling of x_1 and y

The bivariate modelling is carried out as outlined in Figure 2.6 of the previous section. As before, we do not consider feedback i.e. the case where the test statistic $r(u_1, u) \neq 0$ for both positive and negative values of the lag k . Consequently we examine only the case of dynamic regression for which $r(u_1, u) \neq 0$ for $k \geq 0$.

STEP 5 Univariate modelling of x_2

STEP 6 Testing the relationship between x_1 and x_2

Using the residuals u_2 from the univariate model fitted to x_2 we examine first the relation between the inputs x_1 and x_2 . This is analogous to the well-known situation in multiple

FIG. 2.7(a)

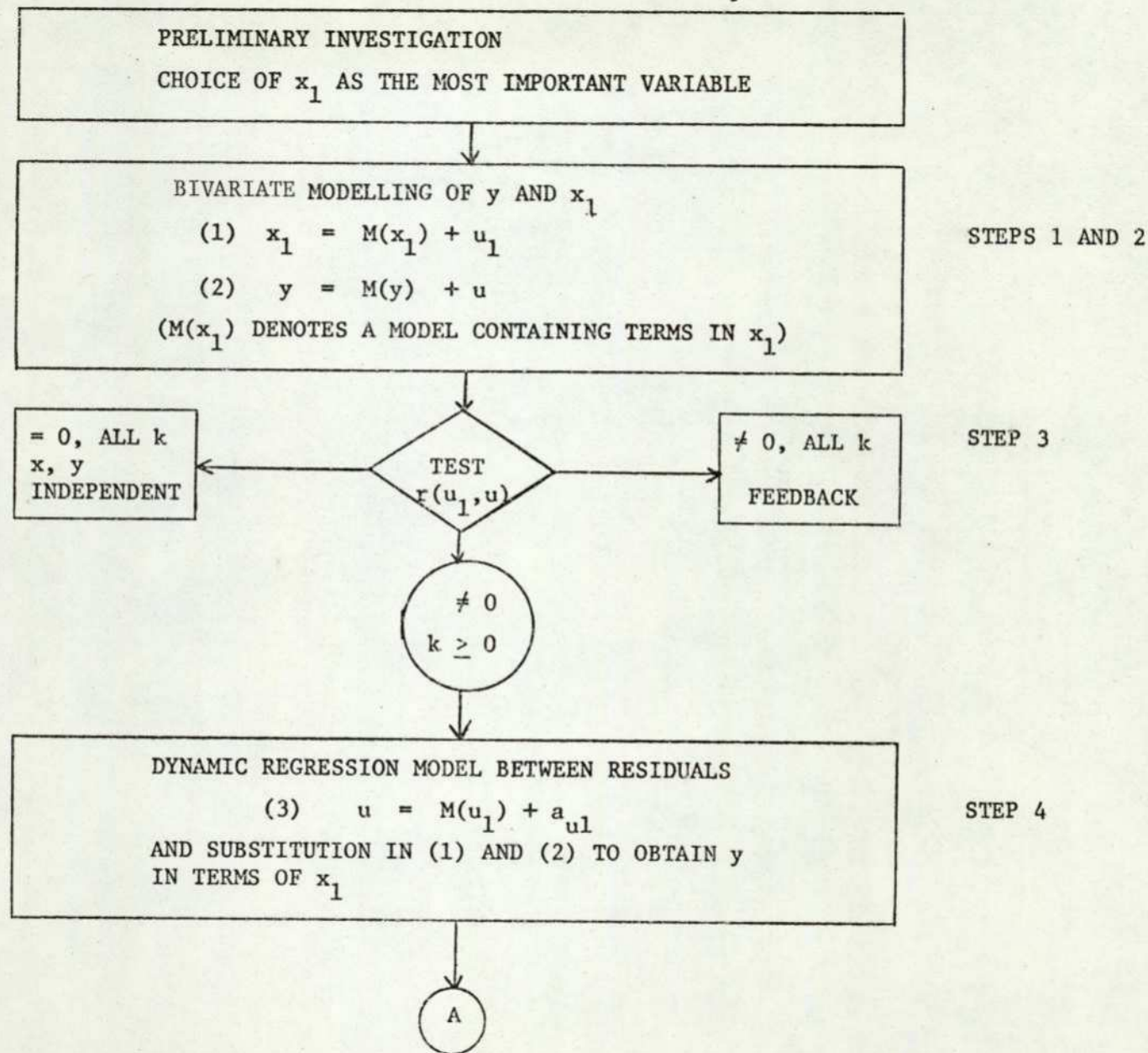


FIG. 2.7(b)

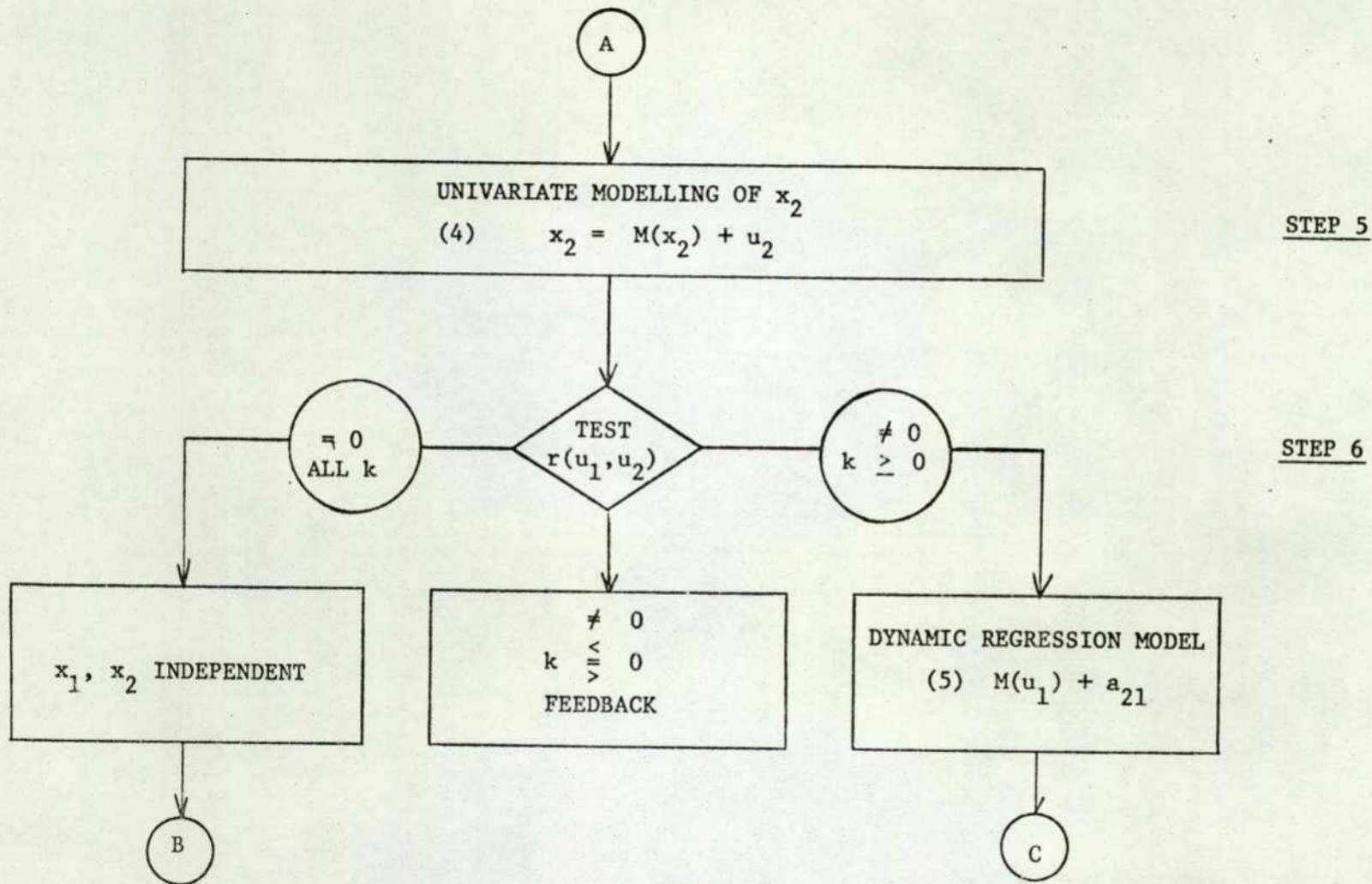
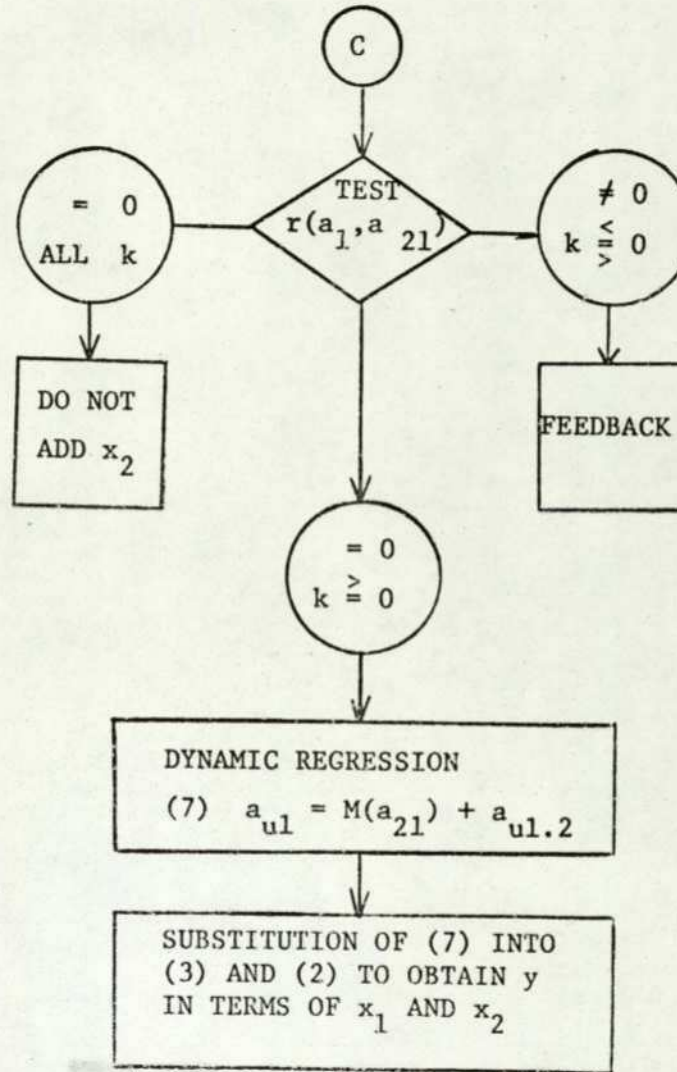
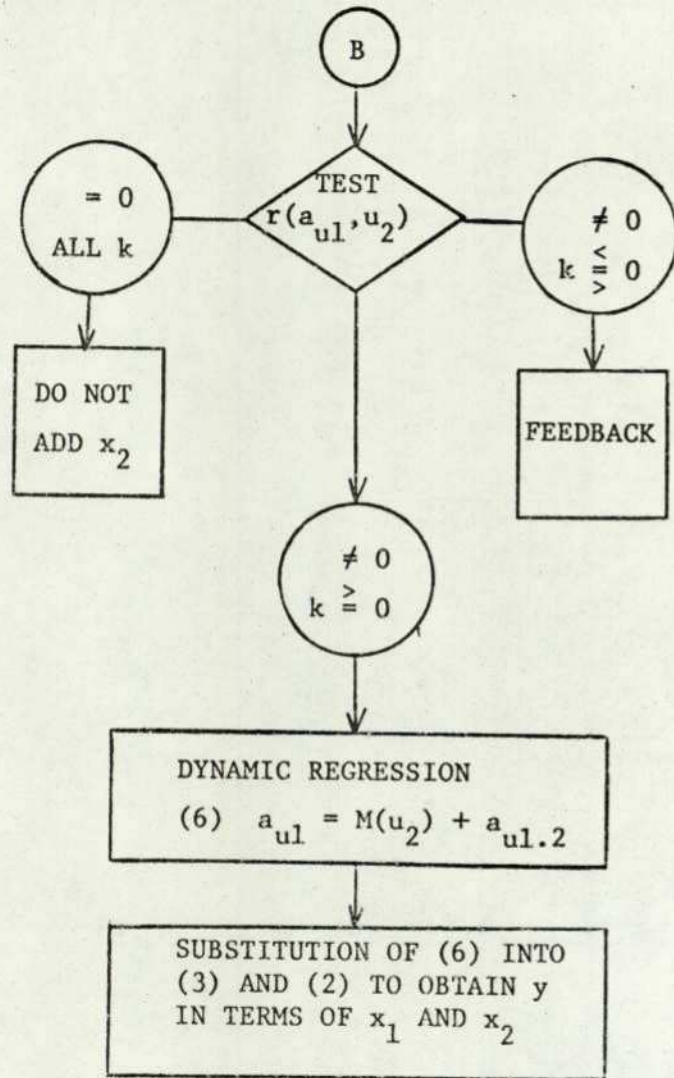


FIG. 2.7(c)



STEP 7

STEP 8

STEP 9

regression where the existence of multicollinearity between independent variables influences the test concerning the inclusion of an additional independent variable in the regression equation.

This preliminary test is based on the value of the cross-correlation $r(u_1, u_2)$ for all lags k . We assume again that $\gamma(u_1, u_2) = 0$ or that $r(u_1, u_2) \neq 0$ for $k \geq 0$ and therefore the two inputs are either independent or their relationship is a dynamic regression.

If x_1 and x_2 are independent we proceed to test whether to add x_2 to the transfer function. If on the other hand $r(u_1, u_2) \neq 0$ for $k \geq 0$ we fit a bivariate dynamic regression model to the residuals u_2 and u_1 and obtain the residuals a_{21} .

STEP 7 Testing whether to include a second input x_2

In the cases of independence and dynamic regression between x_1 and x_2 we base the decision to include the second input x_2 on the values of the correlation coefficients $r(a_{u1}, u_2)$ and $r(a_{u1}, a_{21})$ respectively.

STEP 8 Estimation of the dynamic regression model between residuals

STEP 9 Derivation of transfer function between the inputs x_1 and x_2 and the output y

2.3.2 Multiple input and multiple output

Many hydrological systems have multiple inputs and outputs. Consider for example the basin shown in Figure 2.8 and assume that we are interested in modelling the short-term behaviour of the flows Q_1 and Q_2 on the basis of the two precipitation stations P_1 and P_2 . The natural transfer function representation of this basin is shown in Fig. 2.9 where the system has two inputs and two outputs. However, the usual way to study this problem is to analyse the two-input, one-output systems shown in figures 2.10 and 2.11 using multiple regression.

In the previous section we have examined the alternative method of transfer function modelling with multiple input and single output. These models have the advantage that their estimation procedure is based on the condition that the residuals are random. As we have seen, the intercorrelation between the input variables causes difficulties in the identification and the estimation of multiple input transfer function models. The estimation of multiple input and multiple output systems presents even greater difficulties both because of the complexity of the identification and estimation procedures (Quenouille 1957, Haugh 1972) and particularly because of the difficulty of interpreting physically the various interrelations between the individual variables. It seems therefore desirable, before studying the multiple input, multiple output transfer function

model, to transform the input and possibly also the output variables so as to reduce their intercorrelations. Given the nature of both input and output, significant correlations may occur at all values of the lag k and therefore it is not possible to find a transformation which eliminates all correlations.

In this study we shall use the well known transformation of canonical correlation described in Appendix 2.2 which eliminates the simultaneous correlations between all variables except for two pairs of variables belonging to the input and the output group respectively. The canonical correlation approach to transfer function modelling consists mainly in transforming the original input and output variables, estimating the transfer functions between the canonical variables and then deriving the corresponding transfer functions between the original variables by algebraic substitution. We shall now discuss the successive steps of this procedure in more detail, assuming for simplicity of presentation that first order autoregressive processes are appropriate for both inputs and outputs.

The transfer functions with single input and single output will be fitted according to Haugh's method discussed in the previous section.

STEP 1 We fit the following univariate models to the input and output variables (P_1, P_2) and (Q_1, Q_2) :

$$P_{1t} = \theta_1 P_{1(t-1)} + u_1 \quad ; \quad Q_{1t} = \theta'_1 Q_{1(t-1)} + u'_1$$

$$P_{2t} = \theta_2 P_{2(t-1)} + u_2 \quad ; \quad Q_{2t} = \theta'_2 Q_{2(t-1)} + u'_2$$

We compute the residuals u_1, u_2, u'_1, u'_2 and analyse the system shown in Figure 2.12.

STEP 2 Using the two sets of variables u_1, u_2 and u'_1, u'_2 we derive the two corresponding sets of canonical variables $(\xi_1, \xi_2), (\xi'_1, \xi'_2)$ as described in Appendix 2.2:

$$\xi_1 = a_{11} u_1 + a_{12} u_2$$

$$\xi_2 = a_{21} u_1 + a_{22} u_2$$

$$\xi'_1 = a'_{11} u'_1 + a'_{12} u'_2$$

$$\xi'_2 = a'_{21} u'_1 + a'_{22} u'_2$$

Solving the above equations for u_1, u_2, u'_1, u'_2 we obtain

$$u_1 = b_{11} \xi_1 + b_{12} \xi_2$$

$$u_2 = b_{21} \xi_1 + b_{22} \xi_2$$

$$u'_1 = b'_{11} \xi'_1 + b'_{12} \xi'_2$$

$$u'_2 = b'_{21} \xi'_1 + b'_{22} \xi'_2$$

STEP 3 According to the theory of canonical correlation, the following pairs of variables are uncorrelated at $k=0$: (ξ_1, ξ_2) , (ξ_1, ξ_2') , (ξ_1', ξ_2) , (ξ_1', ξ_2') and therefore the original system of Figure 2.12 has been reduced to the two systems shown in Figure 2.13 which have one input and one output each.

STEP 4 We estimate the two bivariate transfer function models of Figure 2.13 and then we derive the transfer function of Figures 2.10 and 2.11 by algebraic substitutions. The details of this computation will be shown in the example given in the next section.

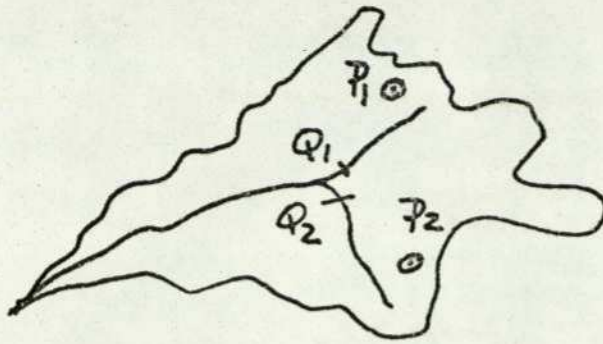


FIG. 2.8

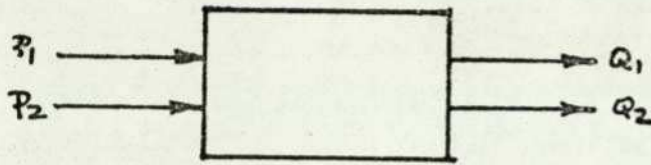


FIG. 2.9

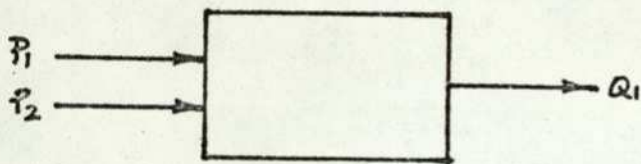


FIG. 2.10

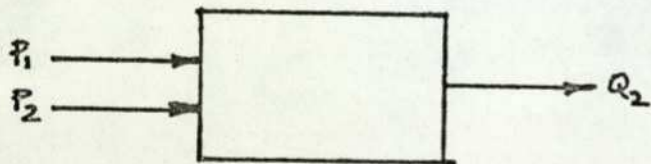


FIG. 2.11

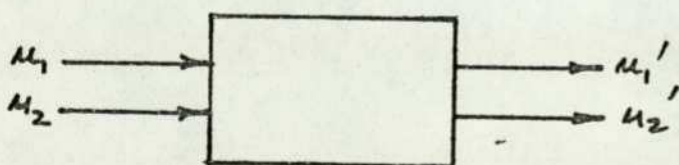


FIG. 2.12

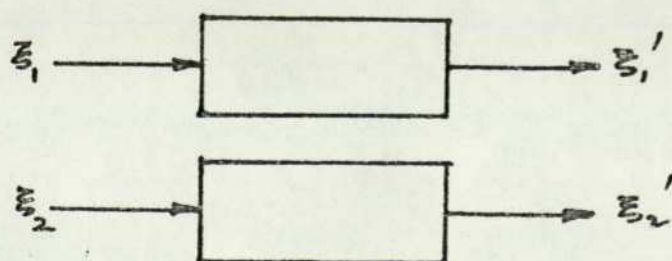


FIG. 2.13

2.4.1 Choice of estimation method on the basis of the partition of the sum of squares of the output

We have studied above various methods of estimating the transfer function between the input and the output of a system. In this section we shall examine the factors which determine the choice of a method of estimation.

An examination of the equation of the transfer function model

$$(2.32) \quad y_t = \left[\sum_{i=1}^p \phi_i y_{t-i} \right] + \left[\sum_{i=0}^{S1} \omega_{1i} x_{1(t-i)} + \dots + \right. \\ \left. \sum_{i=0}^{SM} \omega_{Mi} x_{M(t-i)} \right] + [u_t]$$

shows that the total sum of squares of y_t can be broken down into three components

$$(2.33) \quad SSy = SSy^* + SSx + SSu$$

where:

SSy = Total sum of squares of y_t

SSy^* = Sum of squares attributable to autoregressive terms

SSx = Sum of squares attributable to the input variables

SSu = Sum of squares attributable to the residuals

The relative values of the three ratios in the equation

$$(2.34) \quad \frac{SSy^*}{SSy} + \frac{SSx}{SSy} + \frac{SSu}{SSy} = 1$$

determine whether the behavior of the output y_t is dominated mainly by the previous observations of the output, by the input or by the noise term. All three cases can be represented as points on the sides or in the interior of the equilateral triangle xya of Fig. 2.14, where the three ratios are equal to the lengths of the segments perpendicular to the sides xa , ya and xy respectively. Figures 2.15 and 2.16 show the correspondence between the types of models and the points of the triangle in more detail.

Case y The point is at the apex Y and we have

$$(2.35) \quad y_t = \sum_{i=1}^p \phi_i y_{t-i}$$

i.e the output is a deterministic function of its previous values. The coefficients $\phi_1, \dots, \phi_p$ can be determined by the methods of finite difference equations.

Case x The point is at the apex x and we have

$$(2.36) \quad y_t = \sum_{i=0}^{S1} \omega_{1i} x_{1(t-i)} + \dots + \sum_{i=0}^{SM} \omega_{Mi} x_{M(t-i)}$$

i.e. the output is a deterministic function of current and lagged values of the input x . The appropriate method for determining the coefficients $\omega_{1i}, \dots, \omega_{Mi}$ is the solution of a system of simultaneous equations (Deconvolution).

Case a The point is at the apex a and we have (6) $y = a_t$ i.e y_t is white noise and no further modelling is possible.

Case xy The point is on the side xy and we have the deterministic relation

$$(2.37) \quad y_t = \sum_{i=1}^P \omega_i y_{t-i} + \sum_{i=0}^{S1} \omega_{1i} x_{1(t-i)} + \dots + \sum_{i=0}^{SM} \omega_{Mi} x_{M(t-i)}$$

Methods for determining the coefficients of the deterministic dynamic model of equation (2.37) are given e.g. in chapter 10 of Box and Jenkins (1970).

Case ya The point is on the side ya and we have the univariate stochastic model

$$(2.38) \quad y_t = \sum_{i=1}^P \phi_i y_{t-i} + u_t$$

Methods of estimating such models were discussed in section 2.1.

Case xa The point is on the side xa and we have the equation

$$(2.39) \quad y_t = \sum_{i=0}^{S1} \omega_{1i} x_{1(t-i)} + \dots + \sum_{i=0}^{SM} \omega_{Mi} x_{M(t-i)} + u_t$$

The methods of regression analysis, possibly with autocorrelated residuals, can be used to estimate the coefficients ω . It is noted that among the exogenous variables $x_1, \dots, x_M$. We may include functions of time (trends) and therefore "signal plus noise" models belong to this category.

In contrast to the above, the remaining cases correspond to proper transfer function models where all three ratios in equation (2.34) are different from zero.

Case (xy)a The point is in the interior of the triangle, near the side (xy). The ratio $\frac{SSu}{SSy}$ is relatively small but the two

other ratios $\frac{SSy^*}{SSy}$ and $\frac{SSx}{SSy}$ are comparable i.e. there is

relative symmetry in the respective roles of the input and the autoregressive terms. In this case we start with a deterministic difference equation model and then we estimate a model for the residuals following the methods appropriate for cases (ya)x, (xa)y or (xya) according to whether the predominant factor is the output, the input or both are equally important.

FIGURE 2.14

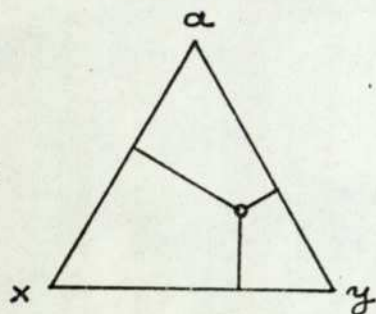


FIGURE 2.15

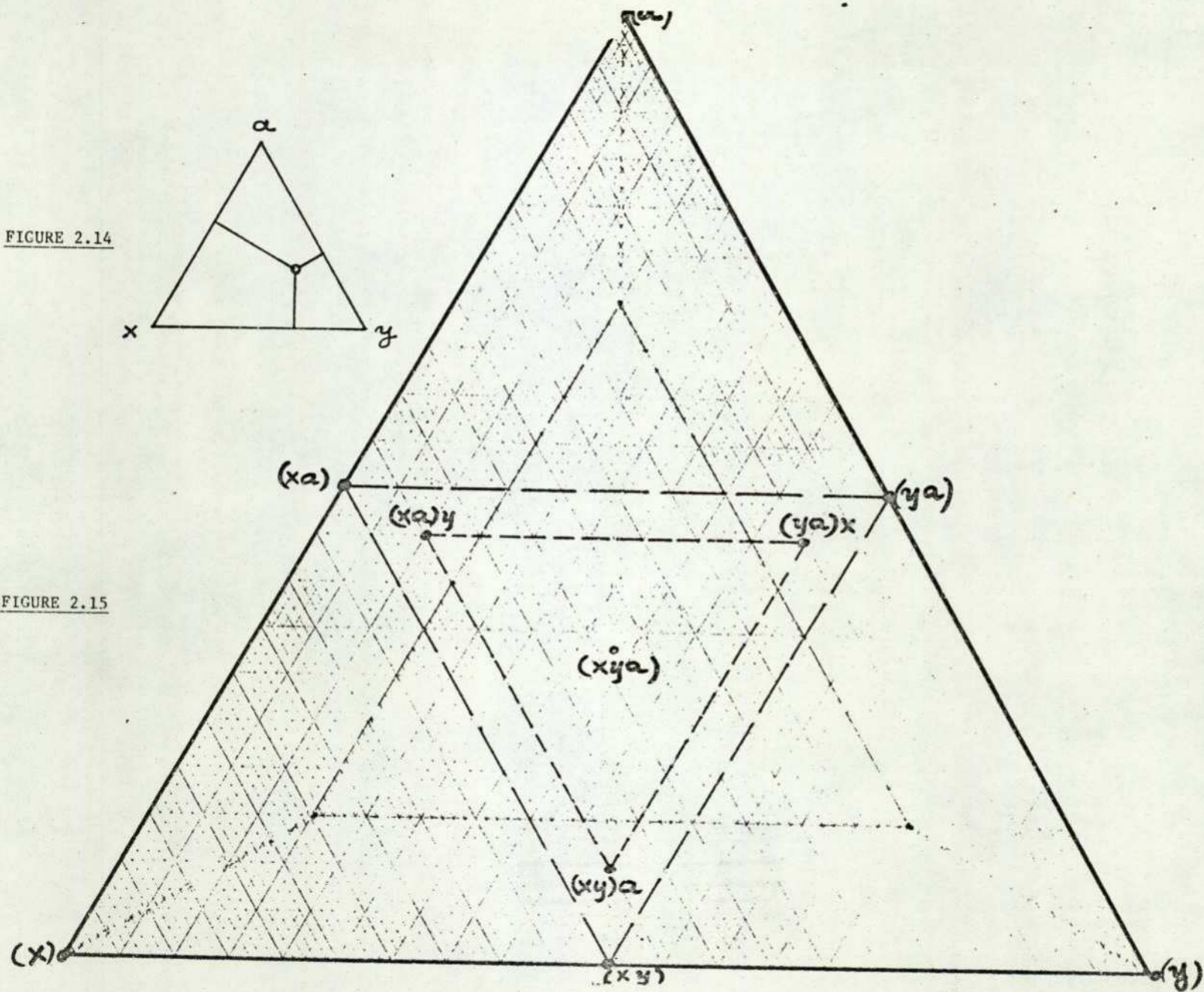
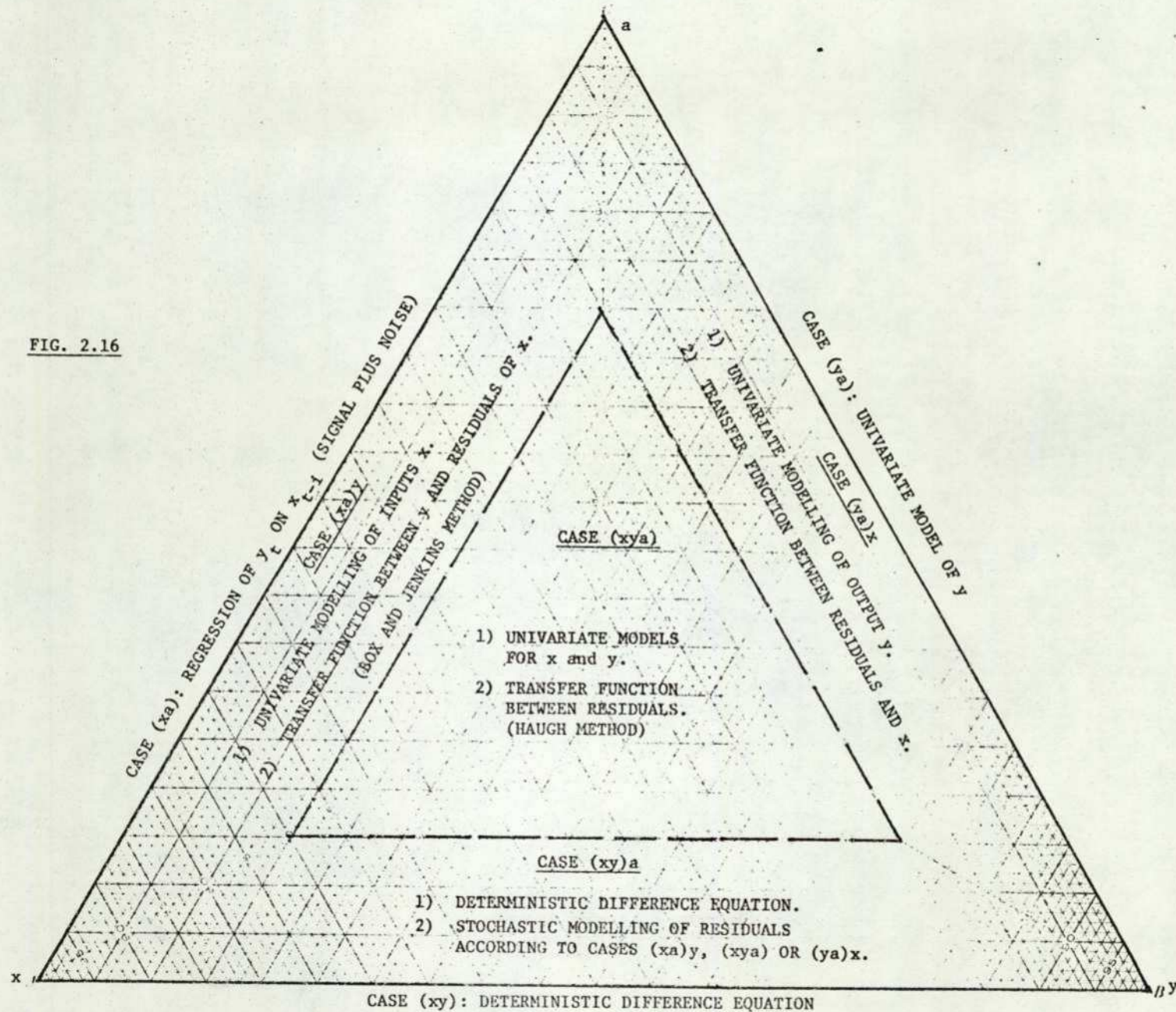


FIG. 2.16



Case (ya)x The point is in the interior of the triangle near the side (ya) and the autoregressive terms predominate. In this case we start by estimating a univariate model of y_t , compute the residuals and then estimate the transfer function between the residuals and the input variables.

Case (xa)y The point is in the interior of the triangle near the side (xa) and the influence of the input predominates. In this case we start with a regression model between y_t and the inputs and then we estimate a univariate model for the residuals. In this case of one input this can be achieved by the application of the Box and Jenkins prewhitening procedure described in the section on transfer function models with one input and one output.

Case (xya) The point is in the central region of the triangle where the relative sizes of the three ratios are comparable. In this case we fit univariate models to both the input and the output variables and then estimate the transfer function between the residuals. This is the method of Haugh described in the section on systems with one input and one output and it is also appropriate for multiple inputs and outputs.

2.4.2 General considerations on the properties of stochastic models

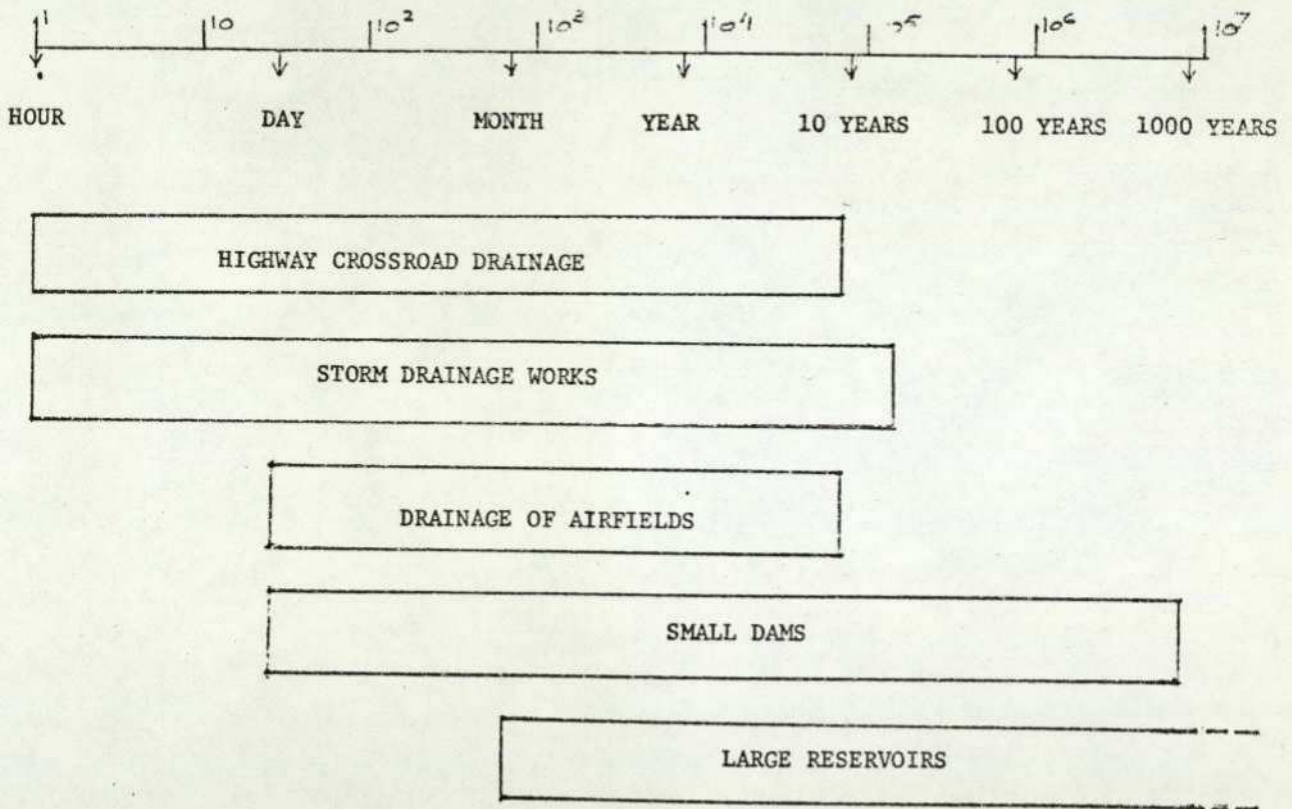
In the previous sections we have discussed several types of stochastic models and, in view of their varying characteristics, it appears necessary to establish criteria for choosing the most appropriate model for a given project. The basis for this choice is that the model should incorporate the properties of the flow record that are most relevant for the project under consideration.

The choice of the appropriate model depends on the conditions of the individual project and therefore a general discussion has limited usefulness. Figure 3.1 shows the time scales appropriate for certain types of water resources projects. An examination of this figure leads to the conclusion that the "perfect" stochastic model, which would incorporate all possible properties of the record, is not only very difficult to formulate but in most cases not even necessary.

Table 2.2 shows various properties of the flow records and the types of processes which incorporate these properties, i.e. for which these properties are taken into account in the estimation of the parameters. An examination of this table shows that:

- (a) The basic properties of the probability distribution such as the mean and the variance are incorporated in all models, but it is only the FCN and broken line processes that attempt to incorporate extreme value properties of the record. On the other hand, no available model takes into account the overall shape of the probability distribution of the observations.

TIME SCALES OF
WATER RESOURCES PROJECTS



LOWER LIMIT = BASIC TIME UNIT

UPPER LIMIT = RETURN PERIOD OF DESIGN FLOOD

FIG. 2.17

PROPERTIES INCORPORATED IN VARIOUS MODELS

TYPE OF PROPERTY OF RECORDED FLOWS	PROPERTY	MODELS INCORPORATING THE PROPERTY
PROPERTIES OF THE PROBABILITY DISTRIBUTION	MEAN VARIANCE SKEWNESS EXTREME VALUES FORM OF THE PROBABILITY DISTRIBUTION	ALL ALL ARIMA, FGN FGN, BROKEN LINE NONE SO FAR
SHORT-TERM PROPERTIES	AUTOCORRELATION CROSS-CORRELATION -INSTANTANEOUS -LAGGED AGGREGATION PROPERTIES SPECTRAL PROPERTIES TIME REVERSIBILITY	AR, ARIMA MULTIVARIATE ARIMA TRANSFER FUNCTION ARIMA WITH CONSTRAINTS ARIMA, FGN NONE SO FAR
LONG-TERM PROPERTIES	HURST'S EXPONENT NUMBER OF RUNS RUN LENGTHS RUN SUMS CROSSING PROPERTIES	FGN, BROKEN LINE, ARIMA BROKEN LINE " " "

TABLE 2.2

- (b) The short-term properties of the record such as autocorrelation and zero-order cross-correlation coefficients are incorporated in the univariate and multivariate ARIMA processes respectively. The lagged cross-correlation properties are taken into account only by the transfer function models which represent associations between stochastic processes in an explicit way. The spectral properties of the record, i.e. the relative importance of harmonic components with different periods, are partly incorporated in the ARIMA and the FGN models with emphasis on the high and low frequency ends of the spectrum respectively. None of these models, however, takes into account the overall shape of the spectrum in the estimation of the model parameters.
- (c) The long-term properties of the record such as the values of Hurst's exponent, the run numbers, lengths and sums and the crossing properties are incorporated only in the estimation of the FGN and broken line processes. It must be noted that the mixed autoregressive-moving average (ARIMA) model has been shown (O'Connell 1971) to simulate the Hurst property a posteriori, and that the incorporation of the above "long-term properties" is a very approximate way to take into account the very complex factors that determine the long-term behavior of geophysical variables.

The choice of an appropriate model for a given project depends also on the practical difficulties of estimating the model and the ability of the user to interpret its properties. From this point of view, the class of ARIMA models is attractive due to the relative ease and interpretability of its parameters. The additional possibility of extending the scope of ARIMA models to consider aggregation properties of the flows and lagged relationships, and even to simulate the Hurst property, suggests that further study of these processes may be very useful.

From this practical point of view, the FGN process is at a disadvantage because it requires a minimum of 30-50 years of record for reliable estimation, and its short-term behavior is very different from that of actual flows. Although the FGN and broken line processes are mathematically equivalent, the broken line process has the advantage of easier computation and interpretability of the parameters. It seems therefore that further work on the properties of these processes would be useful. It must be emphasized however that stochastic models are still in the early stage of their development and the classes of models discussed above represent only the first of a long series of models that will be formulated in the future to take into account specific needs of water resources project.

2.5.1 Hydrological system with one input and one output
(Bell and Megiscane rivers)

The methods of stochastic modelling discussed in Section 2 will now be applied to the monthly flows of the Megiscane and Bell rivers located near the James Bay in subarctic Quebec. The basins of the two gauging stations are shown in Figure 2.18 and the basic data are given in Tables 2.3 and 2.4 taken from two publications of the Government of Quebec (Rivière Megiscane 1969 and Rivière Bell 1969).

The purpose of the analysis is to determine the extent to which the flow of the Bell river can be estimated from the flow of the Megiscane river. The period from May 1924 to April 1939 was used for the computation of this example. The "hydrological year" from May to April was chosen to coincide with the occurrence of the spring flood, in order to minimize the correlation between the last month of one hydrological year and the first month of the next.

Before estimating their statistical association, the two monthly flows were standardized by subtracting the monthly means and dividing by the monthly standard deviations. This transformation is known to eliminate a large part of the seasonal variability of the monthly flows.

The statistical association between the two flows was estimated using different methods, and the results are shown in Table 2.5 which also includes the proportion of the variance of Bell flows (Y) that is explained by each model.

(1) Univariate model for Bell flows

The estimated model E(1) is autoregressive of first order and the U-test on the residuals u_{1t} shows that they are random. The proportion of the variance of Y explained by this univariate model is 0.40.

(2) Simple regression and univariable modelling of the residuals

Before dealing with the proper transfer function models, we estimate the simple regression between Bell (Y) and Megiscane (X). The residuals e_t of the regression equation E(2) are not random and therefore we estimate the univariate model E(3) for the residuals e_t and substitute into equation E(2). The resulting final model E(4) is rather complicated but its residuals e'_t are random. The model E(4) explains 0.69 of the variance of Y.

(3) Transfer function model $X \rightarrow Y$ (Box and Jenkins method of prewhitening X)

The estimated transfer function model is given by equation E(5) which is simpler than E(4) and has random residuals. This equation explains 0.74 of the variance of Y.

(4) Transfer function model $X \rightarrow Y$

(Haugh method of prewhitening both X and Y)

According to this method we estimate first the transfer function model E(7) between the residuals of the univariate models E(1) and E(6). The resulting model E(8) between the standardized Bell and Megiscane flows has random residuals and accounts for 0.73 of the variance of Y.

A comparison of the respective proportions of $\text{var}(Y)$ explained by the various transfer function models indicates that in this case the Box and Jenkins method gives the best results.

In table 2.6 we show the results of the corresponding computations for the system Bell - Megiscane. The above example shows that the methods of transfer function modelling described in this study can be used for the estimation of relationships between hydrological variables even in cases where there is no physical input-output relationship between them.

TABLE 2.5

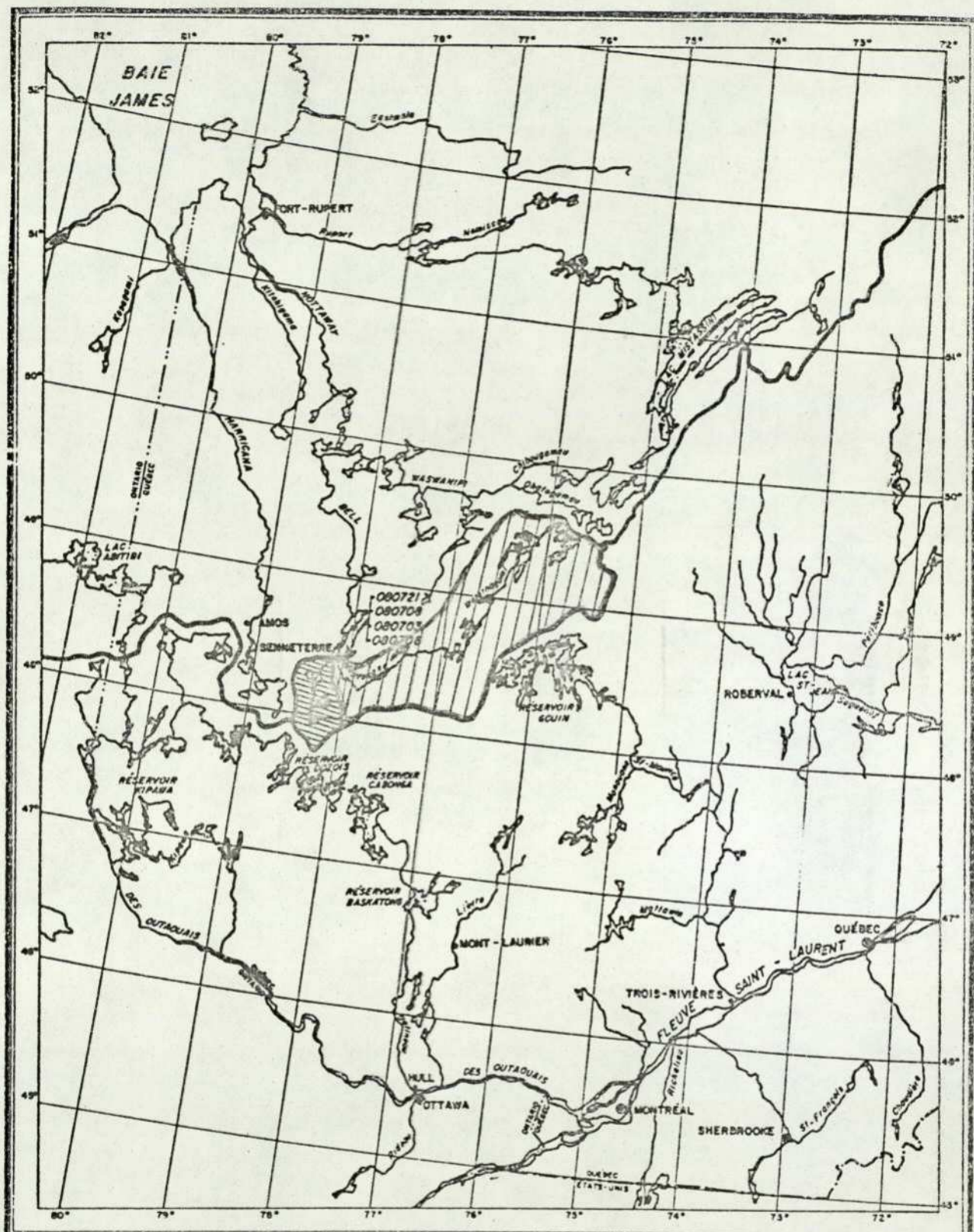
STATISTICAL ASSOCIATION BELL → MEGISCANE*

MODELLING METHOD	MODEL EQUATIONS	RANDOMNESS TEST OF MODEL RESIDUALS	STANDARD DEVIATION OF MODEL RESIDUALS	PROPORTION OF VAR (Y) EXPLAINED
UNIVARIATE MODEL FOR MEGISCANE	$Y_t = 0.52 Y_{t-1} + u_{1t}$	$U_{u1} = 17.77 < 33.9$ $\chi^2_{.05}(22) = 33.9$	0.83	0.27
SIMPLE REGRESSION WITH MODELLING OF RESIDUALS	E(2): $Y_t = 0.79X_t + e_t$ E(3): $e_t = 0.18e'_t + e'_t$ E(4): $Y_t = 0.79X_t + 0.18e'_t + e'_t$	$U_e = 29.5 < 31.4$ $\chi^2_{.05}(20) = 31.4$ $e = \text{RANDOM}$	$s(e) = 0.59$	0.64
		$U'_e = 16.5 < 28.9$ $\chi^2_{0.05}(18) = 28.9$	$s(e') = 0.58$	
TRANSFER FUNCTION (BOX AND JENKINS)	E(5): $Y_t = 0.82X_t - 0.12X_{t-2} + 0.18a_{t-1} + a_t$	$U_a = 13.8 < 30.1$ $\chi^2_{.05}(19) = 30.1$	$s(a) = 0.58$	0.64
TRANSFER FUNCTION (HAUGH)	E(6): $X_t = 0.68 X_{t-1} + u_{2t}$	$U_{u2} = 19.0 < 28.9$ $\chi^2_{.05}(18) = 28.9$	$s(u_2) = 0.75$	0.64
	E(7): $u_{1t} = -30 u_{1(t-1)} + u_{2t} + 0.44 u_{2(t-1)} + 0.06 u_{2(t-2)} + a_t$ E(8): $Y_t = 0.205 Y_{t-1} + 0.79X - 0.16X_{t-1} + a_t$	$U(a) = 14.2 < 30.1$ $\chi^2_{.05}(19) = 30.1$	$s(a) = 0.58$	

TABLE 2.6

STATISTICAL ASSOCIATION MEGISCANE → BELL

MODELLING METHOD	MODEL EQUATIONS	RANDOMNESS TEST OF MODEL RESIDUALS	STANDARD DEVIATION OF MODEL RESIDUALS	PROPORTION OF VAR (Y) EXPLAINED
BELL UNIVARIATE MODEL	$E(1): Y_t = 0.68Y_{t-1} + u_{1t} ; S(Y) = 0.97$	$U_{u1} = 19 < 29$ $\chi^2_{.05}(18) = 29$	$s(u_1) = 0.75$	0.40
SIMPLE REGRESSION WITH MODELLING OF RESIDUALS	$E(2): Y_t = 0.79X_t + e_t$ $E(3): e_t = 0.29e_{t-1} + 0.21e_{t-2} + e'_t$ $E(4): Y_t = 0.29Y_{t-1} + 0.79X_t - 0.23X_{t-1} - 0.17X_{t-2} + e'_t$	$U_e = 110 < 31$ $\chi^2_{.05}(20) = 31$ $e = \text{NON RANDOM}$ $U'_{e'} = 15.7 < 29$ $\chi^2_{.05}(18) = 29$	$s(e) = 0.59$ $s(e') = 0.54$	0.69
TRANSFER FUNCTION (BOX AND JENKINS)	$E(5): Y_t = 0.36Y_{t-1} + 0.60X_t + a_t - 0.36a_{t-1}$	$U_a = 19 < 30$ $\chi^2_{.05}(19) = 30$	$s(a) = 0.51$	0.74
TRANSFER FUNCTION (HAUGH)	$E(6): X_t = 0.52X_{t-1} + u_{2t}$ $E(7): u_{1t} = -0.32u_{1(t-1)} + 0.64u_{2t} + 0.20u_{2(t-1)} + a'_t$ $E(8): Y_t = 0.37Y_{t-1} + 0.21Y_{t-2} + 0.63X_t - 0.23X_{t-1} - 0.13X_{t-2} + a'_t$	$U(a) = 14 < 29$ $\chi^2_{.05}(18) = 29$	$s(a') = 0.52$	0.73



MINISTÈRE DES RICHESSES NATURELLES DU QUÉBEC.

BASSIN VERSANT DES STATIONS DE JAUGEAGES DE LA BELL

ÉCHELLE

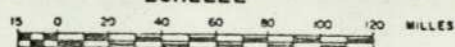


FIG. 2.18

Table 2.3

Rivière Mégiscane

DEBITS MOYENS MENSUELS

	JANVIER	FÉVRIER	MARS	AVRIL	MAI	JUIN	JUILLET	AOÛT	SEPTEMBRE	OCTOBRE	NOVEMBRE	DÉCEMBRE
1924	--	--	1.790E	2.730E	17.600	13.700	5.170	3.690	6.950	5.600	4.780	4.440E
1925	2.540E	2.090E	1.870E	5.730E	16.500	12.200	8.340	8.920	3.820	4.570	4.140	3.710E
1926	2.580E	1.940E	1.650E	1.610E	16.900E	14.900	9.790	5.500	3.010	3.010	6.560E	5.250E
1927	3.140E	2.190E	2.300E	4.800E	13.800	9.390	6.310	7.360	4.610	6.390	5.290E	3.820E
1928	2.800E	2.130	1.880	5.460E	26.500	16.800	11.200	6.450	7.200	10.600	9.750	4.650E
1929	3.300E	2.540	2.080	4.580E	18.400	15.200	9.690	5.050	5.760	5.480	5.910	4.140E
1930	2.820	2.170	2.100	5.610E	16.400	14.400	12.900	6.190	5.280	4.870	4.380	3.290E
1931	2.020E	1.600	1.380	5.020	8.570	5.460	4.460	4.530	2.980	4.050	6.050	4.810E
1932	3.390E	2.930E	2.010E	4.120E	9.390	5.530	9.700	8.070	8.460	19.100	9.880	4.510E
1933	3.200E	2.640E	2.240E	9.510E	18.200	8.200	4.250	2.630	1.770	2.220	3.480E	2.490E
1934	1.780	1.380	1.380	4.890	20.900	9.360	6.410	3.260	5.390	7.260	5.260	4.340E
1935	2.840	2.150	1.880	4.740	12.200	7.320	6.470	3.130	1.940	5.260	8.060	3.980E
1936	2.160E	1.600	2.350E	4.850E	28.300	15.800	5.760	5.490	4.270	7.840	8.700	5.030
1937	3.660	2.680E	2.260E	5.540E	16.000	7.260	3.770	2.410	3.690	7.620	10.200	6.120E
1938	3.300E	2.420E	2.890E	10.500E	18.500	5.710	4.460	8.500	8.650	6.540	5.690	3.780
1939	3.070	2.560	2.010	3.060E	17.500	9.840	9.850	5.350	3.760	5.700	6.220	3.500E
1940	2.390E	1.880E	1.540E	2.190E	17.700	16.100	7.240	3.620	5.900	4.010	4.800E	3.760E
1941	--	--	--	--	--	4.080	5.210	4.400	11.000	10.000	5.180	--
1942	--	--	--	--	--	--	--	2.960	2.650	6.370	6.490	--
1943	--	--	--	--	14.100	12.700	5.310	3.740	4.310	3.100	3.470	--
1944	--	--	--	--	12.700	8.320	3.650	3.300	2.930	5.050	5.750	--
1945	--	--	--	--	--	--	4.810	3.210	4.000	5.470	7.230	--
1946	--	--	--	--	--	--	--	--	3.680	4.920	5.550	--
1947	--	--	--	--	--	--	--	3.900	3.020	4.480	2.930	--
1948	--	--	--	--	--	5.070	3.010	3.750	2.560	2.300	2.820	--
1949	--	--	--	--	20.800	12.300	6.650	2.660	2.680	2.340	2.060E	--
1951	--	--	--	--	18.300	4.890	3.140	1.990	2.180	7.070	9.060	--
1952	--	--	--	--	15.200	9.120	7.840	9.030	5.210	4.300	4.840	--
1953	--	--	--	--	--	4.670	2.610	1.550	3.020	3.330	2.600	--
1954	--	--	--	--	12.700	8.460	4.450	4.320	3.350	6.340	5.630	5.220
1955	2.780	1.840	1.450	5.230E	9.640	4.750	3.690	2.490	2.080	4.440	5.480	3.580
1956	2.330	1.570	1.400	2.870E	12.300	11.000	6.510	5.680	8.510	7.040	4.190	3.040E
1957	2.280E	1.850E	1.630E	5.070	11.000	6.000	9.990	4.200	6.360	6.630	8.430	5.120E
1958	4.120E	2.760	2.270	4.510	7.620	6.990	6.680	4.490	6.970	8.140	5.520	3.350E
1959	2.070	1.560	1.340	3.570E	12.200	9.160	4.190	4.780	7.020	7.400	8.720	4.930E
1960	2.690E	2.400E	2.120E	5.610E	25.600	9.610	10.500	9.110	6.120	4.080	4.230	3.780
1961	2.250	1.620E	1.470E	3.560E	13.500	7.390	3.900	3.470	6.720	6.050	5.220	4.520
1962	2.800E	2.220E	1.790E	3.060E	14.300	9.140	3.200	2.180	2.750	3.250	2.470	1.870
1963	1.370	1.160	1.120E	3.330E	9.530	6.120	7.850	7.710	5.350	3.300	4.290	3.580E
1964	2.320	1.940	1.730	6.660	13.800	9.100	6.630	--	--	--	--	--
1965	2.560	1.770	1.400	1.620	16.200	8.450	4.350E	9.380	8.870	10.700	5.930	3.600

Table 2.4

Rivière Bell

Débits moyens mensuels

	JANVIER	FÉVRIER	MARS	AVRIL	MAI	JUIN	JUILLET	AOÛT	SEPTEMBRE	OCTOBRE	NOVEMBRE	DÉCEMBRE
1914	---	---	---	---	---	---	---	---	---	202	356	435
1915	325	219	204	715E	2,580	1,920	1,080	539	501	1,090	1,000	819
1916	382	234	205	1,120	4,010	2,130	853	451	406	487	1,000	1,050
1917	700	523	432	379	2,880	3,080	2,240	1,960	1,440	1,290	1,210	873
1918	459	434	316	723	2,530	2,550	1,540	1,100	1,240	1,300	1,560	1,280
1919	549	328	211	871	3,950	2,950	883	656	1,280	2,450	1,430	954
1920	422	342	295	1,480	2,610	1,430	804	724	541	601	678	653
1921	488	284	269	2,060	3,400	1,150	375	565	1,000	1,920	1,280	773
1922	439	244	198E	1,190E	3,100	1,810	1,980	1,290	1,210	1,090	1,320	714E
1923	380E	314E	257E	392	2,440E	1,800	797	447	752	833	774	642E
1924	598E	487E	394E	444E	2,730E	2,960E	1,470	1,030	1,540	1,400	1,030	1,100E
1925	728E	569E	546E	1,110E	2,750	2,400	1,810	1,850	903	936	860	783
1926	712	559	431	362	2,530	3,030	1,870	1,130	623	593	1,150	1,220
1927	754	539	488	883	2,190	2,080	1,340	1,280	936	1,290	1,340	998
1928	732	546	420	712	4,750	3,970	2,420	1,740	1,730	2,510	2,480	1,250
1929	834	605	489	644	3,020	2,780	2,450	1,130	1,100	1,310	1,370	905
1930	683	517	445	744	2,590	2,960	2,650	1,350	1,210	1,080	860	713
1931	524	395	311	705	1,460	1,320	799	882	538	682	1,030	956
1932	628	514	385	585	1,620	1,350	1,800	1,660	1,450	3,700	2,670	1,170
1933	771	647	521	1,540	3,170	1,970	990	812	529	555	779	611
1934	411	319	303	524	2,910	1,950	1,480	954	1,270	1,450	1,060	896
1935	602	451	385	658	1,630	1,520	1,220	821	512	966	1,420	961
1936	597	352	365	577	3,170	2,790	1,270	1,120	986	1,560	1,890	1,050
1937	779	601	481	630	2,620	1,650	934	821	893	1,420	1,990	1,290
1938	742	599	562	1,720	3,120	1,440	1,130	1,570	1,690	1,250	1,160	934
1939	721	550	455	516	2,690	2,570	1,730	1,080	730	1,150	1,370	937
1940	626	429	353	457	2,340	2,510	1,690	933	1,210	922	985	878
1941	671	547	466	1,380	2,510	1,140	1,080	1,160	2,120	2,420	2,490	1,660
1942	584	456	602	1,130	2,390	1,690	937	691	472	856	912	741
1943	516	401	325	420	2,080	2,530	1,540	974	1,230	1,050	1,090	777
1944	523	372	317	379	2,170	1,540	870	788	613	746	848	706
1945	456	332	472	1,910	1,660	1,850	991	738	607	1,100	1,420	904
1946	593	455	513	1,180	1,950	1,860	1,120	995	934	883	933	872
1947	454	457	376	620	3,500	4,660	1,670	857	596	750	545	445
1948	311	276	238	560	2,070	1,340	814	667	599	513	607	657
1949	530	460	412	1,270	2,960	2,280	1,580	753	752	758	725	626
1950	573	523	414	737	2,190	1,680	1,150	627	377	404	758	857
1951	585	378	338	1,950	2,460	1,350	1,130	867	707	1,330	1,390	1,540
1952	561	556	486	1,080	2,680	2,090	1,500	1,650	1,170	1,020	990	1,140
1953	825	687	731	2,460	1,810	1,100	694	458	436	514	582	695
1954	516	450	512	1,070	2,320	2,520	1,900	1,540	1,150	1,650	1,530	1,340
1955	815	601	524	1,330	1,860	1,240	800	555	470	892	1,310	1,070
1956	645	506	418	708	2,450	2,290	1,380	1,440	1,390	1,380	1,010	875
1957	643	497	359	619	1,510	1,210	1,530	948	1,190	1,270	1,640	1,370
1958	1,020	671	493	839	1,320	1,290	1,250	970	1,260	1,640	1,410	916
1959	422	479	386	751	2,440	1,860	1,020	876	1,140	1,210	1,550	1,170
1960	801	653	563	1,110	5,520	2,220	2,370	2,090	1,360	931	893	845
1961	568	352	344	677	1,810	1,700	1,060	781	1,120	1,240	1,070	1,240
1962	878	555	441	817	2,000	1,340	730	492	574	539	440	394
1963	325	275	251	675	1,340	1,310	1,450	1,170	1,240	781	794	807
1964	541	438	374	1,140E	2,380E	1,900	1,130	781	726	847	742E	579
1965	556E	524E	428E	406E	2,990	2,080	928	1,530	2,120	2,460	1,580	1,040
1966	453	522	485	880	1,960	2,010	1,570	1,940	1,310	1,400	1,500	2,200
1967	1,150	765	536	1,210	3,980	2,240	1,320	732	514	1,210	2,310	1,460
1968	816	538	416	1,680	1,970E	1,230E	1,320	1,410	1,010	1,120E	1,130	854

2.5.1 A Hydrological Example of a Transfer Function Model with Multiple Input and Output

In this section we give an application of the methods of transfer function estimation for systems with multiple input and output. The data for this example are the levels of four ground water wells L_1 , L_2 , L_4 , L_5 given in Table 2.12 p.2.76 and shown in Figure 2.19. The purpose of the study is to express the two lower levels L_4 and L_5 in terms of the upper levels L_1 and L_2 . (Isbister 1966)

An examination of Figure 2.19 shows that the fluctuations of the upper wells L_1 and L_2 lead the corresponding fluctuations

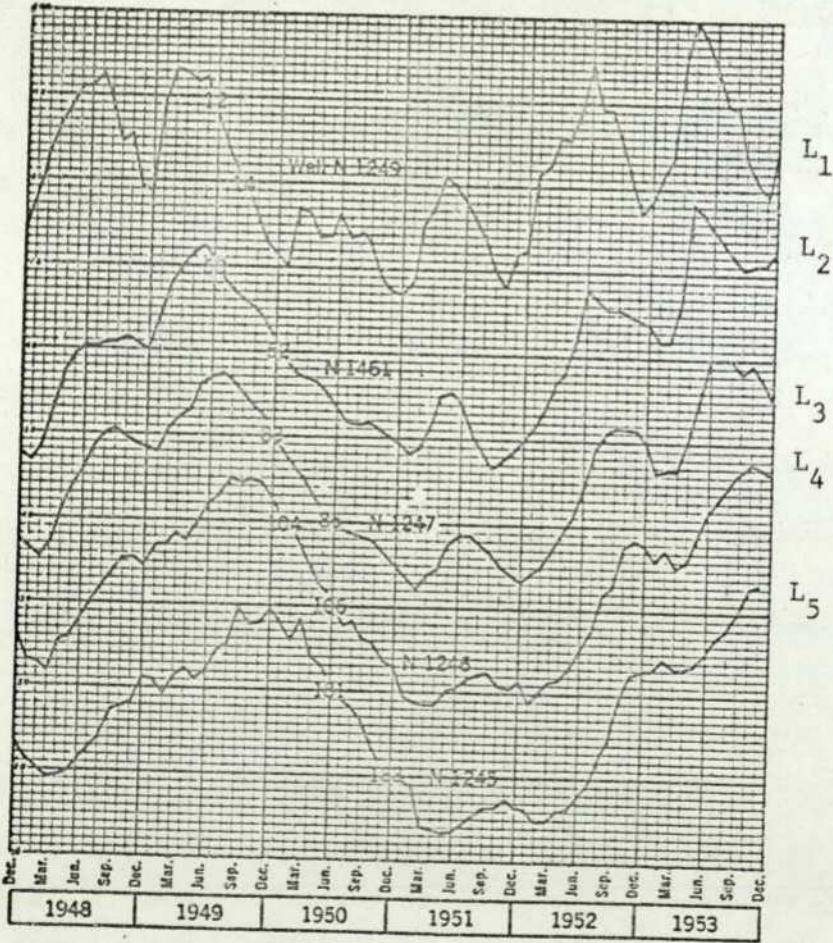


FIGURE 1

of the lower wells L_4 and L_5 and therefore we may consider the system shown in Figure 2.20.

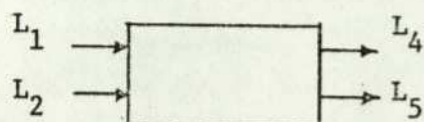


FIGURE 2.20

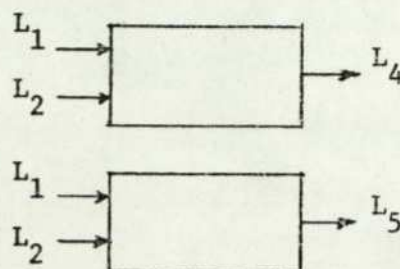


FIGURE 2.21

Before analyzing the multivariate system of Figure 2.20 we shall study the elementary properties and the statistical relationships of the four variables. The results of the preliminary computations are given in Table 2.5 which shows that the multiple regressions with L_1 and L_2 reduce the standard deviations of the lower levels L_4 and L_5 from 1.81 ft. and 1.78 ft. to 0.771 ft. and 0.983 ft. respectively. However, the residuals of both regressions are not random and therefore the relationships between the two pairs of levels must be analyzed using transfer function methods, the first step of which is the identification and estimation of univariate models for the four ground water series. The results of these computations are shown in Table 3 which shows that all fitted models are of the integrated moving average type and the standard deviations of the residuals are substantially smaller than the standard deviations of the original variables.

TABLE 2.5

STATISTICAL PROPERTIES OF GROUNDWATER VARIABLES

VARIABLES	STATISTICAL PROPERTIES																																	
LEVEL L_1	<u>MEAN</u> = 13.64 ft; <u>S. DEVIATION</u> = 1.65 ft																																	
L_2	51.94	1.69																																
L_4	105.62	1.81																																
L_5	181.58	1.78																																
L_1, L_2, L_4, L_5	<table><tr><td></td><td>L_1</td><td>L_2</td><td>L_4</td><td>L_5</td></tr><tr><td><u>CORRELATION MATRIX</u></td><td>L_1</td><td>L_2</td><td>L_4</td><td>L_5</td></tr><tr><td></td><td>1.00</td><td>0.719</td><td>0.280</td><td>0.117</td></tr><tr><td></td><td></td><td>1.00</td><td>0.800</td><td>0.661</td></tr><tr><td></td><td></td><td></td><td>1.00</td><td>0.949</td></tr><tr><td></td><td></td><td></td><td></td><td>1.00</td></tr></table>					L_1	L_2	L_4	L_5	<u>CORRELATION MATRIX</u>	L_1	L_2	L_4	L_5		1.00	0.719	0.280	0.117			1.00	0.800	0.661				1.00	0.949					1.00
	L_1	L_2	L_4	L_5																														
<u>CORRELATION MATRIX</u>	L_1	L_2	L_4	L_5																														
	1.00	0.719	0.280	0.117																														
		1.00	0.800	0.661																														
			1.00	0.949																														
				1.00																														
$(L_1, L_2), L_4$	<u>MULTIPLE REGRESSION</u> $(L_1, L_2) \rightarrow L_4$ $R = 0.907, r_{41.2} = -0.711, r_{42.1} = 0.898$ $s_{4.12} = 0.771$ ft																																	
$(L_1, L_2), L_5$	<u>MULTIPLE REGRESSION</u> $(L_1, L_2) \rightarrow L_5$ $R = 0.839, r_{51.2} = -0.687, r_{52.1} = 0.836$ $s_{5.12} = 0.983$ ft.																																	

One way of dealing with the multiple input and output system of Fig.2.20 would be to consider the two systems shown in Fig.2.21 following the procedure described in section 2.3.1.

In this example we shall use the canonical correlation method of section 2.3.2 which takes into account the multivariate nature of the system of Figure 2.20.

The canonical transformation can be applied either to the residuals of the univariate models or to the original variables. In this example we follow the second approach. The results of the canonical correlation computations are shown in Table 2.7 together with the results of the regressions between the pairs of variables (ξ_1, ξ_4) and (ξ_2, ξ_5) . The pairs of canonical variables (ξ_1, ξ_4) and (ξ_2, ξ_5) are plotted in Figs.2.22 & 2.23. An examination of these figures shows that the first pair of canonical variables tends to represent the long-term fluctuations of water supply whereas the second pair represents the seasonal variations. Consequently the canonical transformation of the original variables is useful not only because it simplifies the transfer function computations but also because it contributes to the understanding of the fluctuations of the groundwater levels. The next two steps are the estimation of the univariate models of the canonical variables $\xi_1, \xi_2, \xi_4, \xi_5$ and the estimation of the transfer functions of the two pairs of canonical variables. The fitted univariate models and the transfer functions are shown in

Tables 2.8 and 2.9 respectively. The standard deviations of the residuals shown in Table 2.9 indicate that the transfer function model between ξ_1 and ξ_4 is only a slight improvement over the univariate model of ξ_4 given in Table 2.7, whereas the coefficients of the transfer function model between ξ_2 and ξ_5 are not significant.

An examination of Figure 2.19 shows that the random fluctuations of groundwater levels in this example are relatively small. This suggests that the problem could also be approached from the point of view of deterministic transfer function estimation or equivalently deterministic deconvolution. It is well known that in the presence of noise, it is preferable to use the least squares version of deconvolution, which can be described briefly as follows:

Consider a linear system with input $\underline{x}' = (x_1, \dots, x_m)'$ and impulse response $\underline{v}' = (v_1, \dots, v_n)$ where $m \geq n+1$. The output $\underline{y}$ is given by the matrix equation

$$\underline{x}\underline{v} = \underline{y}$$

where $\underline{x}$ is the following $(m + n - 1) \times (n)$ matrix.

TABLE 2.6

UNIVARIATE MODELLING OF WATER LEVEL SERIES

VARIABLE	M O D E L		STANDARD DEVIATION OF RESIDUALS
	TYPE	EQUATION	
L_1	IMA(6)	$L_{1t} = L_{1(t-1)} + a_t + 0.274a_{t-1} - 0.316a_{t-6}$	0.712 ft
L_2	IMA(1)	$L_{2t} = L_{2(t-1)} + a_t + 0.480a_{t-1}$	0.437 ft
L_4	IMA(2)	$L_{4t} = L_{4(t-1)} + a_t + 0.371a_{t-2}$	0.320 ft
L_5	IMA(1)	$L_{5t} = L_{5(t-1)} + a_t + 0.166a_{t-1}$	0.386 ft

TABLE 2.7

RESULTS OF CANONICAL CORRELATION

CANONICAL VARIABLE	EQUATION	MEAN	STANDARD DEVIATION	CANONICAL CORRELATION
ξ_1	$-\dot{0}.3670Z_1 + 0.8529Z_2$	0	0.640	0.916
ξ_4	$0.9447Z_4 - 0.3278Z_5$	0	0.640	
ξ_2	$-0.3129Z_1 + 0.1309Z_2$	0	0.237	0.452
ξ_5	$-0.6682Z_4 + 0.7440Z_5$	0	0.237	
ξ_1, ξ_4	$\hat{\xi}_4 = 0.916 \xi_1$		$s_{4.1} = 0.26$	
ξ_2, ξ_5	$\hat{\xi}_5 = 0.452 \xi_2$		$s_{5.2} = 0.21$	

NOTE: Z_1, Z_2, Z_3, Z_4 are standardized variables given by the formula

$$Z_i = \frac{L_i - \bar{L}_i}{s_i}$$

TABLE 2.8

TRANSFER FUNCTION MODELLING OF CANONICAL VARIABLES

VARIABLES	M O D E L		STANDARD DEVIATION OF RESIDUALS
	TYPE	EQUATION	
ξ_1	IMA(6)	$\xi_{1t} = \xi_{1(t-1)} + a_t + 0.241a_{t-6}$	0.019
		FIRST DIFFERENCES	0.020
ξ_2	IMA(6)	$\xi_{2t} = \xi_{2(t-1)} + a_t + 0.241a_{t-6}$	0.012
		FIRST DIFFERENCES	0.013
ξ_4	IMA(2)	$\xi_{4t} = \xi_{4(t-1)} + a_t + 0.311a_{t-2}$	0.015
		FIRST DIFFERENCES	0.016
ξ_5	IMA(2)	$\xi_{5t} = \xi_{5(t-1)} + a_t + 0.310a_{t-2}$	0.014
		FIRST DIFFERENCES	0.015

TABLE 2.9

TRANSFER FUNCTION MODELLING OF CANONICAL VARIABLES

VARIABLES	MODEL	STANDARD DEVIATION OF RESIDUALS
ξ_1, ξ_4	$\Delta \xi_{4t} = 0.559 \Delta \xi_{4(t-2)} + 0.241 \Delta \xi_{1t}$ $+ 0.187 \Delta \xi_{1(t-1)} + a_t$ $+ 1.365 a_{t-1} + 0.451 a_{t-2}$ FIRST DIFFERENCE $\Delta \xi_4$	0.013 0.016
ξ_2, ξ_5	(TRANSFER FUNCTION COEFFICIENTS NOT SIGNIFICANT) FIRST DIFFERENCE	0.015

TABLE 2.10DETERMINISTIC TRANSFER FUNCTION (DECONVOLUTION)

LAG OF L_4	IMPULSE RESPONSE WEIGHT	t VALUE
0	-0.211	-2.88
1	0.043	0.40
2	0.213	2.04
3	0.056	0.54
4	0.187	1.86
5	0.189	1.93
6	0.083	0.87
7	0.174	1.81
8	0.152	1.53
9	0.179	1.70
10	0.080	0.75
11	0.006	0.06
12	0.091	0.82
13	0.273	3.73
MULTIPLE CORRELATION COEFFICIENT R	0.986	
STANDARD DEVIATION OF RESIDUALS $S_{4.1}$	0.317 ft	

TABLE 2.11
ANALYSIS OF GROUNDWATER LEVEL L₄

ANALYSIS	SOURCE OF VARIATION	STANDARD DEVIATION	SUM OF SQUARES SS	RATIO TO INITIAL SS
Univariate	Original observations	1.81 ft	222.00	1.000
	First differences		178.50	0.804
	Univariate model		215.10	0.966
	Residuals	0.32	6.90	0.034
Deterministic transfer function	Original observations	1.81	222.00	1.000
	Transfer function		215.10	0.966
	Residuals	0.32		0.034

$$x = \begin{bmatrix} x_1 & 0 & 0 & 0 \\ x_2 & x_1 & 0 & 0 \\ : & x_2 & x_1 & 0 \\ : & : & x_2 & x_1 \\ x_m & : & : & x_2 \\ 0 & x_m & : & : \\ 0 & 0 & x_m & : \\ 0 & 0 & 0 & x_m \end{bmatrix}$$

The transfer function or impulse response weights are given by the equation

$$\underline{y} = (x'x)^{-1} x'y$$

i.e they are the multiple regression coefficients of the output y on the "independent variables" given by the columns of the matrix x .

This method was applied to the relation between L_1 and L_4 and the results are shown in Tables 7 and 8. The latter table also shows the results of the univariate modelling of the water level L_4 . The final results of the computations are also displayed in Figure 6 which shows the decomposition of the sum of squares of the variable L_4 according to the method described in section 2.4.1. This figure indicates that the analysis of L_4 can proceed in two different ways:

(1) Univariate modelling

Taking the first differences ΔL_4 reduces the sum of squares to 19.6% of the original value. The univariate model of ΔL_4 reduces the sum of squares to 3.4% of the original value but it is not very useful for forecasting. Moreover, the small size of the sum of squares indicates that we are near the base of the SS-triangle of section 2.4.1 and therefore it is appropriate to estimate the impulse response weights of the system by deterministic methods.

(2) Deterministic deconvolution

The results of this computation are shown in Table 2.10 & Fig. 2.24 which indicate a further slight reduction in the sum of squares of residuals.

The various approaches examined in this example lead to the following conclusions:

1. The boundaries between the stochastic and deterministic methodologies for transfer function estimation are not well defined.
2. The choice between univariate and transfer function modelling depends on the relative importance of understanding the nature of the series or forecasting its future.

Year	L ₁	L ₂	L ₃	L ₄	L ₅
1948	14.400	54.600	85.700	107.400	182.700
	13.500	54.200	85.900	107.400	182.900
	12.800	53.400	85.600	107.600	183.700
	12.300	52.800	84.800	106.900	183.100
	11.800	52.200	84.200	106.800	183.000
	11.800	51.900	83.800	106.400	182.700
	11.500	51.900	83.300	106.000	182.500
	12.100	51.800	83.000	105.600	182.200
	13.100	51.800	82.900	105.200	181.500
	12.900	51.700	83.100	104.900	181.400
	14.100	51.900	83.200	104.900	181.300
	14.200	52.000	83.300	105.100	180.800
1949	12.100	51.200	83.400	104.600	180.300
	11.200	50.400	82.800	104.600	181.100
	11.300	50.000	82.600	104.300	180.700
	11.400	49.700	82.400	104.500	180.500
	11.300	49.500	81.300	104.100	180.200
	12.400	49.800	81.700	103.600	180.600
	13.100	50.200	81.600	103.400	180.200
	13.700	50.600	81.700	103.000	180.000
	14.200	50.800	82.000	103.100	179.000
	14.800	51.000	82.300	103.000	179.500
	15.400	51.300	82.500	103.100	179.100
	15.700	51.600	83.000	103.500	179.100
1950	16.000	52.100	83.400	103.900	179.400
	14.600	52.500	83.700	104.200	179.800
	14.700	52.600	84.000	104.700	179.300
	15.200	52.700	84.400	105.100	180.200
	15.200	53.000	84.800	105.600	180.400
	14.700	53.300	85.000	105.900	180.800
	15.300	53.700	85.200	106.400	181.200
	15.200	53.700	85.300	106.300	181.400
	15.500	53.600	85.400	106.800	181.600
	16.200	53.800	85.500	106.800	182.200
	16.500	54.000	85.800	107.300	182.700
	16.600	54.100	86.000	107.400	183.000
1951	16.300	54.400	86.300	108.100	183.100
	15.000	54.200	86.600	108.200	183.200
	14.500	53.800	86.200	108.300	184.200
	13.800	53.000	86.100	108.300	184.300
	14.000	52.900	85.500	108.000	184.400
	14.400	53.200	85.300	107.900	184.300
	14.800	53.800	85.300	107.700	184.100
	15.200	54.200	85.500	107.600	183.900
	16.000	54.700	85.700	107.500	183.700
	16.400	54.500	86.000	107.800	183.700
	15.600	54.300	86.200	107.900	183.500
	15.500	54.000	86.300	107.700	183.700
1952	13.700	53.700	86.100	108.200	183.800
	13.500	53.300	86.000	108.000	184.000
	12.800	52.600	85.600	107.700	184.000
	12.900	52.400	85.200	107.700	183.800
	12.300	51.600	84.800	107.400	183.800
	11.000	50.400	84.200	106.900	183.400
	12.100	50.700	83.200	106.500	183.100
	12.200	50.900	82.900	105.600	182.500
	13.000	50.900	82.700	105.400	182.000
	13.900	51.000	82.700	104.500	181.000
	14.600	51.100	82.700	104.300	180.500
	14.300	51.200	83.000	104.500	180.400
1953	13.700	51.600	83.700	104.800	180.400
	13.300	51.600	83.700	104.500	180.100
	10.800	50.700	83.600	104.900	180.300
	10.000	48.300	82.900	104.800	180.400
	10.500	48.600	82.000	104.100	180.200
	11.000	49.000	81.000	103.600	179.900
	12.000	49.200	81.000	103.300	179.600
	12.100	49.600	81.000	103.000	179.400
	13.200	49.900	81.300	102.700	179.000
	13.800	49.800	81.100	102.500	178.400
	14.100	49.800	81.500	102.600	178.300
	13.000	49.500	82.000	102.700	178.400

TABLE 1. GROUNDWATER LEVELS

ND

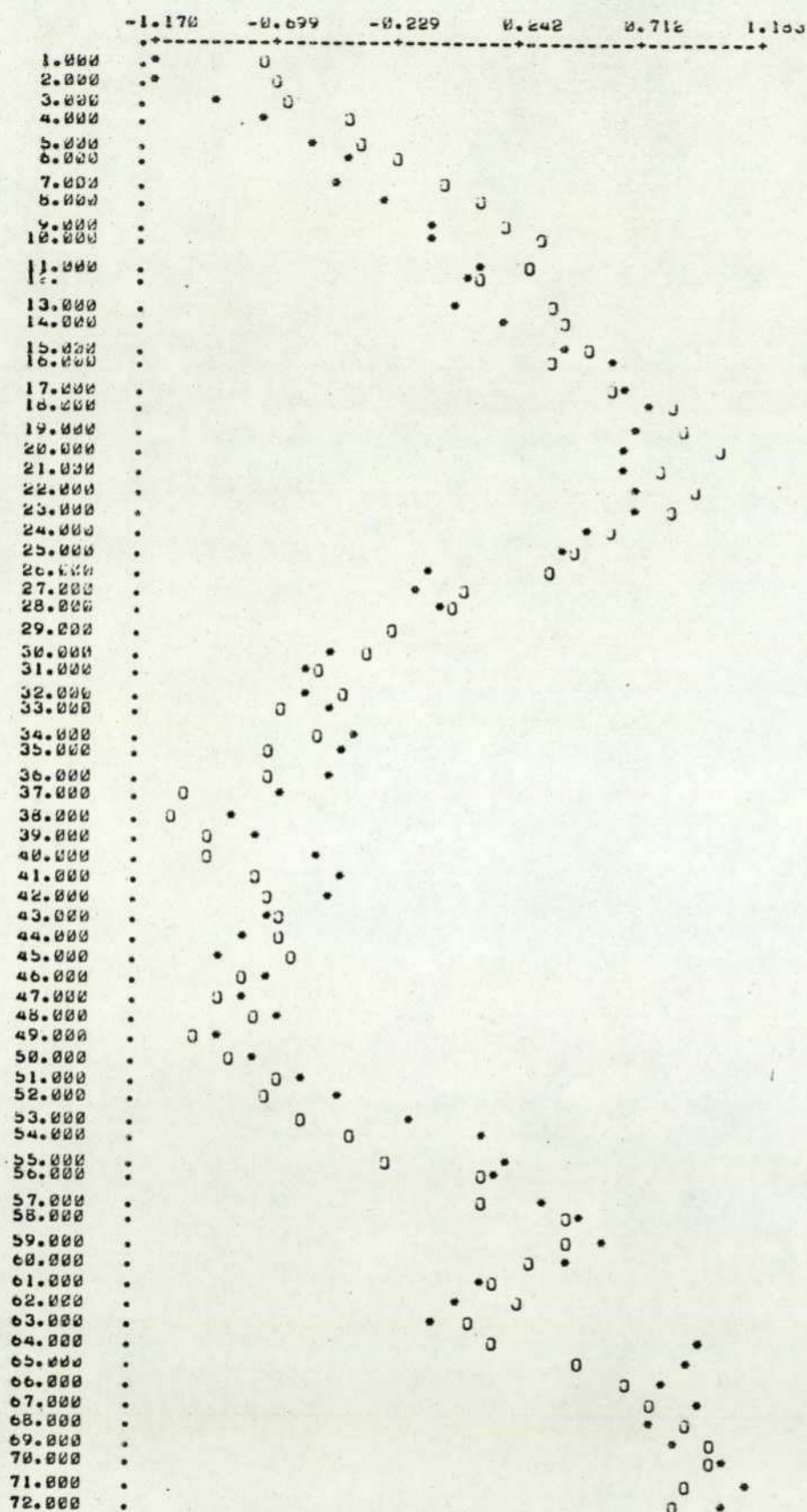


FIG. 2.22

* = ξ_1 0 = ξ_4

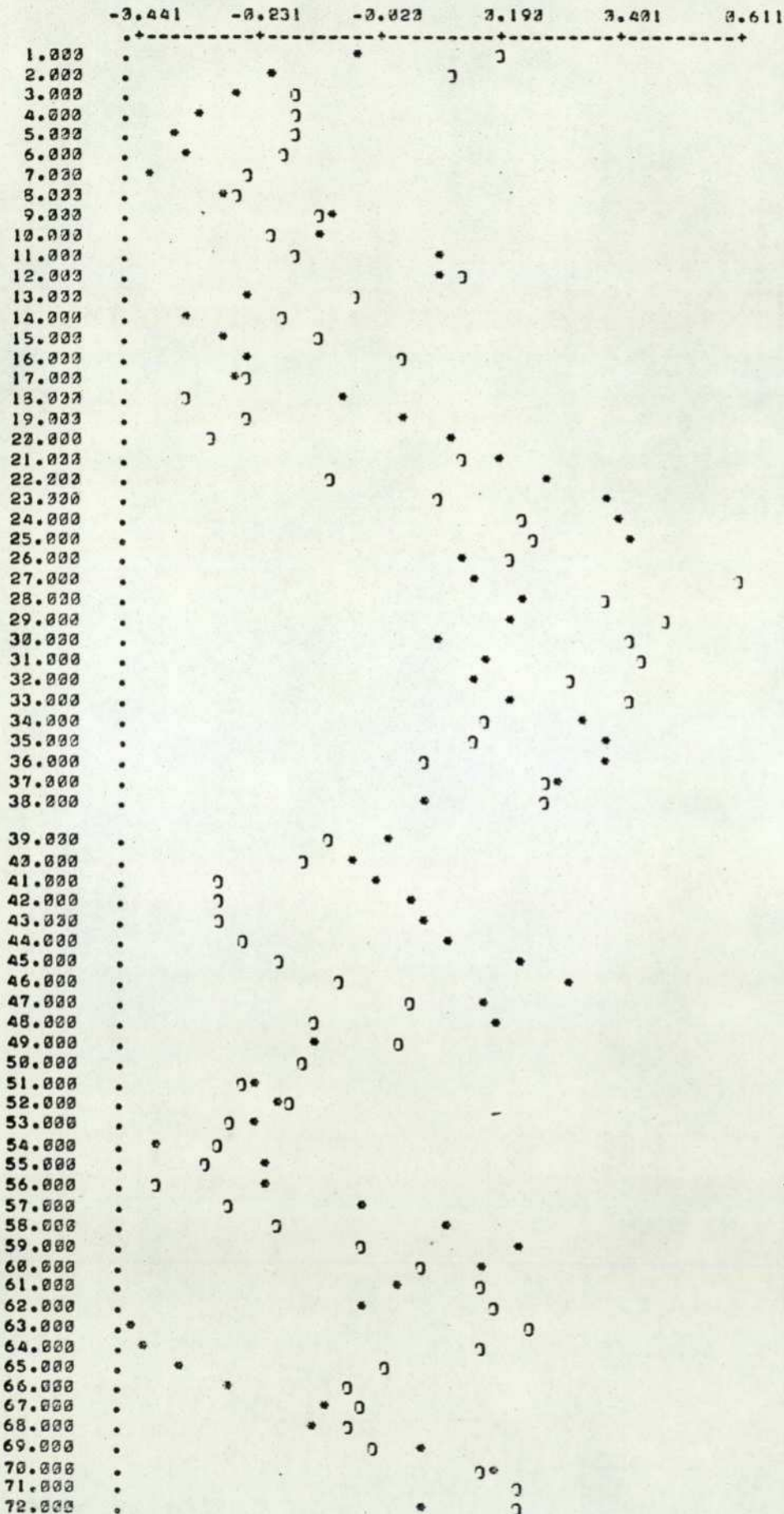
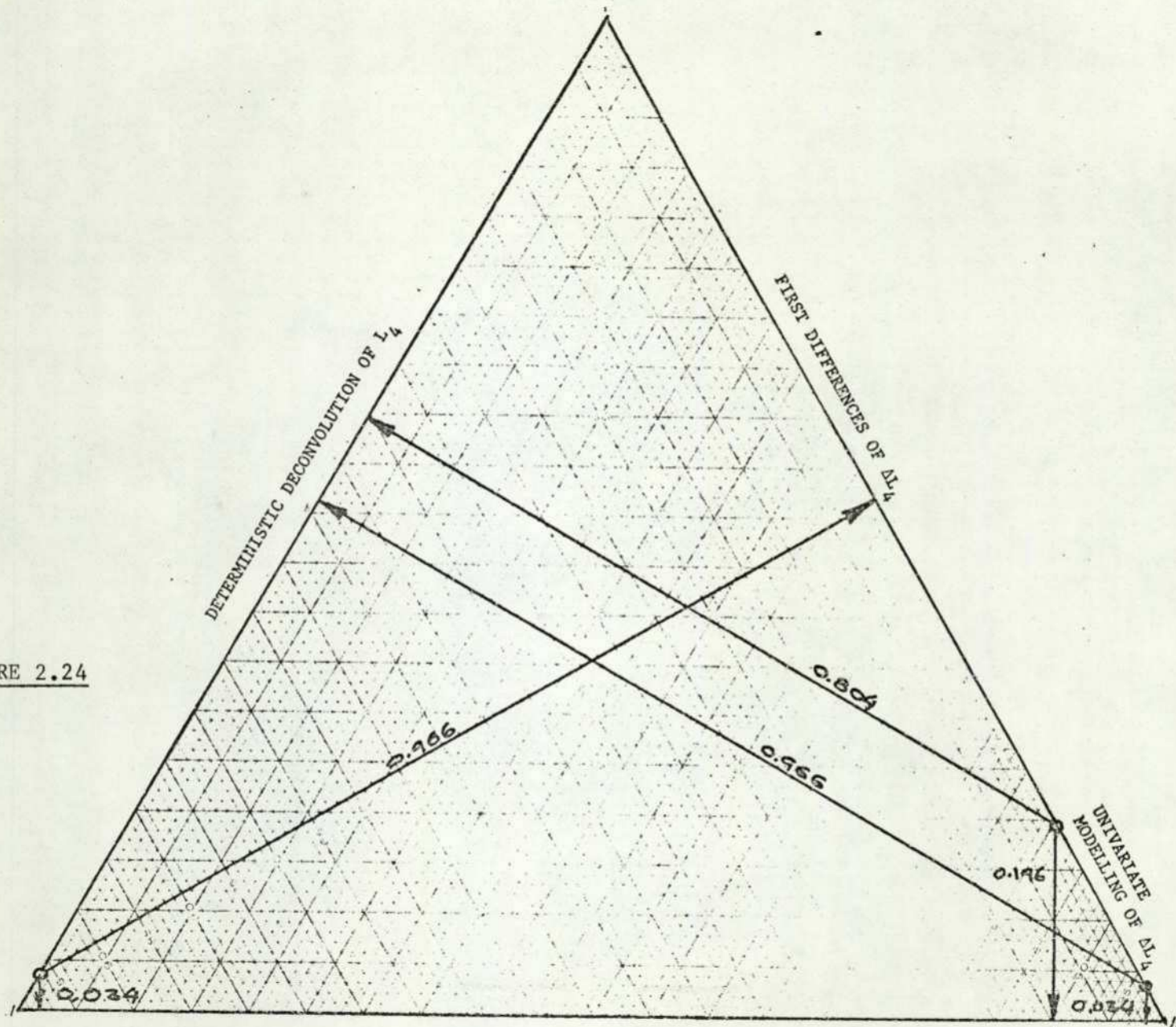


FIG. 2.23

FIGURE 2.24



Appendix 2.1

Estimation of ARIMA Models

Appendix 2.1

Estimation of ARIMA models

The first step in fitting an ARIMA model to a time series is to identify the process using the correlogram of the observations. The next step is to estimate the parameters of the process using iterative least squares, in the following way:

Consider the ARIMA process

$$(2.40) y_t = \phi y_{t-1} + a_t - \theta a_{t-1}$$

where we assume $E(y_t) = 0$

For a given choice of $\phi = \phi_0$ and $\theta = \theta_0$

we can compute the residuals a_t recursively:

$$(2.41) a_t(\phi_0, \theta_0) = a_{t,0} = y_t - \phi_0 y_{t-1} + \theta_0 a_{t-1,0}$$

The optimal values of the parameters ϕ, θ are estimated by minimising the sum of squares $\sum_t a_t^2$ (Fig.2.25). These optimal values can also be obtained by the following iterative computation:

Step 1 Expand $a_t = f(\phi, \theta)$ in a Taylor series about its value at the point (ϕ_0, θ_0) retaining only linear terms:

$$(2.42) a_t = a_{t,0} + (\phi - \phi_0) \left(\frac{\partial a_t}{\partial \phi} \right) + (\theta - \theta_0) \left(\frac{\partial a_t}{\partial \theta} \right)$$

This equation can be written

$$(2.43) \quad a_{t,o} = (\phi - \phi_o) \left(-\frac{\partial a_t}{\partial \phi} \right) + (\theta - \theta_o) \left(-\frac{\partial a_t}{\partial \theta} \right) + a_t$$

Consequently, to minimise $\sum_t a_t^2$ we regress $a_{t,o}$ on

$$\left(-\frac{\partial a_t}{\partial \phi} \right) = y_{t-1} \text{ and } \left(-\frac{\partial a_t}{\partial \theta} \right) = -a_{t-1,o} \text{ so the equation becomes:}$$

$$(2.44) \quad a_{t,o} = (\phi - \phi_o)(-a_{t-1,o}) + (\theta - \theta_o)y_{t-1} + a_t$$

The new values are $\phi_1 = \phi_o + (\phi - \phi_o)$ and $\theta_1 = \theta_o + (\theta - \theta_o)$, where

$(\phi - \phi_o)$, $(\theta - \theta_o)$ are the regression coefficients of the above equation.

Step 2 Repeat the calculation with the adjusted values ϕ_1 , and θ_1 and proceed with further s iterations until the corrections $(\phi - \phi_s)$, $(\theta - \theta_s)$ remain below a preassigned level.

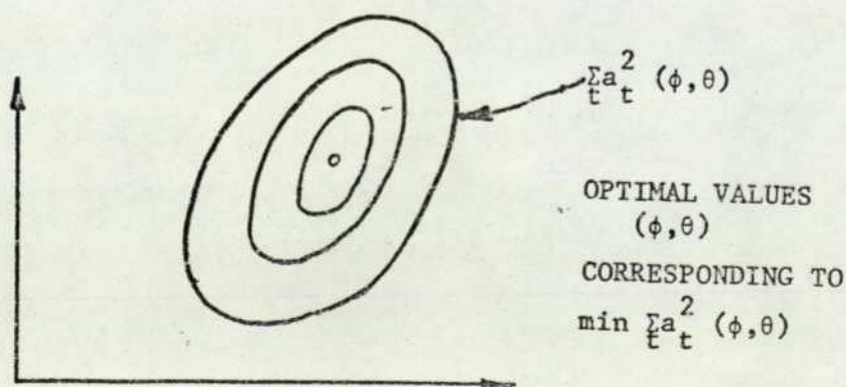


FIGURE 2.25

Appendix 2.2

An Outline of Canonical Correlation

Appendix 2.2

An Outline of Canonical Correlation

Assume that we have n simultaneous observations of two sets of variables (u_1, u_2) and (u'_1, u'_2) and we are interested in the relationship between the two sets. For example we may have a new pair of values of the variables (u_1, u_2) and we want to estimate the corresponding values of the variables (u'_1, u'_2) . Figs. 2.26 and 2.27 show the two scatter diagrams in the planes (u_1, u_2) and (u'_1, u'_2) with the corresponding points labelled with the serial numbers of the observations.

An examination of these figures shows that the two patterns of n points are similar but they have different orientations. Given an additional $(n + 1)^{\text{th}}$ point in the plane (u_1, u_2) , we expect intuitively that the corresponding $(n + 1)^{\text{th}}$ point in the plane (u'_1, u'_2) will have a position within the pattern of Fig. 2.27 that resembles the position of the $(n + 1)^{\text{th}}$ point in the plane (u_1, u_2) . It is however difficult to predict the coordinates (u'_1, u'_2) of this point from the corresponding coordinates (u_1, u_2) , due to the different orientations of the patterns of points in the two planes.

One way to avoid this difficulty is to select two new pairs of coordinates (ξ_1, ξ_2) , (ξ'_1, ξ'_2) which will make the

similarity of the two patterns more apparent. Figures 2.28 and 2.29 show two new pairs of coordinates that seem to satisfy this requirement, re-oriented so as to make the similarity of the two patterns more obvious. Given the position of the $(n + 1)^{\text{th}}$ point in the plane (ξ_1, ξ_2) , it is now easier to estimate the position of the corresponding point in the plane (ξ'_1, ξ'_2) .

The problem of defining mathematically the desirable properties of the two new pairs of coordinates and of determining the coefficients of the corresponding linear transformations was formulated and solved by Hotelling (1935) who called the new variables (ξ_1, ξ_2) , (ξ'_1, ξ'_2) "canonical variables" and the method "canonical correlation".

Assume that we have a set of p variables $(u_1, \dots, u_p)$ and a second set of q variables $(u'_1, \dots, u'_q)$ where $p \geq q$. We want to estimate the coefficients of the equations

$$(2.45) \quad \xi_i = a_{i1} u_1 + \dots + a_{ip} u_p \quad i = 1, 2, \dots, p$$

$$(2.46) \quad \xi'_i = a'_{i1} u'_1 + \dots + a'_{iq} u'_q \quad i = 1, 2, \dots, q$$

from the following set of conditions (Kendall 1961):

1. All ξ and ξ' have zero mean and unit variance.
2. Any ξ_i is independent of any ξ_j and any ξ'_i is independent of any ξ'_j where $i \neq j$.
3. The correlation coefficients between any ξ_i and any ξ'_j are zero except for q correlations $r_1, \dots, r_q$ between the pairs $(\xi_1, \xi'_1), (\xi_2, \xi'_2), \dots, (\xi_q, \xi'_q)$.

The computation of the coefficients of ξ_i and ξ'_i is based on the partitioned correlation matrix of the original variables.

$$(2.47) \quad R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

where:

R_{11} = Intercorrelations among the p variables u_i

R_{12} = Intercorrelations between the variables in the first and second sets

R_{21} = Transpose of R_{12}

R_{22} = Intercorrelations among the q variables u'_i

The computation of the coefficients of the canonical variables

ξ_i and ξ'_i starts by finding the characteristic roots

$\lambda_1, \dots, \lambda_q$ of the $(q \times q)$ matrix $R_{22}^{-1} R_{21} R_{11}^{-1} R_{12}$ i.e. by

solving the equation

$$(2.48) \quad |R_{22}^{-1} R_{21} R_{11}^{-1} R_{12} - \lambda I| = 0$$

The coefficients $a'_{i1}, \dots, a'_{iq}$ of equation 2.46 for $i = 1, 2, \dots, q$

are the corresponding characteristic vectors of R and are computed

by solving the following system of homogeneous equations

$$(2.49) \quad (R_{22}^{-1} R_{21} R_{11}^{-1} R_{12} - \lambda_i I) \begin{pmatrix} a'_{i1} \\ \vdots \\ a'_{iq} \end{pmatrix} = 0$$

subject to the normalising conditions $\sum_{i=1}^q a_i'^2 = 1$.

The corresponding coefficients $a_{i1}, \dots, a_{ip}$ of equation (2.45), for $i = 1, 2, \dots, q$ are computed from the equation

$$(2.50) \quad \begin{pmatrix} a_{i1} \\ \vdots \\ a_{ip} \end{pmatrix} = \frac{1}{\sqrt{\lambda_i}} (R_{11}^{-1} R_{12} \begin{pmatrix} a'_{i1} \\ \vdots \\ a'_{iq} \end{pmatrix})$$

where again $\sum_{i=1}^q a_i^2 = 1$

The square root of the largest characteristic root $\lambda_{\max}$ can be interpreted as the maximum possible correlation between any two linear combinations of the two sets of variables (u_1, u_2) and (u'_1, u'_2) . The following numerical example will clarify the concepts of canonical correlation.

The values of the variables (u_1, u_2) , (u'_1, u'_2) are shown in table 2.13 and the standardised values of these variables and the canonical variables (ξ_1, ξ_2) , (ξ', ξ') are given in tables 2.14 and 2.15.

TABLE 2.13

Point	u_1	u_2	u'_1	u'_2
1	4	7	8	11
2	6	9	11	10
3	9	9	12	8
4	11	12	15	7
5	14	13	17	5
6	11	7	10	6
7	8	5	8	7
8	6	5	7	9
9	4	3	4	10
10	2	4	4	12
Mean	7.50	7.40	9.60	8.50
Standard Deviation	3.78	3.34	4.30	2.27

TABLE 2.14

Point	u_1	u_2	u'_1	u'_2
0001	-0.926	-0.120	-0.372	1.101
0002	-0.397	0.479	0.326	0.661
0003	0.397	0.479	0.558	-0.220
0004	0.926	1.377	1.256	-0.661
0005	1.720	1.677	1.721	-1.542
0006	0.926	-0.120	0.093	-1.101
0007	0.132	-0.719	-0.372	-0.661
0008	-0.397	-0.651	-0.605	0.220
0009	-0.926	-1.317	-1.302	0.661
0010	-1.455	-1.018	-1.302	1.542

TABLE 2.15

Point	ξ_1	ξ'_1	ξ_2	ξ'_2
0001	-0.380	-0.387	1.296	1.223
0002	0.233	0.317	1.318	1.341
0003	0.482	0.561	0.059	0.290
0004	1.317	1.265	0.510	0.407
0005	1.789	1.742	-0.317	-0.385
0006	0.201	0.108	-1.641	-1.530
0007	-0.495	-0.363	-1.242	-1.391
0008	-0.610	-0.608	-0.306	-0.342
0009	-1.272	-1.311	-0.425	-0.459
0010	-1.215	-1.323	0.844	0.847

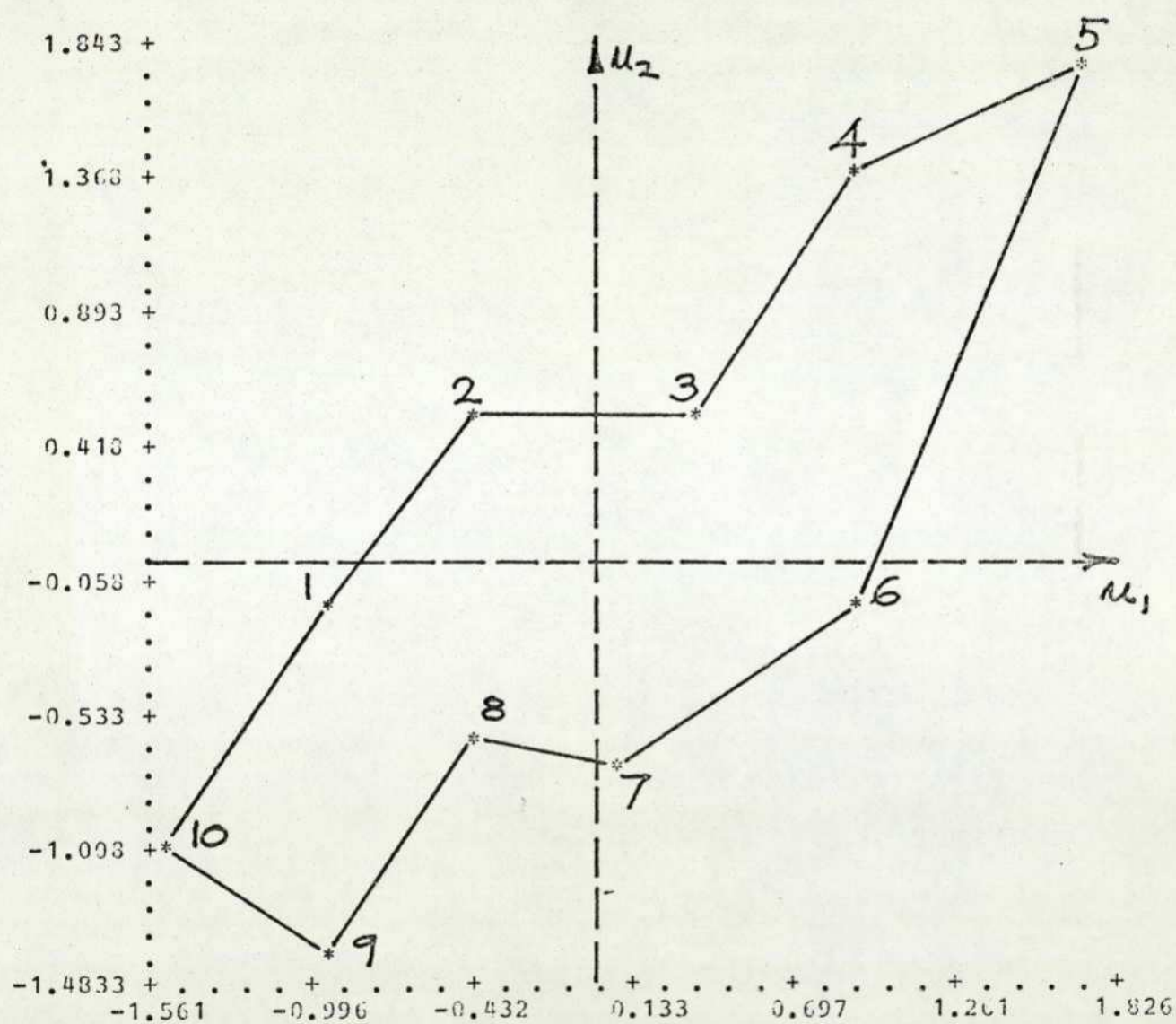


FIGURE 2.26

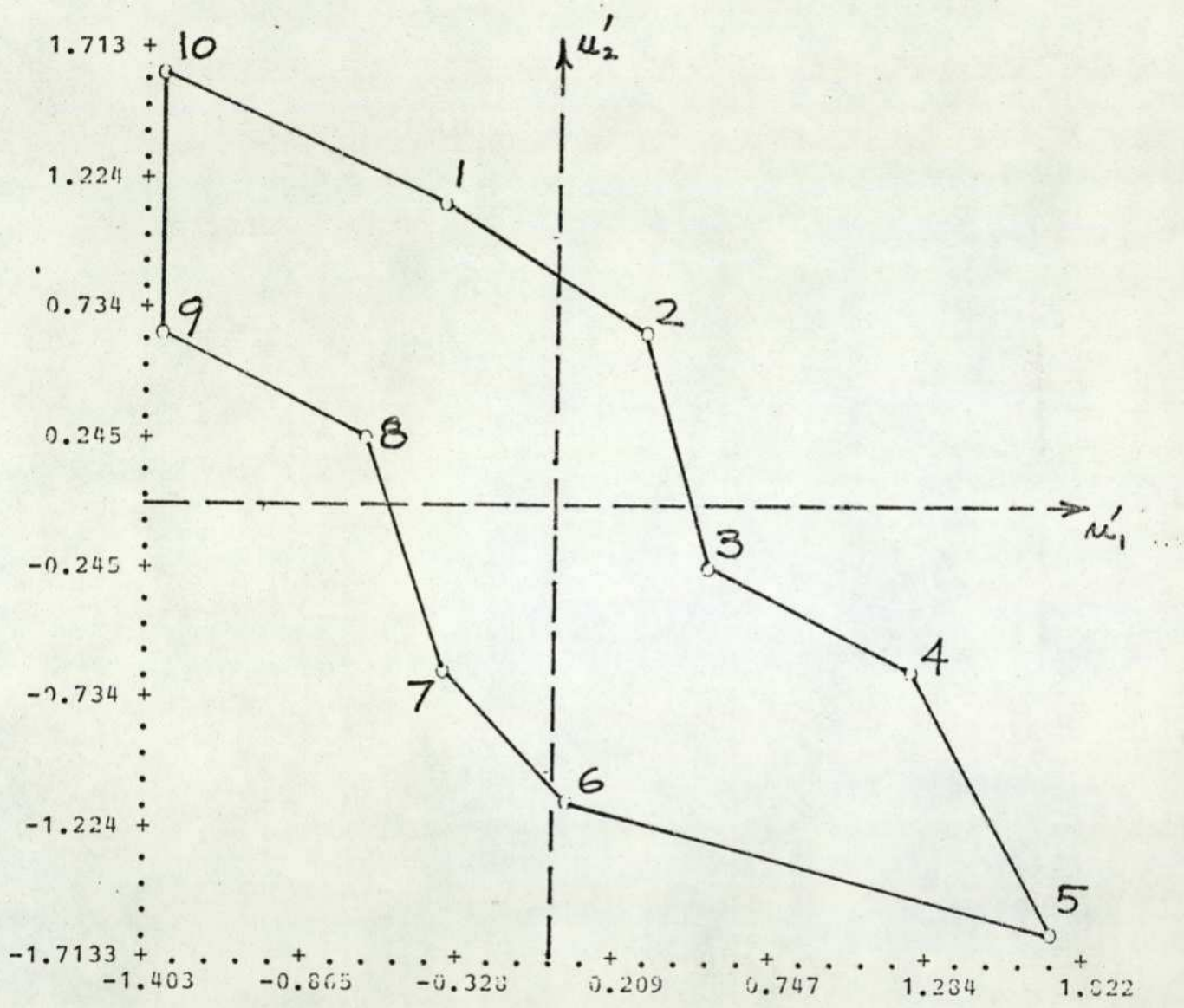


FIGURE 2.27

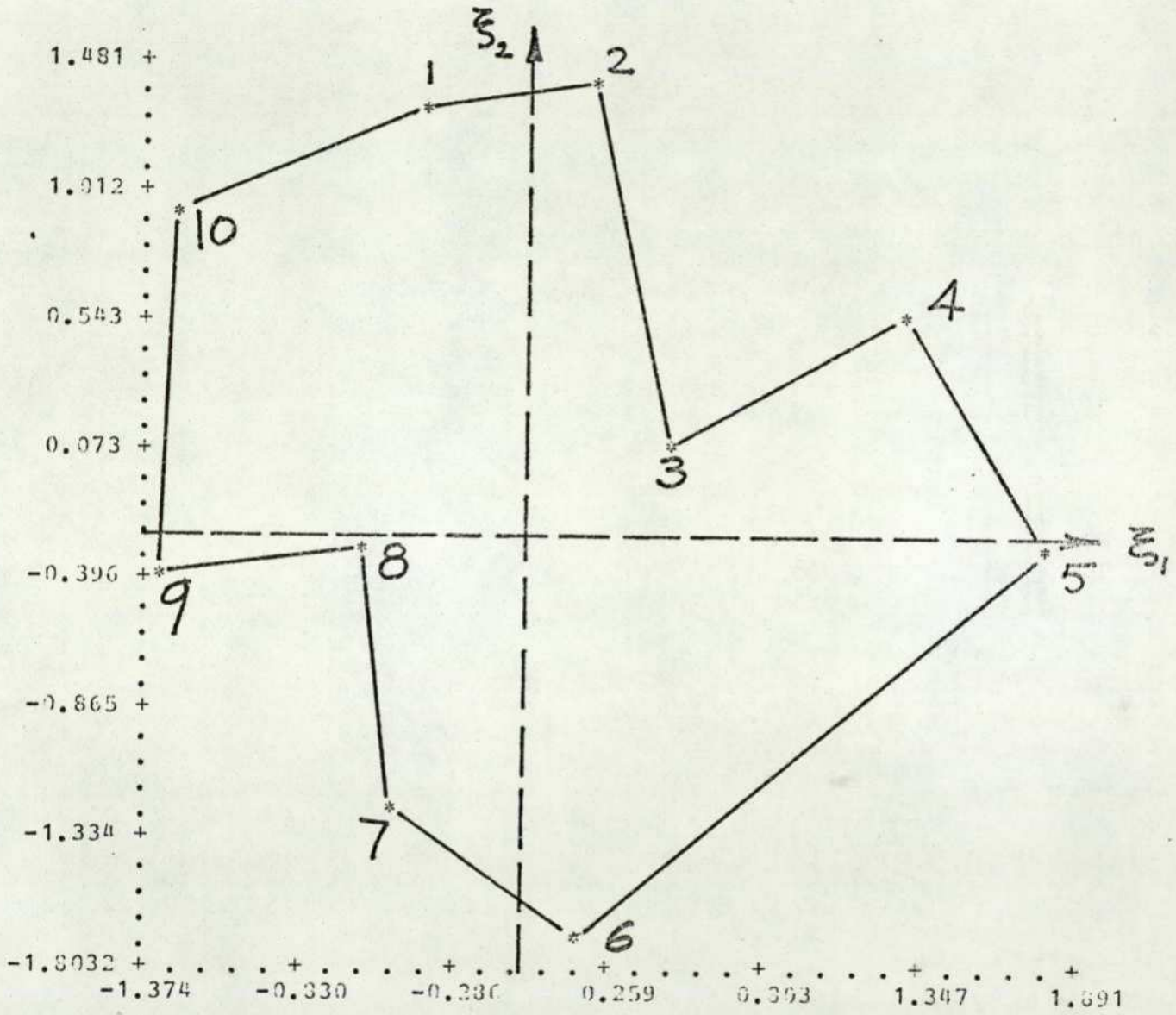


FIGURE 2.28

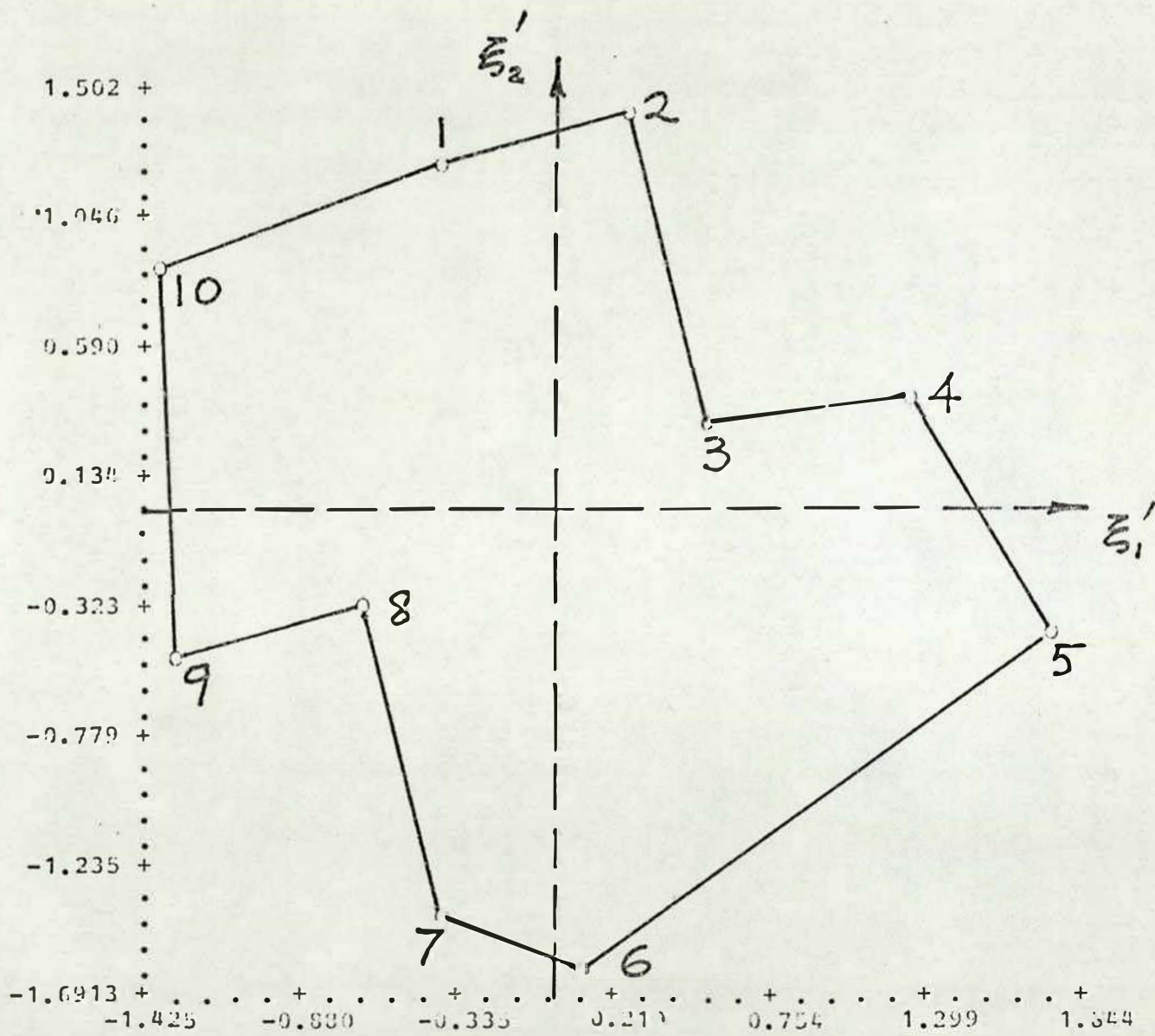


FIGURE 2.29

3. REGRESSION WITH STOCHASTIC CONSTRAINS (MIXED REGRESSIONS)

3. REGRESSION WITH STOCHASTIC CONSTRAINTS (MIXED REGRESSION)

The method of mixed regression deals with the following problem: Given n observations of k independent variables $x_1, \dots, x_k$ and the corresponding n observations of the dependent variable y , estimate the parameters $b_0, b_1, \dots, b_k$ of the multiple regression equation (Main regression):

$$y = b_0 + b_1 x_1 + \dots + b_k x_k + e$$

where e is a disturbance term, assuming that the coefficients $b_0, b_1, \dots, b_k$ satisfy the following equation (stochastic constraint):

$$W = b_0 + b_1 z_1 + \dots + b_k z_k + v$$

where $z_1, \dots, z_k$ and W are sets of q observations of the k independent variables z_1, z_k and the dependent variable W and v is a disturbance term.

This situation could arise e.g. when two separate sets of observations on the dependent and of the dependent variable W and v available and we want to obtain a combined estimate of the regression coefficients and to assess the significance of each set of data for the combined regression. The method was introduced by Theil (1963) and is also summarized in Theil (1971) and Johnston (1972).

In order to make the presentation self-contained and to clarify the notation, we begin by formulating the multiple linear regression model:

$$3.1 \quad y_i = b_0 + b_1 x_{1i} + \dots + b_k x_{ki} + e_i$$

where $i = 1, 2, \dots, n$ = number of observations and the following assumptions are made about the disturbance term e .

$$(2) \quad E(e_i) = 0$$

$$(3) \quad \text{Var}(e_i) = \sigma^2 \text{ for all } i$$

$$(4) \quad \text{Cov}(e_i, e_j) = 0 \text{ for } i \neq j$$

The above assumptions mean that the disturbance term e is a random variable with mean zero, constant variance σ^2 and no autocorrelation.

In dealing with many variables, it is more convenient to write the equations in matrix notation and therefore we shall use the following matrix formulation of the above multiple linear regression model:

$$(5) \quad y = Xb + e$$

where X is a $n \times k$ ($k + 1$) matrix of observations on the k independent variables and the first column of X consists of 1's.

The assumptions about the disturbance term e can now be written:

$$(6) \quad E(e) = 0$$

$$(7) \quad E(ee') = \sigma^2 I$$

where σ^2 is an unknown positive parameter and I is the Unit matrix.

The vector b of the regression coefficients is given by:

$$(8) \quad b = (X'X)^{-1}X'y$$

The variance of the residuals and the covariance matrix of the vector b are given respectively by:

$$(9) \quad \sigma^2 = \frac{1}{n - k - 1} (y'y - b'X'Xb)$$

and

$$(10) \quad V = E(bb') = \sigma_e^2 (X'X)^{-1}$$

The assumptions (3) and (4) of the linear regression model i.e. that the disturbances have the same variance and are uncorrelated are not always realistic. When these assumptions are not satisfied we use Aitken's Generalised least squares model (Aitken (1935), (Theil 1971)) in which it is assumed that the variance of the disturbances may vary for different observations and that the disturbances may be autocorrelated. Under Aitken's modified assumptions the problem can be stated as follows:

The regression equation is

$$(3.11) \quad y = Xb + e$$

where the assumptions about the disturbance term e can be written:

$$(3.12) \quad E(e) = 0$$

$$(3.13) \quad E(ee') = \sigma^2 V$$

where σ^2 is an unknown positive parameter and V is a known symmetric positive definite matrix whose trace is equal to n .

The following formulas (3.14), (3.15), and (3.16) give respectively the vector b of the regression coefficients, the covariance matrix $V(b)$ of this vector (3.14) and the variance σ^2 of the residuals e .

$$(3.14) \quad b = (X'V^{-1}X)^{-1}X'V^{-1}y$$

$$(3.15) \quad E(bb') = \sigma^2 (X'V^{-1}X)^{-1} = V(b)$$

where

$$(3.16) \quad \hat{\sigma}^2 = \frac{1}{n - k - 1} (y - Xb)' \cdot V^{-1} \cdot (y - Xb)$$

Using the notation of generalised regression we can now give the solution of the problem of constrained (mixed) regression as stated above.

Given n sets of observations we want to estimate the regression vector b in the equation

$$(3.17) \quad y = Xb + e \quad (\text{Main regression})$$

where the assumptions about e are given by equations (18) and (19)

$$(3.18) \quad E(e) = 0$$

$$(3.19) \quad E(ee') = \sigma_m^2 V_m$$

The vector b of the regression coefficients must also satisfy the q constraints (3.20) where the assumptions about the disturbance term v are given by equations (3.21) and (3.22).

$$(3.20) \quad w = Zb + v$$

$$(3.21) \quad E(v) = 0$$

$$(3.22) \quad E(vv') = \sigma_c^2 V_c$$

The combined estimator b_M of the regression vector and its covariance matrix are given by equations (23) and (24) respectively

$$(3.23) \quad b_M = \left[\frac{1}{\sigma_m^2} X'V_m^{-1} X + \frac{1}{\sigma_c^2} Z'V_c^{-1} Z \right]^{-1} \cdot \left[\frac{1}{\sigma_m^2} X'V_m^{-1} y + \frac{1}{\sigma_c^2} Z'V_c^{-1} w \right]$$

$$(3.24) \quad V(b_M) = E(b_M b_M') = \left[\frac{1}{\sigma_m^2} X'V_m^{-1} X + \frac{1}{\sigma_c^2} Z'V_c^{-1} Z \right]^{-1}$$

The variance of the disturbances of the combined equation is given by:

$$(3.25) \quad \sigma_M^2 = \frac{1}{n + q - k - 1} \left[(y - Xb_M)' V_m^{-1} (y - Xb_M) + (w - Zb_M)' \cdot \frac{\sigma_m^2}{\sigma_c^2} V_c^{-1} (w - Zb_M) \right]$$

The above formulas are rather complicated and do not show directly the contributions of the main equation and of the constraints to the combined estimator b_M . In order to clarify this point we shall now derive a relation between b_M and the two estimates b_m and b_c of the regression vector, based separately on the main regression and on the constraints. Consider first the Main Regression (3.17) and apply the formula (3.14) of generalised regression. We obtain the estimate b_m :

$$(3.26) \quad b_m = (X'V_m^{-1} X)^{-1} X'V_m^{-1} y \quad \text{and therefore}$$

$$(3.27) \quad (X'V_m^{-1} X) b_m = X'V_m^{-1} y$$

The covariance matrix of b_m is:

$$(3.28) \quad V(b_m) = E(b_m b_m') = \sigma_m^2 (X'V_m^{-1} X)^{-1} \quad \text{and therefore}$$

$$(3.29) \quad \frac{1}{\sigma_m^2} (X'V_m^{-1} X) = V_m^{-1}(b_m)$$

Applying the same formula (3.14) to the constraints we obtain the estimate b_c :

$$(3.30) \quad b_c = (Z'V_c^{-1} Z)^{-1} Z'V_c^{-1} w \quad \text{and therefore}$$

$$(3.31) \quad (Z'V_c^{-1} Z) b_c = Z'V_c^{-1} w$$

The covariance matrix of b_c is:

$$(3.32) \quad V(b_c) = E(b_c b_c') = \sigma_c^2 (Z' V_c^{-1} Z)^{-1} \quad \text{and therefore}$$

$$(3.33) \quad \frac{1}{\sigma_c^2} (Z' V_c^{-1} Z) = V^{-1}(b_c)$$

From the above equations we have

$$(3.34) \quad V^{-1}(b_m) b_m = \frac{1}{\sigma_m^2} X' V_m^{-1} y \quad \text{and}$$

$$(3.35) \quad V^{-1}(b_c) b_c = \frac{1}{\sigma_c^2} Z' V_c^{-1} W$$

Using the equations (3.28), (3.32), (3.33), (3.34), & (3.35) we write the combined estimator b_M as follows:

$$(3.36) \quad b_M = \left[V^{-1}(b_m) + V^{-1}(b_c) \right]^{-1} \cdot \left[V^{-1}(b_m) \cdot b_m + V^{-1}(b_c) \cdot b_c \right]$$

The covariance matrix of the estimator b_M can be written:

$$(3.37) \quad V(b_M) = E(b_M \cdot b_M') = \left[V^{-1}(b_m) + V^{-1}(b_c) \right]^{-1}$$

Formulas (36) and (37) can be interpreted in the following way:

The combined regression coefficient vector b_M , based on the information contained in both the main regression and the constraints is a "weighted average" of the individual vectors b_m and b_c where the weights are the inverses of the covariance matrices of these coefficients. It is easier to interpret this result intuitively by writing formulas (3.36) and (3.37) for the case of simple regression where we obtain:

$$(3.38) \quad b_M = \frac{\frac{1}{\sigma^2(b_m)} \cdot b_m + \frac{1}{\sigma^2(b_c)} \cdot b_c}{\frac{1}{\sigma^2(b_m)} + \frac{1}{\sigma^2(b_c)}}$$

$$(3.39) \quad \sigma^2(b_M) = \frac{1}{\frac{1}{\sigma^2(b_m)} + \frac{1}{\sigma^2(b_c)}}$$

Formulas (3.37) and (3.38) show that b_M is a weighted average of b_m and b_c and the weights are the reciprocals of the variances $\sigma^2(b_m)$ and $\sigma^2(b_c)$ of b_m and b_c .

It is to be noted that formulas (3.38) and (3.39) are identical to the formulas in Bayesian analysis giving the posterior regression coefficient as a combination of the prior and the sample regression coefficients in the case of normal prior and sampling distributions. The method of constrained regression and the resulting combination of the information contained in the main regression and in the constraints should be used only if the two sets of data are compatible. Theil (1971) gives the following compatibility test. Under the null hypothesis that the data of the main regression and the constraint come from the same multivariate population, the statistic

$$(3.40) \quad Q = (w - zb_m)' \left[\sigma_m^2 z (X' V_m^{-1} X)^{-1} z' + \sigma_c^2 V_c \right]^{-1} (w - zb_m)$$

follows the chi-square distribution with q degrees of freedom.

Assuming that the two sets of data are compatible, we can compute as follows the contributions θ_m and θ_c of the main regression and the constraint respectively:

$$(3.41) \quad \theta_m = \frac{1}{k} \operatorname{tr} \left[V^{-1}(b_m) \cdot V(b_M) \right]$$

$$(3.42) \quad \theta_c = 1 - \theta_m = \frac{1}{k} \operatorname{tr} \left[V^{-1}(b_c) \cdot V(b_M) \right]$$

In the case of one independent variable we have

$$(3.43) \quad \theta_m = \frac{\frac{1}{\sigma^2(b_m)}}{\frac{1}{\sigma^2(b_m)} + \frac{1}{\sigma^2(b_c)}} ; \quad \theta_c = \frac{\frac{1}{\sigma^2(b_c)}}{\frac{1}{\sigma^2(b_m)} + \frac{1}{\sigma^2(b_c)}} .$$

A slight modification of the formula for the mixed regression coefficient b_M for one independent variable shows clearly the factors which influence the weights of b_m and b_c . We have:

$$(3.44) \quad b_M = \frac{\frac{1}{\sigma^2(b_m)} \cdot b_m + \frac{1}{\sigma^2(b_c)} \cdot b_c}{\frac{1}{\sigma^2(b_m)} + \frac{1}{\sigma^2(b_c)}}$$

and by substituting the value of the standard deviations of the regression coefficients given by the formulas

$$(3.45) \quad \sigma^2(b_m) = \frac{\sigma_m^2}{\sigma_m^2(X)(n-2)} \quad ; \quad \sigma^2(b_c) = \frac{\sigma_c^2}{\sigma_c^2(X)(q-2)}$$

We obtain:

$$(3.46) \quad b_M = \frac{(n-2) \cdot \frac{\sigma_m^2(X)}{\sigma_m^2} \cdot b_m + (q-2) \cdot \frac{\sigma_c^2(X)}{\sigma_c^2} \cdot b_c}{(n-2) \cdot \frac{\sigma_m^2(X)}{\sigma_m^2} + (q-2) \cdot \frac{\sigma_c^2(X)}{\sigma_c^2}}$$

This formula shows that the relative contributions of b_m and b_c are proportional to the respective variances of the independent variables in the main regression and the constraint and inversely proportional to the standard errors of estimate of the two regressions. The corresponding formulas for the contributions θ_m and θ_c of the main regression and the constraint are:

$$(3.47) \quad \theta_m = \frac{(n-2) \cdot \frac{\sigma_m^2(X)}{\sigma_m^2}}{(n-2) \cdot \frac{\sigma_m^2(X)}{\sigma_m^2} + (q-2) \cdot \frac{\sigma_c^2(X)}{\sigma_c^2}} \quad ; \quad \theta_c = 1 - \theta_m$$

Errors of observations

If the dependent variables y and w of the main regression and the constraints are observed with errors which have variances $\sigma_o^2(y)$ and $\sigma_o^2(w)$ respectively, the total variances of the estimated values of the dependent variables y and w are:

$$(3.48) \quad \sigma^2(y) = \sigma_m^2 + \sigma_o^2(y)$$

$$(3.49) \quad \sigma^2(w) = \sigma_c^2 + \sigma_o^2(w)$$

The above formulas are based on the assumption that the errors of observation are independent of the disturbance terms of the regressions.

. In the case of errors of observation, we must substitute $\sigma^2(y)$ and $\sigma^2(w)$ for σ_m^2 and σ_c^2 in equations (3.23), (3.24), (3.40), (3.41) and (3.42).

4. APPLICATION OF MIXED REGRESSION TO THE ESTIMATION OF
PRECIPITATION AND RUNOFF FROM A MULTI-BASIN REGION

4. APPLICATION OF MIXED REGRESSION TO THE ESTIMATION OF PRECIPITATION AND RUNOFF FROM A MULTI-BASIN REGION

The methodology of mixed regression described in section 3 will now be used for the estimation of the regional distribution of precipitation and runoff. An application to the region of Gaspé in Quebec will be given at the end of the present section.

The estimation of the regional distribution of precipitation is particularly difficult in regions where the meteorological stations are located mostly in inhabited areas whereas a substantial part of the precipitation occurs in ungauged, remote sub-basins. For example, in the case of the Gaspé peninsula (Figures 4.1 and 4.2), the meteorological stations are located mostly on the coast and therefore the data are not representative of the meteorological conditions of the interior regions where a large part of the flow originates. The gauging stations are also located mainly in the coastal regions but the recorded flows include the contributions of the interior sub-basins whose precipitation is not observed.

The problem here is to use the information contained in the flow measurements to improve the estimates of precipitation, particularly in the interior basins. A solution of this problem is given in a paper by Solomon et al. (1968), where the authors establish a square grid covering the area of study, which is

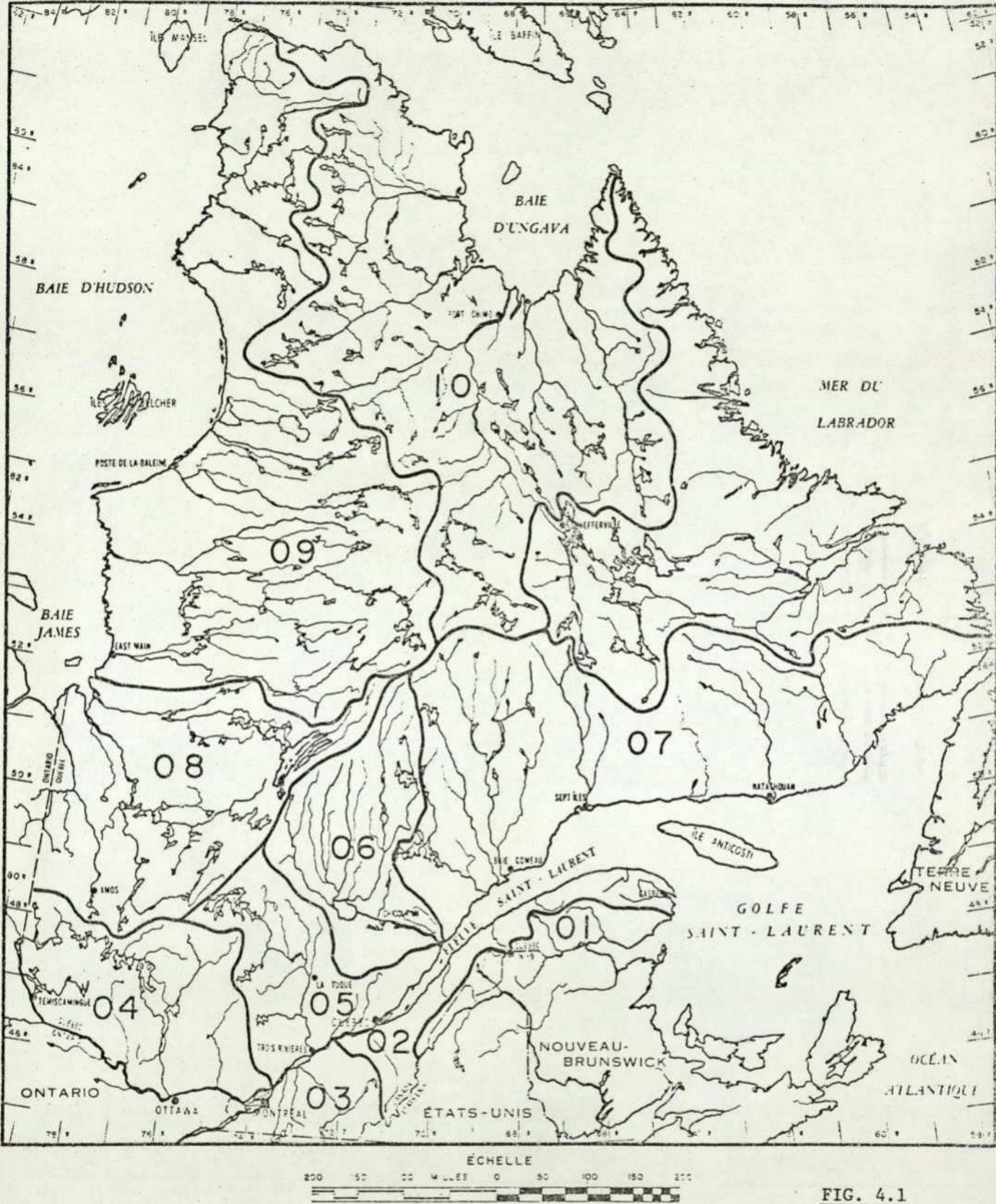


FIG. 4.1

PROVINCE DE QUÉBEC RÉGIONS HYDROGRAPHIQUES

۵/۲

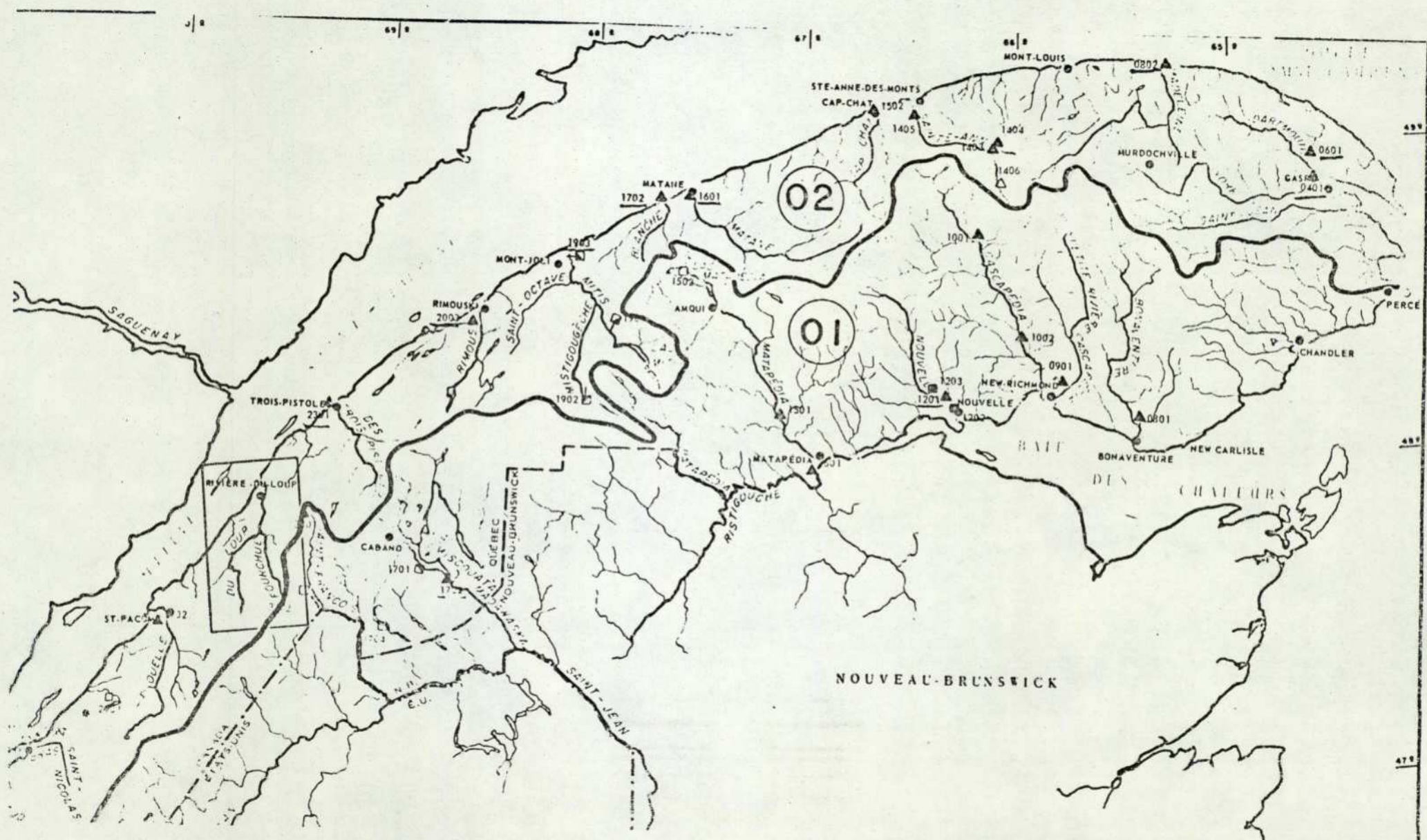
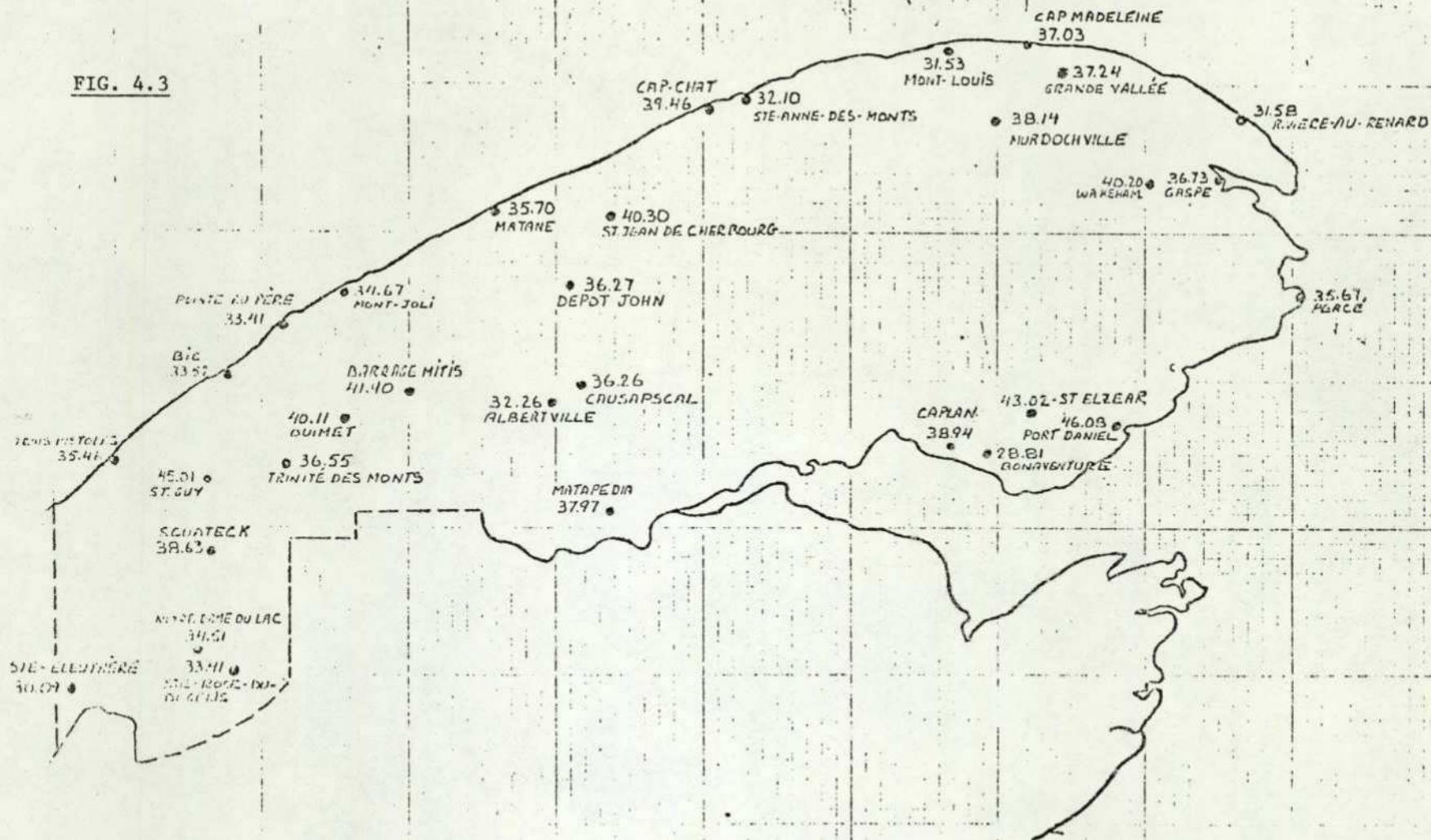


FIG. 4.3



LONG TERM AVERAGE PRECIPITATION (1937-1967) @ CLIMATOLOGICAL STATIONS

used to store and retrieve information on the temperature, precipitation, evaporation, runoff and physiographic characteristics of the region.

More specifically the square grid is used in the following way to correct iteratively the values of precipitation taking into account the observed values of temperature and flow. As a first step, the temperature, precipitation and runoff values are regressed on several physiographic variables by a step-wise regression program. The runoff of the gauged basins is then estimated by the quantity $(P - E)$ where P is the precipitation and E is the evapotranspiration given by Turc's formula:

$$E = \frac{P}{(0.9 + (\frac{P}{L})^2)^{\frac{1}{2}}}$$

where:

E = Mean annual evaporation in mm.

P = Mean annual precipitation in mm.

$L = 300 + 25T + 0.05T^3$

T = Mean annual temperature $^{\circ}\text{C}$

The next step is to compute the ratio $k = \frac{Q}{P - E}$ where Q is the recorded runoff of the gauged basins. mm p.a. the computed values of the coefficients k are then used to modify the precipitation values for all squares of the grid in the gauged

basins and the regression on the physiographic variables is repeated with the new precipitation values. This iterative process stops when further changes in k remain below a preassigned level and the regression equations are used to compute the precipitation and runoff values for all squares of the grid.

A detailed flow chart of the computations of the iterative method is given on the left side of Table 4.1. The successive steps of the computation are described by Solomon et al (1968) and therefore only a brief summary will be given here:

STEP 1. Step-wise regression of the mean annual temperature on the physiographic variables for the m squares of the grid which have meteorological stations.

STEP 2. Initial stepwise regression of the mean annual precipitation on the physiographic variables for the m squares of the grid which have meteorological stations.

STEP 3. Computation of the mean annual temperatures for all squares of the grid based on the regression equation of Step 1.

STEP 4. Computation of the mean annual precipitation for all squares of the grid based on the regression equation of Step 2.

TABLE 4.1

ITERATIVE METHOD					MIXED REGRESSION METHOD				
STEP	EQUATION	SQUARES USED	REGRESSION	COMPUTATION	STEP	EQUATION	SQUARES USED	REGRESSION	COMPUTATION
1	$\hat{T} = f(\underline{G})$	m	✓		1	AS IN ITERATIVE METHOD			
2	$\hat{P}_O = f(\underline{G})$	m	✓		2				
3	$T = f(\underline{G})$	ALL		✓	3				
4	$P_O = f(\underline{G})$	ALL		✓	4				
5	E	ALL		✓	5				
6	$Q_O = \hat{P}_O - E$	B		✓	6	UNDER THE STOCHASTIC CONSTRAINTS $\sum_{B_1} (\hat{P}_1 - E) = Q_1 + v$ $\sum_{B_q} (\hat{P}_1 - E) = Q + v$	✓		
7	$k = Q/Q_O$	B		✓					
8	$\hat{P}_1 = kQ_O + E$	b		✓					
9	$\hat{P}_1 = f(\underline{G})$. . .	cm+b	✓				b		
10	$\hat{P}_s = f(\underline{G})$	cm+b	✓						
11	$P_s = f(\underline{G})$	ALL		✓	7	AS IN ITERATIVE METHOD			
12	$Q = P_s - E$	ALL		✓	8				
13	$\hat{Q} = f(\underline{G}, \underline{S})$	b	✓		9				
14	$Q = f(\underline{G}, \underline{S})$	ALL		✓	10				

T = TEMPERATURE; P = PRECIPITATION; Et = EVAPOTRANSPIRATION; Q = FLOW

G = PHYSIOGRAPHIC VARIABLES; S = SOIL VARIABLES

m = SQUARES OF MET. STATIONS; b = SQUARES IN GAUGED BASINS; B = GAUGED BASINS

c = 10 = COEFFICIENT OF SQUARES CONTAINING METEOROLOGICAL STATIONS

$\hat{P}$ = Precipitation estimated by regression

STEP 5. Computation of mean annual evaporation by Turc's formula for all squares of the grid, using the values of mean annual temperature and precipitation derived in Steps 3 and 4.

The above five steps constitute the first phase of the computation and form the basis of the iterations of the following Steps 6 to 10.

STEP 6. The mean annual runoff Q_o of the gauged basins is estimated as the difference between the precipitation $\hat{P}_o$ and the mean annual evapotranspiration E .

STEP 7. Computation of the ratios $k = \frac{Q}{Q_o}$ for the gauged basins where Q = observed runoff.

STEP 8. Computation of the corrected value $\hat{P}_1$ of the precipitation of the b squares of the gauged basins as the sum of the observed mean annual runoff and the estimated evapotranspiration E . In equation form we have:

$$P_1 = Q + E = k Q_o + E$$

STEP 9. Stepwise regression between the precipitation $\hat{P}_1$ estimated in Step 8 and the physiographic variables. The regression is computed for the b squares of the gauged basins and the m squares with meteorological stations. The number of the latter squares is multiplied by $c = 10$ to give more

weight to the observed precipitation. The iterative procedure consists in using the precipitations $\hat{P}_1$ estimated by the regression equation to compute new values of the ratios k and then to arrive at new corrected values $\hat{P}_2$ of the precipitation of the gauged basins, which is in turn regressed on the physiographic variables.

STEP 10. The last regression equation after s iterations gives the estimated precipitation $\hat{P}_s$ as a linear combination of the physiographic variables.

STEP 11. Computation of the mean annual precipitation for all squares of the grid on the basis of the multiple regression equation of Step 10.

STEP 12. Computation of the runoff for all squares.

STEP 13. Regression of the runoff on the physiographic and soil variables for the b squares of the gauged basins.

STEP 14. Computation of the runoff for all squares, based on the regression equation of Step 13.

As mentioned earlier, the iterative procedure was used to estimate the distribution of precipitation and runoff in the Gaspé peninsula in Quebec. A detailed description of the data for this application and a partial listing of the results of iterative method are given in Appendices 4.1 and 4.2 respectively.

A summary of the computations of the iterative method is given in Table 4.2. An examination of this table indicates the following facts concerning the iterative method:

- (1) The subset of physiographic characteristics ($X_2, X_8, X_{11}, X_{14}, X_{18}$) selected in the final regression of precipitation is similar to the subset of variables (X_2, X_8, X_{14}, X_{19}) selected in the regression of runoff and differs substantially from the variables ($X_7, X_{10}, X_{15}, X_{19}, X_{20}$) selected in the initial precipitation regression (step 2).
- (2) The multiple correlation coefficient of the final precipitation regression is very small because the iterative process, due to the correction $P_1 = k Q_0 + E$ tends to choose the physiographic variables that are more relevant for runoff than for precipitation.

The above two statements are confirmed by an examination of the following scatter diagrams (Figures 4.12 and 4.13):

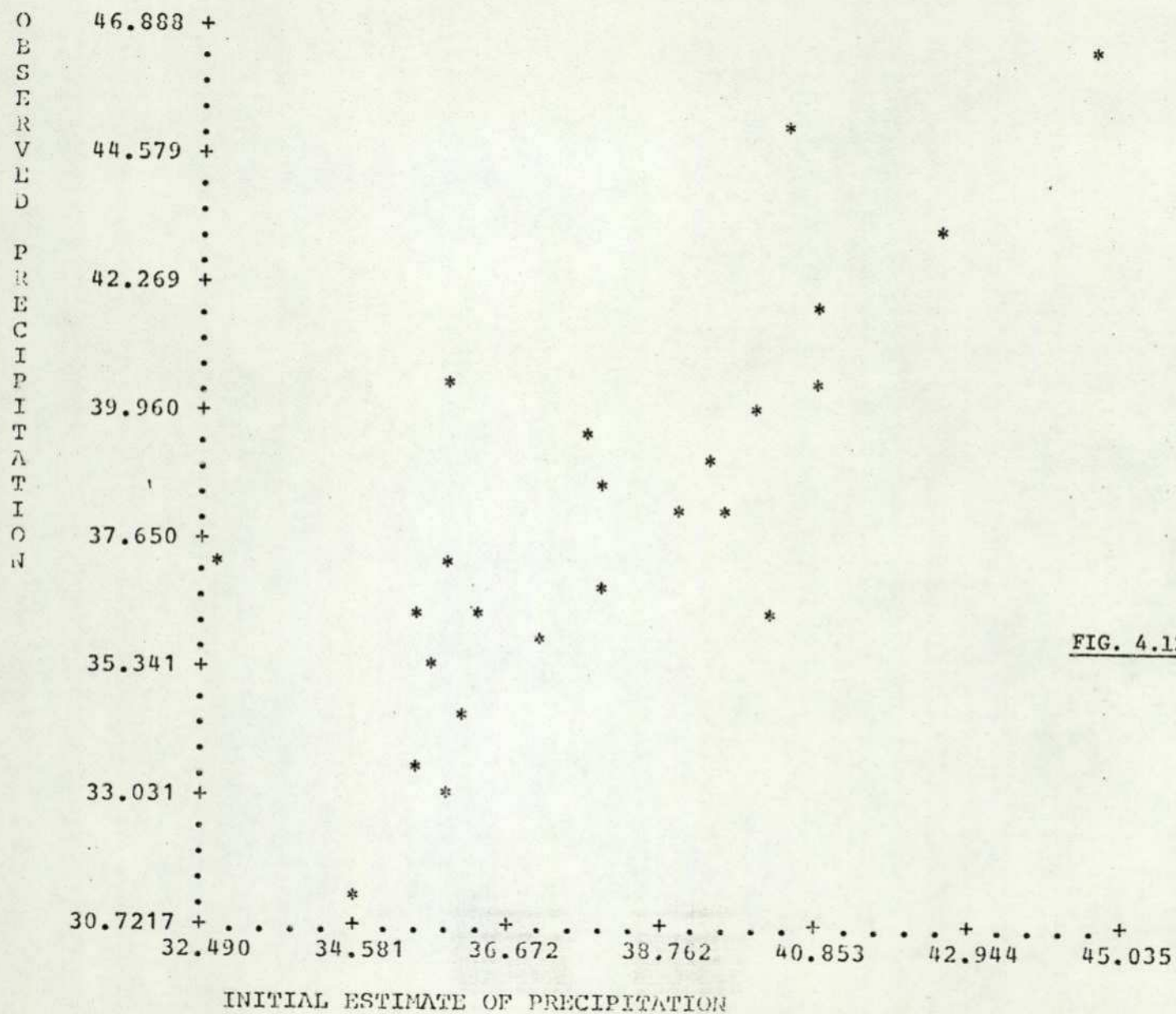
- (1) Observed vs estimated precipitation (First iteration, Step 2).
- (2) Observed vs estimated precipitation (Last iteration, Step 10).

We conclude from the above that the approach suggested by Solomon et al for improving the estimates of precipitation in inaccessible

TABLE 4.2
ITERATIVE METHOD
SUMMARY OF COMPUTATION RESULTS

INDEPENDENT VARIABLES	<u>STEP 2</u> INITIAL PRECIPITATION STEPWISE REGRESSION	<u>STEP 10</u> FINAL PRECIPITATION STEPWISE REGRESSION	<u>STEP 13</u> FINAL RUNOFF STEPWISE REGRESSION
X(1) = Longitude index			
X(2) = Latitude index		0.508046	0.642980
X(3) = Average elevation			
X(4) = Slope of square			
X(5) = Azimuth of slope			
X(6) = Shortest distance to sea			
X(7) = Distance north	0.271499		
X(8) = Distance N-E		0.061618	0.079732
X(9) = Distance East			
X(10) = Barrier height North	0.040127		
X(11) = Barrier height N-E		0.021875	
X(12) = Barrier height East			
X(13) = $X(3)^2$			
X(14) = $X(6)^2$		-0.002375	-0.002292
X(15) = $X(7)^2$	-0.001997		
X(16) = $X(8)^2$			
X(17) = $X(9)^2$			
X(18) = $X(10)^2$		-0.000020	
X(19) = $X(11)^2$	-0.000186		0.000074
X(20) = $X(12)^2$	0.000038		
Standard deviation of precipitation	3.551		
Coefficient $\bar{R}$	0.786	-0.086	0.959
Standard error of estimate $S_{y.x}$	2.244		

*/



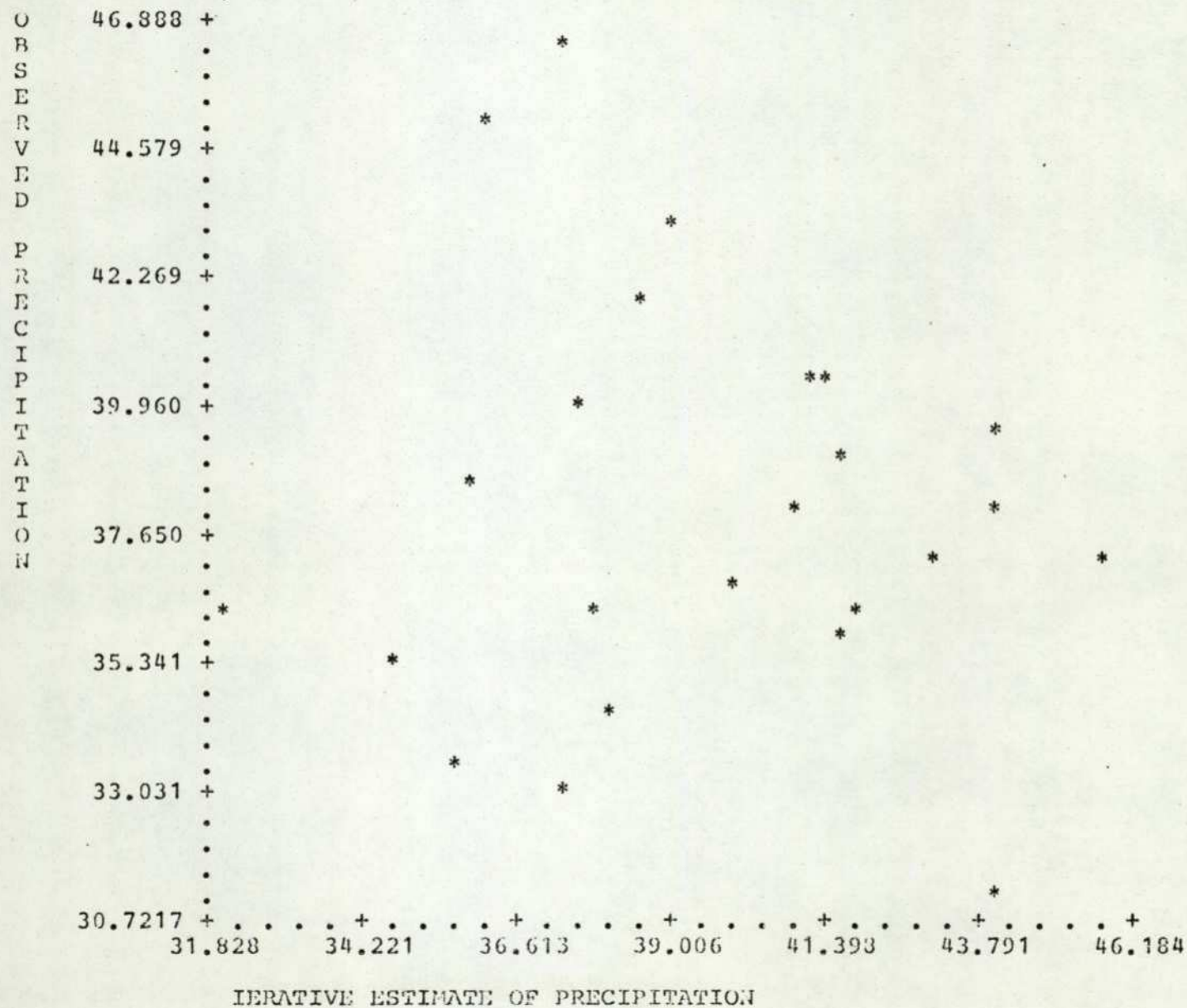


Figure 4.13

areas by using information on the river flow from those areas and appropriate estimates of evaporation is very useful.

It appears however that the iterative computation may converge to a regression equation that is excessively influenced by the relationship of flow to physiographic variables, regardless of the relative magnitudes of the measurement errors of precipitation and runoff.

An alternative to the iterative method is the mixed regression discussed in the previous section. The basic idea is to consider the regression of the precipitation on the physiographic variables as the main regression whereas the constraints will be given by runoff and evaporation data from the gauged basins. The details of the computations are as follows.

Assume that we have n meteorological stations, k physiographic variables $G_1, \dots, G_k$ and q gauged basins. The regression equation of precipitation P on the physiographic factors $G_1, \dots, G_k$ shown in Step 6 for each meteorological station is

$$(4.1) \quad P_i = \beta_0 + \beta_1 G_1 + \dots + \beta_k G_k + e_i \quad \text{for } i = 1, 2, \dots, n$$

where P_i is the measured precipitation at the station i and the residuals e_i include the random errors of measurement of the precipitation P_i .

Due to the fact that the mean annual temperature in the

Gaspé peninsula is about 35°F , the evaporation E computed by Turc's formula is almost independent of relatively small changes in P . Consequently we shall assume that the mean annual runoff of the river basin B_j can be estimated by

$$\frac{1}{n_j} \sum_{i=1}^{n_j} (P_i - E_i)$$

where $j = 1, 2, \dots, q$

$i = 1, 2, \dots, n_j$

and n_j is the number of squares in the basin j .

Equating now the estimated and the measured annual runoff of the basin B_j we have

$$\frac{1}{n_j} \sum_{i=1}^{n_j} (P_i - E_i) = Q_j$$

and substituting the value of P_i from equation (4.1) we obtain

$$(4.2) \quad Q_j + \frac{1}{n_j} \sum_{i=1}^{n_j} E_i = \beta_0 + \beta_1 \bar{G}_1 + \dots + \beta_k \bar{G}_k + v_j$$

where Q_j is the measured annual flow of the basin B_j , $\bar{G}_1, \dots, \bar{G}_k$ are the mean values of the physiographic variables for the basin and the residuals v_j include the random errors of measurement of the flow Q_j . The coefficients $\beta_0, \beta_1, \dots, \beta_k$ of the above equations can be estimated by the method of mixed regression described in Chapter 2.

The variances $\sigma^2(e_i) = \sigma_m^2$ and $\sigma^2(v_i) = \sigma_c^2$ of the residuals of the main regression and the constraint can be computed by formula (3.16) of Chapter 3. Assuming that the errors of observation are independent of the random residuals we can use formulas (3.48) and (3.49) of Chapter 3:

$$\sigma_{PT}^2 = \sigma_m^2 + \sigma_{PE}^2$$

$$\sigma_{QT}^2 = \sigma_c^2 + \sigma_{QE}^2$$

where

$$\sigma_{PT}^2 = \text{Total variance of precipitation}$$

$$\sigma_m^2 = \sigma^2(e_i) = \text{Variance of the residuals of the main regression}$$

$$\sigma_{PE}^2 = \text{Variance of observation errors of precipitation}$$

$$\sigma_{QT}^2 = \text{Total variance of flow}$$

$$\sigma_c^2 = \sigma^2(v_i) = \text{Variance of the residuals of the constraint}$$

$$\sigma_{QE}^2 = \text{Variance of observation errors of flow}$$

In this case, the covariance matrices used in the computations are given by:

$$E(ee') = \sigma_{PT}^2 \cdot v_m$$

$$E(vv') = \sigma_{QT}^2 \cdot v_c$$

Detailed results of the mixed regression computations are given in Appendix 4.2. These results are summarized in Table 4.3 which shows that the precipitation and the runoff information for the Gaspé peninsula cannot be readily combined for estimating the distribution of precipitation, since the compatibility statistic exceeds the critical value at the 1% level.

In order to interpret this result we must examine the basic purpose of both the iterative and the mixed regression computations. Given that the precipitation network is inadequate, we are justified in attempting to improve the estimate of the regional distribution of precipitation by using available data on regional runoff and evaporation. In trying to use this additional information, the iterative method proceeds on the implicit assumption that, in case of conflict between the two kinds of data, we should use the data on runoff and evaporation.

On the other hand, the mixed regression method uses the compatibility test to decide whether the two types of information should be combined by means of the mixed regression equation or whether the incompatibility between the two types of information makes necessary a further examination of the data in order to discover the reason for this discrepancy. From this point of view the method of mixed regression appears to be preferable.

TABLE 4.3
MIXED REGRESSION METHOD
SUMMARY OF COMPUTATION RESOURCES

INDEPENDENT VARIABLES	MAIN REGRESSION	CONSTRAINTS	MIXED REGRESSION	MIXED REGRESSION
	Precipitation	Runoff + Evapotrans- piration	$\sigma_{PE}^2 = 0$ $\sigma_{QE}^2 = 0$	$\sigma_{PE}^2 = 100$ $\sigma_{QE}^2 = 16$
X(1) = Longitude index				
X(2) = Latitude index	-0.22202	0.14842	0.23304	0.24545
X(3) = Average elevation				
X(4) = Slope of square				
X(5) = Azimuth of slope				
X(6) = Shortest distance to sea				
X(7) = Distance North	0.280134	-0.55883	0.07752	0.05250
X(8) = Distance N-E				
X(9) = Distance East				
X(10) = Barrier height North	0.01889	0.02405	0.01590	0.01620
X(11) = Barrier height N-E				
X(12) = Barrier height East	0.09687	0.02114	0.04711	0.04826
X(13) = $X(3)^2$				
X(14) = $X(6)^2$				
X(15) = $X(7)^2$	-0.00210	0.00453	-0.00026	0.00008
X(16) = $X(8)^2$				
X(17) = $X(9)^2$	-0.00012	0.00001		
X(18) = $X(10)^2$				
X(19) = $X(11)^2$	-0.00025	0.00025	-0.00011	-0.00009
X(20) = $X(12)^2$	-0.00022	-0.00013	-0.00017	-0.00018
Standard deviation of dependent variable	3.551	3.227	3.551	3.551
Multiple correlation coefficient $\bar{R}$	0.762	0.948	0.435	0
Standard error of estimate $S_{y.x}$	2.756	1.240	3.20	3.551
Compatibility statistic = 23.59 $\chi^2(9)_{0.01} = 21.67$			23.59	151.00

An additional advantage of this method is that the underlying assumptions can be stated explicitly and therefore probability statements can be made concerning the regression coefficients and the values of the dependent variable.

Appendix 4.1

Data1. Physiographical data

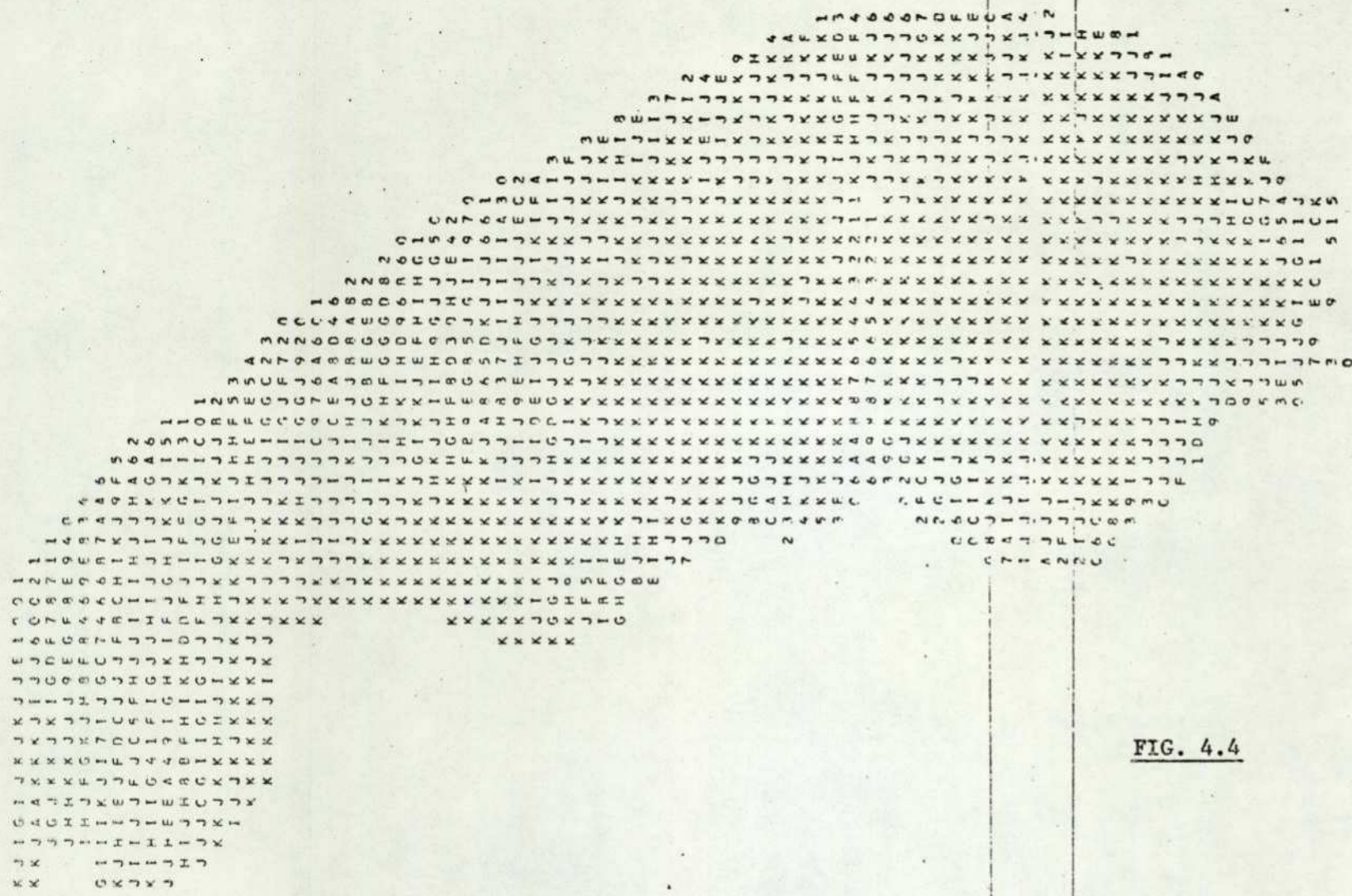
The basic physiographic data for each square of the grid were taken from the 1: 250 000 maps of the Surveys and Mapping Branch, Department of Energy, Mines and Resources, Canada.

Specifically, the following data were extracted:

- (1) Lake area (Percentage of the area of the square).
- (2) Forest area (Percentage of the area of the square).
- (3) Area of bogs and swamps (Percentage of the area of the square).
- (4) Elevation of the south-east corner of the square.
- (5) Minimum distance to the sea or the St. Lawrence River.

These data were used to produce schematic maps, an example of which is shown in Figure 4.4 (Area of forests as a percentage of the square). On the basis of the above information, the following physiographic variables were computed for each square of the grid:

- (1) Mean elevation
- (2) Mean slope
- (3) Azimuth of slope
- (4) Distance North to the sea or the St. Lawrence River
- (5) Distance North-east to the sea or the St. Lawrence River



AREA OF FOREST (PERCENT)

0	=	0	-	4
1	=	5	-	9
2	=	10	-	14
3	=	15	-	19
4	=	20	-	24
5	=	25	-	29
6	=	30	-	34
7	=	35	-	39
8	=	40	-	44
9	=	45	-	49
A	=	50	-	54
B	=	55	-	59
C	=	60	-	64
D	=	65	-	69
E	=	70	-	74
F	=	75	-	79
G	=	80	-	84
H	=	85	-	89
I	=	90	-	94
J	=	95	-	99
K	=	100	-	104
L	=	105	-	109
M	=	110	-	114
N	=	115	-	119
O	=	120	-	124
P	=	125	-	129
Q	=	130	-	134
R	=	135	-	139
S	=	140	-	144
T	=	145	-	149
U	=	150	-	154
V	=	155	-	159
W	=	160	-	164
X	=	165	-	169
Y	=	170	-	174
Z	=	175	-	179

FIG. 4.4

SLOPE

0	0	0
1	5	9
2	10	14
3	15	19
4	20	24
5	25	29
6	30	34
7	35	39
8	40	44
9	45	49
A	50	54
B	55	59
C	60	64
D	65	69
E	70	74
F	75	79
G	80	84
H	85	89
I	90	94
J	95	99
K	100	104
L	105	109
M	110	114
N	115	119
O	120	124
P	125	129
Q	130	134
R	135	139
S	140	144
T	145	149
U	150	154
V	155	159
W	160	164
X	165	169
Y	170	174

FIG. 4.5

- (6) Distance East to the sea or the St. Lawrence River
- (7) Height of barrier in a Northern direction
- (8) Height of barrier in a North-eastern direction
- (9) Height of barrier in an Eastern direction

The computed values of these variables were also used to produce schematic maps, an example of which is shown in Figure 4.5 (Mean Slope).

2. Meteorological data

Monthly precipitation and temperature data were taken from the meteorological maps of the Department of Natural Resources of the Province of Quebec. Missing data for the period October 1937 to September 1967 were estimated by means of regressions with neighbouring stations. The list of the twenty-four meteorological stations used in the study is given in Table 4.4.

3. Flow data

The river flows at eleven stations were used in the study. For the stations whose records was less than thirty years the missing flows were estimated by regression. Table 4.5 gives the eleven gauging stations used in the study. Drainage areas were obtained from the 1: 250 000 maps of the Department of Natural Resources. For each square of the grid a file was created with the corresponding values of precipitation, temperature, evaporation and runoff.

TABLE 4.4

METEOROLOGICAL STATIONS

STATION #	NAME	PERIOD OF RECORD	
		FROM	TO
7058560	Trois Pistoles	Oct. 51	Sept. 67
7057304	St. Guy	Oct. 48	Sept. 67
7058220	Squateck	May 46	Sept. 67
7057720	Ste. Rose du Degelis	Oct. 37	Sept. 67
7058520	Trinité des monts	July 51	Sept. 67
7050760	Bic	Oct. 37	Sept. 67
7056067	Pointe-au-Père	Oct. 37	Sept. 67
7055770	Ouimet	Sept. 49	Sept. 67
7050463	Barrage Mitis	Oct. 37	June 66
7055120	Mont-Joli A	Oct. 43	Sept. 67
7054640	Matane	Oct. 51	Sept. 53
7057395	St. Jean-de-Cherbourg	Nov. 50	Sept. 67
7051200	Causapscal	Oct. 37	Sept. 67
7054680	Matapedia	Feb. 37	July 62
7051120	Caplan	Oct. 47	Sept. 67
7056120	Port Danial	Oct. 37	Sept. 67
7057160	St. Elzear	Apr. 50	Sept. 67
7051040	Cap Chat	Oct. 37	Sept. 67
7055380	Murdochville	Oct. 52	Sept. 67
7055200	Mont. Louis	Oct. 37	Sept. 67
7051160	Cap Madelaine	Oct. 39	Sept. 67
7052865	Grande-Vallée	May 44	Sept. 67
7052600	Gaspé	Oct. 37	Sept. 67
7058840	Wakeham	Oct. 53	Sept. 67

TABLE 4.5

GAUGING STATIONS

STATION #	NAME	PERIOD OF RECORD		Drainage Area
		FROM	TO	
011702	Madawaska	1918	1967	1050 sq. mi.
022301	Trois Pistoles	1923	1967	373
022003	Rimouski	1961	1967	612
021903	Mitis	1923	1967	702
021702	Blanche	1968	1967	80.4
021601	Matane	1923	1967	636
011201	Nouvelle	1964	1967	440
010901	Petite Riviere Cascapedia	1961	1967	536
020802	Madeleine	1953	1967	470
020401	York	1945	1967	390
020601	Dartmouth	1945	1967	287

Appendix 4.2

In this appendix we give the main computation results for both the iterative and mixed regression methods.

ITERATIVE METHOD

- FIG. 4.6 Stepwise regression of temperature (Step 1)
- FIG. 4.7 Initial stepwise regression of precipitation (Step 2)
- FIG. 4.8 Initial evapotranspiration (Step 5)
- FIG. 4.9 Initial k-factors (Step 7)
- FIG. 4.10 Final stepwise regression of precipitation (Step 10)
- FIG. 4.11 Final stepwise regression of runoff (Step 13)

MIXED REGRESSION METHOD

- FIG. 4.14 Mixed regression computations ($\sigma_{PE} = 0$, $\sigma_{QE} = 0$)
- FIG. 4.15 Observed vs. estimated precipitation (main regression)
- FIG. 4.16 Observed vs. estimated runoff plus evapotranspiration (constraints)
- FIG. 4.17 Observed vs. estimated precipitation (mixed regression)
- FIG. 4.18 Observed vs. estimated runoff plus evapotranspiration (mixed regression)

FOR 4 VARIABLES ENTERED

MULTIPLE CORRELATION COEFFICIENT... 0.9242328E 00
 (ADJUSTED FOR D.F.)..... 0.9123253E 00
 F-VALUE FOR ANALYSIS OF VARIANCE... 0.2783025E 02
 STANDARD ERROR OF ESTIMATE..... 0.5917572E 00
 (ADJUSTED FOR D.F.)..... 0.6238648E 00

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
3	-0.285772E-01	0.300557E-02	-0.951471E 01
19	0.659412E-04	0.174576E-04	0.377931E 01
14	0.514995E-03	0.175845E-03	0.292659E 01
11	-0.919569E-02	0.343571E-02	-0.267650E 01
CONSTANT	0.333664E 02		

X(1)=LONGITUDE INDEX
 X(2)=LATITUDE INDEX
 X(3)=AVERAGE ELEVATION - FT/10
 X(4)=SLOPE OF SQUARE (PERCENT/100)
 X(5)=DIRECTION OF SLOPE
 X(6)=SHORT-DIST. TO SEA -KM
 X(7)=DISTANCE NORTH KM.
 X(8)=DISTANCE N-E KM.
 X(9)=DISTANCE EAST KM.
 X(10)=CARRIER HEIGHT NORTH FT X 10
 X(11)=CARRIER HEIGHT N-E FTX10
 X(12)=CARRIER HEIGHT EAST FTX10
 X(13)=X(3)**2
 X(14)=X(6)**2
 X(15)=X(7)**2
 X(16)=X(8)**2
 X(17)=X(9)**2
 X(18)=X(10)**2
 X(19)=X(11)**2
 X(20)=X(12)**2
 X(21)=TEMPERATURE DEG.FAR.

SELECTION..... 1

TABLE OF RESIDUALS

CASE NO.	Y VALUE	Y ESTIMATE	RESIDUAL	PERCENT	STATION
1	37.200	34.128	-0.928	-2.5	7056067
2	39.200	37.774	0.426	1.1	7055120
3	37.300	37.495	-0.395	-1.1	7054640
4	39.500	37.597	0.903	2.3	7058560
5	36.100	36.300	-0.200	-0.6	7037304
6	36.900	36.851	0.049	0.1	7058220
7	38.100	39.056	0.044	0.1	7057720
8	35.900	35.895	0.005	0.0	7058520
9	39.800	39.094	0.416	1.1	7050760
10	35.400	35.351	0.549	1.5	7055770
11	35.500	36.960	-0.560	-1.6	7050453
12	34.000	35.278	-1.278	-3.8	7057395
13	36.200	36.251	-0.051	-0.1	7051200
14	37.500	37.540	-0.040	-0.1	7054680
15	39.400	39.105	0.295	0.7	7051120
16	39.100	39.785	-0.685	-1.8	7056120
17	37.000	36.939	0.061	0.0	7057180
18	37.900	37.665	0.135	0.4	7051040
19	33.900	33.485	0.415	1.2	7053380
20	38.700	38.704	0.496	1.3	7055200
21	37.500	37.613	-0.113	-0.3	7051160
22	38.100	38.369	-0.269	-0.7	7052865
23	39.100	38.243	-0.143	-0.4	7052600
24	37.600	36.679	0.921	2.5	7058840

FIG. 4.6

STEP 5

VARIABLE ENTERED 20

SUM OF SQUARES REDUCED IN THIS STEP.. 0.1220024E 02
 PROPORTION REDUCED IN THIS STEP..... 0.4206903E-01
 CUMULATIVE SUM OF SQUARES REDUCED.... 0.1791779E 03
 CUMULATIVE PROPORTION REDUCED..... 0.6178436E 00 3F 0.2900051E 03

FOR 5 VARIABLES ENTERED

MULTIPLE CORRELATION COEFFICIENT... 0.7960303E 00
 (ADJUSTED FOR D.F.)..... 0.7330689E 00
 F-VALUE FOR ANALYSIS OF VARIANCE... 0.5920227E 01
 STANDARD ERROR OF ESTIMATE..... 0.2491345E 01
 (ADJUSTED FOR D.F.)..... 0.2730072E 01

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
7	0.271499E 00	0.701725E-01	0.386902E 01
17	-0.146036E-03	0.509755E-04	-0.165668E 01
15	-0.139726E-02	0.617104E-03	-0.313491E 01
10	0.401269E-01	0.212843E-01	0.172333E 01
20	0.344471E-04	0.273412E-04	0.146766E 01
CONSTANT	0.333326E 02		

X(1)=LONGITUDE INDEX
 X(2)=LATITUDE INDEX
 X(3)=AVERAGE ELEVATION - FT/10
 X(4)=SLOPE OF SQUARE (PERCENT/100)
 X(5)=DIRECTION OF SLOPE
 X(6)=SHORT.DIST.TO SFA -KM
 X(7)=DISTANCE NORTH KM.
 X(8)=DISTANCE N-E KM.
 X(9)=DISTANCE EAST KM.
 X(10)=BARRIER HEIGHT NORTH FT X 10
 X(11)=BARRIER HEIGHT N-E FTX10
 X(12)=BARRIER HEIGHT EAST FTX10
 X(13)=X(3)**2
 X(14)=X(6)**2
 X(15)=X(7)**2
 X(16)=X(8)**2
 X(17)=X(9)**2
 X(18)=X(10)**2
 X(19)=X(11)**2
 X(20)=X(12)**2
 X(21)=PRECIPITATION INCH/YEAR

TABLE OF RESIDUALS

CASE NO.	Y VALUE	Y ESTIMATE	RESIDUAL	PERCENT	STATION
1	33.410	35.976	-2.566	-7.7	7056067
2	34.670	36.053	-1.393	-4.0	7055120
3	35.700	37.106	-1.406	-3.9	7054640
4	35.460	35.787	-0.327	-0.9	7058560
5	45.010	40.452	4.558	10.1	7057304
6	38.630	39.084	0.546	1.4	7058220
7	36.200	36.425	-0.225	-0.6	7057720
8	36.550	40.270	-3.720	-10.2	7058520
9	33.520	35.578	-2.058	-6.1	7050760
10	40.110	40.065	0.045	0.1	7055770
11	41.900	40.318	0.342	2.3	7050463
12	40.300	35.925	4.375	10.9	7057395
13	36.260	35.447	0.813	2.2	7051200
14	37.970	39.022	-1.052	-2.8	7054690
15	38.940	39.576	-0.639	-1.6	7051120
16	45.080	43.643	1.437	3.1	7056120
17	43.020	42.724	0.296	0.7	7057160
18	39.460	37.756	1.704	4.3	7051040
19	38.140	39.730	-1.590	-4.2	7055330
20	31.530	34.730	-3.200	-10.2	7055200
21	37.030	32.882	4.148	11.2	7051160
22	37.240	35.274	1.966	5.4	7052805
23	36.730	37.976	-1.246	-3.4	7052600
24	40.200	40.944	-0.744	-1.9	7058840

FIG. 4.7

0	=	0	-	199
1	=	200	-	399
2	=	400	-	599
3	=	600	-	799
4	=	800	-	999
5	=	1000	-	1199
6	=	1200	-	1399
7	=	1400	-	1599
8	=	1600	-	1799
9	=	1800	-	1999
A	=	2000	-	2199
B	=	2200	-	2399
C	=	2400	-	2599
D	=	2600	-	2799
E	=	2800	-	2999
F	=	3000	-	3199
G	=	3200	-	3399
H	=	3400	-	3599
I	=	3600	-	3799
J	=	3800	-	3999
K	=	4000	-	4199
L	=	4200	-	4399
M	=	4400	-	4599
N	=	4600	-	4799
P	=	4800	-	4999
Q	=	5000	-	5199
R	=	5200	-	5399
S	=	5400	-	5599
T	=	5600	-	5799
U	=	5800	-	5999
V	=	6000	-	6199
W	=	6200	-	6399
X	=	6400	-	6599
Y	=	6600	-	6799
Z	=	6800	-	6999

FIG. 4.8

K FACTORS FOR RUNOFF STUDY

INITIAL RUNOFF R&P-E

PAGE 1

STATION	--RUNOFF (INCHES)--		(ACT./PRED.)
	ACTUAL	PREDICTED	
C11702	21.840	27.496	0.794
022301	22.460	26.612	0.844
022201	24.210	26.528	0.917
021203	23.510	26.753	0.879
021701	30.460	23.830	1.278
021601	29.270	26.194	1.117
C11201	29.400	31.318	0.942
C10201	26.100	31.493	0.832
020402	29.550	27.546	1.073
020401	26.280	28.745	0.914
020601	31.330	24.304	1.263

MADAWASKA AVAL BARRAGE LAC TEMISCOUATA
 TROIS-PISTOLES EN AMONT DU PONT-ROUTE 10
 PIMOUSKI EN AMONT DU PONT-ROUTE 10
 MITIS - CENTRALE DE PRICE
 BLANCHE EN AMONT DU PONT-ROUTE 6 ST-ULRIC
 MATANE EN AMONT DU PONT-ROUTE 6 A MATANE
 NOUVELLE AU PONT-ROUTE 6 A NOUVELLE
 PETITE RIVIERE CASCADES EN AMONT DU PONT CN
 MADELEINE EN AMONT DU PONT-ROUTE 6 RIV.MDL
 YORK EN AMONT DU PONT-ROUTE 6 A SUNNY-BANK
 DARTMOUTH EN AMONT DU PONT-ROUTE 6

FIG. 4.9

MADAWASKA AVAL BARRAGE LAC TEMISCOUATA

STEP 5

VARIABLE ENTERED 11

SUM OF SQUARES REDUCED IN THIS STEP.. 0.3146067E 03

PROPORTION REDUCED IN THIS STEP..... 0.3182869E-01

CUMULATIVE SUM OF SQUARES REDUCED.... 0.8011965E 04

CUMULATIVE PROPORTION REDUCED..... 0.8615027E 00 OF 0.9299996E 04

FOR 5 VARIABLES ENTERED

MULTIPLE CORRELATION COEFFICIENT... 0.9281717E 00

(ADJUSTED FOR D.F.)..... 0.9278275E 00

F-VALUE FOR ANALYSIS OF VARIANCE... 0.1077359E 04

STANDARD ERROR OF ESTIMATE..... 0.1219563E 01

(ADJUSTED FOR D.F.)..... 0.1222372E 01

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	0.508046E 00	0.145516E-01	0.349134E 02
19	-0.197578E-04	0.705305E-05	-0.280131E 01
14	-0.237472E-02	0.622379E-04	-0.381555E 02
9	0.616178E-01	0.278557E-02	0.221203E 02
11	0.218749E-01	0.150407E-02	0.145439E 02
CONSTANT	0.177060E 02		

TABLE OF RESIDUALS

CASE NO.	Y VALUE	Y ESTIMATE	RESIDUAL	PERCENT	STATION
1	33.410	37.564	-4.154	-12.4	7056067
2	34.670	35.182	-3.512	-10.1	7055120
3	35.700	41.762	-6.062	-17.0	7054640
4	35.460	34.772	0.688	1.9	7058560
5	45.010	36.280	8.730	19.4	7057304
6	38.630	35.978	2.652	6.9	7058220
7	36.200	32.277	3.923	10.0	7057720
8	36.550	37.883	-1.333	-3.6	7058520
9	33.520	35.745	-2.225	-6.6	7050760
10	40.110	37.681	2.429	6.1	7055770
11	41.900	38.657	3.243	7.7	7050463
12	40.330	41.457	-1.127	-2.9	7057395
13	35.260	42.033	-5.773	-15.9	7051200
14	37.970	44.151	-6.181	-16.3	7054660
15	38.940	41.831	-2.891	-7.4	7051120
16	46.080	37.565	8.515	18.5	7056120
17	43.020	39.067	3.953	9.2	7057160
18	39.460	44.237	-4.777	-12.1	7051040
19	38.140	40.951	-2.811	-7.4	7055380
20	31.530	44.075	-12.545	-39.8	7055200
21	37.030	45.735	-8.705	-23.5	7051160
22	37.240	43.127	-5.887	-15.9	7052865
23	36.730	40.102	-3.372	-9.2	7052600
24	40.200	41.220	-1.020	-2.5	7058840

X(1)=LONGITUDE INDEX

X(2)=LATITUDE INDEX

X(3)=AVERAGE ELEVATION - FT/10

X(4)=SLOPE OF SQUARE (PERCENT/100)

X(5)=DIRECTION OF SLOPE

X(6)=SHORT.DIST.TO SEA -KM

X(7)=DISTANCE NORTH KM.

X(8)=DISTANCE N-E KM.

X(9)=DISTANCE EAST KM.

X(10)=BARRIER HEIGHT NORTH FT X 10

X(11)=BARRIER HEIGHT N-E FTX10

X(12)=BARRIER HEIGHT EAST FTX10

X(13)=X(3)**2

X(14)=X(6)**2

X(15)=X(7)**2

X(16)=X(8)**2

X(17)=X(9)**2

X(18)=X(10)**2

X(19)=X(11)**2

X(20)=X(12)**2

X(21)=PRECIPITATION INCH/YEAR

FIG. 4.10

VARIABLE ENTERED 19

SUM OF SQUARES REDUCED IN THIS STEP.. 0.1651943E 03
PROPORTION REDUCED IN THIS STEP..... 0.1598887E-01

CUMULATIVE SUM OF SQUARES REDUCED.... 0.9566023E 04
CUMULATIVE PROPORTION REDUCED..... 0.9200890E 00 OF 0.103968E 05

FOR 4 VARIABLES ENTERED

MULTIPLE CORRELATION COEFFICIENT... 0.9592127E 00
(ADJUSTED FOR D.F.)..... 0.9590647E 00
F-VALUE FOR ANALYSIS OF VARIANCE... 0.2426542E 04
STANDARD ERROR OF ESTIMATE..... 0.9927543E 00
(ADJUSTED FOR D.F.)..... 0.9945172E 00

VARIABLE NUMBER	REGRESSION COEFFICIENT	STD. ERROR OF REG. COEFF.	COMPUTED T-VALUE
2	0.642980E 00	0.116162E-01	0.553520E 02
14	-0.229240E-02	0.442308E-04	-0.518280E 02
8	0.797321E-01	0.217398E-02	0.366755E 02
19	0.738285E-04	0.570254E-05	0.129466E 02
CONSTANT	-0.840724E 00		

X(1)=LONGITUDE INDEX
X(2)=LATITUDE INDEX
X(3)=AVERAGE ELEVATION - FT/10
X(4)=SLOPE OF SQUARE (PERCENT/100)
X(5)=DIRECTION OF SLOPE
X(6)=SHORT.DIST.TO SEA -KM
X(7)=DISTANCE NORTH KM.
X(8)=DISTANCE N-E KM.
X(9)=DISTANCE EAST KM.
X(10)=BARRIER HEIGHT NORTH FT X 10
X(11)=BARRIER HEIGHT N-E FTX10
X(12)=BARRIER HEIGHT EAST FTX10
X(13)=X(3)**2
X(14)=X(6)**2
X(15)=X(7)**2
X(16)=X(8)**2
X(17)=X(9)**2
X(18)=X(10)**2
X(19)=X(11)**2
X(20)=X(12)**2
X(21)=AREA LAKE/AREA LAND
X(22)=AREA FOREST/AREA LAND
X(23)=AREA HOGS/AREA LAND
X(24)=RUNOFF INCHES/YEAR

K FACTORS FOR RUNOFF STUDY FINAL RUNOFF FROM REGRESSION EQUATION

STATION	--RUNOFF(INCHES)--		(ACT./PRED.)
	ACTUAL	PREDICTED	
011702	21.840	21.685	1.007
022301	22.460	22.373	1.004
022021	24.210	23.620	1.025
021903	23.510	24.805	0.948
021701	30.460	26.865	1.134
021601	29.270	28.762	1.018
011201	29.690	29.374	1.004
010901	26.190	26.493	0.989
020802	29.550	29.942	0.987
020401	26.280	27.327	0.962
020601	31.330	29.705	1.055

WADAWASKA AVAL BARRAGE LAC TENISQUATA
TROIS-PISTOLES EN AMONT DU PONT-ROUTE 10
RIMOUSKI EN AMONT DU PONT-ROUTE 10
VITIS - CENTRALE DE PRICE
BLANCHE EN AMONT DU PONT-ROUTE 6 ST-ULRIC
MATANE EN AMONT DU PONT-ROUTE 6 A MATANE
NOUVELLE AU PONT-ROUTE 6 A NOUVELLE
PETITE RIVIERE CASCAPIEDIA AMONT DU PONT CH
MADELEINE EN AMONT DU PONT-ROUTE 6 RIV.MOL
YORK EN AMONT DU PONT-ROUTE 6 A SUNNY-BANK
DARTMOUTH EN AMONT DU PONT-ROUTE 6

FIG. 4.11

*/

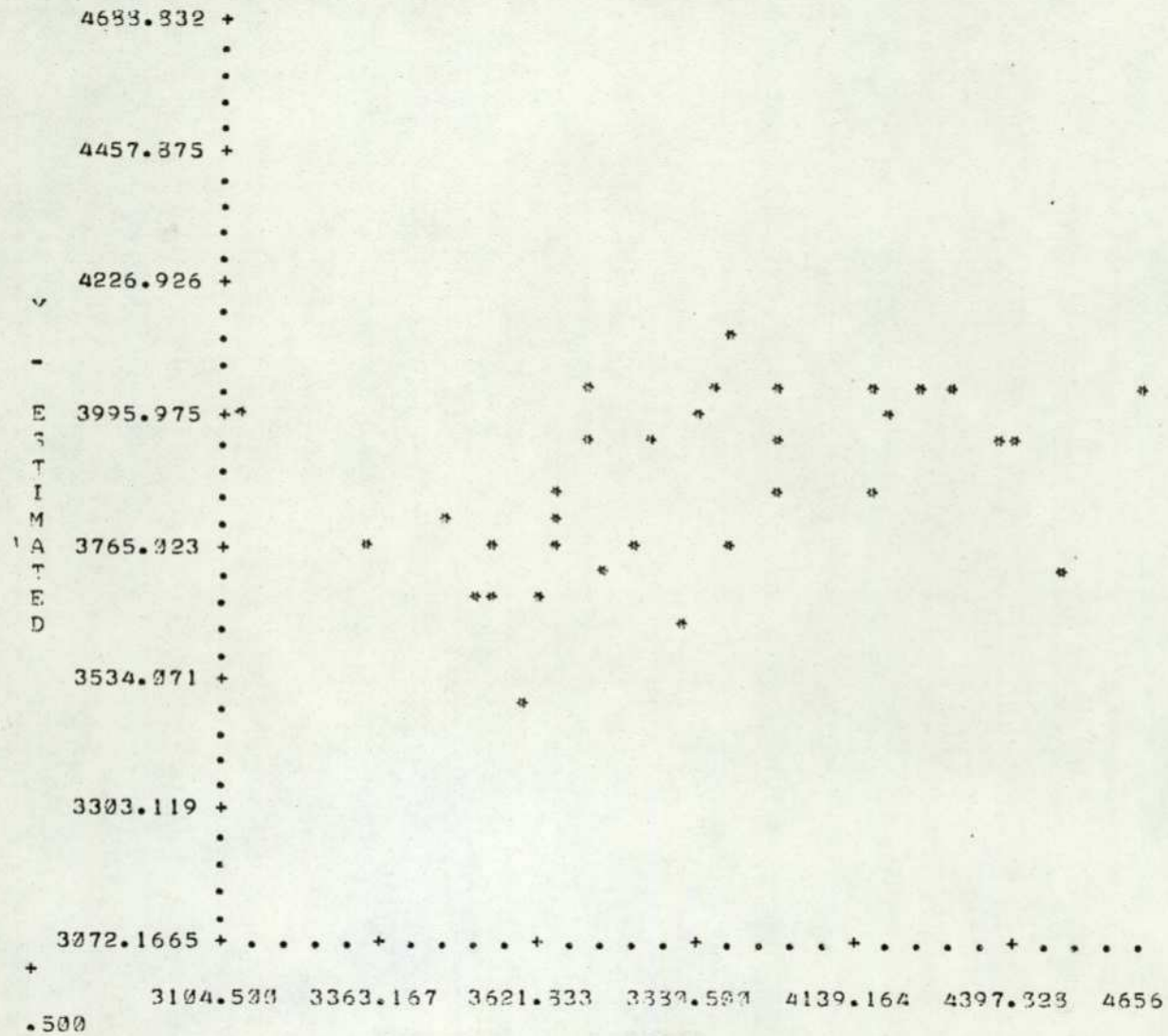
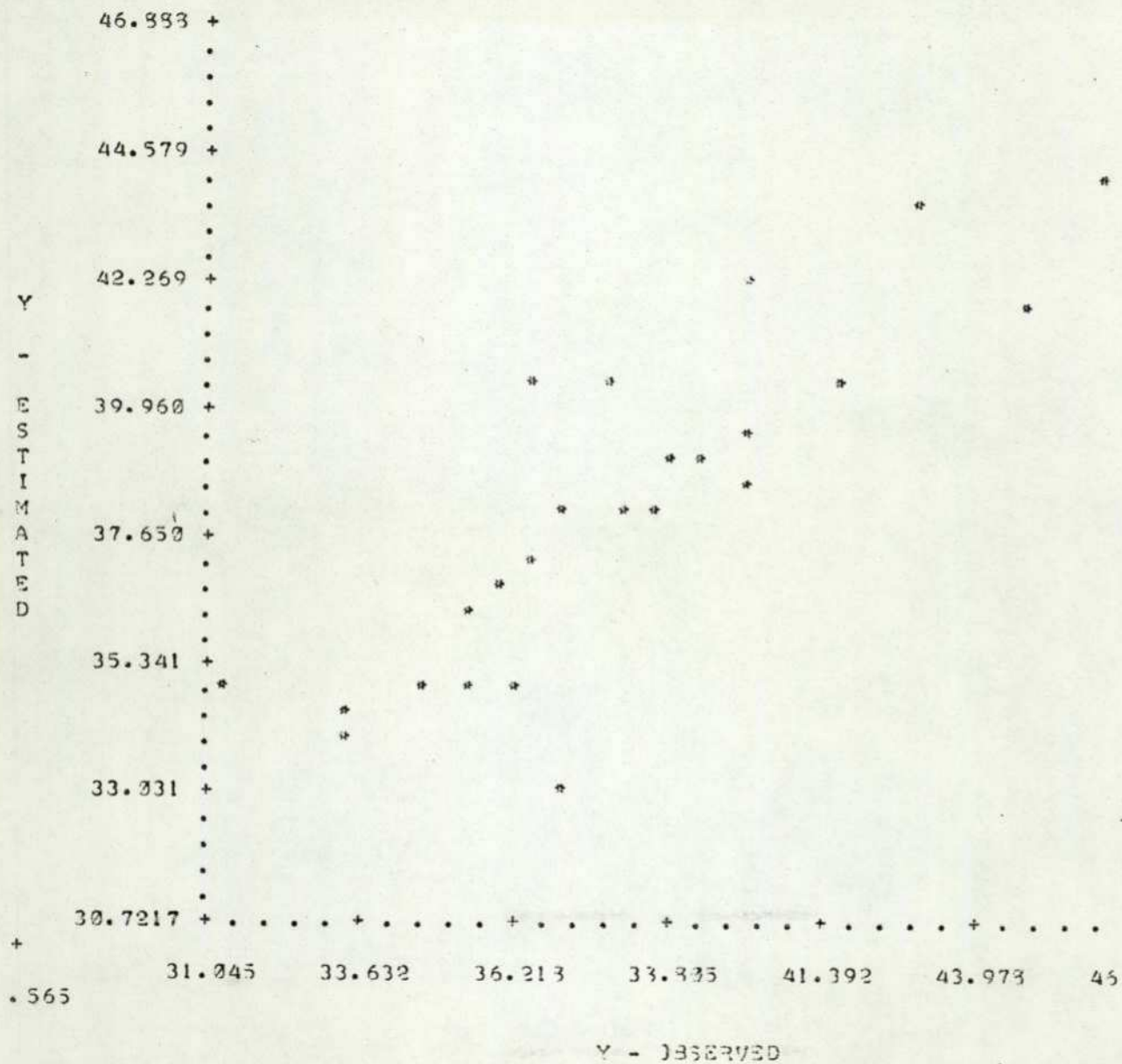
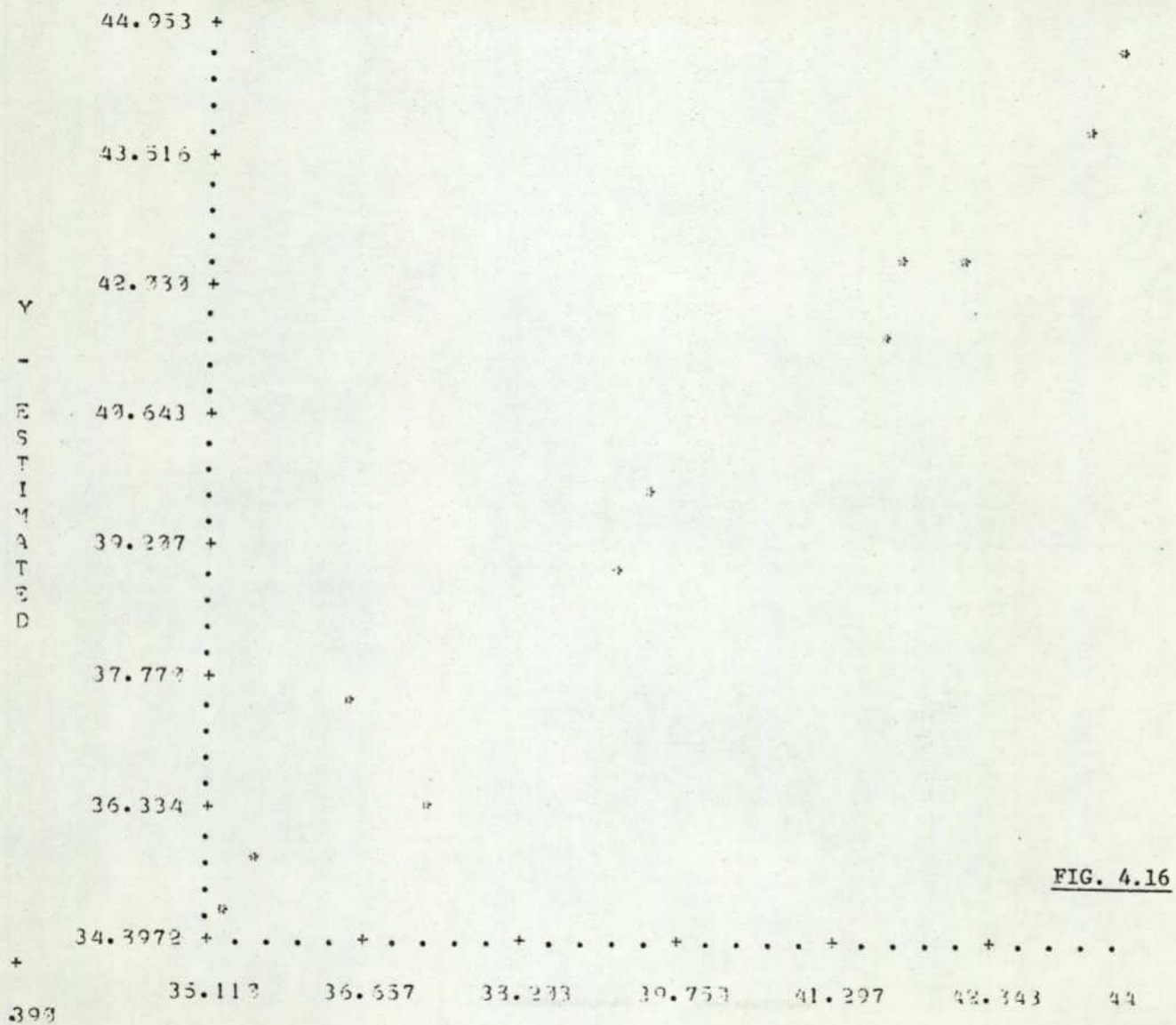


FIG. 4.14

* /



*/



*/

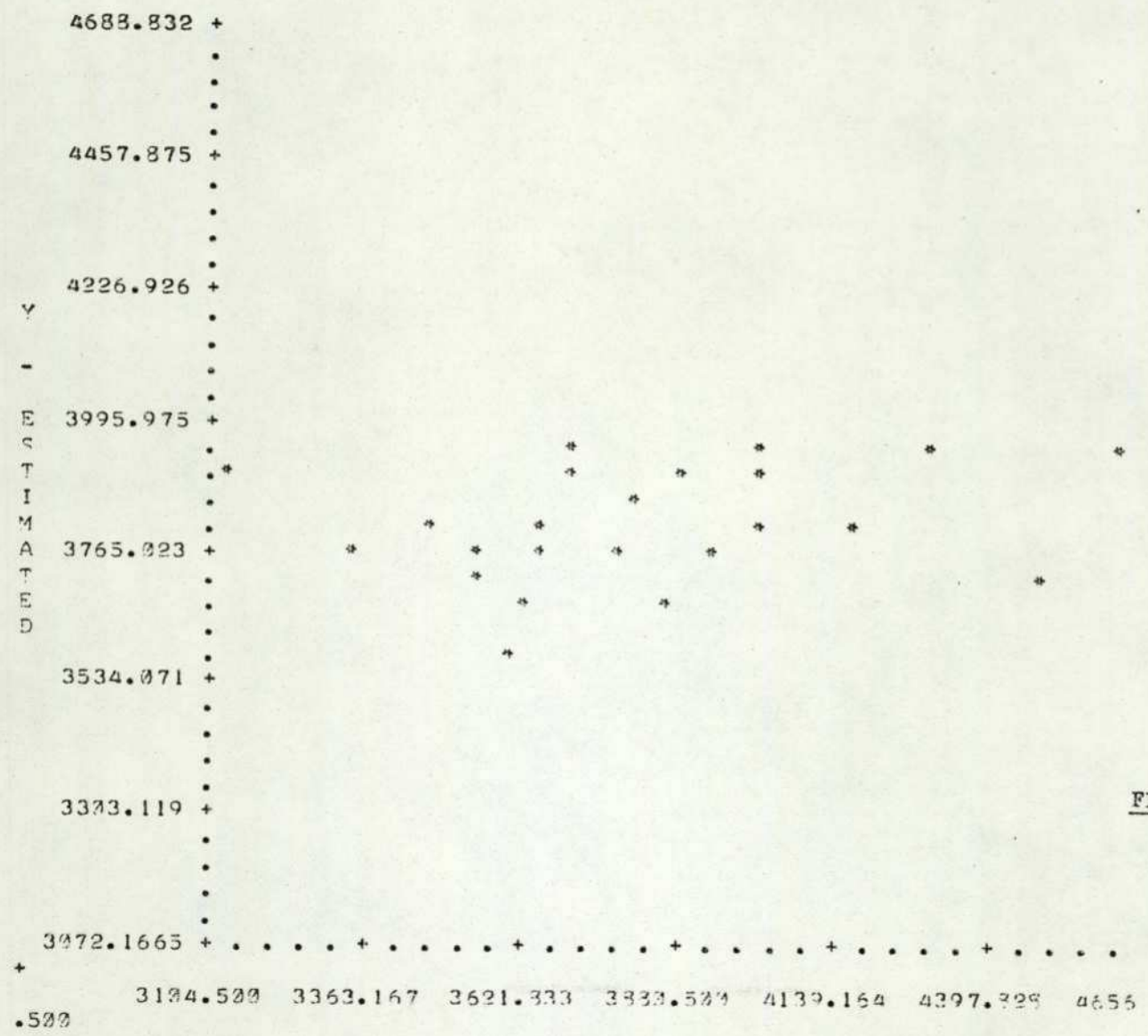
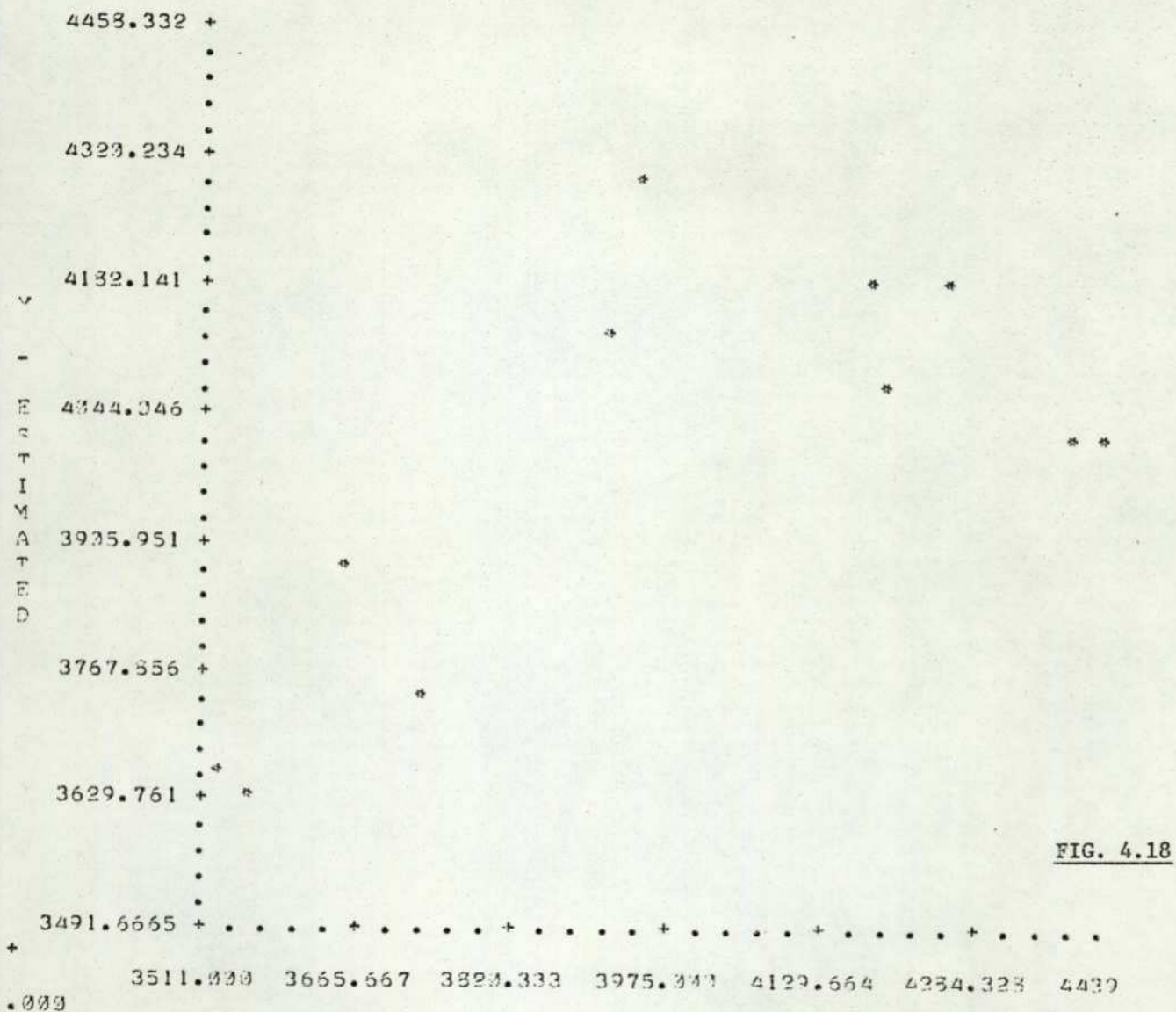


FIG. 4.17

Y - OBSERVED

*/



5. AGGREGATION PROPERTIES OF STOCHASTIC MODELS

5. AGGREGATION PROPERTIES OF STOCHASTIC MODELS

In this section we shall deal with the statistical properties of aggregated flows (See also table 2.2 of section 2.4.2). This problem can be defined in the context of monthly and annual flows but it is relevant in all situations where we are interested in sequences of short-term flows and also in the sequences of non-overlapping sums of the same flows.

The two main contributions to this problem in the hydrological literature are the papers by Moreau and Pyatt (1970) and Valencia and Schaaake (1972).

In the paper by Moreau and Pyatt the assumption is made that the four weekly flows follow a multivariate normal distribution. The generation model proposed is mathematically identical to the model proposed by Matalas (1967) and depends on a large number of parameters which are estimated by smoothing the observed series, using moving averages. This semi-empirical estimation of parameters is a disadvantage of the approach.

The paper by Valencia and Schaaake discusses a disaggregation model. The authors start by fitting a model to the long-term (e.g. annual) flows and then estimate a number of parameters which they use in disaggregating the annual flows already generated into a set of short-term flows which have the same properties as the corresponding historical flows. The main difficulty of this approach is again that a large number of parameters must

be estimated. For example, if a model has been fitted to the annual flows and we want to disaggregate the generated annual flows into monthly flows, we need to estimate a total of 156 parameters. It is apparent with the records usually available for hydrological studies it will be impractical to estimate such large numbers of parameters.

A knowledge of the statistical properties of aggregated flows is important in hydrology in two different situations:

1. In flow simulation we may be interested in reproducing properties of the long-term flows which are not necessarily derived from the record such as for example the Hurst property. In this case the model of the long-term flows as derived analytically from the short-term model is not adequate and we must incorporate the desired properties of the long-term flows in the estimation of the parameters of the short-term model.
2. The observations of certain variables of the hydrological cycle are sums of these variables over fixed periods of time. An example of such variable are the daily observations of precipitation by means of a non-recording gauge. In this case we may be interested in reproducing the properties of the unobserved process from the available aggregated observations. This problem is very common in Economics and therefore many references can be found in the economic literature.

5.1. Aggregation properties of univariate models

Before dealing with the problem in the hydrological context we consider the analytical derivation of the properties of aggregated observations given the stochastic model fitted to the short-term observations. The basic theoretical results concerning the aggregation problem can be found in Theil (1965), Telser (1967), Wu (1971), Theil (1971), Zellner and Montmarquette (1971), Amemiya and Wu (1972), Tiao (1972), Brewer (1973) and Tiao and Wei (1973). In this study we shall restrict our attention to the aggregation properties of auto-regressive models which are sufficient to establish the nature of the relationship between the properties of the short and long term observations.

The main result of interest in our analysis (Amemiya and Wu 1972) is that, if the short-term observations follow an autoregressive process of order p , non-overlapping sums of m terms follow a mixed autoregressive - moving average process of order (p,p) the parameters of which can be computed from those of the original process (See also Appendix 5.1). It appears therefore that if we have a sufficiently long record we can fit a stochastic model to the long-term flows and then test whether the aggregated values of the simulated flows satisfy this process. The successive steps of this procedure can be described as follows:

Step 1 Fit a stochastic model to the short-term flows.

Step 2 Determine analytically the form and the values of the parameters of the long-term flows.

- Step 3 Using the form of the model identified in Step 2 and assuming that the number of years of record is sufficient, estimate the parameters of the annual model using the observed annual values and determine their 95% confidence intervals.
- Step 4 If the model fitted to the short-term flows is adequate, we would expect that the values of the parameters derived in Step 2 are within the confidence intervals computed in Step 3.
- Step 5 If the two models are not compatible we must re-examine the model fitted to the short-term flows in Step I and try to take into account the properties of the observed annual flows in the estimation of the parameters of the short-term model. This can be done by using the methods of Valencia and Schaake or the mixed regression method described in the next sub-section.

5.2. Aggregation properties of transfer function models

In this study we have dealt extensively with the identification and estimation of transfer functions between hydrological variables. In this context we may be interested in the form of the transfer function between aggregated variables given the transfer function between the short-term variables. This problem has been studied by Theil (1965), Telser (1967) Theil (1971), and Tiao and Wei (1973). The main result of these studies (Tiao and Wei 1973) is that given that the transfer function between the short-term variables is a "dynamic regression" i.e. the direction of predictability is one sided (See also fig.2.6 of section 2), the transfer function between the respective non-overlapping aggregated variables includes feedback effects except in the case where the input variable consists of independent, identically distributed random variables. The result is discussed in more detail in Appendix I

This fact raises a fundamental question concerning the estimation of transfer functions between hydrological variables, which can be formulated as follows: If we know the direction of causality as for example in the case of precipitation and flow and we can assume that consecutive precipitation observations are independently distributed, then we can assert that the transfer function between any aggregated values of precipitation and flow is also a dynamic regression. On the other hand, if the input follows an autoregressive or moving average process as for example the relationship between the flows of a river at an upstream and a downstream site, the transfer function between aggregated flows includes feedback effects and therefore it cannot be estimated using the dynamic regression methodology.

5.3. The application of mixed regression to the aggregation problem

If the characteristics of the long-term flows cannot be derived from the corresponding characteristics of the short-term flows, we must use them in the estimation of the parameters of the model of the short-term flows. As an alternative to the method of Valencia and Schaake we may use the more parsimonious mixed regression method discussed in previous sections. This method enables us to use prior information in estimating the coefficients of a multiple regression. The application of this method is based on the fact that ARIMA (1,1) models have been shown to reproduce the Hurst property in a satisfactory way. Consequently this type of model is fitted to the annual flows and the values of

the parameters are used as stochastic constraints in the estimation of the parameters of the monthly model. The details of this computation are given in appendix 2.

It must be noted however that in hydrology we know very seldom anything more about the behavior of the annual flows than the information contained in the monthly flows. Consequently, the practical hydrologist does not have many opportunities to use combined methods of estimation of aggregation properties such as mixed regression or the method of Valencia and Schaake.

A HYDROLOGICAL EXAMPLE

We now apply the theory of flow aggregation discussed in this section to the modelling of monthly and annual flows of the Niagara River at Queenston Ontario from 1947 to 1970. The data are taken from the publication of the Department of the Environment, Ottawa, Canada, entitled "Discharge of Selected Rivers in Canada (1972)".

The monthly flows of the Niagara River at Queenston are shown in Table 5.1. Using these data the following calculations were carried out:

1. Computation of the monthly standardised flows w_t
2. Computations of the annual standardised flows $\bar{w}_t$
3. Fitting of ARIMA models to the flows w_t and $\bar{w}_t$ with the following results:

Standardised monthly flows: $w_t = 0.954w_{t-1} + a_t$ (Var $a_t = 0.089$)

Standardised annual flows: $\bar{w}_t = 0.602\bar{w}_{t-1} + a'_t + 0.469a'_{t-1}$ (Var $a'_t = 0.383$)

and the 95% confidence intervals for the coefficients of $\bar{w}_{t-1}$ and a'_{t-1} are (0.174, 1.00) and (0, 0.977) respectively.

4. Using the first order autoregressive process fitted to the monthly flows and the formulas given in Amemiya and Wu (1972) we compute the parameters of the annual model and the result of these computations is the following model:

$$\bar{w}_t = 0.56\bar{w}_{t-1} + a'_t + 0.121a'_{t-1} \quad (\text{Var } a'_t = 0.394)$$

We see the computed parameters are within the 95% confidence intervals of the parameters of the fitted annual model. Consequently the null hypothesis that the two processes are identical cannot be rejected.

Given the fitted and computed models do not differ significantly it is not necessary for this example to modify the computed values of the monthly parameters in order to take into account the properties of aggregated flows.

TABLE 5.1

NIAGARA RIVER AT QUEENSTON

Monthly and Annual Mean Discharges (m^3/s)

Year/ Année	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	Annual/ Annuel
1943	5520	5490	5860	6310	6600	6480	6290	6140	5830	5660	5720	5580	5950
1949	5720	5780	5750	5800	5800	5630	5550	5410	5380	5150	5100	5120	5520
1950	5580	5660	5750	5970	6200	6120	5860	5660	5580	5520	5660	5890	5780
1951	5830	5720	5890	6460	6570	6480	6290	6120	6000	5890	6030	6170	6120
1952	6480	6600	6790	7050	7130	6850	6600	6510	6460	6230	6030	6090	6570
1953	6060	6260	6290	6370	6400	6340	6290	6260	6120	5800	5800	6060	6170
1954	5580	5550	5970	6370	6510	6290	6030	6090	5970	6230	6340	6340	6120
1955	6460	6170	6570	6770	6740	6460	6170	6120	5860	5950	6090	5800	6260
1956	5270	5180	5440	5780	6310	6170	6000	6090	6060	5630	5630	5520	5750
1957	5460	5270	5350	5860	5920	5890	6000	5720	5630	5320	5490	5490	5610
1958	5410	5010	5100	5150	5180	5180	5270	5240	5180	5040	5070	4870	5150
1959	4700	4870	5210	5410	5610	5580	5290	5120	4950	5070	5210	5210	5180
1960	5520	5440	5270	5580	5920	6030	5950	5750	5660	5520	5630	5460	5630
1961	5120	5180	5460	5920	5400	6260	6000	5890	5830	5520	5380	5460	5690
1962	5010	4840	5270	5380	5520	5290	5180	5120	5040	5070	4870	5180	5150
1963	4700	4450	4980	5100	5270	5270	5010	4900	4560	4450	4590	4640	4810
1964	3960	3960	4620	4930	5180	5070	4810	4700	4590	4470	4330	4280	4590
1965	4360	4560	4810	4900	5100	5180	4980	4840	4810	4930	4980	4930	4870
1966	4980	5070	5270	5290	5630	5580	5440	5380	5180	5040	4950	5440	5270
1967	5380	5270	5180	5690	5860	5610	5580	5490	5320	5490	5630	5630	5520
1968	5460	6090	5800	5950	6000	6060	6060	5890	5800	5720	5610	5920	5860
1969	5660	5920	5890	6140	6710	6710	6600	6510	6290	6170	6060	6000	6230
1970	5580	5660	5720	6030	6200	6140	6140	5950	5970	5830	6000	6060	5950

APPENDIX 5.1

DERIVATION OF AGGREGATION PROPERTIES

1. Aggregation properties of univariate models

Following Amemiya and Wu (1972) we assume that the short-term observations Y_t follow the p^{th} order autoregressive process

$$(1) \quad \sum_{j=1}^p \phi_j y_{t-j} + a_t$$

where we assume as usual that the roots of the equation

$$(2) \quad \sum_{j=1}^p \phi_j z^j = 1$$

lie outside the unit circle (Box and Jenkins, 1970). If we denote

the roots of this equation by $\frac{1}{G_1}, \dots, \frac{1}{G_p}$, equation (1) can be written

$$(3) \quad (1 - G_1 B) \cdot (1 - G_2 B) \dots (1 - G_p B) y_t = a_t$$

where B is the backward shift operator such that $By_t = y_{t-1}$

Define now the aggregated variable by the equation:

$$(4) \quad Y_t = \sum_{j=0}^{m-1} y_{t-j}$$

The basic theorem of Amemiya and Wu (1972) is that the aggregated variable Y_t satisfies the mixed autoregressive - moving average process

$$(5) \quad (1 - G_1^m B^m) \dots (1 - G_p^m B^m) Y_t = e_t - \sum_{j=1}^q \phi_j e_{t-j^m}$$

where e_t are independent identically distributed random variables

with $Ee_t = 0$, $Ee_t^2 = \sigma_e^2$, the constants ϕ_j and σ_e^2

are determined as functions of ϕ 's and q is the largest

integer satisfying the inequality

$$(6) \quad qm < (p + 1) (m - 1) + 1$$

and consequently N is at most equal to p and if $m > p+1$ then $q = p$.

The formulas giving σ_c^2 and θ_j in terms are given in Amemiya and Wu (1972). Since these formulas are rather complicated they will not be reproduced here.

2. Aggregation in transfer function models

Following Tiao and Wei (1973) we consider the simple

dynamic regression model

$$(1) \quad y_t = \omega_0 x_t + \omega_1 x_{t-1} + u_t$$

where u_t are assumed independent and $N(0, \sigma_u^2)$

Define the aggregated variables

$$Y_T = \sum_{j=0}^{m-1} y_{mT-j}; \quad X_T = \sum_{j=0}^{m-1} x_{mT-j} \quad \text{where } t = mT$$

Further let

$$W_T = \sum_{j=0}^{m-1} x_{mT-1-j}$$

$$E_T = \sum_{j=0}^{m-1} e_{mT-j}$$

Summing the terms of (1) we obtain

$$Y_T = \omega_0 X_T + \omega_1 W_T + U_T$$

where the U_t are independent and $N(0, m \sigma_u^2)$.

Tiao and Wu studied the following cases:

- (1) $x_t = a_t$ (x_t are independent random variables)
- (2) $x_t = a_t - \phi a_{t-1}$ (x_t follow a MA process)
- (3) $x_t = \phi x_{t-1} + a_t$ (x_t follow an AR process)

- Case 1 (1) X_t are a series of normal variables $N(0, \sigma_a^2)$
- (2) Y_t can be represented by a univariate moving average model of order 1.
- (3) Y_t depends only on X_t and X_{t-1} and therefore the relationship can be described as a dynamic regression as the basic model (1).
- (4) The residuals U_t are not uncorrelated and follow a univariate moving average process of order one.
- Case 2 (1) The series X_t and Y_t follow univariate moving average process of order one and these series are dependent.
- (2) Unlike the basic model and the previous case there does not exist a dynamic regression between X_{t-j} and Y_t for $j=0,1,2,\dots$ and therefore the transfer function between these two variables includes feedback from Y_t to X_t .
- Case 3 (1) X_t is an ARMA (1,1) process
- (2) Same remark as for Case 2.

APPENDIX 5.2

MIXED REGRESSION APPLIED TO THE AGGREGATION PROBLEM

In order to establish the notation we begin by arranging the short-term flows in a $M \times N$ table where M = number of short-term periods and N = number of long-term periods.

	1				...	n			...	N
s_1										
s_m						y_{mn}				
s_M										
$\bar{s}$										
						$\bar{y}_n$				$\bar{y}$

where $t = 1, 2, \dots, MN$ = Index for short-period flows
numbered sequentially

$m = 1, 2, \dots, M$ = Index for short-period flows
(for monthly flows $M = 12$)

$n = 1, 2, \dots, N$ = Index for long-period flows
(e.g. annual)

Let the standardized flow be defined by: $w_{mn} = \frac{y_{mn} - \bar{y}_m}{s_m}$

We have,

$$\bar{y}_n = \frac{1}{M} \sum y_{mn} = \frac{1}{M} \sum_{m=1}^M (\bar{y}_m + s_m w_{mn})$$

$$= \frac{1}{M} \sum \bar{y}_m + \frac{1}{M} \sum s_m \cdot \frac{1}{\sum s_m} \sum s_m w_{mn}$$

$$\text{or } \bar{y}_n = \bar{y} + \bar{s} \cdot \bar{w}_n$$

where the symbols in the last equation are defined as follows

$\bar{y}_n$ = Mean flow for the n^{th} year

$\bar{y}$ = Long term mean flow

$\bar{s}$ = Mean of the monthly standard deviations

$\bar{w}_n$ = Weighted average of the standardized monthly flows

Assume that we have fitted a mixed ARIMA process to the weighted average of the standardized monthly flows:

$$(1) \quad \bar{w}_n = \phi \bar{w}_{n-1} + A_n - Q A_{n-1}$$

where $\bar{w}_n$ is defined by:

$$\bar{w}_n = \frac{\bar{y}_n - \bar{y}}{\bar{s}} = \frac{\sum_m s_m w_{mn}}{\sum_m s_m}$$

and we want to fit a mixed process to the standardized monthly flows

$$(2) \quad w_{mn} = \phi w_{(m-1)n} + a_{mn} - \theta a_{(m-1)n}$$

Note that if we have fitted the process (1) to the weighted average of the standardized monthly flows $\bar{w}_n$, we have

$$\bar{y}_n = \bar{y} + \bar{s} \bar{w}_n = \bar{y} + \bar{s}(\phi \bar{w}_{n-1} + A_n - Q A_{n-1}) \quad \text{or}$$

$$(3) \quad \bar{y}_n - \bar{y} = \phi(\bar{y}_{n-1} - \bar{y}) + \bar{s} A_n - Q \bar{s} A_{n-1}$$

Equation (3) shows that the centered annual flows $(\bar{y}_n - \bar{y})$ follow a mixed ARIMA process the characteristics of which can be computed from those of the process fitted to the quantity $\bar{w}_n$.

Solving (2) for a_{mn} we obtain

$$(4) \quad a_{mn}(\phi, \theta) = w_{mn} - \phi w_{(m-1)n} + \theta a_{(m-1)n}(\phi, \theta)$$

The iterative estimation of the parameters ϕ and θ starts with an initial pair of values (ϕ_o, θ_o) . Expanding the function $a_{mn}(\phi, \theta)$ in a Taylor series and retaining only the linear terms we obtain (Appendix 2.1):

$$(5) \quad a_{mn}(\phi, \theta) = a_{mn}(\phi_o, \theta_o) + (\phi - \phi_o) \frac{\partial a_{mn}}{\partial \phi} + (\theta - \theta_o) \frac{\partial a_{mn}}{\partial \theta}$$

Solving for $a_{mn}(\phi_o, \theta_o)$ we have:

$$(6) \quad a_{mn}(\phi_o, \theta_o) = (\phi - \phi_o) \left(-\frac{\partial a_{mn}}{\partial \phi} \right) + (\theta - \theta_o) \left(-\frac{\partial a_{mn}}{\partial \theta} \right) + a_{mn}(\phi, \theta)$$

or substituting the values of the partial derivatives:

$$(7) \quad a_{mn}(\phi_o, \theta_o) = (\phi - \phi_o) w_{(m-1)n} + (\theta - \theta_o) \cdot \left[-a_{(m-1)n}(\phi_o, \theta_o) \right] + a_{mn}(\phi, \theta)$$

This can be considered as a multiple regression equation with dependent variable $a_{mn}(\phi_o, \theta_o)$ and independent variables

$$w_{(m-1)n} \text{ and } \left[-a_{(m-1)n}(\phi_o, \theta_o) \right]$$

We shall now write equation (7) for M consecutive months and multiply it with the appropriate weights in order to form the weighted average $\bar{w}_n$:

YEAR	MONTH	MULTIPLIER	EQUATION
n	2	$s_1 / \Sigma s$	$a_{2n}(0) = (\phi - \phi_o) w_{1n} + (\theta - \theta_o) (-a_{1n}(0)) + a_{2n}$
	$\vdots$		
	m	$s_{m-1} / \Sigma s$	$a_{mn}(0) = (\phi - \phi_o) w_{(m-1)n} + (\theta - \theta_o) (-a_{(m-1)n}(0)) + a_{mn}$
	$\vdots$		
	M	$s_M / \Sigma s$	$a_{Mn}(0) = (\phi - \phi_o) w_{(M-1)n} + (\theta - \theta_o) (-a_{(M-1)n}(0)) + a_{Mn}$
n+1	1	$s_m / \Sigma s$	$a_{1(n+1)}(0) = (\phi - \phi_o) w_{Mn} + (\theta - \theta_o) (-a_{Mn}(0)) + a_{1(n+1)}$

Multiplying and summing the M equations we obtain:

$$\begin{aligned}
 (8) \quad & \frac{1}{\Sigma s} \sum_{m=2}^M s_{m-1} a_{mn}(0) + \frac{1}{\Sigma s} s_M a_{1(n+1)}(0) = \\
 & = (\phi - \phi_o) \sum_{m=1}^M \frac{s_m w_{mn}}{\Sigma s} + (\theta - \theta_o) \sum_{m=1}^M \frac{s_m (-a_{mn}(0))}{\Sigma s} + \\
 & + \frac{1}{\Sigma s} \sum_{m=2}^{M-1} s_{m-1} a_{mn} + \frac{1}{\Sigma s} s_M a_{1(n+1)}
 \end{aligned}$$

Letting:

$$(9) \quad C_{no} = \frac{1}{\Sigma s} \left[\sum_{m=1}^{M-1} s_m a_{(m+1)n}(0) + s_M a_{1(n+1)}(0) \right]$$

$$(10) \quad D_{no} = \frac{1}{\Sigma s} \sum_{m=1}^M s_m (-a_{mn}(0))$$

$$(11) \quad F_n = \frac{1}{\Sigma s} \left[\sum_{m=2}^{M-1} s_{m-1} a_{mn} + s_M a_{1(n+1)} \right]$$

equation (8) can be written:

$$(12) \quad C_{no} = (\phi - \phi_o) \bar{w}_n + (\theta - \theta_o) D_{no} + F_n$$

The same equation for the period (n-1) is:

$$(13) \quad C_{(n-1)o} = (\phi - \phi_o) \bar{w}_{n-1} + (\theta - \theta_o) D_{(n-1)o} + F_{n-1}$$

Multiplying (13) by ϕ and subtracting from (12) we obtain:

$$\begin{aligned}
 (14) \quad & C_{no} - \phi C_{(n-1)o} = (\phi - \phi_o) \left[\bar{w}_n - \phi \bar{w}_{n-1} \right] + (\theta - \theta_o) \left[D_{no} - \phi D_{(n-1)o} \right] \\
 & + F_n - \phi F_{n-1}
 \end{aligned}$$

or

$$(15) \quad Y_n = (\phi - \phi_o) \cdot X_{1n} + (\theta - \theta_o) \cdot X_{2n} + V_n$$

where

$$(16) \quad Y_n = C_{no} - \phi C_{(n-1)o}; \quad X_{1n} = \bar{w}_n - \phi \bar{w}_{n-1}; \quad (16)$$

$$X_{2n} = D_{no} - \phi D_{(n-1)o}; \quad V_n = F_n - \phi F_{n-1}$$

We can now formulate the problem of fitting a mixed ARIMA process to the monthly flows taking into account the properties of the annual flows, in the following way:

Given the initial values ϕ_o, θ_o , we estimate the corrections $(\phi - \phi_o)$ and $(\theta - \theta_o)$ from the regression equation (7) taking into account the $n-1$ constraints (15) by the method of mixed regression described in Chapter 2.

The details of the computations are as follows:

Given the initial values (ϕ_o, θ_o) we can use equation (4) to compute the residuals

$$a_{mn}(\phi_o, \theta_o) = w_{mn} - \phi_o w_{(m-1)n} + \theta_o a_{(m-1)n}(\phi_o, \theta_o)$$

for $m = 1, 2, \dots, M$ and $n = 1, 2, \dots, N$. and then compute C_{no} and D_{no} and the quantities Y_n and X_{2n} which appear in the constraints (15).

The quantity X_{1n} can be computed directly from the weighted average $\bar{w}_n$ of the standardised monthly flows. In order to compute the mean and variance of this quantity we use the fact that

$$X_n = \bar{w}_n - \phi \bar{w}_{n-1} = A_n - QA_{n-1}$$

where ϕ and Q are known quantities (parameters of a fitted process) and A_n is a series of random variables with mean zero and known variance σ_A^2 . We have $E(x_n) = 0$ and:

$$(17) \text{ Var } X_{1n} = \text{Var } (A_n - QA_{n-1}) = (1 + Q^2) \sigma_A^2$$

$$(18) \text{ Cov } (X_{1n}, X_{1(n-1)}) = -Q \sigma_A^2$$

The random term V_n of equation (15) has expected value, variance and autocovariance given by:

$$E(V_n) = E(F_n - \phi F_{n-1}) = 0$$

$$(19) \text{ Var}(V_n) = (1 + \phi^2) \text{Var } F_n = (1 + \phi^2) \sigma_F^2$$

$$(20) \text{ Cov}(V_n, V_{n-1}) = -\phi \sigma_F^2$$

and therefore the covariance matrix of the residuals V_n is given by:

$$(21) \quad \sigma_F^2 \cdot V = \sigma_F^2$$

	1	2	...	...	N - 1	
	$1 + \phi^2$	$-\phi$	0	0	0	1
	$-\phi$	$1 + \phi^2$	$-\phi$	0	0	2
	0	$-\phi$	$1 + \phi^2$	$-\phi$	0	...
	0	0	$-\phi$	$1 + \phi^2$	$-\phi$	...
	0	0	0	$-\phi$	$1 + \phi^2$	N - 1

Instead of inverting the matrix V we shall use the approximate inverse of V given in Kendall's "The Advanced Theory of Statistics", Volume III, Page 474:

$$V^{-1} \approx \frac{1}{1-\phi^2} \begin{array}{|c|c|c|c|c|} \hline 1 & \phi & \phi^2 & \dots & \phi^{N-2} \\ \hline \phi & 1 & \phi & \phi^2 & \dots \\ \hline \phi^2 & \phi & 1 & \phi & \phi^2 \\ \hline \dots & \phi^2 & \phi & 1 & \phi \\ \hline \phi^{N-2} & \dots & \phi^2 & \phi & 1 \\ \hline \end{array}$$

We can now summarise the sequence of the computations in the following way:

- Step (1) Estimate model (1) for $\bar{w}_n$ (Parameters ϕ, Q, σ_A^2)
- (2) Estimate model (2) for w_{mn} (Parameters $\phi_0, \theta_0, \sigma_0^2$)
- (3) Using ϕ, Q compute X_{1n} and V from (16) and (21)
- (4) Using the initial values ϕ_0, θ_0 compute the residuals a_0 from (4)
- (5) Compute C_{no}, D_{no} from (9) and (10)
- (6) Compute Y_n, X_{2n} from (16)
- (7) Form constraints (15) using Y_n, X_{1n}, X_{2n}, V
- (8) Estimate $(\phi_1 - \phi_0), (\theta_1 - \theta_0)$ using constraints alone
- (9) IS CONVERGENCE CRITERION SATISFIED?
- (10) If YES Print ϕ_1, θ_1
- (11) If NO Using ϕ_1, θ_1 compute residuals a_1 from (4)

- (12) Compute C_{nl} , D_{nl} from (9) and (10)
- (13) Compute Y_n , X_{2n} from (16)
- (14) Form constraints (15) using Y_n , X_{1n} , X_{2n} , V
- (15) Estimate the corrections $(\phi_2 - \phi_1)$, $(\theta_2 - \theta_1)$

from the regression equation (7)

$$a_{mn}(\phi_1, \theta_1) = (\phi_2 - \phi_1)w_{(m-1)n} + (\theta_2 - \theta_1) \left[-a_{(m-1)n}(\phi_1, \theta_1) \right]$$

$$+ a_{mn}(\phi, \theta)$$

and the constraint (15)

$$Y_n = (\phi_2 - \phi_1)X_{1n} + (\theta_2 - \theta_1)X_{2n} + V_n$$

- (16) Go to step (9) and repeat steps 9 - 16 as many times as necessary to obtain convergence.

Appendix 5.3

A Computer Program for Mixed Regression

**This content has been removed for
copyright protection reasons**

REFERENCES

REFERENCES

1. Aitken A.C. (1934)
On least squares and linear combinations of observations.
Proc. Royal Soc. Edinburgh.
Vol. 55, pp. 42-48, 1934.
2. Amemiya T. and R.Y. Wu (1972)
The effect of aggregation on prediction in the
autoregressive model.
Journal of the American Statistical Association.
Vol. 67, No. 339, September 1972.
3. Barnard G. (1956)
Discussion of the paper by Hurst: Methods of using
long-term storage in reservoirs.
Institution of Civil Engineers Proc.
Vol. 5, No. 5, pp. 552-553.
4. Barton B.M.J. (1973)
Synthetic monthly run-off records for ungauged
British catchments.
Journal of the Institution of Water Engineers.
Vol. 27, No. 7, October 1973
5. Brewer K.R.W. (1973)
Some consequences of temporal aggregation and systematic
sampling for ARMA and ARMAX models.
Journal of Econometrics.
6. Beard L.R. (1965)
Use of interrelated records to simulate streamflow.
Am. Soc. Civil Engineers Proc.
Vol. 91, No. HY5, pp. 13-22.
7. Blum J.R. (1954)
Approximation methods which converge with probability one.
Annals of Math. Stat., 25 (1954).
8. Box G.E.P. and J.M. Jenkins (1970)
Time series analysis, forecasting and control.
California, Holden Day, p. 553.
9. Cavadias G. (1966)
River flow as a stochastic process.
Proceedings, Symposium No. 5, National Research
Council of Canada (1966).
10. Clarke R.T. (1973)
A review of some mathematical models used in hydrology.
Journal of Hydrology 19 (1973) pp. 1-20.

11. Cormary Y. and A. Guilbot
The adjustment of deterministic models and the setting
of their basic parameters.
La Houille Blanche, No. 2-1971.
12. Crawford N.H. and R.K. Linsley (1964)
A conceptual model of the hydrologic cycle.
In Publication de l'A.I.H.S., no. 63 (1964).
13. Dansgaard W., S.J. Johnson, N. Reeh, N. Gundestrup, H.B. Clausen
and C.U. Hammer.
Climatic changes, Norsemen and modern man.
Nature, Vol. 255, May 1, 1975.
14. Dawdy D.R. and R.W. Lichty (1968)
Methodology of hydrologic model building. Symposium
of uses of digital computers in hydrology.
Tucson, Arizona, 1968.
IAASH 81: 347-355.
15. Dawdy D.R. and T. O'Donnell (1965)
Mathematical models of catchment behavior.
Journal of the Hydraulics Division.
Proc. A.S.C.E., 91, no. HY4 (1965)
16. Delleur J.W. and R.A. Rao (1971)
Linear systems analysis in hydrology.
The transform approach, the kernel oscillations and
the effect of noise.
Systems approach to hydrology.
V. Yevjevich Ed.
Fort Collins, 1971.
17. Dooge J.C.I. (1972)
Mathematical models of hydrologic systems.
International symposium on mathematical modelling techniques.
Ottawa, 1972.
18. Draper N. and H. Smith (1966)
Applied regression analysis.
J. Wiley and Sons, New York, 1966.
19. Fiering M.B. (1967)
Streamflow synthesis.
Massachusetts, Harvard University Press, p. 139.
20. Freeze R.A. and R.L. Harlan (1969)
Blueprint for a physically based, digitally simulated
response model.
Journal of Hydrology, Vol.9, 1969.

21. Garcia D., D.R. Dawdy and J.M. Mejia
Long memory monthly streamflow simulation by a broken
line model.
Water Resources Research, Vol. 8, No. 4, August 1972.
22. Garrison J.M., J. Granju and J.T. Price (1969)
Unsteady flow simulation in rivers and reservoirs.
Proceedings A.S.C.E.
Journal of Hydr. Div., September 1969.
23. Granger C.W.J. (1969)
Investigating causal relations by econometric models
and cross-spectral methods.
Econometrica 37, 424-438.
24. Hamlin M.J. (1971)
A study of synthetic flow generation techniques using
Elan valley data.
Journal of the Institution of Water Engineers.
Vol. 25, No. 7, October 1971.
25. Harbaugh T.E. (1967)
Numerical techniques for spatially varied unsteady flow.
University of Missouri Water Resources Center.
Report No. 3, 1967.
26. Haugh L. (1972)
The identification of time series interrelationships
with special reference to dynamic regression models.
University of Wisconsin, Ph.D. Thesis, 1972.
27. Hazen A. (1914)
Storage to be provided in impounding reservoirs for
municipal water supply.
Am. Soc. Civil Engineers Trans., Vol. 77, pp. 1539-1640.
28. Hurst H.E. (1951)
Long-term storage capacities of reservoirs.
Am. Soc. Civil Engineers Proc., Vol. 5, No. 5, pp. 519-543.
29. Isbister J. (1966)
Geohydrology of northeastern Nassau county, Long
Island, New York.
U.S. Geol. Survey, Water Supply paper 1825.
30. Johnston J. (1972)
Econometric methods.
2nd edition, McGraw-Hill, New York, 1972.
31. Kendall M.G. (1961)
A course in multivariate analysis.
Griffin's statistical monographs.
London 1961.

32. Kiefer J. and J. Wolfowitz (1952)
Stochastic estimation of the maximum of a regression function.
Annals of Math. Stat., 23 (1952).
33. Korfage H. (1974)
The transfer function model as a forecasting tool.
M.B.A. Thesis, McGill University, 1974.
34. Kottegoda N.T. (1972)
Stochastic five daily streamflow model.
Journal of the Hydraulics Division.
Proceedings of the American Society of Civil Engineers. September 1972.
35. Lavi A. and T.P. Vogl (1965)
Recent advances in optimization techniques.
Wiley 1965.
36. Linsley R.K, M.A. Kohler and J.L.H. Paulhus (1975)
Hydrology for Engineers.
Second edition. McGraw-Hill, 1975.
37. Mandelbrot B.B. and J.W. Van Ness (1968)
Fractional Brownian motions, fractional gaussian noises and applications.
Siam Review of Applied Mathematics, Vol. 10, No. 4, pp. 422-437.
38. Matalas N.C. (1967)
Mathematical assessment of synthetic hydrology.
Water Resources Research.
Vol. 3, No. 4, pp. 937-945, 1967.
39. Megerian E. and R. Pentland
Simulation of Great Lakes Basin Water Supplies.
Water Resources Research.
Vol. 4, No. 1, February 1968.
40. Mejia J.M., I. Rodriguez-Iturbe, and D.R. Dawdy (1972)
Streamflow simulation, 2. The broken line process as a potential model for hydrologic simulation.
Water Resources Research.
Vol. 8, No. 4, August 1974.
41. Moran P.A.P. (1959)
The theory of storage.
London, Methuen, p. 111.
42. Moreau D.H. and E.E. Pyatt (1970)
Weekly and monthly flows in synthetic hydrology.
Water Resources Research.
Vol. 6, No. 1, pp. 53-61, 1970.

43. Nash J.E. (1959)
Systematic determination of unit hydrograph parameters.
Journal of Geophysical Research.
Vol. 64, pp. 111-115 (1959)
44. Nash J.E. (1969)
Conceptual models of the runoff process.
Proceedings of IASH symposium on floods.
Rome 1969.
45. O'Connell P.E. (1971)
A simple stochastic modelling of Hurst's Law.
International symposium on mathematical models in
hydrology.
Warsaw, 1971.
46. Pierce D.A. and L.D. Haugh
The assessment and detection of causality in temporal
systems.
Unpublished report, May 1975.
47. Quenouille M.H. (1957)
The analysis of multiple time series.
Griffins Statistical monographs.
Charles Griffin and Co., London 1957.
48. Quimpo R.G. (1967)
Stochastic models of daily rivers flow sequences.
Hydrology Paper No. 18.
Colorado State Univ., p. 30.
49. Rivière Mégiscane, Révision des débits journaliers de 1924
à 1965.
Ministère des Richesses Naturelles.
50. Rivière Bell, Révision des débits journaliers de 1914 à 1968.
Ministère des Richesses Naturelles, Gouvernement du
Québec, 1969.
51. Roche M. and M. Slivitzky (1967)
Modèle mathématique pour une crue de fonte de neige.
In Cahiers O.R.S.T.O.M. série hydrologie, Vol. IV,
no. 1 (1967).
52. Roesner L.A. and V. Yevjevich (1966)
Mathematical models for time series of monthly precipitation
and monthly runoff.
Hydrology Paper No. 15, Colorado State Univ., p. 56.
53. Sittner W.T., C.E. Schauss and J.C. Monro (1969)
Continuous hydrograph synthesis with an API-type
hydrologic model.
Water Resources Res., Vol. 5, pp. 1007-1022,
October 1969.

54. Solomon S.I., J.P. Denouivilliez, E.J. Chart, J.A. Woolley and C. Cadou (1968)
The use of the square grid system for computer estimation of precipitation, temperature and runoff.
Water Resources Research 4(5): 919-929.
55. Stephenson G.R. and R.A. Freese
Mathematical simulation of subsurface flow contributions to snowmelt runoff, Reynolds Creek Watershed, Idaho.
Water Resources Research, Vol. 10, No. 2, April 1974.
56. Sudler C.E. (1927)
Storage required for the regulation of streamflow.
Am. Soc. Civil Engineers Trans., no. 91. pp. 622-704.
57. Telser L.G. (1967)
Discrete samples and moving sums in stationary stochastic processes.
Journal of the American Statistical Association.
pp. 484-499, June 1967.
58. Theil H. (1973)
On the use of incomplete prior information in regression analysis.
Journal of the American Statistical Association
Vol. 58, 401-414, 1963.
59. Theil H. (1971)
Principles of Econometrics.
North Holland Publishing Company, Amsterdam-London, 1971.
60. Thomas H.A. and M.B. Fiering (1962)
Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation.
In Design of Water Resources System.
Harvard University Press, pp. 459-493.
61. Tiao G.C. and W.S. Wei (1973)
Effect of temporal aggregation on a simple dynamic time series model.
Paper presented at the 1973 Annual Conference of the American Statistical Association.
New York, December, 1973.
62. Valencia D.R. and J.C. Schaake (1972)
A disaggregation model for time series analysis and synthesis.
Ralph Parsons laboratory for Water Resources and Hydrodynamics, Report No. 149, M.I.T., 1972.
63. Ven Te Chow (1959)
Open channel hydraulics.
McGraw-Hill 1959.

64. Wu R.Y. (1971)
The implication of linear aggregation on predictions
from econometric models.
Unpublished Ph.D. Thesis, Stanford University, 1971.
65. Young P.C. and P. Whitehead (1975)
A recursive approach to time series analysis for
multivariate systems.
Computer simulation of Water Resources Systems.
G.C. Vansteenkiste (Ed.)
North Holland Publishing Company, Amsterdam
Oxford (1975)
66. Yevjevich V. (1963)
Fluctuations of wet and dry years. Part I.
Research data assembly and mathematical models.
Hydrology Paper No. 1, Colorado State Univ., p. 55.
67. Zellner A. and C. Montmarquette (1971)
A study of some aspects of temporal aggregation
problems in econometric analysis.
Rev. Econ. Stat. 53, 335-342.