

Developments in Atmospheric Science, 13

STATISTICAL CLIMATOLOGY

Further titles in this series

1. F. VERNIANI (Editor)
Structure and Dynamics of the Upper Atmosphere
2. E.E. GOSSARD and W.H. HOOKE
Waves in the Atmosphere
3. L.P. SMITH
Methods in Agricultural Meteorology
4. O. ESSENWANGER
Applied Statistics in Atmospheric Science
5. G.W. PALTRIDGE and C.M.R. PLATT
Radiative Processes in Meteorology and Climatology
6. P. SCHWERDTFEGER
Physical Principles of Micro-Meteorological Measurements
7. S. TWOMEY
Atmospheric Aerosols
8. E. FUKUI (Editor)
The Climate of Japan
9. A.L. FYMAT and V.E. ZUEV (Editors)
Remote Sensing of the Atmosphere: Inversion Methods
and Applications
10. W. BACH, J. PANKRATH and W. KELLOGG (Editors)
Man's Impact on Climate
11. A. LONGHETTO (Editor)
Atmospheric Planetary Boundary Layer Physics
12. C. MAGONO
Thunderstorms

Developments in Atmospheric Science, 13

Statistical Climatology

*Proceedings of the First International Conference on Statistical
Climatology (a Satellite Meeting to the 1979 Session of the ISI),
held at the Inter-University Seminar House, Hachioji, Tokyo (Japan),
November 29–December 1, 1979*

Edited by

S. IKEDA (Editor-in-Chief)

Soka University, Institute of Information Sciences, 1-236 Tangi-cho, Hachioji, Tokyo, Japan

E. SUZUKI, E. UCHIDA and M.M. YOSHINO



ELSEVIER SCIENTIFIC PUBLISHING COMPANY
Amsterdam — Oxford — New York 1980

ELSEVIER SCIENTIFIC PUBLISHING COMPANY
335 Jan van Galenstraat
P.O. Box 211, Amsterdam, The Netherlands

Distributors for the United States and Canada:

ELSEVIER NORTH-HOLLAND INC.
52 Vanderbilt Avenue
New York, N.Y. 10017

LIBRARY / BIBLIOTHÈQUE
ATMOSPHERIC ENVIRONMENT SERVICE
SERVICE DE L'ENVIRONNEMENT ATMOSPHERIQUE
6000 RUE DUNDAS STREET
TORONTO, ONTARIO, CANADA
M3J 1K4

0-444-41923-3

ISBN: 0-444-41923-3 (Vol. 13)

ISBN: 0-444-41710-9 (Series)

© Elsevier Scientific Publishing Company, 1980.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher, Elsevier Scientific Publishing Company, P.O. Box 330, Amsterdam, The Netherlands

Printed in The Netherlands

FOREWORD

Since some time ago the need for a meeting on statistical climatology was strongly felt both in those circles of statisticians interested in climatology as well as of climatologists who are applying the statistical methodology in their branch.

Our feeling is that this was the time for such a session. Indeed, it is not a pure coincidence that one of the positive contributions of the Climate Conference in Geneva last February was to make a clear distinction between weather and climate, that is between the dynamical and the statistical aspects of the meteorological phenomenon.

More precisely, if weather may be characterized mathematically by the joint values of the parameters defining the thermodynamical state at each point of the atmosphere, values which are in evolution with the time; climate may be considered as being the statistical population of all the possible joint values for these parameters at any point of the atmosphere. Moreover, the study of climate may even be more ambitious since it has been found that stochastic models such as autoregressive schemes or Markov chains are able to describe the random character of the evolution of weather.

There were thus some changes that a meeting grouping statisticians and climatologists would lead to fruitful exchanges of ideas and we are grateful to Dr. Fred Leone, Executive Director of the American Statistical Association, for his endeavour to realize the basis of such a meeting, to the Organizing and Executive Committee in Japan involving Dr. M. Hirose, Prof. S. Ikeda, Prof. E. Suzuki, Prof. Y. Suzuki and Dr. E. Uchida who solved for the best all the practical problems and to the sponsors who covered the material needs.

About thirty papers were presented to the fifty participants to the meeting and if in one sense we hoped a still larger participation, we believe that, in spite of the duration limited to two and a half day, easy discussion remained possible which after all remains the main purpose of such events.

Likewise, the presented papers did not cover all the wanted topics as well as we do not think that the discussions exhausted the subject. It seems thus that henceforth another meeting on the same subject has to be considered as useful in a near future with the hope that this time some encouragement will come from the World Meteorological Organization.

Again our gratitude goes to the Organizing Committee in Japan and especially to Prof. S. Ikeda and his staff who made everything smooth and nice so that every one could enjoy the most during his stay in Hachioji and in Tokyo.

Dr. R. Sneyers
Chairman

EDITOR'S PREFACE

The present volume is the Proceedings from the 1-st International Conference on Statistical Climatology, held at Inter-University Seminar House in Hachioji, Tokyo, Japan, from Nov.29 through Dec.1, 1979, as a satellite meeting to the 1979 session of the ISI. The conference, promoted by Dr. Fred C. Leone, the Executive Director of the American Statistical Association, was sponsored by the Bernoulli Society for Mathematical Statistics and Probability, the Japan Statistical Society and the Meteorological Society of Japan.

It should be acknowledged that the conference was almost fully supported financially by the U.S.Office of Naval Research and the Institute of Information Sciences at Soka University.

Under the chairmanship of Dr. R. Sneyers (Institut royal météorologique de Belgique) and Prof. M.M.Yoshino (Tsukuba Univ.), the conference was arranged and executed by the Organizing Committee involving Drs. M.Hirose and E.Uchida (Met.Res.Inst.,Tokyo), E.Suzuki (Aoyama-Gakuin Univ.), Y.Suzuki (Tokyo Univ.) and S.Ikeda (Soka Univ.), keeping close contact with Dr.F.C.Leone and Prof. S.S.Gupta (purdue Univ.).

Topics originally planned for the scientific program of the conference were :
(1) Time series - Assessment of randomness (with corresponding fields of application in Climatology: Homogeneity of series - climatic change), (2) Theoretical distributions - single values, extreme values, continuous or discrete variables, Markov chains (Statistical prediction, simple random climatic models), (3) Joint (multivariate) distributions - continuous or discrete variates, estimation when one marginal is known, factor analysis, multivariate analysis (Statistical prediction, simple random climatic models, statistical description), (4) Statistical quality control (Outliers in series of observations, quality of predictions), (5) Stochastic models of meteorological fields (Estimation of lacking points, optimal density of networks), (6) Discriminant analysis (Climatic classifications - climates, weather types,etc.), (7) Stochastic models - autoregressive models (Climatic models, stochastic dynamic prediction), and (8) Circular distributions - harmonic analysis, spectral analysis, cross test of significance (Climatic models).

Preparations for the publication of the present volume have been done by the Editorial Committee with the aid of referees and a group of reviewers of English usage.

Although the conference was not completely successful in some points, I believe that it should serve as the first step-stone for developing a mutual cooperation between the both fields of Statistics and Climatology to cope with a fastly increasing necessity of statistical methods in climatological researches.

On behalf of the Organizing and Editorial Committees, I would like to express my sincere gratitude equally to all those people who helped us in many occasions and in various ways, including Mr.S.Iida, the General Director of the Seminar House, and his staff; Prof. K. Takamatsu, the Chancellor of Soka Univ.; Dr. H.Hudimoto at the Inst. Stat. Math.; Prof. M.Fukushima, one of my colleague; Mr.V.S.Rao, Miss. K.Nishida, Mrs. R.Ikeda, Mrs.A.Takeda, Mr.Y.Nonaka and the students who helped us at the meeting, and all other people who helped us anonymously.

June 15,1980

Sadao Ikeda
Conf. Secretary
and Chief Editor

CONTENTS

Foreword	V
Editor's Preface	VII
1. A summarized review of theoretical distributions fitted to climatic factors and Markov chain models of weather sequences, with some examples. E. Suzuki	1
2. A generalized circular distribution. R. Sneyers and J. Van Isacker	21
3. Some properties of a family of generalized logistic distributions R.R. Davidson	27
✓ 4. Some statistical techniques for climatological data. S.S. Gupta and S. Panchapakesan	35
5. Asymptotic theory of estimation of the location and scale parameters based on a set of small number selected sample quantiles. J. Ogawa	49
6. Asymptotics for the multisample, multivariate Cramér-von Mises statistic with some possible applications. D.S. Cotterill and M. Csörgő.....	67
7. The behavior of Bayes decision for normal mean under non-standard prior: unknown precision. A.K. Bansal	85
8. Some results on exchangeability and majorization in statistics. A.M. Abouammoh	99
9. Prediction of a future ordered observation based on a sample from the exponential population. E.H. Gan, M. Safiul Hak and M.M. Ali	109
10. Testing homogeneity of variances of a series of linear models. Y.P. Chaubey	119
11. Some properties of generalized ridge estimators in linear models. T.D. Dwivedi, J.M. Lowerre and V.K. Srivastava	125
12. Selection of the number of regression parameters in small sample cases. R. Shibata	137
13. Modelling weather data as a Markov chain. L. Billard and M.R. Meshkani	149
✓ 14. On red noise and quasi-periodicity in the time series of atmospheric temperature. O.M. Essenwanger	165
✓ 15. Detection of changes in the parameters of periodic or pseudo-periodic systems when the change times are unknown. I.R. MacNeill	183

16.	Precipitation simulation process with Markov chain modeling. O.P. Bishnoi and K.K. Saxena	197
✓ 17.	Statistical prediction of climatological extreme values and return period in the case of small samples. E. Suzuki, M. Miyata and S. Hongo	207
18.	On the use of exponential smoothing for the estimation of climatic elements. M. Ogawara	217
19.	An optimum linear restriction in the estimation problem for a generalized linear model and its application to climatic data. E. Suzuki, T. Oohashi and S. Hongo	229
20.	Application of the discriminant analysis in meteorology. G. Der-Megreditchian	241
21.	Regional classification of East African rainfall stations into homogeneous groups using the method of principal component analysis. L. Ogallo	255
✓ 22.	On a mathematical model of carbon dioxide concentration in the mid troposphere. J. Gould, F.A. Ahrens and C.S. Hong	267
23.	Some new worldwide cloud cover models. S.T. Bean and P.N. Somerville	279
✓ 24.	Variability of northern hemisphere mean surface air temperature during the recent two hundred years. R. Yamamoto	307
✓ 25.	The four-year cycle in atmospheric and solar phenomena. K. Takahashi	325
26.	Classification of monsoon climates and stability of their moisture regime. V.P. Subrahmanyam and H.S. Ram Mohan	335
27.	Suitable probability model for severe cyclonic storms striking the coast around the Bay of Bengal. D.A. Mooley	349
28.	Rainfall intensity-duration-return period equations and nomographs of India. Ram Babu, K.G. Tejwani, M.C. Agarwal and L.S. Bhushan	359
✓ 29.	Probability model for the calamitous behavior of the summer monsoon over India. D.A. Mooley and B. Parthasarathy	375
✓ 30.	Problems in statistical climatology - Concluding remarks of the conference. M.M. Yoshino	383

A SUMMARIZED REVIEW OF THEORETICAL DISTRIBUTIONS FITTED TO CLIMATIC FACTORS AND
MARKOV CHAIN MODELS OF WEATHER SEQUENCES, WITH SOME EXAMPLES

E. SUZUKI

Inf. Sci. Res. Center, Aoyama-Gakuin Univ., Shibuya, Tokyo (Japan)

ABSTRACT

Suzuki, E. A summarized review of theoretical distributions fitted to climatic factors and Markov chain models of weather sequences, with some examples. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

Up to this time, various functional expressions have been proposed as theoretical probability distribution models of climatic factors and weather states by many meteorologists and statisticians to clarify statistical situations of data sets.

To estimate the parameters of these distributions, maximum likelihood method has often been used, instead of moment method or others, in which some difficulties of solving transcendental equations were involved, except for the cases of normal and exponential type. Therefore, some simplified iterative procedures have occasionally been necessary for computing the asymptotic maximum likelihood estimates.

Several authors have worked with the asymptotic distribution of climatic extreme values, and Markov chain modelings have been applied to the sequences of two-states under several innovative trials and extended to the sequential process of a generalized category sets of weather types. AIC was certified to be a powerful and reasonable criterion in determining the order of Markov chain.

Reviewing these theoretical distributions, the author will give some notices.

INTRODUCTION

With the systematic accumulation of various climatic data and weather records for long period, analytical distribution models which fit the observed distributions well have been proposed by many climatologists and statisticians. The following theoretical distribution models have been proposed:

- (a) Temperature ... Normal, Pearson I types.
 - (b) Precipitation ... Gamma, Log-Normal, Kappa types.
 - (c) Relative humidity ... Beta type.
 - (d) Wind speed ... Gamma, Weibull, Log-Normal types.
 - (e) Wind-rose ... Circular distribution model and an empirical non-negative p.d.f.
 - (f) Some other climatic elements ... Poisson, Negative binomial and binomial types.
- Case (b) has been studied by many researchers, but in contrast little attention has been given to cases (e) and (f).

In earlier times the moments method of parameter estimation was used but soon the maximum likelihood method replaced it as the preferred method for the estimation of parameters contained in the theoretical models. Except for the normal distribution,

the exponential and other simple distribution models, the maximum likelihood estimators must be obtained by solving transcendental equations with the aid of computer, and the asymptotic variances of these estimators are not always easily obtainable in analytical forms for most of these theoretical models.

Furthermore, the double exponential model has been more frequently applied in the analysis of climatic extreme value than any other model. However, this limiting distribution model (i.e., Fisher-Tippett type I or Gumbel's model) is sometimes not suitable for the case of a finite single sample as Jenkinson (1955,1975) and the author (1968) have pointed out. A generalized 3-parameter distribution model was proposed by Jenkinson who also demonstrated the application of the computational technique by giving examples.

Finally, the Markov chain model has recently been applied to weather sequences (such as dry and wet days, etc.) in place of traditional persistence indices, and several innovative trials have been made under the assumption of an ergodic Markov chain as a statistical background. In such applications of the Markov chain, the Akaike Information Criterion (AIC) and the loss function are used to determine the order γ of the ergodic Markov chain for the two-state weather sequences. For example, Gates and Tong (1976) have shown the effective usefulness of AIC, and Chin (1977) has made a contour map of the chain order γ .

As a statistical test procedure of the goodness of fit, both the Kolmogoroff-Smirnoff statistic and the Kimball statistics are well suited for fitting the theoretical model to the observed frequency.

1. THEORETICAL DISTRIBUTION MODELS OF PRECIPITATION

At present, the Gamma distribution has a long history as a suitable theoretical model for frequency distributions precipitation.

The ordinary Gamma distribution can be written by either one of the following two expressions for the p.d.f.:

$$f(x;v,\beta) = \begin{cases} \frac{\beta^v}{\Gamma(v)} e^{-\beta x} x^{v-1}, & x \geq 0, \\ 0, & x < 0, \end{cases} \quad (1.1)$$

$$f(x;p,a) = \begin{cases} (x/a)^{p-1} e^{-x/a} / a\Gamma(p), & x \geq 0 \\ 0, & x < 0, \end{cases} \quad (1.2)$$

where both $v = p$ and $\beta = a^{-1}$ are positive parameters defining the shape and scale, respectively.

Let x_1, x_2, \dots, x_n be a random sample of size n , then we have the moments method of estimating the parameters:

$$\tilde{v} = \tilde{p} = A^2 / s^2, \quad \tilde{\beta} = \tilde{a}^{-1} = A / s^2 \quad (1.3)$$

where

$$A = \frac{\sum_{i=1}^n X_i}{n}, \quad s^2 = \frac{\sum_{i=1}^n (X_i - A)^2}{n}.$$

Moreover, ML estimates would be obtained by solving the following equations:

$$\hat{p} - \psi(\hat{p}) = \ln(A/G), \quad \hat{p} \cdot \hat{a} = A \quad (1.4)$$

where

$$\psi(\hat{p}) = \left[\frac{d \ln \Gamma(x)}{dx} \right]_{x=\hat{p}}, \quad G = \left(\prod_{i=1}^n X_i \right)^{1/n}.$$

In order to have approximate solutions of p and a , Thom (1958) proposed the following Formulas:

$$p^* = \frac{1 + \sqrt{1 + 4 \ln(A/G)/3}}{4 \ln(A/G)}, \quad a^* = A/p^* \quad (1.5)$$

Greenwood and Durand (1960) derived the numerical computation formulas as follows:

$$\hat{p}_c = \frac{0.5000876 + 0.1648852y - 0.0544276y^2}{y}, \quad 0 \leq y \leq 0.5772,$$

$$\hat{p}_c = \frac{8.898919 + 9.059950y + 0.9775373y^2}{y(17.79728 + 11.968477y + y^2)}, \quad 0.5772 < y \leq 7, \quad (1.6)$$

$$\hat{a}_c = A/\hat{p}_c$$

where $y = \ln(A/G)$.

For correcting a bias of the ML estimators, Bowman and Shenton (1970) proposed a simple correction factor $(1 - 3n^{-1})$ for the computed \hat{p} , i.e., $(1 - 3n^{-1})\hat{p}$ is almost unbiased as is seen in Table 1. Moreover, in 1973 they derived nearly unbiased estimators:

$$\hat{p} = \sum_{i=-1}^2 c_i(n) y^i, \quad \hat{a} = A \sum_{i=1}^4 b_i(n) y^i \quad (1.7)$$

where $c_i(n)$ and $b_i(n)$ are coefficients depending on the sample size n .

TABLE 1.

Examples for the application of the correction factor for the MLE of p . (Bowman and Shenton, 1970)

n	p	$E(\hat{p})$	$(1-3n^{-1})E(\hat{p})$
6	1.0	1.813	0.906
6	2.0	3.795	1.898
10	0.5	0.651	0.456
20	3.0	3.492	2.968
30	0.5	0.540	0.486
30	1.0	1.091	0.983
30	2.0	2.200	1.979
50	0.2	0.208	0.196

On the other hand, Suzuki (1964) proposed a 3-parameter hypergamma distribution model (a provisional name) as a generalized model of the Gamma distribution which is suitable for precipitation data of various time intervals as follows:

$$f(x; \alpha, \beta, \nu) = \begin{cases} [\alpha \beta^{\nu/\alpha} / \Gamma(\nu/\alpha)] \exp(-\beta x^\alpha) x^{\nu-1}, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (1.8)$$

This general model has the following properties:

- (i) If $\alpha = 1$, then it is the ordinary Gamma type.
- (ii) If $\alpha = \nu$, then it is the Weibull type.
- (iii) If $\alpha = 2$ and $\nu = 1$, then it is the ordinary quasi-normal type.
- (iv) If $\alpha = 2/3$ and $\nu = 1/3$, then it is the cube-root quasi-normal type.
- (v) If $\alpha = -1$, then it is the Pearson V type.

The moment generating function (m.g.f.) and the k -th moment of this general model are given by

$$\begin{aligned} \phi(\theta) &= \frac{1}{\Gamma(\nu/\alpha)} \sum_{k=0}^{\infty} \frac{\theta^k \Gamma((\nu+k)/\alpha)}{k! \beta^{k/\alpha}} \\ \mu'_k &= [\partial^k \phi(\theta) / \partial \theta^k]_{\theta=0} = \frac{\Gamma((\nu+k)/\alpha)}{\Gamma(\nu/\alpha) \beta^{k/\alpha}}, \quad k = 0, 1, 2, \dots \end{aligned} \quad (1.9)$$

Hence, one can obtain the estimators of the parameters by the moments method. The ML estimators can be obtained by solving the set of transcendental equations:

$$\begin{aligned} \alpha \log G - \psi(\nu/\alpha) + \log \beta &= 0, \\ \alpha \beta \sum_{i=1}^n X_i - n\nu &= 0, \\ \alpha \beta \sum_{i=1}^n X_i \log X_i - n - n\nu \log G &= 0. \end{aligned} \quad (1.10)$$

An alternative method of computing the ML estimators and their asymptotic variances was given by Suzuki (1964) with the aid of nomographs.

Both the di-gamma and tri-gamma functions must be computed in order to calculate the ML estimators and their asymptotic variances, respectively. Mielke (1975, 1976) verified the following series expansions:

$$\begin{aligned} \psi(z) &= \frac{d \ln \Gamma(z)}{dz} = -\gamma + (z-1) \sum_{j=1}^{\infty} [j(j+z-1)]^{-1}, \\ \tilde{\psi}(z; s) &= -\gamma + (z-1) \sum_{j=1}^s [j(j+z-1)]^{-1} + \ln\left(\frac{s+z-1/2}{s+1/2}\right), \\ \psi'(z) &= \frac{d\psi(z)}{dz} = \sum_{j=1}^{\infty} (j+z-1)^{-2}, \\ \tilde{\psi}'(z; s) &= \sum_{j=1}^s (j+z-1)^{-2} + (s+z-1/2)^{-1}, \end{aligned} \quad (1.11)$$

where $\gamma = 0.577215665$ and s is a certain large value.

The log-normal distribution model is often fitted to the amount of precipitation for short time intervals caused by such factors as cumulus clouds or weather modification experiments. (Johnson and Mielke (1973), Biondini (1975), Crow (1977), etc.). The p.d.f. of the log-normal distribution and the k -th moment about the origin are written as

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma x}} \exp[-(\ln x - \mu)^2 / 2\sigma^2], \quad x > 0, \quad (1.12)$$

$$\mu_k'(X) = \exp(k\mu + k^2\sigma^2/2), \quad k = 0, 1, 2, \dots,$$

from which $E(X)$ and $V(X)$ can be easily obtained. ML estimators of the parameters and their expectation and asymptotic variances can also be readily obtained.

Biondini (1975) studied the log-normal model and pointed out the following two characteristics: (a) the reproductive property of the original variables, and (b) the relation to the central limit theorem of Lindeberg and Levy.

Recently, a positively skewed 2-parameter distribution model was proposed by Mielke (1973) to explain the long-tailed property of rainfall amount distributions:

$$F(x; \alpha, \beta) = \left[\frac{(x/\beta)^\alpha}{\alpha + (x/\beta)^\alpha} \right]^{1/\alpha}, \quad x \geq 0, \quad (1.13)$$

$$f(x; \alpha, \beta) = (\alpha/\beta) [\alpha + (x/\beta)^\alpha]^{-(\alpha+1)/\alpha}, \quad x \geq 0,$$

where $\alpha (> 0)$ and $\beta (> 0)$ denote the shape and scale parameters respectively. In this case, the moments estimators are obtainable from (Mielke (1973)):

$$g(\alpha) = \alpha B\left(\frac{2}{\alpha}, \frac{\alpha-1}{\alpha}\right) / B\left(\frac{3}{\alpha}, \frac{\alpha-2}{\alpha}\right)^2 = \tilde{\mu}_2' / \tilde{\mu}_1'^2, \quad (1.14)$$

$$h(\alpha) = \alpha^{(\alpha-1)/\alpha} / B\left(\frac{2}{\alpha}, \frac{\alpha-1}{\alpha}\right) = \beta / \tilde{\mu}_1',$$

where $\tilde{\mu}_1'$ and $\tilde{\mu}_2'$ are the sample mean and the squared sample mean, respectively; $g(\alpha)$ and $h(\alpha)$ are approximated with the aid of the Beta-function $B(p, q)$.

ML estimators are computable through an iterative procedure with Newton-Raphson's method:

$$\alpha_{i+1} = \alpha_i' - A(\alpha_i, \beta_i) / C(\alpha_i, \beta_i), \quad \beta_{i+1} = \beta_i - D(\alpha_i, \beta_i) / C(\alpha_i, \beta_i), \quad i = 0, 1, \dots \quad (1.15)$$

where $A(\alpha_i, \beta_i)$, $C(\alpha_i, \beta_i)$ and $D(\alpha_i, \beta_i)$ are given explicitly by $\partial \ln L(\alpha, \beta) / \partial \alpha$ etc., $L(\alpha, \beta)$ being the likelihood function:

$$L(\alpha, \beta) = (\alpha/\beta)^n \prod_{i=1}^n [\alpha + (x_i/\beta)^\alpha]^{-(\alpha+1)/\alpha}, \quad \min(x_1, \dots, x_n) > 0.$$

Mielke (1973) indicated that this model is better suited than the Gamma type for fitting the long drawn-out tailend of the precipitation amount distribution, by showing the skewness characteristic:

$$\gamma = (\mu_3' - 3\mu_2'\mu_1' + 2\mu_1'^3) / (\mu_2' - \mu_1'^2)^{3/2} \longrightarrow \infty, (\alpha \longrightarrow 0).$$

Mielke also proposed a generalized model, the 3-parameter Kappa distribution:

$$F(x; \alpha, \beta, \theta) = [(x/\beta)^{\alpha\theta} / \{ \alpha + (x/\beta)^{\alpha\theta} \}]^{1/\alpha}, \quad x \geq 0, \quad (1.16)$$

$$f(x; \alpha, \beta, \theta) = (\alpha\theta/\beta) (x/\beta)^{\theta-1} \{ \alpha + (x/\beta)^{\alpha\theta} \}^{-(\alpha+1)/\alpha}, \quad x \geq 0,$$

where α , β and $\theta > 0$. (See also Essenwanger (1976)).

Among other several papers the following two studies also deal with fitting the heavy tails of precipitation amounts:

(a) Bryson (1973) proposed a conditional mean exceedance defined by

$$CME_{(x)} = E\{ X - x \mid X > x \}$$

which is a measure of heavy tailedness of the precipitation distribution. According to Bryson if $CME_{(x)}$ is an increasing function of x for sufficiently large x , then such a distribution model is considered to be heavy tailed. He calculated the actual rate of increase of $CME_{(x)}$ ($ICME_{(x)}$) for several theoretical distribution models (i.e., Pareto, Kappa, exponential, Gamma, and one-sided normal, etc.) and concluded that the Kappa type is the one producing the longest drawn-out tailend.

(b) Phonsombat and Leduc (1977) tried to fit three theoretical distribution models (i.e., Gamma, 2-parameter and 3-parameter Kappa types) to the actual frequency distribution of weekly precipitation amounts obtained for the period 1954-1973 at 140 observation points in Thailand, and computed the numerical values of the test statistic T' which was proposed by Bryson (1973):

$$T' = \bar{X} \cdot X_{(n)} / \{ (n-1) \bar{X}_{GA}^2 \} \quad (1.17)$$

where $\{ X_{(1)}, X_{(2)}, \dots, X_{(n)} \}$ is a sample of increasing order, $\bar{X} = \sum_{i=1}^n X_{(i)} / n$, $\bar{X}_{GA} = [\prod_{i=1}^n (X_{(i)} + A)]^{1/n}$ and $A = X_{(n)} / (n-1)$.

After testing the heavy-tailed distributions by this statistic T' , he classified the cases by best fit; The number of stations in each category is seen in Table 2.

TABLE 2.

Number of cases in each category for the best fitting model of weekly rainfall frequency (Phonsombat and Leduc (1977))

Distribution	Number of best fit criteria
Kappa (3)	82
Gamma	37
Kappa (2)	21
Total	140

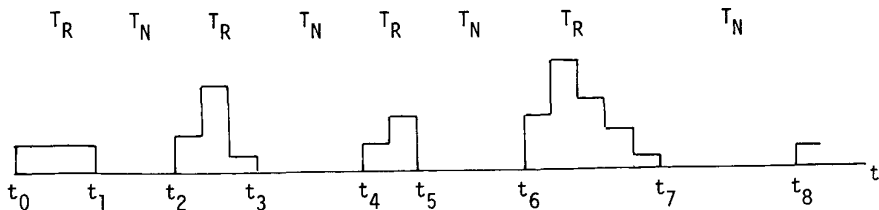


Fig. 1. A tentative model of rainfall process to explain the distributions of intervals of rain and their precipitation amounts.

I would like to propose the following tentative modeling of rainfall distribution (see Suzuki (1967)): Let us consider a simple model to the queueing process as shown in Fig. 1, in which the sets $\{t_0, t_2, t_4, t_6, t_8, \dots\}$ and $\{t_1, t_3, t_5, t_7, \dots\}$ designate the time of the beginning and ending of rain, and T_R and T_N the time-interval of rain and no-rain, respectively.

The following two assumptions can be deduced from the statistical background.

(a) The time points t_v ($v = 0, 1, 2, \dots$) are mutually independent random points, and therefore the frequency model of their occurrences within a unit interval is of the Poisson type.

(b) An empirical (experimental) relation between the rainfall amount R and the interval of rain T_R can be written as $R = aT_R^b + \epsilon$, where a and b are parameters depending mainly on the rainfall characteristics and ϵ is a random error having the normal distribution with zero mean and a finite variance.

Under the above assumptions, the following results are straightforward.

(c) Both T_R and T_N are distributed as the exponential distribution, whose parameters being readily determined from the assumption (a).

(d) The distribution of R can be formulated by making convolution of the two distributions of T_R^b and ϵ , and this convolution is fundamentally equivalent to the hyper-gamma distribution previously proposed.

2. THEORETICAL DISTRIBUTIONS OF THE OTHER CLIMATIC ELEMENTS.

(a) Theoretical models of wind speed.

Several models have been proposed for the distribution of wind speed. For example, the Pearson III type (i.e., Gamma type) has been proposed by Sherlock (1951). The Rayleigh type (a special case of the Gamma model) which is equivalent to the type of 2 degrees of freedom has been tried by several authors. The 3-parameter Planck distribution has been taken by Wentink (1974), and the Weibull distribution model has been proposed by Justus, Hargraves and Yacilin (1976), Stewart and Essenwanger (1978), and others.

Hennessey (1977) studied the theoretical distribution model of wind power density

$p = \rho V^3/2$ under the assumption that the wind speed V follows the Weibull distribution where ρ is a constant representing the air density.

The p.d.f. of the 2-parameter Weibull distribution can be written as

$$f(x; a, c) = acx^{c-1} \exp(-ax^c) \quad , \quad x \geq 0 \quad (2.1)$$

where $c (> 0)$ is the shape parameter and $a^{-1/c} (> 0)$ is the scale parameter. The Rayleigh distribution

$$f(x; a, 2) = 2ax \exp(-ax^2) \quad , \quad x \geq 0 \quad , \quad a > 0 \quad (2.2)$$

is apparently a special case of the Weibull distribution.

Moments estimators of a and c can be obtained from the relations

$$\begin{aligned} \mu &= E(X) = (1/a)^{1/c} \Gamma(1 + 1/c) \quad , \\ \sigma^2 &= V(X) = (1/a)^{2/c} \{ \Gamma(1 + 2/c) - \Gamma^2(1 + 1/c) \} \quad . \end{aligned} \quad (2.3)$$

It has been pointed out by Johnson and Kotz (1970) that Kotel'nikov's nomogram is very useful for finding moments estimator \tilde{c} from the sample mean $\tilde{\mu}$ and the variance of the sample s^2 . Then the moments estimator \tilde{a} is easily computed from $\tilde{a} = \{ \Gamma(1 + \tilde{c}^{-1}) / \tilde{\mu} \}^{\tilde{c}}$.

As a distribution model of the wind power $p = \rho V^3/2$ we may consider the distribution of V^3 . If $X (= V)$ follows the above Weibull distribution, then the c.d.f., p.d.f., mean and variance of $Y = X^3$ are given by

$$\begin{aligned} F(y; a, c) &= 1 - \exp(-ay^{c/3}) \quad , \\ f(y; a, c) &= a(c/3) y^{c/3-1} \exp(-ay^{c/3}) \quad , \\ E(Y) &= a^{-3/c} \Gamma(1 + 3/c) = \mu_3 \{ \Gamma(1 + 1/c) \}^{-3} \Gamma(1 + 3/c) \quad , \\ V(Y) &= a^{-6/c} \{ \Gamma(1 + 6/c) - \Gamma^2(1 + 3/c) \} \\ &= \mu_6 \{ \Gamma(1 + 1/c) \}^{-6} \{ \Gamma(1 + 6/c) - \Gamma^2(1 + 3/c) \} \end{aligned} \quad (2.4)$$

where μ_3 and μ_6 are the 3rd and 6th moments of X , respectively.

Hennessey (1977) studied the application of this distribution model in practical work.

(b) Theoretical models of the surface relative humidity.

Yao (1969) tried to fit the Beta-distribution model to the so-called R index defined by

$$R \text{ index} = \frac{\text{evapotranspiration } E_t}{\text{potential evapotranspiration } E_p} \quad .$$

He also considered the fitting of a similar Beta model to the relative humidity

(see Yao (1974)).

In general, the p.d.f. of the Beta-distribution is given by

$$f(x; \alpha, \beta) = \{B(\alpha, \beta)\}^{-1} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 \leq x \leq 1, \quad (2.5)$$

where α and β are positive parameters, and $B(\alpha, \beta)$ is the Beta-function. It is wellknown that the moments estimators of α and β are given by

$$\tilde{\alpha} = \frac{\tilde{\mu}'_1 (\tilde{\mu}'_1 - \tilde{\mu}'_2)}{\tilde{\mu}'_2 - \tilde{\mu}'_1{}^2}, \quad \tilde{\beta} = \frac{(1 - \tilde{\mu}'_1) (\tilde{\mu}'_1 - \tilde{\mu}'_2)}{\tilde{\mu}'_2 - \tilde{\mu}'_1{}^2} \quad (2.6)$$

where $\tilde{\mu}'_1$ and $\tilde{\mu}'_2$ are the sample moments of the 1st and 2nd order, respectively.

Furthermore, the values of the incomplete Beta-function

$$I_x(\tilde{\alpha}, \tilde{\beta}) = B(\tilde{\alpha}, \tilde{\beta})^{-1} \int_0^x t^{\tilde{\alpha}-1} (1-t)^{\tilde{\beta}-1} dt$$

are available by a numerical method, though the computing procedure is omitted here. Yao (1974) gave many examples of computing distributions for the daily, 5-day, 15-day and monthly mean data of the relative humidity.

Mielke (1975) has shown an iterative procedure of computing the ML estimators $\hat{\alpha}$ and $\hat{\beta}$ of the parameters α and β : Reparametrizing $\gamma = \alpha + \beta$, $p/(1-p) = \alpha/\beta$ or equivalently $\alpha = p\gamma$, $\beta = (1-p)\gamma$, the p.d.f. (2.5) can be rewritten as

$$f(x; p, \gamma) = B(p\gamma, (1-p)\gamma)^{-1} x^{p\gamma-1} (1-x)^{(1-p)\gamma-1}, \quad 0 \leq x \leq 1, \quad (2.7)$$

where $0 < p < 1$ and $\gamma > 0$, from which, by using the series expansion (1.11), the following iteration formulas are finally obtained.

$$\begin{aligned} \hat{\alpha}_k &= \frac{G + \ell n \frac{s + \hat{\alpha}_{k-1} + \hat{\beta}_{k-1} - 1/2}{s + \hat{\alpha}_{k-1} - 1/2} + \sum_{j=1}^s \frac{\hat{\beta}_{k-1} (j + \hat{\alpha}_{k-1})}{j (j + \hat{\alpha}_{k-1} - 1) (j + \hat{\alpha}_{k-1} + \hat{\beta}_{k-1} - 1)}}{\hat{\beta}_{k-1} + \sum_{j=1}^s [j (j + \hat{\alpha}_{k-1} - 1) (j + \hat{\alpha}_{k-1} + \hat{\beta}_{k-1} - 1)]^{-1}} \\ \hat{\beta}_k &= \frac{H + \ell n \frac{s + \hat{\alpha}_k + \hat{\beta}_{k-1} - 1/2}{s + \hat{\beta}_{k-1} - 1/2} + \sum_{j=1}^s \frac{\hat{\alpha}_k (j + \hat{\beta}_{k-1})}{j (j + \hat{\beta}_{k-1} - 1) (j + \hat{\alpha}_k + \hat{\beta}_{k-1} - 1)}}{\hat{\alpha}_k + \sum_{j=1}^s [j (j + \hat{\beta}_{k-1} - 1) (j + \hat{\alpha}_k + \hat{\beta}_{k-1} - 1)]^{-1}} \end{aligned} \quad (2.8)$$

with initial values $\hat{\alpha}_0 = \tilde{\alpha}$ and $\hat{\beta}_0 = \tilde{\beta}$ (moments estimators), where

$$G = \sum_{i=1}^n \ell n x_i / n, \quad H = \sum_{i=1}^n \ell n (1 - x_i) / n, \quad s = 25.$$

Mielke (1975) referred to various versions of the likelihood ratio tests for testing several statistical hypotheses.

(c) Theoretical models of other climatic elements.

A circular distribution model has been considered as suitable for the wind direction frequency (also called the wind-rose in climatology). An empirical distribution of m segments of the wind direction θ can be written as

$$p_k = n_k / \sum_{k=1}^m n_k, \quad k = 1, 2, \dots, m; \quad m = 8 \text{ or } 16, \quad (2.9),$$

where n_k is the sample frequency of the k -th segmented directional intervals, (θ_{k-1}, θ_k) .

As a model to be fitted to the above empirical expression, Jones (1976) first proposed the truncated Fourier series model:

$$f(\theta) = \frac{1}{360} \left[1 + \sum_{v=1}^{(m-2)/2} \left\{ a_v \cos \frac{2\pi v \theta}{360} + b_v \sin \frac{2\pi v \theta}{360} \right\} + a_{m/2} \cos \frac{\pi m \theta}{360} \right] \quad (2.10)$$

where θ is the wind direction (in units of degrees). The following three relations are straightforward:

$$\int_0^{360} f(\theta) d\theta = 1, \quad a_v = \int_0^{360} f(\theta) \cos \frac{2\pi v \theta}{360} d\theta, \quad b_v = \int_0^{360} f(\theta) \sin \frac{2\pi v \theta}{360} d\theta,$$

$$v = 1, 2, \dots, (m-2)/2.$$

In order to avoid negative values of the probability density, Jones (1976) studied an area matching method and a weighting function method; He found that the latter is more suitable than the former, and the following weighting functions were chosen as the weights w_v of the Fourier coefficients a_v and b_v : The simplest weight function is Bartlett's Window $w_v = 1 - v/m$, and another one is Parzen's Window defined by $w_v = 1 - 6(v/m)^2(1 - v/m)$ for $v < m/2$, $= 2(1 - v/m)^3$ for $m/2 \leq v < m$. These two window functions are verified to produce probability densities with non-negative values.

In this meeting on statistical climatology, R. Sneyers presented a quite ingenious methodology of utilizing the circular distribution by generalizing the von Mises circular normal distribution, and a testing of the suitability by an example (see the article in this volume).

Falls (1971) fitted the negative binomial distribution model to monthly thunderstorm events at Cape Kennedy, Florida, for the period 1957-1967, and reported reasonable results for the goodness-of-fit. In general, the negative binomial model is expressed as

$$P\{X = k\} = \binom{r+k-1}{k} p^r q^k, \quad k = 0, 1, \dots, r+k-1, \quad (2.11)$$

where r and p are parameters. It is wellknown that this model is a generalization of the geometric distribution $P(X = k) = pq^k$, $k = 0, 1, \dots$. The moments estimators of r and p and their efficiencies have been studied already by Fisher (1950).

3. THEORETICAL DISTRIBUTION MODELS OF CLIMATIC EXTREMES

It is wellknown that three types of limiting distribution of the extreme value (sample maximum) statistic have been derived by Fisher and Tippet (1928). The double exponential distribution model or Gumbel's distribution, which is one of the above three types, is often applied to the estimation of return periods in Japan.

On the other hand, Jenkinson (1955) proposed the following 3-parameter extreme value distribution as a generalization of the above three types:

$$F(x; k, \alpha, x_0) = \exp[-\{1 - k(x - x_0)/\alpha\}^{1/k}] \quad (3.1)$$

where k , α and x_0 are parameters and $F(x; k, \alpha, x_0)$ represents the probability that an annual maximum value, say, is not greater than x . Putting

$$y = -(1/k) \log \{1 - k(x - x_0)/\alpha\}, \quad (3.2)$$

one can see a simple relationship between y and F :

$$y = -\log \log \{1/F(x; k, \alpha, x_0)\}. \quad (3.3)$$

The relation between Fisher-Tippet models (F-T Type I, II, III) and Jenkinson's general extreme value distribution (G.E.V.) is such that F-T Type I, II and III are the special cases of G.E.V. with $k = 0$, $k < 0$ and $k > 0$, respectively. For example, if we make a linear transformation $x = x_0 + \alpha y$ of a F-T type I variable y , then the resulting double exponential model

$$F(x) = \exp[-\exp\{-(x - x_0)/\alpha\}] \quad (3.4)$$

appears to be the case $k = 0$ of the G.E.V. distribution.

Jenkinson (1969) proposed an iterative method to compute the ML estimators of α , x_0 and k , and later in 1975 he showed several computational scheme and actual examples of the above iterative procedure: A transformed sample of size n is written as

$$y_i = -(1/k) \log \{1 - k(x_i - x_0)/\alpha\}, \quad i = 1, 2, \dots, n, \quad (3.5)$$

for the original ordered sample $x_1 < x_2 < \dots < x_n$. Then the iterative procedure is as follows: Let the joint correction form be

$$\alpha_{i+1} = \alpha_i + \Delta\alpha_i, \quad x_{0i+1} = x_{0i} + \Delta x_{0i}, \quad k_{i+1} = k_i + \Delta k_i, \quad (i = 0, 1, 2, \dots),$$

then the set of simultaneous equations is obtained to compute the correction terms:

$$\begin{bmatrix} \Delta\alpha_i / \alpha_i \\ \Delta x_{0i} / x_{0i} \\ \Delta k_i \end{bmatrix} = \begin{bmatrix} a(k) & h(k) & g(k) \\ h(k) & b(k) & f(k) \\ g(k) & f(k) & c(k) \end{bmatrix} \begin{bmatrix} -U(y_i, k) \\ -Q(y_i, k) \\ V(y_i, k) \end{bmatrix} \quad (3.6)$$

where the elements of the matrix on the right-hand side are obtained from a table

for each value of k $[-0.6(0.2)0.6]$, and

$$P(y) = 1 - \exp(-y), \quad R(y) = 1 - y + y \exp(-y),$$

$$Q(y,k) = \exp(-y + ky) - (1 - k)\exp(ky), \quad U(y,k) = \{P(y) + Q(y,k)\} / k, \text{ and}$$

$$V(y,k) = \{U(y,k) - R(y)\} / k.$$

Started from the 2-parameter distribution model of the case $k = 0$ of G.E.V., Jenkinson noted the following four points: (i) The standard error of x ($= x_0 + ay$) is expressed through y and the sample size n . (ii) As $k \rightarrow 0$, we have $-U(y,k) \rightarrow 0$, $-Q(y,k) \rightarrow 0$, $V(y,k) \rightarrow S(y) = y + \{y^2 \exp(-y) - y^2\} / 2$. (iii) Starting from a set of initial estimates, one can arrive at the optimum estimates of the two parameters. (iv) Quartile means (QM) are helpful to form the initial estimates for the possible domain $(-0.3, +0.3)$ of k . (For example, the initial estimates of the parameters can be chosen as $k_0 = 0$, $\alpha_0 = (QM3 - QM1)/1.57$ and $x_{00} = QM2$, where QM1, QM2 and QM3 denote the lower, middle and upper quartile means of the variables y_i , respectively.)

A specific distribution model of the m -th extreme value of an ordered sample of size n , $X_1 > X_2 > \dots > X_m > \dots > X_n$, was studied by Stevens (1975): The theoretical distribution of X_m can not always be expressed an analytical formula for all original distribution models. Under the following two assumptions: (i) the original distribution is of the exponential type, and (ii) the sample size n is sufficiently large and m is rather small in comparison to n , he derived the p.d.f. and the d.f. of X_m as

$$f(x, \alpha_m, \beta_m) = m^m / \beta_m^{m(m-1)!} \exp[-m(x - \alpha_m) / \beta_m - m \exp\{-(x - \alpha_m) / \beta_m\}] , \quad (3.7)$$

$$F(x, \alpha_m, \beta_m) = \exp[-m \exp\{-(x - \alpha_m) / \beta_m\}] \sum_{v=0}^{m-1} m^v \exp\{-(x - \alpha_m) / \beta_m\} / v! , \quad (3.8)$$

where m is the rank from the top, and α_m and β_m stand for the location and the shape parameters, respectively. ML estimators of α_m and β_m are obtained by solving the simultaneous equations:

$$\begin{aligned} \alpha_m + \beta_m \ln \left[\sum_{i=1}^n \exp(X_i / \beta_m) / n \right] &= 0, \\ \beta_m - \sum_{i=1}^n X_i / n + \{ \sum_{i=1}^n X_i \exp(-X_i / \beta_m) \} / \{ \sum_{i=1}^n \exp(-X_i / \beta_m) \} &= 0. \end{aligned} \quad (3.9)$$

An example of application of the above study to daily temperature data at Columbia, Missouri in the period 1890-1974 has been demonstrated in detail, in which case $n = 85$ and $m = 12$: Let X_m be highest daily temperature in the ranked data set, $m = 12$ in each year. The ML estimates of the parameters were computed by Newton-Raphson's method as an iteration process with a relative estimation error ≤ 0.0005 . One example reported shows that $\hat{\alpha}_{12} = 94.5$ and $\hat{\beta}_{12} = 10.74$. A graphical comparison

has been shown between the empirical distribution function and the 12th highest extreme value distribution with the estimated parameters by ML method. Similar iterative computations were performed for other 10 station locations, with the values of the Kolmogoroff-Smirnov D-statistic. The following results were obtained: $N = 53-56$, $\hat{\alpha} = 93-96$, $\hat{\beta} = 10-13$ and $D = 0.10-0.17$.

High skewness is a general nature of extreme value data. Simson, Rosenzweig and Biondini (1975) studied the skewness of probability distributions by computing conditional probabilities for the case of log-normal, Weibull and Gamma types as the postulated distribution to be fitted for the various precipitation data of weather modification experiments.

Chin and Miller (1977) indicated that the stochastic limiting to Fisher-Tippett type I distribution is very slow, though the limiting distribution has often been utilized.

The G.E.V. model proposed by Jenkinson is certainly useful, but the variance of the ML estimators and the reliability of the extreme value extrapolation will still be principal problems in the application of the model to practical data.

In addition to the recent contributions mentioned above, one of the Japanese innovative studies should be referred to, whose applications are very common in climatological and hydrological fields in Japan.

Ogawara et al. (1954) proposed the following four steps to derive the stochastic limits for the maximum possible amounts of precipitation in Japan: (a) Estimation of a normalizing transformation curve for logarithms of the observed variate, (b) Extrapolation of this transformation curve by utilizing the Fisher-Tippett type limiting distribution, (c) Computation of the stochastic limits for the transformed variates by making use of the two-sample theory, (d) Transformation of the above stochastic limits inversely. These four steps have already been verified to be useful for several locations in Japan under the assumption of a stationary time series. The same authors have also studied the stationarity and the case of a dependent series.

Moreover, Kikuchi-hara verified the various plotting rules proposed by various authors for the purpose of computing return periods appearing in wind speed or gust. The theoretical assumptions and verifications of the plotting rules in practical work have been summarized in his excellent Technical Note No.98 (WMO 1972), and further details will be omitted here because these are out of the scope of the present discussion.

4. THE MARKOV CHAIN MODEL FOR A TIME SEQUENCE OF WEATHER OBSERVATIONS.

Gabriel and Neuman (1962) studied an application of a Markov chain model to the time sequence of weather situations which may be classified into either one of two categories, i.e., wet and dry days, etc.

Feyerherm and Bark (1965,1967) proposed a first order Markov chain model to be

applied to weather changes.

Bayne and Weber (1973) tried the following formulation:

$$P(D_t, \dots, D_{t+n}) = P(D_{t+1}|D_t)P(D_{t+2}|D_{t+1}) \cdots P(D_{t+n}|D_{t+n-1}), \quad (4.1)$$

$$P_t = U + \sum_{h=1}^{12} A_h \sin(2\pi ht/365) + \sum_{h=1}^{12} B_h \cos(2\pi ht/365) + e_t$$

where D_t represents "dry" for the t -th day within a one year period, P_t the initial or transition probability, and e_t denotes a small random error following a normal distribution. They applied this model to North Carolina precipitation data observed during the period 1952-79 at 48 observation points.

Katz (1974) derived a model related to the recurrence

$$\begin{pmatrix} W_0(k;N) \\ W_1(k;N) \end{pmatrix} = \begin{pmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{pmatrix} \begin{pmatrix} W_0(k;N-1) \\ W_1(k;N-1) \end{pmatrix} \quad (4.2)$$

to be applied to a stationary weather time sequence, where $X_n = 1$ or 0 according as the n -th day is wet or dry; $P_{ij} = P(X_n = j | X_{n-1} = i)$ ($i, j = 0, 1$) are assumed to be independent of n ; $S_N = \sum_{n=1}^N X_n$ is the number of wet days within N consecutive days; $W_0(k;N) = P(S_N = k | X_0 = 1)$ and $W_1(k;N) = P(S_N = k | X_0 = 0)$.

Two interesting models have been proposed: A chain-dependent process by Katz (1975) to describe the consecutive sequence of precipitation days, and a Markov chain exponential model by Woolhier (1975) to present the n -day precipitation amounts analytically. Details of these two models will be omitted here because of space limitation.

Gates and Tong (1976) studied an effective use of Akaike's Information Criterion (AIC) for the optimum determination of the chain order γ . Several studies have been offered to hypothesis testing, after a sequential procedure was proposed by Bartlett (1951) and Hoel (1954) to test the null hypothesis $H_{\gamma-1}$ (chain is γ -ldependent) against the alternative H_γ (chain is γ dependent).

Akaike (1972) proposed the AIC as a powerful test criterion to select the optimum order in a multi-variate regression model:

$$AIC = (-2)\ln(\text{max.likelihood}) + 2(\text{no. of independent parameters}). \quad (4.3)$$

Gates and Tong (1976) postulated the following theorem as a recommendation for an optimum procedure in determining the order γ of the chain :

Theorem

For the testing problem $(H_{\gamma-1}; H)$, the logarithmic likelihood ratio test statistic is given by

$$-2\ln\lambda_{\gamma-1} = 2 \sum_{i, \dots, l} n_{ij \dots l} \left(\ln \frac{n_{ij \dots kl}}{n_{ij \dots k}} - \ln \frac{n_{j \dots kl}}{n_{j \dots k}} \right), \quad (4.4)$$

which follows the χ^2 -distribution with $\nabla^2 S^{\gamma+1}$ degrees of freedom under $H_{\gamma-1}$.

where

$$H_{\gamma-1} : P_{i \underset{\gamma+1}{j} \dots k \underset{\gamma}{l}} = P_{j \dots k \underset{\gamma}{l}} , \quad i = 1, 2, \dots, S,$$

$n_{ij \dots kl}$ designates the number of transitions from $i \rightarrow j \rightarrow \dots \rightarrow k \rightarrow l$ in the observed sequence (γ -step), and $\nabla^2 S^{\gamma+1} = \nabla S^{\gamma+1} - \nabla S^{\gamma} = S^{\gamma+1} - 2S^{\gamma} + S^{\gamma-1} = (S-1)2S^{\gamma-1}$;
The basic model N has a sequence $\{x_1, \dots, x_n\}$ with a basic space $S = \{1, \dots, S\}$.

Tong (1975) defined the following loss function $R(k)$ basing upon the AIC approach:

$$R(k) = k \eta_M - 2(\nabla S^{M+1} - \nabla S^{k+1}) \quad (4.4)$$

where

$$k \eta_M = -2 \sum_{m=k}^{M-1} \ell n \lambda_{m,m+1} , \quad k < M, \quad k = 0, 1, \dots , \quad M = 1, 2, \dots .$$

The $\min_k R(k)$ is a reasonable criterion to determine the order of the chain, and he computed the loss function for a set of rain sequence data at Manchester and Liverpool over the period Nov.1973 - Feb.1974. An example is given below:

k	0	1	2	3
R(k)	7.56	-8.32	-4.17	0

from which he was able to conclude that the data are fitted best by a chain of order 1. He gave many other examples of AIC values (eqn. 4.3) for the case $k = 0 \sim 3$.

Ozaki and Tong (1975) proposed the pooled AIC defined by $\min_k R(k)_M + \min_{k'} R(k')_N$ in the determination of non-stationarity, where M and N represent the former time interval and the present time interval under consideration, respectively. Numerical examples are omitted here.

Chin (1977) tried a Markov chain modeling of daily precipitation occurrences by determining the chain order γ for about 100 observing locations in the U.S. during the period 1948-1973, and plotted maps of the order of the chains. An example of a map pattern is displayed in Fig.2. The numbers 1, 2 and 3 show the first, second and the third order dependencies, respectively.

Recently, Katz (1979) has compared two procedures in chain modeling of daily rainfall occurrence, in which he treated the AIC and the SBC (Schwarz Bayesian Criterion) from both the theoretical and practical point of view: The likelihood ratio test statistic for a hypotheses testing problem (null H_0 : k -th order, alt. H_1 : m -th order) can be easily written down by $\eta_{k,m} = -2 \log \lambda_{k,m}$ using Katz's notation, which is fundamentally equivalent to the first member on the right hand side of (4.4). The AIC and the SBC are formulated as

$$AIC(k) = \eta_{k,m} - 2(s^m - s^k)(s-1),$$

$$SBC(k) = \eta_{k,m} - 2(s^m - s^k)(s-1) \log n ,$$

where s and n are the number of states and the sample size, respectively.

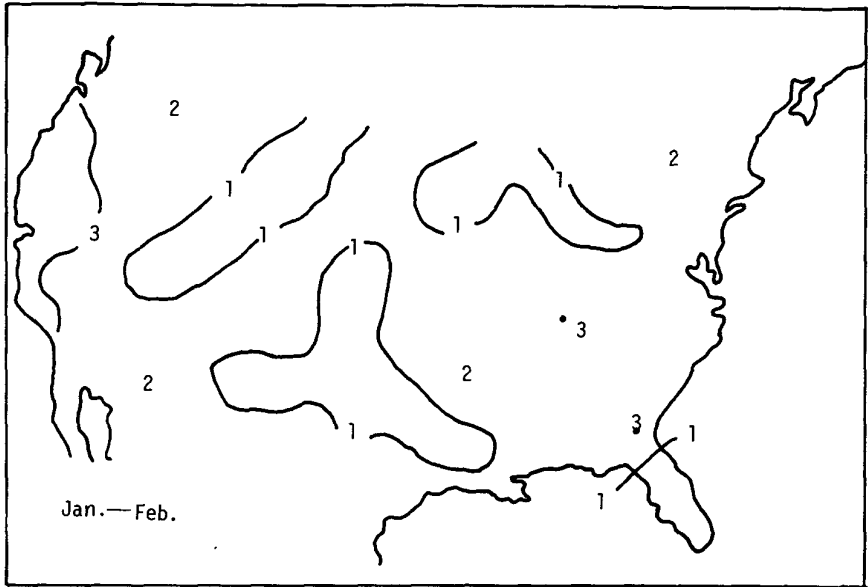


Fig. 2. Map pattern of the Markov chain order of the daily precipitation occurrence process for January - February (after Chin (1977)).

In order to determine the order of a Markov chain, $\min_k AIC(k)$ and $\min_k SBC(k)$ are to be computed. It is evident that the difference between AIC and the SBC, $AIC(k) - SBC(k) = 2(2^m - 2^k)(\log n - 1)$, is positive in the case of 2-state ($s = 2$) and $k < m$ in a finite sample $n > 10$; Katz (1979) stressed that the SBC procedure gives a consistent estimator of the Markov chain order in comparison with the overestimated inconsistency derived by the AIC procedure. An optimum order of the Tel Aviv data is 1, contrary to the reexamined results obtained previously by Gates and Tong(1976).

The AIC procedure has originally been considered as a certain fundamental measure of statistical model building, and therefore several devices would sometimes be desirable for application in practical work of a specific modeling. In this sense, the work by Katz (1979) would really be of precedence in Markov chain modeling of weather sequences in the present and in future.

At the present, the author proposes the following four selection statistics as best suited for model building: AIC (Akaike,1972,1973), C_p (Mallows,1973), S_3 (Bhansali and Downham,1977) and BIC (Schwarz,1978 and Akaike,1978).(see Shibata,1979)

In the author's opinion, the following three points are left open: (1) Ergodic and stationary properties should be better verified in connection with the seasonal variation of the weather, because a singular change of the climate often occurs

actually. (2) The generalization of theoretical background and the verification of an expansion of the Markov chain model would be necessary for the case of the state sequences ($S \geq 3$), i.e., a transition into finite possible categories as in real climatic situations. (3) A stability test of the transition probability matrix $P = [p_{ij}]$ would be desirable, because a sample estimator \tilde{P} of P is a set of stochastic variables.

5. CONCLUDING REMARKS.

Numerous studies of theoretical distribution models and the Markov chain modeling have been summarized in convenient form from the author's rather subjective point of view; Some important and interesting papers and valuable reports have been omitted or have not been precisely presented because of space limitations. Some examples have been simplified in each section.

In concluding this article, the author would like to stress the following points in a synthetic form:

(1) The main purpose of utilizing theoretical distribution models is to describe the observed statistical or stochastic properties by the most suitable and reasonable function for each climatic element. The different characteristics caused by different locations and/or different periods may be explained by discrepancies of estimated parameters under the same distribution model. Hence, many trial and error proposals of applying different types of distribution models to a specific climatic element will not always be appropriate in comparing or interpreting the climatological data, even if the goodness of fit test shows the acceptance of the hypothesis in every model.

(2) In estimation of the parameters of applied distribution model, the maximum likelihood method should be more emphasized than the others, but the ML estimators are biased very often and the asymptotic sample variances of the ML estimators are not always sufficiently clear in statistical climatology field.

The author believes that in building a theoretical distribution model fitted to a climatic element the actual value of a goodness of fit test statistic is rather important. Accordingly, other methods of parameter estimation, say, the least square method or mini-max procedure, would be necessary to be taken into consideration in connection with the use of a suitable test statistic.

(3) In Markov chain and Markov process modeling, the main problem is to determine the order γ under the assumption of stationarity in climatic time series. Statistical treatment of annual trend and seasonal variation should be taken into consideration before the modeling. Thus, the three statistical tests, significance test of trend, test of harmonic coefficients and test of ergodic Markov model, are necessary for an exact modeling of the climatic fluctuations or sequences.

ACKNOWLEDGEMENTS

The author is deeply grateful to Dr. R. Katz for his giving me timely information of his recent result and to Drs. Oskar M. Essenwanger and Dorathy A. Stewart for revision of the English text. Thanks are also due to Prof. S. Ikeda for his friendly suggestions. The author is greatly indebted to Mr. S. Hongo for his assistance in drafting and also to Miss K. Karasawa for typing the manuscript.

REFERENCES

- Akaike, H., 1972. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N. and Csaki, F. (eds.) Proc. 2nd Intern. Symp. Information Theory, Akademiai Kiado, Budapest : 267-281.
- Bartlett, M.S., 1951. The frequency goodness of fit test for probability chains. Proc. Camb. Phil. Soc. 47: 86-95.
- Bayne, C.K. and A.H. Weber, 1973. Statistical analysis of North Carolina precipitation data. Third Conf. on Prob. and Stat. in Atm. Sci.: 250-251.
- Biondini, R.W., 1975. The log-normal distribution and cumulus clouds. Fourth Conf. on Prob. and Stat. in Atm. Sci.: 76-79.
- Bowman, K.O. and Shenton, L.R., 1970. Small sample properties of estimators for the Gamma distribution. Report CTC-28, Union Carbide Corp., Nuclear Div.
- Brooks, C.E.P. and Carruthers, N.C., 1953. Handbook of Statistical Method in Meteorology. H.M.S.O.: 413 pp.
- Bryson, W.C., 1973. Describing and testing for heavy-tailed distributions. Third Conf. on Prob. and Stat. in Atm. Sci. : 118-121.
- Chin, E.H., 1977. Modeling daily precipitation occurrence process with Markov chain. Water Resour. Research 13: 949-956.
- Chin, E.H. and Miller, F.F., 1977. On the estimation of daily precipitation extremes. Fifth Conf. on Prob. and Stat. in Atm. Sci.: 217-220.
- Choi, S.C. and Wette, R., 1969. Maximum likelihood estimation of the parameters of the gamma distribution and their bias. Technometrics 11: 683-690.
- Crow, E.L., 1977. Confidence limits for seeding effect in large-area weather modification experiments. Third Conf. on Prob. and Stat. in Atm. Sci.: 206-211.
- Essenwanger, O.M., 1976. Applied Statistics in Atmospheric Science. Part A: Frequencies and Curve Fitting. Elsevier, 412 pp.
- Falls, L.E., Williford, W.O. and Cater, M.C., 1971. Probability distributions for thunderstorm activity at Cape Kennedy, Florida. J. Appl. Met. 10: 97-104.
- Feyerherm, A.M. and Bark, L.D., 1965. Statistical methods for persistent precipitation patterns. J. Appl. Met. 4: 320-328.
- Feyerherm, A.M. and Bark, L.D., 1967. Goodness of fit of a Markov chain model for sequences of wet and dry days. J. Appl. Met. 6: 770-773.
- Fisher, R.A. and Tippett, L.H.C., 1928. Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Camb. Phil. Soc. 24, Part 2: 180-190.
- Fisher, R.A., 1950. Contributions to Mathematical Statistics. Wiley, New York : 182-187.
- Gabriel, K.R. and Neumann, J., 1962. A Markov chain model for daily rainfall occurrence at Tel Aviv. Quart. J. Roy. Met. Soc. 88: 90-95.
- Gates, P. and Tong, H., 1976. On Markov chain modeling to some weather data. J. Appl. Met. 15: 1145-1151.
- Greenwood, J.A. and Durand, D., 1960. Aids for fitting the gamma distribution by maximum likelihood. Technometrics 2: 55-65.
- Green, J.R., 1964. A model for rainfall occurrences. J. Roy. Stat. Soc. B26: 345-353.
- Green, J.R., 1970. A generalized probability model for sequences of wet and dry days. Mon. Weather Rev. 98: 238-241.
- Grinorten, Irving I., 1971. Modeling of conditional probability. J. Appl. Met. 10: 646-657.
- Gumbel, E.J., 1958. Statistics of Extremes. Columbia Univ. P., New York.

- Hennessey, J.J., 1977. Some aspect of wind power statistics. *J. Appl. Met.* 16:119-128.
- Hoel, P.G., 1954. A test for Markov chains. *Biometrika* 41:430-433.
- Ison, N.T., Feyerherm, A.M. and Bark, L.D., 1971. Wet period precipitation and the Gamma distribution. *J. Appl. Met.* 10:658-665.
- Jenkinson, A.F., 1955. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Quart. J. R. Met. Soc.* 81:158.
- Jenkinson, A.F., 1969. Chpt. 5 - Statistics of Extremes. Tech. Note 98, Estimation of Maximum Floods. WMO, Geneva:183-228.
- Jenkinson, A.F., 1975. Extreme value analysis in meteorology. Fourth Conf. on Prob. and Stat. in *Atm. Sci.* :83-89.
- Johnson, N.L. and Kotz, S., 1970. Continuous Univariate Distributions. Vol. 2., Houghton Mifflin, 306pp.
- Jones, R.H., 1976. Fitting a circular distribution to a histogram. *J. Appl. Met.* 15: 94-98.
- Justus, C.G., Hargraves, W.R. and Yaclin, Y., 1976. Nationwide assessment of potential output from wind powered generators. *J. Appl. Met.* 15:673-678.
- Katz, R.W., 1974. Computing probabilities associated with the Markov chain model for precipitation. *J. Appl. Met.* 13:953-954.
- Katz, R.W., 1975. Precipitation as a chain-dependent process. Fourth Conf. on Prob. and Stat. in *Atm. Sci.* : 131-134.
- Katz, R.W., 1979. Estimating the order of a Markov chain - Another look at the Tel Aviv rainfall data - . Sixth Conf. on Prob. and Stat. in *Atm. Sci.*, Oct. 9-12, Banff, Alberta, Canada. (will be published by Amer. Met. Soc.)
- Kikuchi, H., 1969. On the comparison of several methods calculating the return period of wind velocity. (some problems in applying double exponential distribution for observed data.) *Tenki* 18:21-34.
- Langley, R.W., 1953. The length of dry and wet periods. *Quart. J. Roy. Met. Soc.* 79: 520-527.
- Mielke, P.W. and Johnson, E.S., 1973. Three-parameter kappa distribution maximum likelihood estimates and likelihood ratio tests. *Mon. Weather Rev.* 101:701-707.
- Mielke, P.W., 1973. Another family of distributions for describing and analysing precipitation data. *J. Appl. Met.* 12:275-280.
- Mielke, P.W., 1975. Convenient bivariate distribution likelihood techniques for describing and comparing meteorological data. *J. Appl. Met.* 14:985-990.
- Mielke, P.W., 1976. Simple iterative procedures for two-parameter gamma distribution maximum likelihood estimates. *J. Appl. Met.* 15:181-183.
- Nordø, J., 1975. Some applications of Markov chains. Fourth Conf. on Prob. and Stat. in *Atm. Sci.* : 125-130.
- Ogawara, M. et al., 1954. Stochastic limits for maximum possible amount of precipitation. *Papers in Met. and Geoph.* 5:8-21.
- Ozaki, T. and Tong, H., 1975. On the fitting of non-stationary autoregressive models in time series analysis. *Proc. 8-th Hawaii Intern Conf. System Sci.*:225-226.
- Phonsombat, V. and LeDuc, S.K., 1977. Comparison of kappa and gamma distributions for weekly rainfall amounts in Thailand. Fifth Conf. on Prob. and Stat. in *Atm. Sci.*: 221-224.
- Serfling, R.J., 1975. The Poisson distribution for the frequency of rare levels of persistent meteorological elements. Fourth Conf. on Prob. and Stat. in *Atm. Sci.*: 135-138.
- Schickelanz, P.T. and Krause, G.F., 1970. A test for the scale parameters of two gamma distributions using the generalized likelihood ratio. *J. Appl. Met.* 9:13-16.
- Shenton, L.R. and Bowman, K.O., 1973. Comments on the Gamma distribution and uses in rainfall data. Third Conf. on Prob. and Stat. in *Atm. Sci.* : 122-127.
- Sherlock, R.H., 1951. Analyzing winds for frequency and duration. *Met. Monograph* 4: 72-79.
- Shibata, R., 1979. Selection of the number of regression parameters in small sample cases. (presented at this meeting and will be included in this Volume)
- Simpson, J., 1972. Use of the gamma distribution in single-cloud rainfall analysis. *Mon. Weather Rev.* 100:309-312.
- Simpson, J., Rosenzweig, P. and Biondini, R., 1975. On the role of highly-skewed distributions in weather modification evaluation. Fourth Conf. on Prob. and Stat. in *Atm. Sci.*:70-75.

- Sneyers, R. and Van Isacker, J., 1979. A generalized circular distribution. (in this Volume).
- Stevens, R. and Van Isacker, J., 1975. Maximum likelihood estimates for parameters of the m th extreme value distributions. Fourth Conf. on Prob. and Stat. in Atm. Sci.: 194-196.
- Stewart, D.A. and Essenwanger, O.M., 1973. Frequency distribution of wind speed near the surface. *J. Appl. Met.* 17:1633-1642.
- Stidd, C.K., 1953. Cube-root-normal precipitation distributions. *Trans. Amer. Geophy. Union* 34:31-38.
- Suzuki, E., 1964. Hyper gamma distribution and its fitting to rainfall data. *Papers in Met. and Geoph.* 15:31-35.
- Suzuki, E., 1967. A statistical and climatological study on the rainfall in Japan. *Papers in Met. and Geoph.* 18:103-181.
- Suzuki, E., 1968. *Statistical Meteorology*. Chijinshokan Co. Ltd., Tokyo. 314pp. (in Japanese).
- Thom, H.C., 1958. A note on the gamma distribution. *Mon. Weather Rev.* 86:117-122.
- Todorovic, P. and Woolheiser, D.A., 1975. A stochastic model of n -day precipitation. *J. Appl. Met.* 14:17-24.
- Tong, H., 1975. Determination of the order of a Markov chain by Akaike's information criterion. *J. Appl. Prob.* 12:488-497.
- Walker, S.H. and Duncan, D.B., 1967. Estimation of the probability of an events as functions of several independent variables. *Biometrika* 54:315-327.
- Wentink, Jr., 1974. Wind power potential of Alaska. Part 1, *Sci. Rep.*, NSF/RANN Grant GI-43098. (not yet seen).
- Wilk, M.B., Gnanadesikan, R. and Huyett, M.J., 1962. Estimation of parameters of the gamma distribution using order statistics. *Biometrika* 49:525-545.
- William, C.B., 1952. Sequences of wet and of dry days considered in relation to the logarithmic series. *Quart. J. Roy. Met. Soc.* 78:91-96.
- WMO, 1972. Estimation of maximum floods. *World Met. Org. Tech. Note* 98: 281pp.
- Yakowitz, D., 1976. Small sample hypothesis tests of Markov order, with application to simulated and hydrologic chains. *J. Amer. Stat. Assoc.* 71:132-136.
- Yao, A.Y.M., 1960. The R index for plant water requirement. *Agr. Met.* 6:259-273.
- Yao, A.Y.M., 1974. A statistical model for the surface relative humidity. *J. Appl. Met.* 13:17-21.

A GENERALIZED CIRCULAR DISTRIBUTION

R. SNEYERS and J. Van ISACKER

Institut royal météorologique de Belgique, Bruxelles (Belgique)

ABSTRACT

Sneyers, R. and Van Isacker, J., A generalized circular distribution. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1,1979

Mises' circular normal distribution is generalized by replacing the cosine function by a finite development in Fourier series, making it possible to adjust any circular frequency distribution. The adjustment is realized by applying the method of the steepest descent to the likelihood function and such a distribution is adjusted to the mean frequencies of the wind directions for January in Uccle (Brussels). Pearson's test of goodness of fit leads to the rejection of the distribution with the first harmonic component, but to the acceptance of the distribution with the first and the second harmonic component. The standard errors of estimation of the adjusted frequencies are computed.

1. INTRODUCTION

Gumbel et al. (1953) have given the theory of the circular normal distribution derived by Mises (1918) and defined by the density function

$$f(x) = C^{-1} \exp \phi(x), \quad (1)$$

with

$$\phi(x) = r \cos(x - x_0), \quad (2)$$

the constant C being fixed by the condition

$$\int_0^{2\pi} f(x) dx = 1. \quad (3)$$

Moreover they computed the tables enabling to adjust such a distribution to a series of observations.

The broadest generalization consists in considering in (1) a continuous function $\phi(x)$ such that

$$\phi(2\pi) = \phi(0). \quad (4)$$

We consider here the function defined by a finite development in Fourier series

$$\phi(x) = \sum_{j=1}^N (a_j \sin jx + b_j \cos jx). \quad (5)$$

2. SUFFICIENCY

From 1.(1) and 1.(5) follows that the distribution has sufficient statistics (cf. Kendall and Stuart (1967), p.28). In particular, if x_1, x_2, \dots, x_n is a random sample having this distribution, the likelihood function being

$$L = C^{-n} \exp \left\{ \sum_{j=1}^N (a_j \sum_{i=1}^n \sin j x_i + b_j \sum_{i=1}^n \cos j x_i) \right\}, \quad (1)$$

the quantities

$$A_j = \left(\sum_{i=1}^n \sin j x_i \right) / n \quad \text{and} \quad B_j = \left(\sum_{i=1}^n \cos j x_i \right) / n \quad (2)$$

are sufficient statistics.

Hence, if θ is the vector of the set of parameters (a_j, b_j) , the value $\hat{\theta}$ maximizing L in (1) is a sufficient estimate of θ . This estimate is unique and has MVB covariance matrix.

3. COMPUTATION OF THE SOLUTION

With

$$C = \int_0^{2\pi} \exp \phi(x) dx, \quad r = [\sum (a_j^2 + b_j^2)]^{1/2}, \quad u_j = a_j/r, \quad v_j = b_j/r, \quad (1)$$

$$g(x) = \sum [u_j (\sin j x - A_j) + v_j (\cos j x - B_j)],$$

we have

$$L^{-1/n} = I(r) = \int_0^{2\pi} \exp[r g(x)] dx. \quad (2)$$

Hence, maximizing L is equivalent to minimizing $I(r)$. Moreover, if x_0 maximizes $g(x)$, we have $g'(x_0) = 0$, $g''(x_0) = -1 < 0$ and, obviously $g(x_0) = K > 0$ as soon as the x_i are not all equal to x_0 .

Hence, we have

$$g(x) = K - (x - x_0)^2 \frac{\ell}{2} + \dots \quad (3)$$

and

$$I(r) = \frac{2\pi}{2\ell r} \exp Kr. \quad (4)$$

It follows that $I(r)$ increases with r beyond any limit; it has thus a finite minimum. Moreover, $\theta_0 \neq \theta_1$ may not be two minimums. In fact, with $\theta = (1-\lambda)\theta_0 + \lambda\theta_1$, we would have two minimums, one for $\lambda = 0$ and one for $\lambda = 1$, which is impossible since $\partial^2 I(r) / \partial \lambda^2$ is strictly positive. This confirms the uniqueness of the solution.

For the computation of the solution, let z_1, z_2, \dots, z_{2N} be the $2N$ components

a_j, b_j of under estimation. The method used is the method of steepest descent.

Therefore, θ being given, we put

$$\omega \equiv \omega_j = - \left[\frac{\partial I}{\partial z_j} \right]_{\theta}, \quad j = 1, 2, \dots, 2N \quad (5)$$

and we consider $I = I(\theta + \lambda \omega)$. Thus we have

$$\frac{\partial I}{\partial \lambda} = \sum \omega_j \frac{\partial I}{\partial z_j}, \quad \frac{\partial^2 I}{\partial^2 \lambda} = \sum \sum \omega_j \frac{\partial^2 I}{\partial z_j \partial z_k} \omega_k > 0, \quad (6)$$

and the limited Taylor expansion

$$I = I(\theta) - \sum \omega_j^2 \lambda + \frac{1}{2} \sum \sum \omega_j \omega_k \left[\frac{\partial^2 I}{\partial z_j \partial z_k} \right] \cdot \lambda^2, \quad (7)$$

with minimum for

$$\lambda_1 = \frac{\sum \omega_j^2}{\sum \sum \omega_j \omega_k \left[\frac{\partial^2 I}{\partial z_j \partial z_k} \right]} > 0. \quad (8)$$

If $I_1 = I(\theta + \lambda_1 \omega) < I(\theta)$, a new step is made at $\theta_1 = \theta + \lambda_1 \omega$ to compute I_2 . Otherwise, a minimum λ_1^* is computed through quadratic interpolation from $I(\theta)$, I_1 , $[\partial I / \partial \lambda]_{\theta + \lambda_1 \omega}$.

4. ERRORS OF ESTIMATION

From 2.(1) we have

$$\frac{\partial \log L}{\partial a_j} = n [A_j - E(\sin jx)] \quad (1)$$

and an equivalent expression for $\partial \log L / \partial b_j$.

Similarly, we find

$$\frac{\partial^2 \log L}{\partial a_j \partial b_j} = -n \text{Cov}(\sin jx, \cos kx) \quad (2)$$

and, more generally, if we put

$$\Delta \equiv \Delta_{jk} = \frac{\partial^2 \log L}{\partial z_j \partial z_k} = E \left[\frac{\partial^2 \log L}{\partial z_j \partial z_k} \right], \quad (3)$$

the covariance matrix of the estimations becomes

$$\text{Var} \hat{\theta} = (-\Delta)^{-1}. \quad (4)$$

Finally, the computation of

$$\hat{p}_i = \int_{\alpha}^{\beta} C^{-1} \exp \hat{\phi}(x) dx \quad (5)$$

being made through some summation

$$\hat{p}_i = \Sigma C^{-1} \exp \hat{\phi}(x) , \quad (6)$$

we have

$$d\hat{p}_i = \Sigma C^{-1} \exp \hat{\phi}(x) \cdot d\hat{\phi}(x) \quad (7)$$

or, in matrix notation

$$d\hat{p}_i = \psi'_\theta \cdot d\theta . \quad (8)$$

Thus, we have

$$\text{Var } \hat{p}_i = \psi'_\theta (\text{Var } \hat{\theta}) \psi_\theta . \quad (9)$$

5. EXAMPLE

The mean frequencies ϕ of the wind directions for January in Uccle (Brussels) lead (Table 1) to the theoretical frequencies e_1 and e_2 through the adjustment of a circular distribution with $N = 1$ and $N = 2$. The computation of A_j and B_j has been made by considering the observations as being uniformly distributed in each sector.

With the size $n = 669.97$, Pearson's χ^2 gives $X_1 = 82.7$, $X_2 = 11.7$ for respective degrees of freedom = 13 and 11 and critical values at the 0.05 level: $X_c = 22.4$ and 19.7.

The second fit may thus be accepted. In this case, the solution is

$$a_1 = -0.3155 , \quad b_1 = -0.6766 , \quad a_2 = 0.4153 , \quad b_2 = -0.2164 ,$$

and the covariance matrix is

	a_1	b_1	a_2	b_2
a_1	$0.3546 \cdot 10^{-2}$	$-0.6425 \cdot 10^{-3}$	$0.9770 \cdot 10^{-3}$	$-0.7331 \cdot 10^{-3}$
b_1		$0.4756 \cdot 10^{-2}$	$0.4169 \cdot 10^{-3}$	$0.1488 \cdot 10^{-2}$
a_2			$0.3641 \cdot 10^{-2}$	$-0.7842 \cdot 10^{-4}$
b_2				$0.3650 \cdot 10^{-2}$

This leads to the standard errors of estimation s for e_2 given in Table 1. It appears in that way that the fit gives an accuracy equivalent to that given through sample estimation by a size 3.53 times larger (cf. Sneyers (1975)).

6. CONCLUSION

It should be noted that the solution found for the foregoing example, characterized by equations 1(1), 1(3) and 1(5) with $N = 2$, leads to a distribution function which enables the construction of the distribution of the wind direction on a continuous manner from 0° to 360° , the center of the sector N being 0° . It gives thus

an immediate solution for smoothed frequencies as is required in statistical diffusion models used to predict the concentration of pollutants. The solution presented here may thus be considered as more advantageous than the smoothing method by weight functions as used in Jones (1976).

Moreover if the choice of N in 1(5) may have some arbitrary character, it should be kept in mind that this choice may be based on a preliminary selective harmonic analysis of the observed frequencies as described in Sneyers (1975), the components used in 1(5) being then the same as the ones which have been selected.

TABLE 1.

Mean frequencies \bar{o} of the wind directions for January in Uccle (Brussels). Estimated frequencies e_1 and e_2 . Standard error s on e_2 (10^{-3}).

	\bar{o}	e_1	e_2	s
N	18	27.4	21.0	2.2
NNE	20	23.8	27.6	2.8
NE	41	23.4	37.8	3.6
ENE	54	26.1	45.6	4.2
E	56	32.3	46.3	4.2
ESE	28	43.3	43.3	3.9
SE	36	59.7	43.8	3.8
SSE	60	80.8	53.9	4.3
S	90	102.6	79.9	5.8
SSW	120	117.8	121.4	8.5
SW	163	119.1	153.4	10.9
WSW	125	107.9	139.4	10.0
W	92	87.2	90.5	6.7
WNW	50	65.4	48.6	4.0
NW	28	47.4	27.4	2.6
NNW	19	35.0	20.2	2.1

REFERENCES

- Gumbel, E.J., Greenwood, J.A. and Durand, D., 1953. The circular normal distribution : Theory and Tables. J.Amer.Stat.Assoc. 48 : 131-152.
- Von Mises, R., 1918. Über die 'Ganzzahligkeit' der Atomgewichte und verwandte Fragen. Physik.Zeitsch. 19 : 490-500.
- Kendall, M.G. and Stuart, A., 1967. The Advanced Theory of Statistics. Vol.2 (2nd ed.) Griffin, London.
- Sneyers, R., 1975. Sur l'analyse des séries d'observations. O.M.M., Note Technique n° 143, Genève.
- Jones, R.H., 1976. Fitting a circular distribution to a Histogram. J.Appl.Met. 15 : 94-98.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

SOME PROPERTIES OF A FAMILY OF GENERALIZED LOGISTIC DISTRIBUTIONS

R.R. DAVIDSON

Math. Dept., Univ. of Victoria, Victoria, B.C. (Canada)

ABSTRACT

Davidson, R.R., Some properties of a family of generalized logistic distributions.
Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29 - Dec.1, 1979

A family of generalized logistic distributions is presented and a number of its properties are discussed. This family is an extension of a class first introduced by Gumbel who derived the limiting distribution of the reduced m -th midrange when sampling from a large class of symmetrical continuous distributions. Originally defined for m an integer, the resulting distribution is well defined for any $v > 0$ as the distribution of $\log R$ where R has the inverted beta distribution with parameters $(v, v, 1)$. A summary of the basic limit theory for extremes is presented with particular emphasis on the extreme value distribution of type I and its generalization for m -th extremes. It is demonstrated that the generalized logistic distribution does not have a unique representation as the distribution of the difference of independent random variables which are identically distributed up to location. In particular, the logistic distribution does not arise uniquely as the difference convolution of random variables having the extreme value distribution of type I.

1. INTRODUCTION

Since the early part of this century there has been considerable interest in the statistical behaviour of extremes and in extremal processes. Much of the early work has been summarized by E.J. Gumbel (1958) in his book *Statistics of Extremes*. The logistic distribution plays a prominent role in the theory of extreme values and its applications. The work of Gumbel contains a number of examples both of situations in which the logistic distribution arises as an appropriate model, and of the importance of the logistic distribution in the limit theory for statistics based on extremes.

In this paper a family of generalized logistic distributions is presented and a number of its properties are discussed. This family is an extension of a class first introduced by Gumbel (1944). Gumbel derived the limiting distribution of the reduced m -th midrange when sampling from a symmetrical continuous distribution, and called the resulting distribution the generalized logistic. Originally defined for m an integer, the resulting distribution is well defined for any $v > 0$

as the distribution of $\log R$ where R has the inverted beta distribution with parameters $(\nu, \nu, 1)$.

A summary of the basic limit theory for extremes is presented in section 3, with particular emphasis on the extreme value distribution of type I and its generalization for m -th extremes. The final section contains an analysis of the generalized logistic distribution as a difference distribution. It is shown that the generalized logistic distribution does not have a unique decomposition as the distribution of the difference of independent random variables which are identically distributed up to location. In particular, the logistic distribution does not arise uniquely as the difference convolution of random variables having the extreme value distribution of type I.

2. GENERALIZED LOGISTIC DISTRIBUTIONS

The probability distribution with density function

$$f(y; \nu) = \frac{\Gamma(2\nu)}{\Gamma^2(\nu)} \cdot \frac{\exp(-\nu y)}{\{1 + \exp(-y)\}^{2\nu}}, \quad -\infty < y < \infty, \quad \nu > 0$$

is called the *generalized logistic distribution* with parameter ν . The special case $\nu = 1$ is the usual logistic distribution. This family of distributions was introduced by Gumbel (1944) in the case where the parameter ν is an integer. The characteristic function of a random variable Y having the generalized logistic (ν) distribution is

$$\phi_Y(t) = \Gamma(\nu - it) \Gamma(\nu + it) / \Gamma^2(\nu).$$

Since the distribution of Y is symmetric about zero it follows that $\overline{\phi_Y(t)} = \phi_Y(t)$.

In his important paper on ranges and midranges Gumbel (1944) derived the generalized logistic distribution and gave it its name. The ν -th midrange is defined to be the average of the two ν -th extremes in a sample from an underlying distribution. Specifically, if $X_{m;n}$ and $X_{n-m+1;n}$ are the m -th and $(n-m+1)$ -th order statistics in a sample of size n , then the m -th midrange $W_{m,n}$ is given by $W_{m,n} = [X_{m;n} + X_{n-m+1;n}] / 2$. Gumbel demonstrated that when sampling from an unlimited continuous symmetrical distribution with mean zero, the limiting distribution of the reduced m -th midrange is the generalized logistic distribution with parameter m . More precisely, suppose X has symmetrical density function f , $f(-x) = f(x)$, and distribution function F , and let $\alpha_m = 2nf(u_m)/m$ where $u_m(n) = u_m$ is the characteristic m -th extreme value defined by $F(u_m) = 1 - m/n$. Then $U_{m,n} = \alpha_m W_{m,n}$ converges in distribution to the generalized logistic distribution with parameter m as n becomes infinite. As a special case of this result, the reduced midrange has as its limiting

distribution the logistic distribution. This result provides an interesting example of a measure of central tendency whose limiting distribution is not normal. A key feature of Gumbel's derivation is the fact that the m -th extremes are asymptotically independent as the sample size n becomes infinite.

The generalized logistic distribution can be obtained through a straightforward transformation from the inverted beta distribution. Let R be a random variable with the density of the inverted beta $(v, v, 1)$ distribution, namely

$$f(r; v) = \frac{\Gamma(2v)}{\Gamma^2(v)} \cdot \frac{r^{v-1}}{(1+r)^2}, \quad 0 < r < \infty, \quad v > 0.$$

It then follows directly that $Y = \log R$ has the generalized logistic (v) distribution.

It is a well known result that if U is the median of a sample of size $n=2m-1$ from a distribution with density function f and distribution function F , then U has density function

$$f(u; m) = \frac{(2m-1)!}{[(m-1)!]^2} F^{m-1}(u) [1-F(u)]^{m-1} f(u).$$

Using this result it is easily seen that the median of a sample of size $n = 2m-1$ from the inverted beta $(1, 1, 1)$ distribution has the inverted beta $(m, m, 1)$ distribution. Because of the monotonicity of the transformation $Y = \log R$, it then follows that the median of a sample of size $n = 2m-1$ from the logistic distribution has the generalized logistic (m) distribution.

A second family of generalized logistic distributions has been proposed by Dubey (1969), but these bear no relationship to those presented here except in the special case of the logistic distribution.

3. LIMIT THEOREMS FOR EXTREMES

The literature concerning the asymptotic behaviour of the extreme observation in a sample dates back to the 1920's. Significant contributions were made by von Mises, Fréchet, Fisher and Tippett, de Finetti, and Gumbel in papers which were published prior to 1938. The early work on the theory of extreme values and its applications has been summarized by Gumbel (1958), and a comprehensive review of the asymptotic theory of extremes is contained in Galambos (1978).

The main theorem on the limiting distribution of the extreme is generally attributed to Fisher and Tippett (1928). The principal idea used in their argument was ingenious, but the mathematical argument is sketchy. In a later paper Gnedenko (1943) provides the most rigorous and comprehensive treatment of the problem.

Theorem. Let $X_{n;n}$ be the extreme observation in a random sample from a distribution

with distribution function F . If there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such that the limiting distribution

$$H(x) = \lim_{n \rightarrow \infty} F^n(a_n x + b_n)$$

exists and is a proper distribution, then H must be one of the following three types :

$$G_1(x) = \exp[-\exp(-x)] , \quad -\infty < x < \infty$$

$$G_2(x) = \begin{cases} 0 & , \quad x \leq 0 \\ \exp[-x^{-\alpha}] & , \quad x > 0, \quad \alpha > 0 \end{cases}$$

$$G_3(x) = \begin{cases} \exp[-(-x)^\alpha] & , \quad x \leq 0, \quad \alpha > 0 \\ 1 & , \quad x > 0 \end{cases}$$

The distribution G_1 is called the extreme value distribution of type I, and is the type which arises most frequently in practice. The following interrelationships are of interest. If X_3 has distribution function G_3 then $X_2 = -1/X_3$ has distribution function G_2 ; if X_2 has distribution function G_2 then $X_1 = \alpha \log X_2$ has distribution function G_1 .

The key idea used by Fisher and Tippett was the following : the maximum in a sample of nr observations can be viewed as the maximum of n maxima in n samples of size r . If a limiting distribution exists, then both these maxima will have the same limiting distribution as r becomes infinite. Fisher and Tippett implicitly assume the theorem on convergence of types (cf. Gnedenko and Kolmogorov (1968) p. 40 Theorem 1) which Gnedenko (1943) attributes to Khintchine (1938). Use of this result and some elementary ideas from the theory of functions yielded the main theorem.

The limit theorem above has a direct generalization in the case of the m -th extreme, when m is held fixed as n becomes infinite. This generalization was given by Gumbel (1935) and states that if the m -th extreme $X_{n-m+1;n}$ can be suitably normalized so as to have a proper limiting distribution, then the limiting distribution function must be one of the following three types.

$$G_1^{(m)}(x) = \frac{1}{(m-1)!} \int_{e^{-x}}^{\infty} u^{m-1} e^{-u} du , \quad -\infty < x < \infty$$

$$G_2^{(m)}(x) = \begin{cases} 0 & , \quad x \leq 0 \\ \frac{1}{(m-1)!} \int_{x^{-\alpha}}^{\infty} u^{m-1} e^{-u} du & , \quad x > 0, \quad \alpha > 0 \end{cases}$$

$$G_3^{(m)}(x) = \begin{cases} \frac{1}{(m-1)!} \int_{(-x)^\alpha}^{\infty} u^{m-1} e^{-u} du & , \quad x \leq 0, \quad \alpha > 0 \\ 1 & , \quad x > 0 \end{cases}$$

The distribution $G_1^{(m)}$ is called the m -th extreme value distribution of type I. The interrelationships noted following the statement of the limit theorem for extremes also hold in the generalized case. Furthermore, each of the distributions $G_1^{(m)}$, $G_2^{(m)}$, $G_3^{(m)}$ is well defined when the integer valued parameter m is replaced by a positive valued parameter.

4. REPRESENTATION AS A DIFFERENCE DISTRIBUTION

Let U be a random variable having the gamma (θ, ν) distribution with density function

$$f(u; \theta, \nu) = \frac{\theta^\nu}{\Gamma(\nu)} u^{\nu-1} \exp(-\theta u), \quad u > 0, \theta, \nu > 0.$$

It then follows that the transformed variable $X = -\log U$ has density function

$$f(x; \theta, \nu) = \frac{\theta^\nu}{\Gamma(\nu)} \exp(-\nu x) \exp[-\theta \exp(-x)], \quad -\infty < x < \infty.$$

This distribution is called the *generalized extreme value distribution* with parameters (θ, ν) . When the parameter $\theta = 1$, $\psi = \log \theta$ is a location parameter, and the shape parameter $\nu = m$, an integer, the above distribution reduces to the m -th extreme value distribution of type I.

It now follows that the generalized logistic distribution arises as the difference distribution associated with the generalized extreme value distribution.

Theorem. Let X_1 and X_2 be independent copies of a random variable X which has the generalized extreme value distribution with parameters (θ, ν) . Then the difference $D = X_1 - X_2$ has the generalized logistic distribution with parameter ν .

Proof. If U_j are independent random variables with the gamma (θ, ν_j) distribution $j = 1, 2$, then the ratio $R = U_2/U_1$ has an inverted beta $(\nu_2, \nu_1, 1)$ distribution. Under the transformation $X_j = -\log U_j$, the difference $D = X_1 - X_2 = \log R$. When $\nu_1 = \nu_2 = \nu$ it follows from the transformation described in section 2 that D has the generalized logistic distribution with parameter ν .

In the special case $\nu = 1$, this theorem states that the logistic distribution is the distribution of the difference of two independent random variables distributed according to the extreme value distribution of type I. This result is implicit in the work of Gumbel, and is presented explicitly by Davidson (1969) in the context of paired comparisons.

It is of interest to note that the generalized logistic distribution does not have a unique representation as the distribution of the difference of independent

random variables which are identically distributed up to location. This is in contrast to the case of normal distribution where it follows from a theorem of Cramér (1936) (cf. Cramér (1970) p. 53 Theorem 19) that if X_1 and X_2 are independent random variables whose difference has a normal distribution, then X_1 and X_2 are normally distributed. The fact that the generalized logistic distribution does not have a unique difference decomposition is a somewhat unexpected result. The following theorem provides a class of such decompositions which is distinct from the one presented in the preceding theorem.

Theorem. Let T have a beta distribution with parameters (ν, η) , and let $Z = -\log V$ where the distribution of V conditional on $T = t$ is a gamma distribution with parameters $(\theta t, \nu + \eta)$. If Z_1 and Z_2 are independent copies of Z , then the difference $D = Z_1 - Z_2$ has the generalized logistic distribution with parameter ν .

Proof. Let U and U^* be independent random variables which have the gamma distribution with parameters (θ, ν) and (θ, η) respectively. It is a well known property that $S = U + U^*$ has a gamma $(\theta, \nu + \eta)$ distribution, $T = U/S$ has a beta (ν, η) distribution, and that S and T are independent random variables. Hence for arbitrary $\eta > 0$, it follows that if S has a gamma $(\theta, \nu + \eta)$ distribution and T has a beta (ν, η) distribution with S and T independent, then $U = S \cdot T$ has a gamma (θ, ν) distribution. Let S_1, S_2 and T_1, T_2 be independent copies of S and T respectively which correspond to independent copies U_1, U_2 of U . Then the ratios $R = U_2 / U_1$ and $R^* = V_2 / V_1$, where $U_j = S_j \cdot T_j$ and $V_j = S_j / T_j$ for $j = 1, 2$, are each a ratio of independent and identically distributed random variables. Moreover, R and R^* have the same distribution, namely the inverted beta $(\nu, \nu, 1)$ distribution. However, the random variables U and V do not have the same distribution. In particular, the distribution of V conditional on $T = t$ is a gamma $(\theta t, \nu + \eta)$ distribution, and T has a beta (ν, η) distribution. The density of V is given by

$$f(\nu; \theta, \nu, \eta) = \frac{\nu^{(\nu+\eta)-1}}{\Gamma(\nu)\Gamma(\eta)} \int_0^1 (\theta t)^{\nu+\eta} e^{-\theta t \nu} t^{\nu-1} (1-t)^{\eta-1} dt.$$

In other words, the distribution of V is a mixture of gamma distributions where the mixing variable T has a beta distribution. Invoking the transformations $X = -\log U$ and $Z = -\log V$ it follows that $D = X_1 - X_2 = \log R$ and $D^* = Z_1 - Z_2 = \log R^*$ have the same distribution, namely the generalized logistic distribution with parameter ν . Now the distribution of X is the generalized extreme value (θ, ν) distribution. The distribution of Z conditional on $T = t$ is the generalized extreme value $(\theta t, \nu + \eta)$ distribution, and T has a beta (ν, η) distribution. Thus, although $X_1 - X_2$ and $Z_1 - Z_2$ have the same distribution, the distributions

of X and Z are distinct.

The approach taken in the proof of this theorem has been to transform the question of a decomposition of a difference into that of the decomposition of a ratio. The latter problem has been considered in depth in the special case where the ratio has a Cauchy distribution. A general treatment of the question of the joint distribution of random variables whose ratio has a specified distribution is given by Kotlarski (1964) for the case of the Cauchy and Snedecor distributions.

It follows, as a special case of the above theorem, that the logistic distribution does not arise uniquely as the difference convolution of random variables having the extreme value distribution of Type I. The fact that the logistic distribution and its generalization do not arise uniquely as a difference distribution adds to the value of these distributions in applications which involve the range or the midrange. In particular, the mixtures described in the proof of the theorem provide additional models for data whose midrange is asymptotically distributed as the generalized logistic.

ACKNOWLEDGEMENT

The result presented in the final theorem is based on a technique suggested by P. Diaconis of Stanford University, and is part of joint investigations into the decomposition of difference distributions. This research was supported in part by the Canada Council under Leave Fellowship W760142 and by the Natural Sciences and Engineering Research Council of Canada under Operating Grant A-7166.

REFERENCES

- Cramér, H., 1936, Über eine Eigenschaft der normalen Verteilungsfunktion. Math. Zeits., 41 : 405-414.
- Cramér, H., 1970, Random Variables and Probability Distributions. Cambridge Tract. in Math., 36. Cambridge Univ. P.
- Davidson, R.R., 1969, On a relationship between two representations of a model for paired comparisons. Biometrics, 25 : 597-599.
- Dubey, S.D., 1969, A new derivation of the logistic distribution. Nav. Res. Logist. Quart. 16 : 37-40.
- Fisher, R.A. and Tippett, L.H.C., 1928, Limiting forms of the frequency distribution of the largest or smallest member of a sample. Proc. Camb. Phil. Soc. 24 : 180-190.
- Galambos, J., 1978, The Asymptotic Theory of Extreme Order Statistics. Wiley & Sons.
- Gnedenko, B.V., 1943, Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math. 44 : 423-453.
- Gnedenko, B.V. and Kolmogorov, A.N., 1968, Limit Distributions for Sums of Independent Random Variables. Addison-Wesley.
- Gumbel, E.J., 1935, Les valeurs extrêmes des distributions statistiques. Ann. Inst. Henri Poincaré 5 : 115-158.
- Gumbel, E.J., 1944, Ranges and midranges. Ann. Math. Statist. 15 : 414-422.
- Gumbel, E.J., 1958, Statistics of Extremes. Columbia Univ. P.
- Khintchine, A.Y., 1938, Limit theorems for sums of independent random variables. GONTI. Moskow-Leningrad.

Kotlarski, I., 1964, On bivariate random variables where the quotient of their coordinates follows some known distribution . Ann. Math. Statist. 35 : 1673-1684 .

SOME STATISTICAL TECHNIQUES FOR CLIMATOLOGICAL DATA

S.S.GUPTA¹ and S.PANCHAPAKESAN²

¹Dept. Stat., Purdue Univ., Indiana

²Southern Illinois Univ., Illinois

ABSTRACT

Gupta, S.S. and Panchapakesan, S., Some statistical techniques for climatological data. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

The need for and the increasing use of statistical techniques in the analysis of climatological data are amply illustrated in the literature. Some known techniques relating to meteorological problems such as weather modification experiments and objective weather forecasting are briefly reviewed here. Also, selection and ranking approach to multiple decision theory is discussed with emphasis on potential applications.

1. INTRODUCTION

The need for statistical methodology in analyzing meteorological data has long been recognized. For example, weather modification provides, as noted by Braham (1979), a "fertile field of interaction and collaboration between meteorologists and statisticians." Satisfactory models have been found to describe meteorological data (see Section 2). Time series data occur commonly in climatological studies. Some of the important and interesting problems arise in connection with weather modification experiments, objective weather forecasting and classification of meteorological patterns. Studies in meteorology in general and rain simulation in particular have inspired novel developments in probability and statistics. The concept of characteristic functional first developed by Kolmogorov was later reintroduced by Le Cam (1947) motivated by meteorological studies (see Neyman (1979a)). The concepts of outlier-prone and outlier-resistant distributions developed in Neyman and Scott (1971) were motivated by cloud seeding experiments.

The objectives of the present paper are to briefly review some important known applications of statistical techniques to meteorological data and to indicate the potential applications of selection and ranking procedures to these problems. No attempt will be made to be comprehensive in the treatment of either objective. Some important distributions that have been satisfactorily used as models in meteorological problems are described in Section 2. The next section deals with weather modification

experiments and some related asymptotic optimal tests and nonparametric tests. Section 4 discusses techniques used in a variety of situations other than weather modification experiments. The topics include Markov chain models, the biplot technique, selection of the best predictors in forecasting and classification of weather patterns. The last section describes some subset selection procedures and discusses the selection of the best regression model under this formulation.

2. STATISTICAL MODELS

In this section, we briefly discuss several distributions that have been found useful as models for meteorological data. Any discussion of the techniques for inference will be deferred until later sections.

The gamma distribution has been extensively used as a model for precipitation data. Rain simulation experiments indicate (see Neyman and Scott (1971)) has been found a satisfactory model in practice. The distribution of nonzero rainfall per experimental unit is assumed to have the density

$$f(x) = \frac{\theta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta x}, \quad x \geq 0, \theta > 0, \alpha > 0. \quad (2.1)$$

Here, θ is reciprocal of the scale parameter and α is the shape parameter. It is generally assumed (Neyman (1979a)) that the seeding of the clouds can change the value of the scale parameter but has no effect on the shape parameter. The gamma distribution has been used or verified as a model by Barger and Thom (1949), Mooley and Crutcher (1968), Neyman and Scott (1967a), Schickedanz (1967), Schickedanz and Decker (1969), Simpson (1972), and Thom and Vestal (1968).

Mielke (1973) considered for describing precipitation data the two-parameter Kappa distribution with distribution function

$$F(x) = \begin{cases} [(x/\beta)^\alpha / \{\alpha + (x/\beta)^\alpha\}]^{1/\alpha}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2.2)$$

where $\alpha > 0$ and $\beta > 0$ denote the shape and scale parameters, respectively.

Wong (1977) made goodness-of-fit comparisons among the gamma, lognormal, three-parameter kappa ($\alpha\theta$ in the place of α in (2.2)), and Weibull distributions using five sets of Alberta hailfall data. He found the Weibull distribution a reasonable alternative to the lognormal and three-parameter kappa distributions for describing precipitation and streamflow data. It should be noted that the lognormal distribution is outlier resistant (Neyman (1979a)) and that Weibull and gamma distributions can be subsumed under the generalized gamma distribution with density

$$f(x) = \frac{\gamma x^{\gamma\alpha-1}}{\beta^{\gamma\alpha} \Gamma(\alpha)} e^{-(x/\beta)^{\gamma}}, \quad x > 0, \quad (2.3)$$

where α , β , and γ are all positive parameters.

The three-parameter Weibull distribution was used by Stewart and Essenwanger (1978) as a model for wind speed near the surface. Tackle and Brown (1978) have used the distribution function

$$F(x) = \begin{cases} F(0) + (1 - F(0))(1 - \exp\{-(x/\theta)^{\beta}\}), & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2.4)$$

where $F(0)$ is the probability of observing zero wind speed.

Luna and Church (1974) have found the lognormal distribution as a satisfactory model for wind speed at many sites. Yao (1974) found the beta distribution as a satisfactory model for frequency distributions of relative humidity observations. The beta distribution has also been used by Mielke (1975).

Bivariate normal distribution is used by Wu, Williams and Mielke (1972) in the analysis of continued-covariate and cross-over designs that arise in cloud seeding experiments. For some other distributions that have been considered in connection with meteorological data, see Mielke (1979).

Associated with all these distributions are the obvious problems of estimation. The several methods of estimation applied to these distributions are of general interest and not restricted to meteorological problems; as such, relevant references can be amply found in the statistical literature. It suffices here to mention a few recent papers motivated by meteorological applications, namely, Crow (1977, 1978), Flueck and Holland (1976), Mielke (1973, 1976), Mielke and Johnson (1973), and Wong (1977). Other problems of inference are discussed in subsequent sections.

3. WEATHER MODIFICATION EXPERIMENTS

Early scientific weather modification experiments are attributed to Vincent Schaefer (1946) and Barnard Vonnegut (1947) who showed that pellets of Dry Ice and minute particles of silver iodide would nucleate ice crystals in supercooled clouds. Early days of weather modification are discussed by Byers (1974) and Elliot (1974). One of the important experiments, known as Project Whitetop, was carried out by Professor Braham and his colleagues at the University of Chicago during the summers of 1960 through 1964. The data of this experiment have been reanalyzed by Professor Neyman and his associates at Berkeley. The details of Project Whitetop, controversies regarding its conclusions, and relevant references can be found in the paper by Braham (1979) and the comments by Dawkins and Scott (1979) and Neyman (1979b). A

categorized bibliography of weather modification experiments is given by Hanson et al (1979).

Weather modification experiments are getting increasing attention of statisticians as evidenced by the papers in Volume V of the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (University of California Press, 1967) devoted entirely to this subject and two special issues of Communications in Statistics - Theory and Methods (Volume A8, Numbers 10 and 11, 1979). In the rest of this section, we briefly describe some of the problems and techniques.

A class of asymptotic tests that is routinely used in testing hypotheses regarding the effect of seeding the clouds is called the optimal $C(\alpha)$ tests. These tests were developed by Neyman (1959) and are applicable to testing composite hypotheses that are frequently encountered in practice. These tests are applied by Neyman and Scott (1967b) for evaluating single rain simulation experiments. The basic assumption is that, whether or not seeded, there corresponds to each experimental unit (a fixed duration like 24 hours) a positive probability, say $1-p$, of the rainfall being zero. Two mechanisms are introduced, one governing the change in p due to seeding and the other governing the effects of seeding per wet day. The effect of each mechanism is a change in the value of p in either direction. On each experimental day (a day considered 'suitable' for seeding), a randomized decision is made whether or not to proceed with seeding. As a measure of the effect of seeding, Neyman and Scott (1967b) use $\xi = (p_s - p_c)/p_c$, where the subscripts s and c denote "seeded" and "control", respectively. Neyman and Scott (1967b) provide three test criteria, labeled Z_1 , Z_2 , and Z_3 , of which the first two are optimal $C(\alpha)$ tests of hypotheses H_1 and H_2 , that cloud seeding does not affect the frequency of wet days, and that it does not affect the rainfall per wet day, respectively. The criterion Z_3 is not a $C(\alpha)$ test; it is a linear combination of Z_1 and Z_2 so chosen as to be sensitive to departures from H_3 that the seeding does not affect the target precipitation averaged per experimental unit, whether wet or dry. The specialization of the conditional density of the target precipitation given that it is not zero, joint with the predictors if such are available, determines several different cases. For some recent work on the detection of variable response to cloud seeding, see Neyman (1979a).

Efficient methods for summary evaluations of several independent experiments are important in view of "the notorious frequency with which rain simulation experiments fail to yield statistically significant results." Davies and Puri (1967) discuss two related but distinct problems specializing certain earlier results concerning $C(\alpha)$ tests.

Suppose that the distribution of the nonzero precipitation is gamma with density given in (2.1). It is assumed that the effect of seeding is to change θ to $\xi\theta$ (i.e. effect is multiplicative). The interest is to test $H: \xi \geq 1$ against $A: \xi < 1$. Note that $\xi < 1$ corresponds to increased average nonzero rainfall. The results of several cloud seeding experiments indicate (Neyman and Scott (1967c)) a value of

α in the interval (0.45, 0.75). One can use likelihood ratio tests or $C(\alpha)$ tests. However, it is a simplistic assumption that the changes induced by cloud seeding can be adequately represented by a simple scale or location parameter shift. Thus, nonparametric techniques are useful in testing for a change due to seeding in the distribution of precipitation amount. Commonly used nonparametric tests are Wilcoxon, Kolmogorov-Smirnov, and median tests. Another test which is applicable is due to Taha (1964) and is based on the statistic $L = \frac{1}{n} \sum_{i=1}^n s_i^2$, where the s_i are the ranks of the "seeded" observations in the combined sample of $2n$ observations. In the sense of asymptotic efficiency, this L test is found superior to Wilcoxon test. James (1967) has made some numerical comparisons of the Pitman efficiency of Wilcoxon, gamma scores, exponential scores and L tests for small values of α coming out in favor of the exponential scores test.

Tamura (1963) proposed a class of tests based on the statistic $A_r = \sum_{j=1}^N j^r Z_j$, where $r > 0$, and $Z_j = 1$ or 0 if the j th ordered observation in the pooled sample of size N is a seeded or a non-seeded observation. A similar class of two-sample nonparametric tests is considered by Mielke (1972, 1974) to treat the same problem but with the cross-over design.

Multivariate nonparametric and permutation procedures are useful when a number of measured responses are obtained from each experimental unit. Mielke, Berry and Johnson (1976) have considered multi-response permutation procedures, special cases of which have been earlier suggested by Mantel and Valand (1970). For some further discussion of these procedures, see Mielke (1979).

Weather modification experiments are carried out in a natural environment subject to much variability. Covariates are used in analysis in order to reduce the experimental error. Bradley, Srivastava and Lanzdorf (1979) have discussed covariance analyses effected through the use of multiple regression methods. They have also reviewed the original results of an experiment conducted by North American Weather Consultants and discussed a multivariate analysis without use of covariates or transforms.

4. STATISTICAL TECHNIQUES FOR OTHER METEOROLOGICAL PROBLEMS

In this section, we briefly discuss applications of certain statistical techniques to meteorological problems other than the weather modification. The examples are chosen to indicate the scope and the nature of applications.

In several situations we need more sophisticated models than those discussed in Section 2. An important problem in meteorology is the determination of the characteristics of hourly temperatures. Hansen and Driscoll (1977) developed a stochastic model for hourly temperatures for Big Spring, Texas. These temperatures are produced by harmonics representing both diurnal and annual variations, and a Markov chain expression incorporating adjustments for several variations such as seasonal variation

of the serial correlation coefficient.

Markov chain models have been used to describe the daily occurrence of precipitation. Gabriel and Neumann (1962) considered a model for daily rainfall occurrence at Tel Aviv. Another model was introduced by Todorovic and Woolhiser (1975). Recently, Katz (1977) proposed a more general model and discussed the distribution of the maximum amount of daily precipitation and the distribution of the total precipitation.

The biplot is a graphical display of a two-dimensional approximation to a matrix obtained by least squares using the first and second singular value components of the matrix. It is related to principal component analysis and multivariate analysis of variance (MANOVA). Its usefulness in the display and analysis of meteorological data is demonstrated by Gabriel (1972) with two sets of data. In one of the examples, the biplot is an approximation to simultaneous tests of different subhypotheses in the one-way layout MANOVA. For mathematical and computational details of the technique, see Gabriel (1971).

In forecasting the state of atmosphere at grid points, we have the problem of obtaining vector-valued estimates of meteorological parameters at a grid point based on multivariate information from several sources. In other words, our estimator Z , a vector of n components, is given by $Z = A_1 X_1 + \dots + A_m X_m$, where the X_i have some joint distribution. The problem is to find the "best" linear combination of the information vectors. Thiebaut (1974a) has considered the criterion of minimizing the variances of the components of Z . An example of this situation is given in Thiebaut (1973). In another paper, Thiebaut (1974b) has discussed a related problem regarding the estimation of covariances of meteorological parameters using local-time averages.

McCutchan and Schroeder (1973) have used stepwise discriminant analysis of eight meteorological variables to classify the days during their study period at a southern California location into one of five types.

Many examples of statistical prediction schemes in climatology are available. The prediction is based on a number of predictor variables. While the prediction can be made more accurate by bringing in as many relevant predictor variables as possible, some of them may be highly correlated among themselves and the contribution of some may be very marginal. The problem of selecting the best set of predictor variables arise in various situations. Stringer (1972, pp. 132-133) has cited some examples from literature regarding prediction of precipitation and visibility. Martin et al (1963) have considered an example dealing with forecasting of the 24-hour movement and change of central pressures of North American winter anticyclones. Lund (1971) has discussed a problem of estimation of precipitation involving almost 4500 potential predictors.

Several criteria for defining the best set of predictor variables and various techniques for selecting the best set have been discussed in a nice expository paper

by Hocking (1976). Also, a brief review and evaluation of significant methods have been given by Thompson (1978). Martin et al (1963) applied forward type stepwise procedure. Lund (1971) has illustrated a method of blending stagewise and stepwise procedures.

It should be noted that these techniques for selecting the best set of predictor variables are not designed to produce a best set with a guaranteed probability. We will come back to this point in the next section.

5. RANKING AND SELECTION PROCEDURES

In dealing with weather data, one may want to compare different sites (weather stations) on the basis of appropriate characteristics of the meteorological variables involved. For example, we may want to compare these locations on the basis of mean temperature, or mean nonzero precipitation amount, or variability of temperature for a fixed duration. One may be interested in ranking the sites in terms of the values of the characteristic or just in selecting the site with the largest (smallest) value of the characteristic.

Formally speaking, we have k independent populations (sites) π_1, \dots, π_k , where π_i is characterized by the distribution function $F(x; \theta_i)$ and θ_i is an unknown parameter which represents the "worth" of the population. For example, $F(x; \theta_i)$ may be the distribution function of the 24-hour nonzero precipitation amount at the i th site and θ_i may be the mean of the distribution. Let $\theta_{[1]} \leq \dots \leq \theta_{[k]}$ denote the ordered θ_i . To be specific, let us say that π_i is "preferable" to π_j if $\theta_i > \theta_j$ so that the best population is the one associated with the largest θ_i . Ranking and selection problems have been generally formulated using either the *indifference zone approach* or the *subset selection approach*.

Let us consider the simple problem of selecting the best population. Under the indifference zone formulation of Bechhofer (1954), we want a procedure R which will select the best population with a minimum guaranteed probability P^* ($1/k < P^* < 1$) whenever $\delta(\theta_{[k]}, \theta_{[k-1]}) \geq \theta^*$, where $\delta(\theta_{[k]}, \theta_{[k-1]})$ is an appropriate measure of the distance between the populations associated with $\theta_{[k]}$ and $\theta_{[k-1]}$, and the quantities θ^* and P^* are specified in advance. In the cases of location and scale parameters, the natural choices for $\delta(\theta_{[k]}, \theta_{[k-1]})$ are $\theta_{[k]} - \theta_{[k-1]}$ and $\theta_{[k]}/\theta_{[k-1]}$, respectively. Consequently, $\theta^* > 0$ in the first case and $\theta^* > 1$ in the second. Suppose we want a procedure R based on samples of equal sizes. The problem is to determine the minimum sample size needed to meet the probability requirement.

In the subset selection approach, our goal is to select a non-empty subset of the k populations so that the best population is included in the selected subset with a minimum guaranteed probability P^* . Selection of any subset which includes the best population is called a correct selection (CS). The general approach is to evaluate the infimum of $P(\text{CS}|R)$, the probability of a correct selection using

The procedure R , over the parameter space $\Omega = \{\theta: \theta = (\theta_1, \dots, \theta_k)\}$ and obtain the constants involved in defining R so that

$$\inf_{\Omega} P(CS|R) \geq P^*. \quad (5.1)$$

The condition (5.1) is referred to as the P^* -condition or the basic probability requirement. In order to meet this requirement, one determines the parametric configuration θ_0 , the Least Favorable Configuration (LFC), for which the infimum in (5.1) is attained. In general, there may not be a unique LFC. The expected size of the subset selected is one of the measures generally used as performance characteristics of a procedure.

For an extensive survey and bibliography of ranking and selection theory and related topics the reader is referred to the recent book of the authors (1979). Other books in this area are Bechhofer, Kiefer and Sobel (1968), and Gibbons, Olkin and Sobel (1977).

In the rest of this section, we describe briefly subset selection procedures for normal populations in terms of means, for gamma populations in terms of the scale parameter, for multivariate normal populations in terms of multiple correlations coefficients and discuss selection of best predictor variables.

5.1 Normal Populations

Let π_1, \dots, π_k be k independent normal populations with unknown means μ_1, \dots, μ_k , respectively, and a common variance σ^2 . Let \bar{X}_i , $i=1, \dots, k$ be the sample means based on samples of size n . The best population is the one associated with the largest μ_i . When σ^2 is known, the procedure R_1 proposed by Gupta (1956) selects the population π_i if and only if

$$\bar{X}_i \geq \max(\bar{X}_1, \dots, \bar{X}_k) - \frac{d_1 \sigma}{\sqrt{n}} \quad (5.2)$$

where $d_1 = d_1(k, P^*) > 0$ is the smallest constant such that the condition (5.1) is satisfied. The LFC is given by $\mu_1 = \dots = \mu_k$. This implies that d_1 is given by

$$\int_{-\infty}^{\infty} \phi^{k-1}(x + d_1) \Phi(x) dx = P^*, \quad (5.3)$$

where $\Phi(x)$ and $\phi(x)$ are the standard normal cdf and density, respectively. The values of d_1 are tabulated for several values of k and P^* by Gupta (1963a) and Gupta, Nagel and Panchapakesan (1973).

When σ^2 is not known, the procedure R_2 of Gupta (1956) is the same as R_1 with σ replaced by s , where s^2 is the usual pooled estimator of σ^2 based on $v = k(n-1)$

degrees of freedom. Here again, the LFC is given by $\mu_1 = \dots = \mu_k$. The values of the constant d_2 (used in the place of d_1) are tabulated by Gupta and Sobel (1957) for selected values of k , v , and P^* .

The procedures R_1 and R_2 can be modified in the case of the population with the smallest μ_i being defined the best. For procedures involving unequal sample sizes, see Gupta and Huang (1976), and Gupta and Wong (1976).

5.2 Gamma Populations

Let π_i have the associated density

$$f(x, \theta_i) = \begin{cases} \frac{x^{r-1}}{\Gamma(r)\theta_i^r} \exp(-x/\theta_i), & x > 0, \theta_i > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (5.4)$$

As we can see, it is assumed that the populations have the same shape parameter $r(> 0)$. Further, r is assumed to be known. Our interest is selecting the population associated with the largest (smallest) θ_i . The gamma distribution not only serves as a model for certain types of measurement, but also includes the case where the observations come from normal populations and the interest is in selecting the population associated with the smallest variance.

For selecting the population associated with the largest θ_i , Gupta (1963b) investigated the procedure R_3 which selects π_i if and only if

$$\bar{X}_i \geq b \max(\bar{X}_1, \dots, \bar{X}_k) \quad (5.5)$$

where $\bar{X}_1, \dots, \bar{X}_k$ are means based on samples of equal size n , and the constant b ($0 < b < 1$) is chosen so that the P^* -condition is met. Gupta (1963b) has shown that $P(\text{CS}|R_3)$ is minimized when $\theta_1 = \dots = \theta_k$ and that the constant b is given by

$$\int_0^\infty G_v^{k-1}(x/b) g_v(x) dx = P^*, \quad (5.6)$$

where $G_v(x)$ is the cdf of a standardized gamma random variable (i.e. with $\theta = 1$) with parameter $v/2$ where $v = 2nr$. Thus the constant b depends on n and r only through v and its values are tabulated by Gupta (1963b) for selected values of k , P^* , and v .

For selecting the normal population with the smallest variance, an analogous procedure is given by Gupta and Sobel (1962a) and the appropriate constant can be obtained from the tables in their companion paper (1962b).

5.3 Multivariate Normal Populations

Let π_1, \dots, π_k be k independent P -variate normal population where π_i is $N(\underline{\mu}_i, \Sigma_i)$. Let $\underline{X}'_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ be a random observation vector from $\pi_i, i=1, \dots, p$. The populations are ranked in terms of the ρ_i , where ρ_i is the multiple correlation coefficient of X_{i1} with respect to the set (X_{i2}, \dots, X_{ip}) . We are interested in selecting a subset containing the population associated with the largest ρ_i . Let R_i denote the sample multiple correlation coefficient between X_{i1} and (X_{i2}, \dots, X_{ip}) . Two cases arise: (i) The case in which X_{i2}, \dots, X_{ip} are fixed, called the conditional case; (ii) The case in which X_{i2}, \dots, X_{ip} are random, called the unconditional case. In either case, Gupta and Panchapakesan (1969) proposed and studied the rule R which selects π_i if and only if

$$R_i^{*2} \geq c \max_{1 \leq j \leq k} R_j^{*2} \quad (5.7)$$

where $R_i^{*2} = R_i^2 / (1 - R_i^2)$, and $0 < c = c(k, P^*, p, n) < 1$ is chosen to satisfy the P^* -requirement. In this case, the infimum of PCS is attained when $\rho_1 = \rho_2 = \dots = \rho_k = 0$ and the appropriate constant c is given by

$$\int_0^\infty F_{2q, 2m}^{k-1}(x/c) f_{2q, 2m}(x) dx = P^*, \quad (5.8)$$

where $q = \frac{1}{2}(p-1)$, $m = \frac{1}{2}(n-p)$, $F_{r,s}$ denotes the cdf of an F random variable with r and s degrees of freedom, and $f_{r,s}$ denotes the corresponding density. The values of c are tabulated by Gupta and Panchapakesan (1969) for selected values of k, m, q , and P^* .

5.4 Selection of Best Set of Predictor Variables

In Section 4, we referred to the techniques that have been commonly used for selecting the best predictor variables. We pointed out that these procedures are not designed to guarantee a minimum probability of obtaining the best set. Recently this problem has been investigated by Arvesen and McCabe (1973, 1975), McCabe and Arvesen (1974), and Gupta and Huang (1977) under the subset selection formulation described earlier in this section. Investigations along these lines continue to be of interest in view of their practical importance.

5.5 Other Procedures and Related Problems

There are several parametric and nonparametric procedures available in the literature to suit many contexts that commonly arise. There are single-stage, double-stage and sequential procedures. There are several modifications of the basic problem.

Also important are the related problems of estimating the ordered parameters. Many of these are areas of current research. For an extensive survey and bibliography, see Gupta and Panchapakesan (1979).

ACKNOWLEDGEMENT

This research was supported by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

REFERENCES

- Arvesen, J.N. and McCabe, G.P., 1973. Variable selection in regression analysis. In: W.O. Thompson and F.B. Cady (eds.), Proc. of Univ. Kentucky Conf. on Regression with a Large Number of Predictors, Dept. Stat., Univ. Kentucky, Lexington.
- Arvesen, J.N. and McCabe, G.P., 1975. Subset selection problems of variances with applications to regression analysis. *J. Amer. Statist. Assoc.*, 70: 166-170.
- Barger, G.L. and Thom, H.C.S., 1949. Evaluation of drought hazard. *Agron. J.*, 41: 519-526.
- Bechhofer, R.E., 1954. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.*, 25: 16-39.
- Bechhofer, R.E., Kiefer, J. and Sobel, M., 1968. Sequential Identification and Ranking Procedures. The Univ. of Chicago P., Chicago.
- Bradley, R.A., Srivastava, S.S. and Lanzdorf, A., 1979. Some approaches to statistical analysis of weather modification experiment. *Comm. Stat.-Theor. Meth.*, A8(11): 1049-1081.
- Braham, R.E., 1979. Field experimentation in weather modification. *J. Amer. Statist. Assoc.*, 74: 57-68.
- Byers, H.R., 1974. History of weather modification. In: W.N. Hess (ed.), Weather and Climate Modification. John Wiley, New York, pp. 3-44.
- Crow, E.L., 1977. Minimum variance unbiased estimators of the ratio of means of two lognormal variates and of two gamma variates. *Comm. Stat.-Theor. Meth.*, A6(10): 967-975.
- Crow, E.L., 1978. Confidence limits for seeding effect in single-area weather modification experiments. *J. Appl. Meteor.*, 17: 1652-1660.
- Davies, R.B. and Puri, P.S., 1967. Some techniques of summary evaluations of several independent experiments. In: L.M. Le Cam and J. Neyman (eds.), Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. V, Univ. of California P., Los Angeles and Berkeley, pp. 385-388.
- Dawkins, S.M. and Scott, E.L., 1979. Comment on the paper by R.R. Braham. *J. Amer. Statist. Assoc.*, 74: 70-77.
- Elliott, R.D., 1974. Experience of the private sector. In: W.N. Hess (ed.), Weather and Climate Modification. John Wiley, New York, pp. 45-89.
- Flueck, J.A. and Holland, B.S., 1976. Ratio estimators and some inherent problems in their utilization. *J. Appl. Meteor.*, 15: 536-543.
- Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58: 453-467.
- Gabriel, K.R., 1972. Analysis of meteorological data by means of decomposition and biplots. *J. Appl. Meteor.*, 11: 1071-1077.
- Gabriel, K.R. and Neumann, J., 1962. A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quart. J. Roy. Meteor. Soc.*, 88: 90-95.
- Gibbons, J.D., Olkin, I. and Sobel, M., 1977. Selecting and Ordering Populations. John Wiley, New York.
- Gupta, S.S., 1956. On a decision rule for a problem in ranking means. Ph.D. Thesis (Mimeo. Ser. No. 150), Inst. Stat., Univ. North Carolina, Chapel Hill.

- Gupta, S.S., 1963a. Probability of integrals of the multivariate normal and multivariate t . *Ann. Math. Statist.*, 34: 792-828.
- Gupta, S.S., 1963b. On a selection and ranking procedure for gamma populations. *Ann. Inst. Stat. Math.*, 14: 199-216.
- Gupta, S.S. and Huang, D.Y., 1976. Selection procedures for the means and variances of normal populations: unequal sample sizes case. *Sankhya Ser. B*, 38: 112-128.
- Gupta, S.S. and Huang, D.Y., 1977. On selecting an optimal subset of regression variables. *Mimeo. Ser. 501, Dept. Stat., Purdue Univ., West Lafayette, Indiana.*
- Gupta, S.S., Nagel, K. and Panchapakesan, S., 1973. On the order statistics from equally correlated normal random variables. *Biometrika*, 60: 403-413.
- Gupta, S.S. and Panchapakesan, S., 1969. Some selection and ranking procedures for multivariate normal populations. In: P.R. Krishnaiah (ed.), *Multivariate Analysis - II*, Academic P., New York, pp. 475-505.
- Gupta, S.S. and Panchapakesan, S., 1979. *Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations.* John Wiley, New York.
- Gupta, S.S. and Sobel, M., 1957. On a statistic which arises in selection and ranking problems. *Ann. Math. Statist.*, 28: 957-967.
- Gupta, S.S. and Sobel, M., 1962a. On selecting a subset containing the population with the smallest variance. *Biometrika*, 49: 495-507.
- Gupta, S.S. and Sobel, M., 1962b. On the smallest of several correlated F -statistics. *Biometrika*, 49: 509-523.
- Gupta, S.S. and Wong, W.Y., 1976. Subset selection procedures for the means of normal populations with unequal variances: unequal sample sizes case. *Mimeo. Ser. 473, Dept. Stat., Purdue Univ., West Lafayette, Indiana.*
- Hansen, J. and Driscoll, D.M., 1977. A mathematical model for the generation of hourly temperatures. *J. Appl. Meteor.*, 16: 935-948.
- Hanson, M.A., Barker, L.E., Bach, C.L., Cooley, E.A. and Hunter, C.H., 1979. A bibliography of weather modification experiments. *Comm. Stat.-Theor. Meth.*, A8(11): 1129-1153.
- Hocking, R.R., 1976. The analysis and selection of variables in linear regression. *Biometrics*, 32: 1-49.
- James, B.R., 1967. On Pitman efficiency of some tests of scale for the gamma distribution. In: L.M. Le Cam and J. Neyman (eds.), *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. V, Univ. of California P., Los Angeles and Berkeley, pp. 389-393.
- Katz, R.W., 1977. Precipitation as a chain-dependent process. *J. Appl. Meteor.*, 16: 671-676.
- Le Cam, L.M., 1947. Un instrument d'étude des fonctions aleatoires: La fonctionnelle caractéristique. *C. R. Acad. Sci. Paris*, 224: 710-711.
- Luna, R.E. and Church, H.W., 1974. Estimation of long-term concentrations using a "universal" wind speed distribution. *J. Appl. Meteor.*, 13: 910-916.
- Lund, I.A., 1971. An application of stagewise and stepwise regression procedures to a problem of estimating precipitation in California. *J. Appl. Meteor.*, 10: 892-902.
- Mantel, N. and Valand, R.S., 1970. A technique of nonparametric multivariate analysis. *Biometrics*, 26: 547-558.
- Martin, F.L., Borsting, J.R., Steckbeck, F.J. and Manhard, A.H., 1963. Statistical prediction methods for North American winter anticyclones. *J. Appl. Meteor.*, 2: 508-516.
- McCabe, G.P. and Arvesen, J.N., 1974. A subset selection procedure for regression variables. *J. Statist. Comput. Simul.*, 3: 137-146.
- Mielke, P.W., 1972. Asymptotic behavior of two-sample tests based on powers of ranks for detecting scale and location alternatives. *J. Amer. Statist. Assoc.*, 67: 850-854.
- Mielke, P.W., 1973. Another family of distributions for describing and analyzing precipitation data. *J. Appl. Meteor.*, 12: 275-280.
- Mielke, P.W., 1974. Squared rank test appropriate to weather modification cross-over design. *Technometrics*, 16: 13-16.
- Mielke, P.W., 1975. Convenient beta distribution likelihood techniques for describing and comparing meteorological data. *J. Appl. Meteor.*, 14: 985-990.
- Mielke, P.W., 1976. Simple iterative procedures for two-parameter gamma distribution maximum likelihood estimates. *J. Appl. Meteor.*, 15: 181-183.
- Mielke, P.W., 1979. Some parametric, nonparametric and permutation inference procedures resulting from weather modification experiments. *Comm. Stat.-Theor. Meth.*, A8(11):

1083-1096.

- Mielke, P.W., Berry, K.J. and Johnson, E.S., 1976. Multi-response permutation procedures for a priori classifications. *Comm. Stat.-Theor. Meth.*, A5: 1409-1424.
- Mielke, P.W. and Johnson, E.S., 1973. Three parameter kappa distribution maximum likelihood estimates and likelihood ratio tests. *Mon. Wea. Rev.*, 101: 701-707.
- Mooley, D.A. and Crutcher, H.L., 1968. An application of the gamma distribution function to Indian rainfall. ESSA Tech. Report. ESD 5.
- McCutchan, M.H. and Schroeder, M.J., 1973. Classification of meteorological patterns in southern California by discriminant analysis. *J. Appl. Meteor.*, 12: 571-577.
- Neyman, J., 1959. Optimal asymptotic tests of composite statistical hypotheses. In: U. Grenander (ed.), *Probability and Statistics*. John Wiley, New York, pp. 213-234.
- Neyman, J., 1979a. Developments in probability and mathematical statistics generated by studies in meteorology and weather modification. *Comm. Stat.-Theor. Meth.*, A8(11): 1097-1110.
- Neyman, J., 1979b. Comment on the paper by R.R. Braham. *J. Amer. Statist. Assoc.*, 74: 90-94.
- Neyman, J. and Scott, E.L., 1967a. Some outstanding problems relating to rain modification. In: L.M. Le Cam and J. Neyman (eds.), *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. V, Univ. of California P., Los Angeles and Berkeley, pp. 293-325.
- Neyman, J. and Scott, E.L., 1967b. Note on techniques of evaluation of single rain simulation experiments. In: L.M. Le Cam and J. Neyman (eds.), *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. V, Univ. of California P., Los Angeles and Berkeley, pp. 371-384.
- Neyman, J. and Scott, E.L., 1967c. On the use of $C(\alpha)$ optimal tests of composite hypotheses. *Bull. Inst. Internat. Statist.*, 41: 477-496.
- Neyman, J. and Scott, E.L., 1971. Outlier proneness of phenomena and of related distributions. In: J. Rustagi (ed.), *Optimizing Methods in Statistics*. Academic P., New York and London, pp. 413-430.
- Schaefer, V.J., 1946. The production of ice crystals in a cloud of supercooled water droplets. *Science*, 104: 457-459.
- Schickedanz, P.T., 1967. A Monte Carlo method for estimating the error variance and power of the test for a proposed cloud seeding experiment. Ph.D. Thesis, Univ. of Missouri, Columbia.
- Schickedanz, P.T. and Decker, W.L., 1969. A Monte Carlo technique for designing cloud seeding experiments. *J. Appl. Meteor.*, 8: 220-228.
- Simpson, J., 1972. Use of the gamma distribution in single-cloud rainfall analysis. *Mon. Wea. Rev.*, 100: 309-312.
- Stewart, D.A. and Essenwanger, O.M., 1978. Frequency distribution of wind speed near the surface. *J. Appl. Meteor.*, 17: 1633-1642.
- Stringer, E.T., 1972. *Techniques in Climatology*. W.H. Freeman Company, San Francisco.
- Tackle, E.S. and Brown, J.M., 1978. Note on the use of Weibull statistics to characterize wind-speed data. *J. Appl. Meteor.*, 17: 556-559.
- Taha, M.A.H., 1964. Rank test for scale parameter for asymmetrical one-sided distributions. *Publ. Inst. Statist. Univ. Paris*, 13: 169-179.
- Thiebaux, H.J., 1973. Statistical approaches to grid-point estimation of meteorological parameters. In: *Proc. of the Third Conference on Probability and Statistics in Atmospheric Science*. American Meteorological Society, Boston, pp. 202-206.
- Thiebaux, H.J., 1974a. Minimum variance estimation of coefficient matrices in a dependent system. *Biometrika*, 61: 87-90.
- Thiebaux, H.J., 1974b. Estimation of covariances of meteorological parameters using local-time averages. *J. Appl. Meteor.*, 13: 592-600.
- Thom, H.C.S. and Vestal, I.B., 1968. Quartiles of monthly precipitation for selected stations in the contiguous United States. ESSA Tech. Report ESD 6.
- Thompson, M.L., 1978. Selection of variables in multiple regression: Part I. A review and evaluation. *Int. Statist. Rev.*, 46: 1-19.
- Todorovic, P. and Woolhiser, D.A., 1975. A stochastic model of n-day precipitation. *J. Appl. Meteor.*, 14: 17-24.
- Vonnegut, B., 1947. The nucleation of ice formulation by silver iodide. *J. Appl. Phys.*, 18: 593-595.
- Wong, R.K.W., 1977. Weibull distribution, iterative likelihood techniques and hydro-

- meteorological data. J. Appl. Meteor., 16: 1360-1364.
- Wu, S.C., William, J.S. and Mielke, P.W., 1972. Some designs and analyses for temporally independent experiments involving correlated bivariate responses. Biometrics, 28: 1043-1061.
- Yao, A.Y.M., 1974. A statistical model for the surface relative humidity. J. Appl. Meteor., 13: 17-21.

ASYMPTOTIC THEORY OF ESTIMATION OF THE LOCATION AND SCALE PARAMETERS BASED ON A SET OF SMALL NUMBER OF SELECTED SAMPLE QUANTILES

J. OGAWA

Dept. Math. and Statist., Univ. of Calgary, Alberta (Canada)

ABSTRACT

Ogawa, J., Asymptotic theory of estimation of the location and scale parameters based on a set of small number of selected sample quantiles. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

We are concerned with the following situation. Suppose that we are given a large sample of size n from a population whose density function is of the form $f((x-\mu)/\sigma)/\sigma$, where μ is the location parameter, σ is the scale parameter and the functional form $f(\cdot)$ is known. This sample has been arranged as order statistics. For a given spacing $0 < \lambda_1 < \lambda_2 < \dots < \lambda_k < 1$ one selects a set of k sample quantiles $x(n_1) < x(n_2) < \dots < x(n_k)$, $n_i = [n\lambda_{i+1}]$, $i = 1, 2, \dots, k$. (i) Find out the best estimators of μ and σ based on the above set of sample quantiles and (ii) find out the optimal spacings which give the highest efficiencies of the estimators.

The theory will provide an economical method of estimation of population parameters in processing the large quantity of climatological data.

1. ASYMPTOTIC DISTRIBUTION OF A SET OF SELECTED SAMPLE QUANTILES

We consider the distribution whose density function depends only on the location and scale parameters: $(1/\sigma)f(x-\mu)/\sigma$, where μ is the location parameter and σ is the scale parameter, and the function $f(u)$ is a known function. For a given spacing

$$0 < \lambda_1 < \lambda_2 < \dots < \lambda_k < 1 \quad (1.1)$$

one has a set of k sample quantiles

$$x(n_1) < x(n_2) < \dots < x(n_k)$$

where $n_i = [n\lambda_{i+1}]$, $i = 1, 2, \dots, k$.

One defines u_i as the λ_i -quantile of the standardized population, i.e.,

$$\lambda_i = \int_{-\infty}^{u_i} f(t) dt, \quad i = 1, 2, \dots, k, \quad \text{and} \quad f_i = f(u_i), \quad i = 1, 2, \dots, k, \quad (1.2)$$

then the λ_i -quantile x_i of the population can be written as $x_i = \mu + \sigma u_i$, $i = 1, 2, \dots, k$. For the sake of convenience we put

$$\lambda_0 = 0, \quad u_0 = -\infty, \quad f_0 = f(u_0) = 0, \quad x_0 = -\infty, \\ \lambda_{k+1} = 1, \quad u_{k+1} = +\infty, \quad f_{k+1} = f(u_{k+1}) = 0, \quad x_{k+1} = +\infty.$$

It can be shown that the asymptotic joint distribution of $X(n_1), \dots, X(n_k)$ as $n \rightarrow \infty$ has the density function

$$h(x(n_1), \dots, x(n_k); \mu, \sigma) = (2\pi\sigma^2)^{-k/2} f_1 \dots f_k [\lambda_1(\lambda_2 - \lambda_1) \dots (\lambda_k - \lambda_{k-1})]^{-1/2} n^{k/2} \exp\left[-\frac{n}{2\sigma^2} S\right] \quad (1.3)$$

where

$$S = \sum_{i=1}^k \frac{\lambda_{i+1} - \lambda_{i-1}}{(\lambda_{i+1} - \lambda_i)(\lambda_i - \lambda_{i-1})} f_i^2 (x(n_i) - \mu - \sigma u_i)^2 \\ - 2 \sum_{i=2}^k \frac{f_i f_{i-1}}{\lambda_i - \lambda_{i-1}} (x(n_i) - \mu - \sigma u_i)(x(n_{i-1}) - \mu - \sigma u_{i-1}) \quad (1.4).$$

(See Mosteller (1946)).

2. FISHER AMOUNT OF INFORMATION AND THE RELATIVE EFFICIENCIES

We calculate the Fisher amount of information from (1.3) and then based on the whole data as a random sample. Relative efficiencies will be defined by the ratio of them.

Case I. The scale parameter σ is known and only the location parameter μ is to be estimated.

The Fisher amount of information with respect to μ calculated from h is shown to be

$$I_S(\mu) = E\left[\frac{\partial \log h}{\partial \mu}\right]^2 = -\frac{\partial^2 \log h}{\partial \mu^2} = \frac{n}{\sigma^2} K_1 \quad (2.1)$$

where

$$K_1 = \sum_{i=1}^{k+1} \frac{(f_i - f_{i-1})^2}{\lambda_i - \lambda_{i-1}}. \quad (2.2)$$

The Fisher amount of information with respect to μ calculated from the whole data as a random sample is seen to be

$$I_w(\mu) = \frac{n}{\sigma^2} E\left[\frac{f'}{f}\right]^2. \quad (2.3)$$

The relative efficiency of the estimator of μ based on a set of selected sample

quantiles $x(n_1), \dots, x(n_k)$ is defined by

$$\eta(\mu) = \frac{I_S(\mu)}{I_w(\mu)} = K_1 / E\left[\frac{f'}{f}\right]^2. \quad (2.4)$$

For example :

$$\begin{aligned} \text{Normal distribution} \quad f(u) &= \frac{1}{2} e^{-u^2/2} & \eta(\mu) &= K_1, \\ \text{Logistic distribution} \quad f(u) &= \frac{e^{-u}}{(1+e^{-u})^2} & \eta(\mu) &= 3K_1. \end{aligned}$$

Case II. The location parameter μ is known and only the scale parameter σ is to be estimated.

The Fisher amount of information with respect to σ calculated from h is shown to be

$$I_S(\sigma) = E\left[\frac{\partial \log h}{\partial \sigma}\right]^2 = -\frac{\partial \log h}{\partial \sigma^2} = \frac{2k}{\sigma^2} + \frac{n}{\sigma^2} K_2 \quad (2.5)$$

where

$$K_2 = \sum_{i=1}^{k+1} \frac{(f_{i,i} u_i - f_{i-1,i-1} u_{i-1})^2}{\lambda_i - \lambda_{i-1}}. \quad (2.6)$$

The Fisher amount of information with respect to σ calculated from the whole data as a random sample is seen to be

$$I_w(\sigma) = \frac{n}{\sigma^2} \{E\left[\frac{Uf'(U)}{f(U)}\right]^2 - 1\}. \quad (2.7)$$

The relative efficiency of the estimator of σ based on a set of selected sample quantiles $x(n_1), \dots, x(n_k)$ is defined by

$$\eta(\sigma) = \frac{K_2}{\{E\left[\frac{Uf'(U)}{f(U)}\right]^2 - 1\}}. \quad (2.8)$$

For example :

$$\begin{aligned} \text{Normal distribution} \quad \eta(\sigma) &= \frac{K_2}{2}, \\ \text{Exponential distribution} \quad \eta(\sigma) &= K_2. \end{aligned}$$

Case III. Both the location and scale parameters are unknown and we are concerned with joint estimation of μ and σ .

The area of the ellipse of concentration of the maximum likelihood estimators of μ and σ based on the whole data as a random sample is proportional to the inverse square root of

$$\frac{n^2}{\sigma^4} \{ E[\frac{f'(U)}{f(U)}]^2 (E[\frac{Uf'(U)}{f(U)}]^2 - 1) - E^2[\frac{Uf'(U)}{f(U)^2}] \}. \quad (2.9)$$

The greatest lower bound of the area of the ellipse of concentration of the joint unbiased estimators of μ and σ is proportional to the inverse square root of

$$E[\frac{\partial^2 \log h}{\partial \mu^2}] E[\frac{\partial^2 \log h}{\partial \sigma^2}] - E^2[\frac{\partial^2 \log h}{\partial \mu \partial \sigma}] = \frac{n^2}{\sigma^4} (K_1 K_2 - K_3^2) + 2 \frac{nk}{\sigma^4} K_1 \quad (2.10)$$

where

$$K_3 = \sum_{i=1}^{k+1} \frac{(f_i - f_{i-1})(f_i u_i - f_{i-1} u_{i-1})}{\lambda_i - \lambda_{i-1}}. \quad (2.11)$$

Relative efficiency of the joint estimation of μ and σ based on a set of the selected sample quantiles $x(n_1), \dots, x(n_k)$ is defined by

$$\eta(\mu, \sigma) = \frac{K_1 K_2 - K_3^2}{E[\frac{f'}{f}]^2 (E[\frac{Uf'}{f}]^2 - 1) - E^2[\frac{Uf'^2}{f^2}]} \quad (2.12)$$

For example:

$$\text{Normal distribution} \quad \eta(\mu, \sigma) = \frac{1}{2} (K_1 K_2 - K_3^2)$$

3. THE BEST LINEAR UNBIASED ESTIMATORS

Case I. σ is known and only μ is to be estimated.

Applying the Gauss-Markov theorem on least squares, the BLUE μ_0^* should be obtained by solving the equation $(\partial S / \partial \mu)_{\mu=\mu_0^*} = 0$, i.e.,

$$\begin{aligned} \left[\sum_{i=1}^k \frac{\lambda_{i+1} - \lambda_{i-1}}{(\lambda_{i+1} - \lambda_i)(\lambda_i - \lambda_{i-1})} f_i^2 - 2 \sum_{i=2}^k \frac{f_i f_{i-1}}{\lambda_i - \lambda_{i-1}} \right] \mu_0^* \\ = \sum_{i=1}^k \left[\frac{f_i - f_{i-1}}{\lambda_i - \lambda_{i-1}} - \frac{f_{i+1} - f_i}{\lambda_{i+1} - \lambda_i} \right] f_i (x(n_i) - u_i \sigma), \end{aligned}$$

or

$$K_1 u_O^* = X - \sigma K_3 \quad (3.1)$$

where

$$X = \sum_{i=1}^{k+1} \frac{(f_i - f_{i-1})(f_i x(n_i) - f_{i-1} x(n_{i-1}))}{\lambda_i - \lambda_{i-1}} \quad (3.2)$$

Hence one obtains

$$\mu_O^* = \sum_{i=1}^k a_i x(n_i) - \frac{K_3}{K_1} \sigma \quad (3.3)$$

where

$$a_i = \frac{f_i}{K_1} \left[\frac{f_i - f_{i-1}}{\lambda_i - \lambda_{i-1}} - \frac{f_{i+1} - f_i}{\lambda_{i+1} - \lambda_i} \right], \quad i = 1, 2, \dots, k \quad (3.4)$$

and

$$V(\mu_O^*) = \frac{\sigma^2}{n} \frac{1}{K_1} \quad (3.5)$$

For a given spacing $0 < \lambda_1 < \lambda_2 < \dots < \lambda_k < 1$, one can consider the *dual spacing*

$$0 < \lambda_1^* < \lambda_2^* < \dots < \lambda_k^* < 1 \quad (3.6)$$

where $\lambda_i^* = 1 - \lambda_{k-i+1}$, $i = 1, 2, \dots, k$. The self-dual spacing is called *symmetric*. For a symmetric spacing

$$\lambda_i + \lambda_{k-i+1} = 1, \quad i = 1, 2, \dots, k \quad (3.7)$$

If, in addition, the function is even, i.e., $f(-t) = f(t)$, then for a symmetric spacing

$$u_i + u_{k-i+1} = 0, \quad i = 1, 2, \dots, k \quad (3.8)$$

and hence

$$f_i = f_{k-i+1}, \quad i = 1, 2, \dots, k \quad (3.9)$$

and consequently

$$K_3 = 0 \quad (3.10)$$

Thus

$$\mu_O^* = \sum_{i=1}^k a_i x(n_i), \quad (3.11)$$

which is independent of σ .

Case II. μ is known and only σ is to be estimated.

The BLUE σ_0^* of σ should be obtained by solving the equation $(\partial S / \partial \sigma)_{\sigma=\sigma_0^*} = 0$, i.e.,

$$\begin{aligned} & \left[\sum_{i=1}^k \frac{\lambda_{i+1} - \lambda_{i-1}}{(\lambda_{i+1} - \lambda_i)(\lambda_i - \lambda_{i-1})} f_{ii}^2 u_i^2 - 2 \sum_{i=2}^k \frac{f_{ii} u_i - f_{i-1} u_{i-1}}{\lambda_i - \lambda_{i-1}} \right] \sigma_0^* \\ &= \sum_{i=1}^k \left[\frac{f_{ii} u_i - f_{i-1} u_{i-1}}{\lambda_i - \lambda_{i-1}} - \frac{f_{i+1} u_{i+1} - f_{ii} u_i}{\lambda_{i+1} - \lambda_i} \right] f_i (x(n_i) - \mu), \end{aligned}$$

or

$$K_2^* \sigma_0^* = Y - K_3 \mu \quad (3.12)$$

where

$$Y = \sum_{i=1}^{k+1} \frac{(f_{ii} u_i - f_{i-1} u_{i-1})(f_i x(n_i) - f_{i-1} x(n_{i-1}))}{\lambda_i - \lambda_{i-1}}. \quad (3.13)$$

Hence one obtains

$$\sigma_0^* = \sum_{i=1}^k b_i x(n_i) - \frac{K_3}{K_2} \mu \quad (3.14)$$

where

$$b_i = \frac{f_i}{K_2} \left[\frac{f_{ii} u_i - f_{i-1} u_{i-1}}{\lambda_i - \lambda_{i-1}} - \frac{f_{i+1} u_{i+1} - f_{ii} u_i}{\lambda_{i+1} - \lambda_i} \right], \quad i = 1, 2, \dots, k, \quad (3.15)$$

and

$$V(\sigma_0^*) = \frac{\sigma^2}{n} \frac{1}{K_2}. \quad (3.16)$$

If $f(-t) = f(t)$ and the spacing is symmetric, then

$$\sigma_0^* = \sum_{i=1}^k b_i x(n_i), \quad (3.17)$$

which is independent of σ .

Case III. Both μ and σ are unknown and they are to be jointly estimated.

The BLUE's μ^* , σ^* of μ , σ respectively are obtained by solving the equations $(\partial S / \partial \mu)_{\mu=\mu^*, \sigma=\sigma^*} = 0$ and $(\partial S / \partial \sigma)_{\mu=\mu^*, \sigma=\sigma^*} = 0$, i.e.,

i.e.,

$$K_1^* + K_3^* = X, \quad K_3^* + K_2^* = Y. \quad (3.18)$$

Hence by putting

$$\Delta = K_1 K_2 - K_3^2 \quad (3.19)$$

we have

$$\mu^* = \frac{K_2}{\Delta} X - \frac{K_3}{\Delta} Y = \sum_{i=1}^k c_i x(n_i), \quad c_i = \frac{K_2}{\Delta} a_i - \frac{K_3}{\Delta} b_i, \quad i = 1, \dots, k, \quad (3.20)$$

$$\sigma^* = -\frac{K_3}{\Delta} X + \frac{K_1}{\Delta} Y = \sum_{i=1}^k d_i x(n_i), \quad d_i = -\frac{K_3}{\Delta} a_i + \frac{K_1}{\Delta} b_i, \quad i = 1, \dots, k, \quad (3.21)$$

and

$$V(\mu^*) = \frac{\sigma^2}{n} \frac{K_2}{\Delta}, \quad V(\sigma^*) = \frac{\sigma^2}{n} \frac{K_1}{\Delta}, \quad \text{Cov}(\mu^*, \sigma^*) = -\frac{\sigma^2}{n} \frac{K_3}{\Delta}. \quad (3.22)$$

If $f(-t) = f(t)$ and the spacing is symmetric, then

$$\mu^* = \mu_o^*, \quad \sigma^* = \sigma_o^* \quad \text{and} \quad \text{Cov}(\mu^*, \sigma^*) = 0.$$

4. OPTIMAL SPACINGS FOR EXPONENTIAL DISTRIBUTION

In case of exponential distribution, $f(t) = e^{-t}$, $t > 0$, we have

$$\lambda_i = 1 - e^{-u_i}, \quad f_i = e^{-u_i}, \quad i = 1, 2, \dots, k,$$

and

$$K_2 = \frac{u_1^2}{e^{u_1-1}} + \frac{(u_2 - u_1)^2}{e^{u_2-1}} + \dots + \frac{(u_k - u_{k-1})^2}{e^{u_k-1}}. \quad (4.1)$$

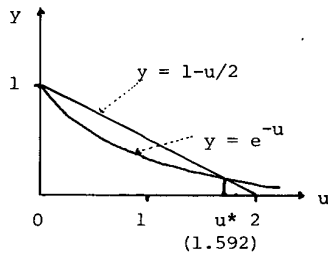
One has to find the spacing $\lambda_i = 1 - e^{-u_i}$, $i = 1, 2, \dots, k$, which maximizes K_2 . This can be done successively as follows. For $k = 1$

$$K_2^{(1)} = \phi(u_1) = u_1^2 / (e^{u_1} - 1), \quad \phi'(u_1) = \frac{u_1}{e^{u_1-1}} \left(2 - \frac{u_1}{1 - e^{-u_1}} \right),$$

hence the optimal spacing should be

$$e^{-u_1^*} = 1 - u_1^*/2, \quad u_1^* = 1.592, \quad \lambda_1^* = 1 - e^{-u_1^*} = 0.795$$

$$\phi''(u_1) = \frac{1}{e^{u_1-1}} \left[2 - 4 \frac{u_1 e^{u_1}}{e^{u_1-1}} - u_1 \frac{u_1 e^{u_1}}{e^{u_1-1}} + 2 \left(\frac{u_1 e^{u_1}}{e^{u_1-1}} \right)^2 \right].$$



Since

$$\frac{u_1^* e^{u_1^*}}{e^{u_1^*} - 1} = 2,$$

$$\phi''(u_1^*) = \frac{2 - 8 - 2u_1^* + 8}{e^{u_1^*} - 1} = \frac{2(1 - u_1^*)}{e^{u_1^*} - 1} < 0,$$

λ_1^* gives the optimal spacing and the maximum value of K_2 is $K_2^{(1)} = \phi(u_1^*) = 0.64761$.

For $k = 2$, K_2 comes out to be

$$K_2^{(2)} = \phi(u_1) + e^{-u_1} \phi(t_1^{(1)}), \quad t_1^{(1)} = u_2 - u_1.$$

Since $\phi(t_1^{(1)})$ is maximized for $t_1^{(1)} = u_1^* = 1.592$ taking on the maximum value 0.64761, the value $K_2^{(2)}$ should be maximized at

$$\phi'(u_1^{(2)}) - e^{-u_1^{(2)}} \times 0.64761 = 0$$

or

$$\psi(u_1^{(2)})^2 - 2\psi(u_1^{(2)}) + 0.64761 = 0,$$

where we have put

$$\psi(u) = ue^u / (e^u - 1).$$

Taking the fact that the function $\psi(u)$ is monotone increasing and $\psi(u) \geq 1$ in $0 \leq u < +\infty$ into account, we have $\psi(u_1^{(2)}) = 1 + \sqrt{1 - 0.64761}$. Hence

$$u_1^{(2)} = 1.02, \quad \lambda_1^{(2)} = 0.63941,$$

$$u_2^{(2)} = u_1^{(2)} + u_1^{(1)} = 2.61, \quad \lambda_2^{(2)} = 0.92647, \quad K_2^{(2)} = 0.82026.$$

For $k = 3$,

$$K_2^{(3)} = \phi(u_1) + e^{-u_1} [\phi(t_1^{(1)}) + e^{-t_1^{(1)}} \phi(t_1^{(2)})]$$

where

$$t_1^{(1)} = u_2 - u_1, \quad t_2^{(1)} = u_3 - u_1, \quad t_1^{(2)} = t_2^{(1)} - t_1^{(1)} = u_3 - u_2.$$

Now $\phi(t_1^{(2)})$ is maximized at $t_1^{(2)} = u_3 - u_2 = 1.59$ and $\phi(t_1^{(2)}) = 0.64761$. Then $(t_1^{(1)} + e^{-t_1^{(1)}} \times 0.64761)$ is maximized at $t_1^{(1)} = u_2 - u_1 = 1.02$; the

maximum value is 0.82026. Finally $\phi(u_1) + e^{-u_1} \times 0.82026$ is maximized at $u_1 = u_1^{(3)}$ satisfying $\psi(u_1^{(3)}) = 1 + \sqrt{1 - 0.82026}$, hence $u_1^{(3)} = 0.75$, $\lambda_1^{(3)} = 0.52763$. Consequently

$$u_2^{(3)} = 1.02 + 0.75 = 1.77, \quad \lambda_2^{(3)} = 0.82967,$$

$$u_3^{(3)} = 1.59 + 1.77 = 3.36, \quad \lambda_3^{(3)} = 0.96527, \quad K_2^{(3)} = 0.89049.$$

In this manner one can continue the calculations and determine the optimum spacings for $k = 4, 5, \dots, 15$. The result is presented in the following Table 4.1. Of course, since the highspeed computer is available nowadays, this whole process can be computerized rather easily. For the applications of this theory, see Greenberg and Sarhan (1958) and Ogawa (1960).

TABLE 4.1 Optimum spacings for estimates of relative efficiencies and the coefficients of best estimates

	1	2	3	4	5	6	7	8
u_1	1.59	1.02	0.75	0.61	0.50	0.43	0.37	0.33
λ_1	.79607	.63941	.52763	.45665	.39347	.34949	.30927	.28108
α_1	.40731	.42835	.39974	.36098	.33051	.29900	.27352	.24989
u_2		2.61	1.77	1.36	1.11	0.93	0.80	0.70
λ_2		.92647	.82967	.74334	.67044	.60545	.55067	.50341
α_2		.14687	.20233	.21719	.21896	.21499	.20654	.19671
u_3			3.36	2.38	1.86	1.54	1.30	1.13
λ_3			.96527	.90745	.84433	.78562	.72747	.67697
α_3			.06938	.10994	.13173	.14242	.14850	.14845
u_4				3.97	2.88	2.29	1.91	1.63
λ_4				.98113	.94387	.89873	.85192	.80407
α_4				.03770	.06668	.08571	.09839	.10677
u_5					4.47	3.31	2.66	2.24
λ_5					.98855	.96348	.93005	.89354
α_5					.02286	.04394	.05919	.07074
u_6						4.90	3.68	2.99
λ_6						.99255	.97478	.94971
α_6						.01487	.02996	.04255
u_7							5.27	4.01
λ_7							.99486	.98187
α_7							.01027	.02154
u_8								5.60
λ_8								.99630
α_8								.00739
u_9								
λ_9								
α_9								
u_{10}								
λ_{10}								
α_{10}								
u_{11}								
λ_{11}								
α_{11}								
\vdots								
K_2	.64761	.82026	.89049	.92691	.94757	.96056	.96926	.97537

(This table was calculated by the Support of the Ordnance Research through Dept. of Biostat., School of Public Health, University of North Carolina, Chapel Hill, N.C., U.S.A.)

TABLE 4.1 (continued)

	9	10	11	12	13	14	15
u_1	0.30	0.27	0.25	0.23	0.21	0.20	0.19
λ_1	.25918	.23662	.22120	.20547	.18942	.18127	.17304
a_1	.23224	.21644	.20181	.19003	.17766	.16759	.16096
u_2	0.63	0.57	0.52	0.48	0.44	0.41	0.39
λ_2	.46741	.43447	.40548	.38122	.35596	.33635	.32294
a_2	.18514	.17729	.16860	.16032	.15409	.14552	.13861
u_3	1.00	0.90	0.82	0.75	0.69	0.64	0.60
λ_3	.63212	.59343	.55956	.52762	.49842	.47271	.45119
a_3	.14569	.14135	.13803	.13398	.13001	.12609	.12031
u_4	1.43	1.27	1.15	1.05	0.96	0.89	0.83
λ_4	.76069	.71917	.68336	.65006	.61711	.58934	.56395
a_4	.10999	.11122	.11012	.10971	.10856	.10644	.10430
u_5	1.93	1.70	1.52	1.38	1.26	1.16	1.08
λ_5	.85485	.81732	.78129	.74842	.71635	.68651	.66040
a_5	.07910	.08396	.08661	.08748	.08887	.08890	.08803
u_6	2.54	2.20	1.95	1.75	1.59	1.46	1.35
λ_6	.92113	.88920	.85773	.82623	.79607	.76776	.74076
a_6	.05239	.06037	.06539	.06880	.07093	.07281	.07350
u_7	3.29	2.81	2.45	2.18	1.96	1.79	1.65
λ_7	.96275	.93980	.91371	.88696	.85914	.83304	.80795
a_7	.03152	.04000	.04702	.05195	.05578	.05802	.06019
u_8	4.31	3.56	3.06	2.68	2.39	2.16	1.98
λ_8	.98657	.97156	.95311	.94144	.90837	.88467	.86193
a_8	.01596	.02407	.03115	.03736	.04213	.04570	.04800
u_9	5.90	4.58	3.81	3.29	2.89	2.59	2.35
λ_9	.99726	.98975	.97785	.96275	.94442	.92498	.90463
a_9	.00547	.01218	.01874	.02475	.03028	.03447	.03778
u_{10}		6.17	4.38	4.04	3.50	3.09	2.78
λ_{10}		.99791	.99201	.98240	.96980	.95450	.93796
a_{10}		.00418	.00948	.01489	.02006	.02479	.02851
u_{11}			6.42	5.06	4.25	3.70	3.28
λ_{11}			.99837	.99365	.98574	.97528	.96237
a_{11}			.00325	.00754	.01207	.01643	.02051
u_{12}				6.65	5.22	4.45	3.89
λ_{12}				.99871	.99486	.98832	.97956
a_{12}				.00258	.00611	.00989	.01358
u_{13}					6.86	5.47	4.64
λ_{13}					.99895	.99579	.99034
a_{13}					.00210	.00500	.00817
u_{14}						7.06	5.66
λ_{14}						.99914	.99652
a_{14}						.00172	.00414
u_{15}							7.25
λ_{15}							.99929
a_{15}							.00142
K_2	.97982	.98316	.98574	.98739	.98939	.99071	.99180

5. OPTIMAL SPACINGS FOR THE NORMAL DISTRIBUTION

In this case

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad f(-t) = f(t),$$

$$\lambda_i = (2\pi)^{-1/2} \int_{-\infty}^{u_i} e^{-t^2/2} dt, \quad f_i = (2\pi)^{-1/2} e^{-u_i^2/2}, \quad i = 1, \dots, k.$$

Case I. σ is known.

Since

$$\frac{\partial K_1}{\partial u_i} = f_i \left[\frac{f_{i+1} - f_i}{\lambda_{i+1} - \lambda_i} - \frac{f_i - f_{i-1}}{\lambda_i - \lambda_{i-1}} \right] \left[2u_i + \frac{f_{i+1} - f_i}{\lambda_{i+1} - \lambda_i} + \frac{f_i - f_{i-1}}{\lambda_i - \lambda_{i-1}} \right], \quad i = 1, \dots, k,$$

and due to the convexity of f as a function of λ

$$\frac{f_i - f_{i-1}}{\lambda_i - \lambda_{i-1}} > \frac{f_{i+1} - f_i}{\lambda_{i+1} - \lambda_i}, \quad i = 1, \dots, k,$$

the optimal spacing must satisfy the following system of equations

$$2u_i + \frac{f_{i+1} - f_i}{\lambda_{i+1} - \lambda_i} + \frac{f_i - f_{i-1}}{\lambda_i - \lambda_{i-1}} = 0, \quad i = 1, \dots, k. \quad (5.1)$$

It is easy to see that the system of equations (5.1) is self-dual, i.e., the dual of the i -th equation is nothing but $(k+i+1)$ -th equation of the system. Higuchi(1956) has shown that the system of equations (5.1) has a unique solution giving the maximum of K_1 . Hence the optimal spacing must necessarily symmetric. Therefore the number of equations of (5.1) reduces to $k/2$ or $(k+1)/2$ according as k is even or odd. Solving those equations numerically we obtained the optimal spacings up to $k = 15$. The result is presented in the following Table 5.1.

TABLE 5.1

Optimal spacing for estimating the mean μ of normal population when σ is known.

	1	2	3	4	5	6	7
u_1		-.612003	-.981599	-1.244357	-1.446850	-1.610758	-1.747928
λ_1		.270268	.163149	.106684	.073969	.053616	.040238
a_1		.500000	.295321	.191840	.132852	.096389	.072472
u_2		.612003	.000000	-.382284	-.658911	-.874362	-1.049957
λ_2		.729732	.500000	.351125	.254976	.190961	.146869
a_2		.500000	.409358	.308160	.232597	.178691	.139991
u_3			.981599	.382284	.000000	-.280288	-.500550
λ_3			.836851	.648875	.500000	.389628	.308344
a_3			.295321	.308160	.269103	.224920	.186259
u_4				1.244357	.658911	.280288	.000000
λ_4				.893316	.745024	.610372	.500000
a_4				.191840	.232597	.224920	.202555
u_5					1.446850	.874362	.500550
λ_5					.926031	.809039	.691656
a_5					.132852	.178691	.186259
u_6						1.610758	1.049957
λ_6						.946384	.853131
a_6						.096389	.139991
u_7							1.747928
λ_7							.959762
a_7							.072472
K_1		.809826	.882518	.920059	.942022	.956000	.965452
$K_2/2$.330079	.532326	.657107	.738358	.793954	.833588
$\Delta/2$.267306	.469787	.604577	.695550	.759019	.804789

TABLE 5.1 (continued)

	8	9	10	11	12	13	14	15
u_1	-1.865528	-1.968218	-2.059193	-2.140732	-2.214552	-2.281837	-2.343673	-2.400804
λ_1	.031054	.024521	.019738	.016148	.013396	.011249	.009547	.008180
a_1	.056053	.044362	.035788	.029342	.024392	.020522	.017449	.014974
u_2	-1.197594	-1.324583	-1.435733	-1.534370	-1.622890	-1.703070	-1.776268	-1.843532
λ_2	.115538	.092655	.075539	.062469	.052306	.044277	.037844	.032626
a_2	.111701	.090604	.074564	.062154	.052399	.044620	.038337	.033205
u_3	-.681217	-.833841	-.965597	-1.081245	-1.184106	-1.276582	-1.360470	-1.437139
λ_3	.247867	.202185	.167123	.139794	.118186	.100857	.086841	.075339
a_3	.154613	.129197	.108831	.092438	.079148	.068284	.059329	.051886
u_4	-.221819	-.404740	-.559913	-.694313	-.812600	-.918039	-1.013009	-1.099286
λ_4	.412227	.342834	.287769	.243743	.208224	.179299	.155528	.135822
a_4	.177633	.154329	.133851	.116305	.101420	.088826	.078157	.069092
u_5	.221819	.000000	-.183729	-.340142	-.476012	-.595882	-.702950	-.799550
λ_5	.587773	.500000	.427113	.366875	.317033	.275627	.241043	.211986
a_5	.177633	.163015	.146966	.131448	.117231	.104540	.093354	.083554
u_6	.681217	.404740	.183729	.000000	-.156887	-.293514	-.414311	-.522404
λ_6	.752133	.657166	.572887	.500000	.437667	.384565	.339323	.300695
a_6	.154613	.154329	.146966	.136626	.125410	.114360	.103965	.094424
u_7	1.197594	.833841	.559913	.340142	.156887	.000000	-.136929	-.258222
λ_7	.884462	.797815	.712231	.633125	.562333	.500000	.445544	.398118
a_7	.111701	.129197	.133851	.131448	.125410	.117696	.109409	.101151
u_8	1.865528	1.324583	.965597	.694313	.476012	.293514	.136929	.000000
λ_8	.968946	.907345	.832877	.756257	.682967	.615435	.554456	.500000
a_8	.056053	.090604	.108831	.116305	.117231	.114360	.109409	.103427
u_9		1.968218	1.435733	1.081245	.812600	.595882	.414311	.258222
λ_9		.975497	.924461	.860206	.791776	.724373	.660677	.601882
a_9		.044362	.074564	.092438	.101420	.104540	.103965	.101151
u_{10}			2.059193	1.534370	1.184106	.918039	.702950	.522404
λ_{10}			.980262	.937531	.881814	.820701	.758957	.699305
a_{10}			.035788	.062154	.079148	.088826	.093354	.094424
u_{11}				2.140733	1.622890	1.276582	1.013009	.799550
λ_{11}				.983852	.947694	.899125	.844472	.788014
a_{11}				.029342	.052399	.068284	.078157	.083554
u_{12}					2.214522	1.703070	1.360470	1.099286
λ_{12}					.986604	.955723	.913159	.864178
a_{12}					.024392	.044620	.059329	.069092
u_{13}						2.281837	1.776268	1.437139
λ_{13}						.988751	.962156	.924661
a_{13}						.020522	.038337	.051886
u_{14}							2.343673	1.843532
λ_{14}							.990453	.967374
a_{14}							.017449	.033205
u_{15}								2.400804
λ_{15}								.991820
a_{15}								.014974
K_1	.972147	.977063	.980780	.983660	.985937	.987768	.989263	.990499
$K_2/2$.862811	.884965	.902156	.915762	.926714	.935660	.943062	.949255
$\Delta/2$.838779	.864667	.884817	.900799	.913682	.924215	.932936	.940236

Case II. μ is known.

In this case,

$$\frac{\partial K_2}{\partial u_i} = f_i \left[\frac{f_{i+1}u_{i+1} - f_i u_i}{\lambda_{i+1} - \lambda_i} - \frac{f_i u_i - f_{i-1} u_{i-1}}{\lambda_i - \lambda_{i-1}} \right] [2u_i^2 - 2 + \frac{f_{i+1}u_{i+1} - f_i u_i}{\lambda_{i+1} - \lambda_i} + \frac{f_i u_i - f_{i-1} u_{i-1}}{\lambda_i - \lambda_{i-1}}].$$

Let

$$G_i = \frac{f_{i+1}u_{i+1} - f_i u_i}{\lambda_{i+1} - \lambda_i} - \frac{f_i u_i - f_{i-1} u_{i-1}}{\lambda_i - \lambda_{i-1}}, \quad H_i = 2u_i^2 + \frac{f_{i+1}u_{i+1} - f_i u_i}{\lambda_{i+1} - \lambda_i} + \frac{f_i u_i - f_{i-1} u_{i-1}}{\lambda_i - \lambda_{i-1}},$$

$i = 1, \dots, k,$

then one has to consider 2^k systems of equations such as

$$G_{i_1} = 0, \dots, G_{i_m} = 0, H_{i_{m+1}} = 0, \dots, H_{i_k} = 0, \quad m = 0, 1, \dots, k,$$

where (i_1, i_2, \dots, i_k) is a permutation of $(1, 2, \dots, k)$. Higuchi (1956) has shown that the optimal spacing must be among the solutions of the system of equations

$$H_i = 0, \quad i = 1, 2, \dots, k. \quad (5.2)$$

This system is also dual. Higuchi (1956) has shown that the system (5.2) has a solution — spacing — for any assigned number of negative u_i , and that gives a local maximum of K_2 . Hence there are $k+1$ local maxima of K_2 . One has to pick up the greatest maximum among them. It turns out that the optimal spacing is symmetric when k is even, whereas the optimal spacings are asymmetric when k is odd. The numerical results are presented in the following Table 5.2.

TABLE 5.2

Optimal spacing for estimating σ of normal population when μ is known.

	1	2	3	4	5	6	7
u_1		-1.482072	-1.452028	-1.995607	-1.982115	-2.313048	-2.305346
λ_1		.069161	.073247	.022988	.023733	.010360	.010574
b_1		-.337365	-.311281	-.115343	-.113609	-.054920	-.054812
u_2		1.482072	1.185513	-1.140138	-1.118931	-1.600190	-1.589698
λ_2		.930839	.882093	.127114	.131585	.054778	.055951
b_2		.337365	.253632	-.236657	-.231751	-.124340	-.123875
u_3			2.024851	1.140138	.983717	-.955753	-.940040
λ_3			.978559	.872886	.837373	.169599	.173598
b_3			.122146	.236657	.185739	-.182056	-.180571
u_4				1.995607	1.619039	.955753	.854805
λ_4				.977012	.947281	.830401	.803670
b_4				.115343	.125884	.182056	.148468
u_5					2.326934	1.600190	1.384815
λ_5					.990016	.945222	.916946
b_5					.055427	.124340	.117101
u_6						2.313048	1.910546
λ_6						.989640	.917968
b_6						.054920	.072965
u_7							2.548769
λ_7							.994595
b_7							.030797
$K_2/2$.652245	.735800	.824396	.858772	.894290	.911681
K_1		.511738	.612725	.712490	.761475	.809198	.836651
$\Delta/2$.333778	.446339	.587374	.653239	.723657	.762591
K_3/K_2		.000000	.064496	.000000	.012691	.000000	.010074

$$(\hat{\sigma}^* = \sum_{i=1}^k b_i x(n_i) - (K_3/K_2)\mu)$$

TABLE 5.2 (continued)

	8	9	10	11	12	13	14	15
u_1	-2.540770	-2.535802	-2.717094	-2.713636	-2.860243	-2.857701	-2.980280	-2.978340
λ_1	.005530	.005610	.003293	.003327	.002117	.002134	.001440	.001449
b_1	-.030817	-.030867	-.019144	-.019194	-.012756	-.012792	-.008960	-.008974
u_2	-1.900308	-1.893940	-2.122213	-2.117929	-2.297395	-2.294318	-2.441524	-2.439212
λ_2	.028696	.029116	.016910	.017091	.010798	.010886	.007313	.007360
b_2	-.072916	-.072976	-.046484	-.046586	-.031516	-.031596	-.022393	-.022449
u_3	-1.371482	-1.363168	-1.652317	-1.467036	-1.864608	-1.860935	-2.034780	-2.032078
λ_3	.085112	.086415	.049235	.049775	.031118	.031377	.020936	.021073
b_3	-.116768	-.116704	-.077448	-.077568	-.053733	-.053849	-.038767	-.038855
u_4	-.835453	-.823296	-1.215945	-1.209191	-1.479335	-1.474894	-1.681780	-1.678611
λ_4	.201731	.205170	.112003	.113295	.069525	.070120	.046306	.046614
b_4	-.147218	-.146606	-.106263	-.106308	-.076514	-.076639	-.056400	-.056512
u_5	.835453	.763439	-.749071	-.739341	-1.101083	-1.095476	-1.349216	-1.345428
λ_5	.798269	.777399	.226907	.229850	.135430	.136654	.088634	.089244
b_5	.147218	.123599	-.123029	-.122708	-.096081	-.096146	-.073381	-.073492
u_6	1.371482	1.225978	.749071	.694439	-.683252	-.675257	-1.011761	-1.007019
λ_6	.914888	.889897	.773093	.756296	.247224	.249756	.155826	.156963
b_6	.116768	.106287	.123029	.105609	-.105284	-.105080	-.086998	-.087058
u_7	1.900308	1.660178	1.215945	1.108968	.683252	.640027	-.631016	-.624304
λ_7	.971304	.951561	.887997	.866278	.752776	.738923	.264015	.266214
b_7	.027916	.077337	.106263	.096031	.105284	.091957	-.091741	-.091592
u_8	2.540770	2.128597	1.652317	1.485589	1.101083	1.018158	.631016	.595750
λ_8	.994470	.983356	.950765	.931306	.864570	.845698	.735985	.724329
b_8	.030817	.046374	.077448	.076371	.096081	.086939	.091741	.081246
u_9		2.722254	2.122213	1.869782	1.479335	1.354334	1.011761	.945072
λ_9		.996758	.983090	.969243	.930475	.912185	.844174	.827689
b_9		.019085	.046484	.053593	.076514	.073249	.086998	.079047
u_{10}			2.717094	2.301732	1.864608	1.686064	1.349216	1.250873
λ_{10}			.996707	.989325	.968882	.954108	.911366	.894510
b_{10}			.019144	.031419	.053733	.056265	.073381	.069350
u_{11}				2.863825	2.297395	2.038437	1.681780	1.546036
λ_{11}				.997907	.989202	.979247	.953694	.938952
b_{11}				.012712	.031516	.038658	.056400	.056498
u_{12}					2.860243	2.444652	2.034781	1.849161
λ_{12}					.997883	.992750	.979064	.967783
b_{12}					.012756	.022324	.038767	.042531
u_{13}						2.982907	2.441524	2.179081
λ_{13}						.998572	.992687	.985337
b_{13}						.008920	.022393	.028814
u_{14}							2.980280	2.565922
λ_{14}							.998560	.994855
b_{14}							.008550	.016459
u_{15}								3.085354
λ_{15}								.998983
b_{15}								.006514
$K_2/2$.929439	.939435	.949577	.955844	.962178	.966365	.970585	.973518
K_1	.863344	.880323	.896848	.908113	.919097	.926975	.934670	.940408
$\Delta/2$.802426	.826953	.851626	.867994	.884335	.895787	.907176	.915500
K_3/K_2	.000000	.005529	.000000	.003369	.000000	.002208	.000000	.001526

Case III. Joint estimation of μ and σ .

There is a conjecture that *the optimal spacing in this case must be symmetric.*

However, we have been able to show that this conjecture is correct only for $k = 2$ (Ogawa (1976a,b)).

For $k = 2$, it turns out that

$$\Delta \equiv K_1 K_2 - K_3^2 = \begin{vmatrix} \frac{\sqrt{\lambda_1}}{f_1} & \frac{\sqrt{\lambda_2 - \lambda_1}}{f_2 - f_1} & \frac{\sqrt{1 - \lambda_2}}{-f_2} \\ \frac{\sqrt{\lambda_1}}{f_1 u_1} & \frac{\sqrt{\lambda_2 - \lambda_1}}{f_2 u_2 - f_1 u_1} & \frac{\sqrt{1 - \lambda_2}}{-f_2 u_2} \end{vmatrix} = \frac{(u_2 - u_1)^2 f_1^2 f_2^2}{\lambda_1 (\lambda_2 - \lambda_1) (1 - \lambda_2)}. \quad (5.3)$$

It follows that

$$\frac{\partial \log \Delta}{\partial u_1} = \frac{-2}{u_2 - u_1} - 2u_1 + \frac{f_1}{\lambda_1} + \frac{f_1}{\lambda_2 \lambda_1} = 0, \quad \frac{\partial \log \Delta}{\partial u_2} = \frac{2}{u_2 - u_1} - 2u_2 - \frac{f_2}{\lambda_2 - \lambda_1} + \frac{f_2}{1 - \lambda_2} = 0. \quad (5.4)$$

Hence the optimal spacing must satisfy the equation

$$\frac{f_1}{\lambda_1} + \frac{f_2 - f_1}{\lambda_2 - \lambda_1} + \frac{-f_2}{1 - \lambda_2} + 2(u_1 + u_2) = 0. \quad (5.5)$$

One can show that this equation has a unique solution such that $u_1 + u_2 = 0$.

Although we have not yet been able to show theoretically that the optimal spacings are symmetric, we present numerical results by computer search in the following Table 5.3. We hope that those results are very near the optimal.

TABLE 5.3

Optimal spacing for estimating μ and σ jointly in case of normal population.

	1	2	3	4	5	6	7
u_1		-1.110591	-1.383403	-1.696102	-1.881729	-2.060013	-2.194565
λ_1		.133372	.083271	.044933	.029936	.019699	.014065
c_1		.500000	.224449	.108404	.067572	.043173	.030123
d_1		-.450211	-.361428	-.201303	-.140604	-.095707	-.070629
u_2		1.110591	.000000	-.689421	-.996941	-1.264729	-1.455219
λ_2		.866628	.500000	.245279	.159396	.102984	.072804
c_2		.500000	.551102	.391596	.234282	.142000	.096637
d_2		.450211	.000000	-.230004	-.236144	-.186276	-.146903
u_3			1.383403	.689421	.000000	-.491807	-.786273
λ_3			.916729	.754721	.500000	.311428	.215854
c_3			.224449	.391596	.396292	.314827	.217038
d_3			.361428	.230004	.000000	-.136750	-.166812
u_4				1.696102	.996941	.491806	.000000
λ_4				.955067	.840604	.688572	.500000
c_4				.108404	.234282	.314827	.312405
d_4				.201303	.236144	.136750	.000000
u_5					1.881729	1.264729	.786273
λ_5					.970064	.897016	.784146
c_5					.067572	.142000	.217038
d_5					.140604	.186276	.166812
u_6						2.060013	1.455219
λ_6						.980301	.927196
c_6						.043173	.096637
d_6						.095707	.146903
u_7							2.195465
λ_7							.985935
c_7							.030123
d_7							.070629
$\Delta/2$.406503	.552681	.682616	.755659	.810322	.847387
K_1		.695217	.853665	.881605	.923516	.938378	.953481
$K_2/2$.584714	.647421	.774288	.818242	.863535	.888729

TABLE 5.3 (continued)

	8	9	10	11	12	13	14
u_1	-2.318807	-2.422614	-2.516547	-2.599516	-2.675132	-2.743687	-2.806783
λ_1	.010203	.007705	.005926	.004668	.003735	.003038	.002502
c_1	.021555	.016102	.012291	.009624	.007665	.006211	.005100
d_1	-.052750	-.040906	-.032229	-.025951	-.021188	-.017554	-.014708
u_2	-1.621826	-1.757898	-1.878343	-1.982883	-2.076834	-2.161028	-2.237765
λ_2	.052420	.050273	.038076	.029643	.023503	.018975	.015535
c_2	.068011	.050273	.038076	.029643	.023503	.018975	.015535
d_2	-.114630	-.091540	-.073747	-.060406	-.050006	-.041902	-.035445
u_3	-1.022305	-1.204642	-1.360159	-1.491503	-1.607191	-1.709222	-1.801028
λ_3	.153308	.114171	.086893	.067915	.054006	.043705	.035849
c_3	.148655	.107737	.080469	.062042	.048838	.039213	.031966
d_3	-.153529	-.132444	-.111862	-.094543	-.080076	-.068272	-.058550
u_4	-.382800	-.649718	-.862079	-1.033132	-1.178842	-1.304219	-1.414928
λ_4	.350934	.257937	.194322	.150771	.119231	.096079	.078545
c_4	.761779	.196759	.145461	.110621	.086107	.068551	.055514
d_4	-.090940	-.123801	-.125005	-.115205	-.102730	-.090599	-.079599
u_5	.382800	.000000	-.314269	-.553385	-.746620	-.906411	-1.043417
λ_5	.649066	.500000	.376658	.290000	.227647	.182359	.148378
c_5	.261779	.258259	.223703	.178023	.138722	.109541	.087960
d_5	.090940	.000000	-.065103	-.095281	-.102565	-.099391	-.092318
u_6	1.022350	.649718	.314269	.000000	-.267072	-.481728	-.658847
λ_6	.846692	.742063	.623342	.500000	.394707	.315000	.254988
c_6	.148655	.196759	.223703	.220096	.195164	.161661	.130930
d_6	.153529	.123801	.065103	.000000	-.048999	-.075443	-.085160
u_7	1.621826	1.204642	.862079	.553385	.267072	.000000	-.232470
λ_7	.947580	.885829	.805678	.710000	.605293	.500000	.408086
c_7	.068011	.107737	.145461	.178023	.195164	.191698	.172994
d_7	.114630	.132444	.125005	.095281	.048999	.000000	-.038237
u_8	2.318807	1.757898	1.360139	1.033132	.746620	.481728	.232470
λ_8	.989797	.960618	.913107	.849229	.772353	.685000	.591914
c_8	.021555	.050273	.080469	.110621	.138722	.161661	.172994
d_8	.052750	.091540	.111862	.115205	.102565	.075443	.038237
u_9		2.422614	1.878343	1.491503	1.178842	.906411	.658874
λ_9		.992295	.969833	.932085	.880769	.817641	.745012
c_9		.016102	.038076	.062042	.086107	.109541	.130930
d_9		.040906	.073747	.094543	.102730	.099391	.085160
u_{10}			2.516547	1.982383	1.607191	1.304219	1.043417
λ_{10}			.994074	.976310	.945994	.903921	.851622
c_{10}			.012291	.029643	.048838	.068551	.087960
d_{10}			.032229	.060406	.080076	.090599	.092318
u_{11}				2.599516	2.076834	1.709222	1.414928
λ_{11}				.995332	.981029	.956295	.921455
c_{11}				.009624	.023503	.039213	.055514
d_{11}				.025951	.050006	.068272	.079599
u_{12}					2.675132	2.161029	1.801028
λ_{12}					.996265	.984653	.964151
c_{12}					.007665	.018975	.031966
d_{12}					.021188	.041902	.058550
u_{13}						2.743687	2.237765
λ_{13}						.996962	.987382
c_{13}						.006211	.015535
d_{13}						.017554	.035445
u_{14}							2.806783
λ_{14}							.997498
c_{14}							.005100
d_{14}							.014708
$\Delta/2$.875282	.895995	.912097	.924703	.934818	.943021	.949780
K_1	.961709	.968858	.973707	.977728	.980772	.983288	.985309
$K_2/2$.910133	.924794	.936727	.945767	.953145	.959048	.963941

For the applications of this theory, the reference is made to Eisenberger and Bosner (1965).

ACKNOWLEDGEMENT

We acknowledge that this research was supported financially by NRC through Grant No.A7683.

REFERENCES

- Eisenberger, I. and Bosner, E.C., 1965. Systematic statistics used for data compression in space telemetry. *J. Amer. Statist. Assoc.* 60: 87-133.
- Greenberg, B.G. and Sarhan, A.I., 1958. Applications of order statistics to health data. *Ann. Journ. Publ. Health* 48: 1388-1394.
- Higuchi, I., 1956. On the solutions of certain simultaneous equations in the theory of systematic statistics. *Ann. Inst. Stat. Math., Tokyo* 5:71-90.
- Mosteller, F., 1946. On some useful "inefficient" statistics. *Ann. Math. Statist.* 19: 377-407.
- Ogawa, J., 1960. Determination of optimum spacings for the estimation of the scale parameter of an exponential distribution based on sample quantiles. *Ann. Inst. Stat. Math., Tokyo* 12: 135-141.
- Ogawa, J., 1976a. A note on the optimal spacing of the systematic statistics — Normal distribution. In: S.Ikeda et al.(eds.), *Essays in Probability and Statistics*. Ogawa Volume: 467-474.
- Ogawa, J., 1976b. Optimal spacings for the simultaneous estimation of the location and scale parameters of a normal distribution based on selected two sample quantiles. *Jour. Statist. Plan. Inf.* 1: 61-72.

ASYMPTOTICS FOR THE MULTISAMPLE, MULTIVARIATE CRAMÉR - VON MISES STATISTIC WITH SOME POSSIBLE APPLICATIONS

D.S.COTTERILL¹ and M.CSÖRGÖ²

1 Dept. of National Defence, Ottawa (Canada)

2 Carleton University, Ottawa (Canada)

ABSTRACT

Cotterill, D.S. and Csörgö, M., Asymptotics for the multisample, multivariate Cramér-von Mises statistic with some possible applications. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29 - Dec. 1, 1979

Let Y_1, \dots, Y_n ($n = 1, 2, \dots$) be independent rv uniformly distributed over the d -dimensional unit cube I^d ($d \geq 1$) and let $\{a_n(y), y \in I^d, n = 1, 2, \dots\}$ be the empirical process based on this sequence of random samples. Let $V_{n,d}(\cdot)$ be the distribution function of the rv $\int_{I^d} a_n^2(y) dy$, and let $V_d(\cdot)$ be that of $\int_{I^d} B^2(y) dy$, where $\{B(y); y \in I^d\}$ is a Brownian bridge, i.e., a separable Gaussian process with $EB(y) = 0$ and $EB(x)B(y) = \prod_{i=1}^d (x_i \wedge y_i) - (\prod_{i=1}^d x_i)(\prod_{i=1}^d y_i)$. Put $\Delta_{n,d} = \sup\{|V_{n,d}(x) - V_d(x)|; 0 \leq x < \infty\}$.

In Cotterill and Csörgö (1980), we proved the rates of convergence for the latter distance; for example, $\Delta_{n,2} = O(n^{-1/2} \log^2 n)$ for $d = 2$. Again in the quoted paper, we also calculated the "usual" levels of significance of the distribution function $V_d(\cdot)$ for $d = 2$ to 50. Previously these were only known for $d = 1$ (Anderson-Darling (1952)), $d = 2$ (Durbin (1970)) and $d = 3$ (Krivyakova, Martynov and Tyurin (1977)).

Preliminaries and earlier results are summarized in Section 1. Section 2 is devoted to some asymptotics for two-sample Cramér-von Mises functionals in terms of empirical measures as measures of integration. Section 3 is on some possible applications.

1. INTRODUCTION

Let Y_1, \dots, Y_n be independent rv uniformly distributed over the d -dimensional unit cube I^d ($d \geq 1$), and let $E_n(y)$ be the empirical distribution function of Y_1, \dots, Y_n , i.e., for $y = (y_1, \dots, y_d) \in I^d$, $E_n(y)$ is n^{-1} times the number of $Y_i = (Y_{i1}, \dots, Y_{id})$, $j = 1, \dots, n$, whose components are less than or equal to the corresponding components of y , conveniently written as

$$E_n(y) = E_n(y_1, \dots, y_d) = n^{-1} \prod_{j=1}^n \prod_{i=1}^d I_{[0, y_i]}(Y_{ij}), \quad (1.1)$$

where, for real numbers $a, u \in [0, 1]$,

$$I_{[0, a]}(u) = \begin{cases} 1 & \text{if } u \leq a \\ 0 & \text{if } u > a \end{cases}. \quad (1.2)$$

Consider the uniform empirical process

$$\alpha_n(y) = n^{1/2} [E_n(y) - \lambda(y)], \quad y \in I^d, \quad d \geq 1, \quad (1.3)$$

where $\lambda(y) = \prod_{i=1}^d y_i$.

It will be convenient for us to also think about $\alpha_n(\cdot)$ in terms of continuous distribution functions F on R^d . Let F be the class of continuous distribution functions on d -dimensional Euclidean space R^d ($d \geq 1$), and let F_0 be the subclass consisting of every member of F which is a product of its associated 1-dimensional marginal distribution functions. Let $F_n(x)$ be the empirical distribution function of X_1, \dots, X_n , i.e., for $x = (x_1, \dots, x_d) \in R^d$, $F_n(x)$ is n^{-1} times the number of $X_j = (X_{j1}, \dots, X_{jd})$, $j = 1, \dots, n$, whose components are less than or equal to the corresponding components of x , namely

$$F_n(x) = F_n(x_1, \dots, x_d) = n^{-1} \prod_{j=1}^n \prod_{i=1}^d I_{(-\infty, x_i]}(X_{ji}), \quad (1.4)$$

where, for all real a, u

$$I_{(-\infty, a]}(u) = \begin{cases} 1 & \text{if } u \leq a \\ 0 & \text{if } u > a \end{cases}. \quad (1.5)$$

Consider the empirical process

$$\beta_n(x) = n^{1/2} [F_n(x) - F(x)], \quad x = (x_1, \dots, x_d) \in R^d, \quad d \geq 1. \quad (1.6)$$

Let $y_i = F_i(x_i)$ be the i th marginal distribution function of $F \in F$ and let $F_i^{-1}(\cdot)$ be its inverse. Now if $F \in F_0$, then

$$\begin{aligned} \beta_n(x) &= n^{1/2} [F_n(x) - \prod_{i=1}^d F_i(x_i)] \\ &= n^{1/2} [F_n(F_1^{-1}(y_1), \dots, F_d^{-1}(y_d)) - \lambda(y)] \\ &= n^{1/2} [E_n(y) - \lambda(y)] = \alpha_n(y), \quad y = (y_1, \dots, y_d) \in I^d, \quad d \geq 1, \end{aligned} \quad (1.7)$$

i.e., if $F \in F_0$, then β_n is distribution-free.

As to $\alpha_n(\cdot)$, the following results are known.

Theorem A. Let X_1, \dots, X_n ($n = 1, 2, \dots$) be independent random d -vectors with a common distribution function $F \in F_0$ and let $\alpha_n(\cdot)$ be as in (1.7). Then one can construct a probability space (Ω, \mathcal{A}, P) with $\{\alpha_n(y); y \in I^d (d \geq 1)\}$, a sequence of Brownian bridges $\{B_n(y); y \in I^d (d \geq 1)\}$ and a Kiefer process $\{K(y, t); y \in I^d (d \geq 1), t \geq 0\}$ on it so that for any $\mu > 0$ there exists a $C > 0$ such that (cf. Csörgő and Révész (1975a)) for each n

$$P \left\{ \sup_{y \in I^d} |\alpha_n(y) - B_n(y)| > C(\log n)^{3/2} n^{-1/(2(d+1))} \right\} \leq n^{-\mu}, \quad d \geq 1, \quad (1.8)$$

and whence

$$\sup_{y \in I^d} |\alpha_n(y) - B_n(y)| \stackrel{a.s.}{=} O[(\log n)^{3/2} n^{-1/(2(d+1))}], \quad d \geq 1, \quad (1.9)$$

$$\sup_{1 \leq k \leq n} \sup_{y \in I^d} |k^{1/2} \alpha_k(y) - K(y, k)| \stackrel{a.s.}{=} O[\log^2 n \cdot n^{(d+1)/(2(d+2))}], \quad d \geq 1. \quad (1.10)$$

Also if $d = 1$, then (cf. Lomlós, Major and Tusnády (1975)) for all n and x

$$P\left\{ \sup_{0 \leq y \leq 1} |\alpha_n(y) - B_n(y)| > n^{-1/2} (C \log n + x) \right\} < Le^{-\lambda x}, \quad (1.11)$$

where C, L, λ are positive absolute constants (e.g., (cf. Tusnády (1977a)) they can be chosen as $C = 100, L = 10, \lambda = 1/50$), and

$$P\left\{ \sup_{1 \leq k \leq n} \sup_{0 \leq y \leq 1} |k^{1/2} \alpha_k(y) - K(y, k)| > (C \log n + x) \log n \right\} < Le^{-\lambda x}, \quad (1.12)$$

where again C, L, λ are positive absolute constants, and where

$$\sup_{0 \leq y \leq 1} |\alpha_n(y) - B_n(y)| \stackrel{a.s.}{=} O(n^{-1/2} \log n), \quad (1.13)$$

$$\sup_{1 \leq k \leq n} \sup_{0 \leq y \leq 1} |k^{1/2} \alpha_k(y) - K(y, k)| \stackrel{a.s.}{=} O(\log^2 n). \quad (1.14)$$

Further, if $d = 2$, then (cf. Tusnády (1977)) for all n and x

$$P\left\{ \sup_{y \in I^2} |\alpha_n(y) - B_n(y)| > n^{-1/2} (C \log n + x) \log n \right\} < Le^{-\lambda x}, \quad (1.15)$$

where C, L, λ are positive absolute constants, and whence

$$\sup_{y \in I^2} |\alpha_n(y) - B_n(y)| \stackrel{a.s.}{=} O(n^{-1/2} \log^2 n). \quad (1.16)$$

The respective a.s. rates of (1.9), (1.10), (1.14) and (1.16) are best available, while that of (1.13) is best possible. For further illuminating comments concerning rates in higher dimensions we refer to Tusnády (1977b).

The Brownian bridges and the Kiefer process of the above theorem are Gaussian processes, defined in terms of a multi-parameter Wiener process as follows:

D1. Wiener process: A separable Gaussian process

$$W(x) = \{W(x_1, \dots, x_d); 0 \leq x_i < \infty \ (i = 1, \dots, d)\}.$$

with $EW(x) = 0$ and covariance function $EW(x_1)W(x_2) = \lambda(x_1 \wedge x_2)$, where $x_1 = (x_{11}, \dots, x_{1d})$, $x_2 = (x_{21}, \dots, x_{2d})$, $x_1 \wedge x_2 = (x_{11} \wedge x_{21}, \dots, x_{1d} \wedge x_{2d})$ and $\lambda(x_1 \wedge x_2) = \prod_{i=1}^d (x_{1i} \wedge x_{2i})$.

D2. Brownian bridge:

$$\begin{aligned} B(x) &= \{B(x_1, \dots, x_d); 0 \leq x_i \leq 1 \ (i = 1, \dots, d)\} \\ &= \{W(x) - \lambda(x)W(1, \dots, 1); x \in I^d\}, \text{ with } \lambda(x) = \prod_{i=1}^d x_i. \end{aligned}$$

$$\text{Whence } EB(x) = 0 \text{ and } EB(x_1)B(x_2) = \lambda(x_1 \wedge x_2) - \lambda(x_1)\lambda(x_2).$$

D3. Kiefer process :

$$\begin{aligned} K(x, t) &= \{K(x, t); x \in I^d, t \geq 0\} \\ &= \{W(x, t) - \lambda(x)W(1, \dots, 1, t); x \in I^d, t \geq 0\}. \end{aligned}$$

$$\text{whence } EK(x, t) = 0 \text{ and } EK(x_1, t_1)K(x_2, t_2) = (t_1 \wedge t_2) \{\lambda(x_1 \wedge x_2) - \lambda(x_1)\lambda(x_2)\}.$$

Given $F \in F_0$, we are interested in the asymptotic distribution of the multivariate Cramér-von Mises statistic

$$W_{n,d}^2 = \int_{R^d} \beta_n^2(x) \prod_{i=1}^d dF_i(x_i) = \int_{I^d} \alpha_n^2(y) \prod_{i=1}^d dy_i, \quad d \geq 1, \quad (1.17)$$

where $\beta_n(x)$, $\alpha_n(x)$, $y_i = F_i(x_i)$ are as in (1.7). Naturally, say by (1.9), we have for $d \geq 1$ that

$$h(\alpha_n(\cdot)) \xrightarrow{\mathcal{D}} h(B(\cdot)), \quad (1.18)$$

for every continuous functional h on the space of real valued functions on I^d endowed with the supremum topology, and whence also that

$$W_{n,d}^2 \xrightarrow{\mathcal{D}} W_d^2 = \int_{I^d} B^2(y) dy \stackrel{\mathcal{D}}{=} \int_{I^d} B_n^2(y) dy = W_d^2(n), \quad d \geq 1, \quad (1.19)$$

with dy standing for $\prod_{i=1}^d dy_i$ from now on. A direct way of seeing (1.19) is via

$$|W_{n,d}^2 - W_d^2(n)| \stackrel{a.s.}{=} \begin{cases} O[r_{1d}(n)(\log \log n)^{1/2}] & \text{if } d \geq 3, \\ O[\rho_2(n)(\log \log n)^{1/2}] & \text{if } d = 2, \\ O[\rho_1(n)(\log \log n)^{1/2}] & \text{if } d = 1, \end{cases} \quad (1.20)$$

or via

$$|W_{n,d}^2 - n^{-1} \int_{I^d} K^2(y, n) dy| \stackrel{a.s.}{=} \begin{cases} O[n^{-1/2} r_{2d}(n)(\log \log n)^{1/2}] & \text{if } d \geq 2, \\ O[\rho_2(n)(\log \log n)^{1/2}] & \text{if } d = 1, \end{cases} \quad (1.21)$$

where

$$r_{id}(n) = \begin{cases} n^{-1/(2(d+1))} (\log n)^{3/2} & \text{if } i = 1, \\ n^{(d+1)/(2(d+2))} \log^2 n & \text{if } i = 2, \end{cases} \quad (1.22)$$

and

$$\rho_i(n) = \begin{cases} n^{-1/2} \log n & \text{if } i = 1 \\ n^{-1/2} \log^2 n & \text{if } i = 2. \end{cases} \quad (1.23)$$

The respective statements of (1.20) follow from (1.9), (1.16) and (1.13) respectively and those of (1.21) by (1.10) and (1.14) respectively when they are also combined with appropriate laws of iterated logarithm. From (1.21), in turn, not only can we deduce that (1.19) is true, but also a law of iterated logarithm for $W_{n,d}^2$ from that of $\int_{\Gamma_d} K^2(y,n) dy$. For a proof of (1.20) and (1.21) we refer to that of Corollary 1 in Csörgő (1979).

In addition to (1.19), from Theorem A we can also prove rates of convergence results for this convergence in distribution. Let $V_{n,d}(x)$ be the distribution function of $W_{n,d}^2$ of (1.17) and let $V_d(x)$ be that of W_d^2 of (1.19). Then (1.19) reads

$$\lim_{n \rightarrow \infty} P\{W_{n,d}^2 \leq x\} = \lim_{n \rightarrow \infty} V_{n,d}(x) = V_d(x), \quad d \geq 1. \quad (1.24)$$

Put $\Delta_{n,d} = \sup_{0 \leq x < \infty} |V_{n,d}(x) - V_d(x)|$. Then we have

Theorem B (Götze (1979)). $\Delta_{n,1} = O(n^{-1+\varepsilon})$ for any $\varepsilon > 0$.

This theorem is the best available such result for $\Delta_{n,1}$ so far. Earlier S. Csörgő (1976) showed that $\Delta_{n,1} = O(n^{-1/2} \log n)$ and, on the basis of his complete asymptotic expansion for the Laplace transform of $W_{n,1}^2$, he conjectured that $\Delta_{n,1}$ is of order $1/n$. This conjecture was further studied by S. Csörgő and L. Stachó (1979).

As to higher dimensions $d \geq 2$, nothing is known about the exact distribution function $V_{n,d}$ (cf. (1.24)), and only the characteristic function of V_d (cf. (1.24)) is known (cf. Dugue (1969), Durbin (1970)), and that (cf. Anderson and Darling (1952), Rosenblatt (1952b)) W_d^2 may be written in the form

$$W_d^2 = \sum_{k=1}^{\infty} \mu_k^{-1} x_k^2, \quad d \geq 1, \quad (1.25)$$

where x_k are independent standard normal random variables and μ_k are the eigenvalues of the integral equation

$$\int_{\Gamma_d} E\{B(x_1)B(x_2)\}f(x_2)dx_2 = \mu f(x_1) \quad (1.26)$$

with eigen functions f and kernel $EB(x_1)B(x_2)$ (cf. D2.). Whence, in order to tabulate V_d ($d \geq 2$), one may try working with a numerical inversion of the characteristic function of V_d , or one may try to calculate a number of necessary eigenvalues for (1.25). Unfortunately both ways turn out to be quite difficult to follow directly. Durbin (1970) succeeded in solving the latter problem for $d = 2$, and Krivyakova, Martynov and Tyurin (1977) for $d = 3$.

Using the characteristic function of Dugue (1969), we obtain in Collerll and

Csörgő (1980) a recursive equation in the cumulants of W_d^2 , and then use the Cornish-Fisher asymptotic expansion to calculate its critical values for $d = 2, 3, \dots, 50$ at various levels of rejection probabilities. These are within 3 % of Durbin's value for $d = 2$ and those of Krivyakova, Martynov and Tyurin for $d = 3$. As far as we know there exist no other tables for $d \geq 4$.

Since nothing is known about the exact distribution function $V_{n,d}$ for $d \geq 2$, it is desirable to have a Theorem B type result also for $\Delta_{n,d}$ when $d \geq 2$. As to the latter we have

Theorem C (Cotterill and Csörgő (1980)).

$$\Delta_{n,d} = \begin{cases} O(n^{-1/2} \log^2 n) & \text{if } d = 2, \\ O(n^{-1/(2(d+1))} (\log n)^{3/2}) & \text{if } d \geq 3. \end{cases} \quad (1.27)$$

As far as we know, the rates of (1.27) are the only available ones for $\Delta_{n,d}$ ($d \geq 2$) and these combined with Theorem B tell the whole story as presently known for $d \geq 1$.

2. ON SOME MULTIVARIATE TWO-SAMPLE ASYMPTOTICS.

Let $X_j = (X_{j1}, \dots, X_{jd})$ ($j = 1, \dots, n$), $Y_j = (Y_{j1}, \dots, Y_{jd})$ ($j = 1, \dots, m$) be two independent random samples with respective distribution functions F and G in F . Let $F_n(x)$ and $G_m(x)$ be the empirical distribution functions of X_j and Y_j , respectively (cf. (1.4)). Given $F = G \in \bar{F}_0$, let $S_{n,m}(x) = (nF_n(x) + mG_m(x))/(n+m)$, i.e., the empirical distribution function of the two independent random samples combined. One is frequently interested in the asymptotic distribution of multi-sample statistic like, e.g., that of

$$W_{nm,d}^2 = \frac{nm}{n+m} \int_{R^d} (F_n(x) - G_m(x))^2 dS_{n+m}(x), \quad d \geq 1. \quad (2.1)$$

Given $F = G \in \bar{F}_0$, let

$$\begin{aligned} \beta_{nm}(x) &= [nm/(n+m)]^{1/2} (F_n(x) - G_m(x)) \\ &= [m/(n+m)]^{1/2} n^{1/2} (F_n(x) - \prod_{i=1}^d F_i(x_i)) - [n/(n+m)]^{1/2} m^{1/2} (G_m(x) - \prod_{i=1}^d F_i(x_i)) \\ &= [m/(n+m)]^{1/2} \beta_n^{(1)}(x) - [n/(n+m)]^{1/2} \beta_m^{(2)}(x), \quad x \in R^d, (d \geq 1), \end{aligned} \quad (2.2)$$

where $\beta_n^{(1)}$ and $\beta_m^{(2)}$ are two independent empirical processes defined as β_n of (1.7), and

$$B_{nm}(y) = [m/(n+m)]^{1/2} B_n^{(1)}(y) - [n/(n+m)]^{1/2} B_m^{(2)}(y), \quad y \in I^d, (d \geq 1), \quad (2.3)$$

where $B_n^{(1)}$ and $B_n^{(2)}$ are two independent sequences of Brownian bridges.

Since $B_{nm}(y) \xrightarrow{\mathcal{D}} B(y)$, $y \in I^d$ ($d \geq 1$), a Brownian bridge for each n and m , $\{B_{nm}\}$ of (2.3) is also a sequence of Brownian bridges. Now the sequence $\{B_n^{(1)}\}$ resp. $\{B_n^{(2)}\}$ can be so constructed that it approximates $\beta_n^{(1)}$ resp. $\beta_n^{(2)}$ à la Theorem A. Hence, given $F = G \in F_0$, a version of Theorem A can be stated immediately also in terms of β_{nm} and B_{nm} . For example with $y_i = F_i(x_i)$ as in (1.7), for

$$\beta_{nm}(x) = [m/(n+m)]^{1/2} \alpha_n^{(1)}(y) - [n/(n+m)]^{1/2} \alpha_m^{(2)}(y) = \alpha_{nm}(y), \quad (2.4)$$

$$y = (y_1, \dots, y_d) \in I^d \quad (d \geq 1),$$

where $\alpha_n^{(1)}$ and $\alpha_n^{(2)}$ are defined as α_n of (1.7), we have the following version of (1.8):

There exists a sequence of Brownian bridges $\{B_{nm}\}$ so that for any $\mu > 0$ there is a $C > 0$ such that for each n and m

$$P\left\{\sup_{y \in I^d} |\alpha_{nm}(y) - B_{nm}(y)| > C(r_{1d}(n) \vee r_{1d}(m))\right\} \leq (n^{-\mu} \vee m^{-\mu}), \quad (2.5)$$

where $r_{1d}(\cdot)$ is as in (1.22). Whence

$$\sup_{y \in I^d} |\alpha_{nm}(y) - B_{nm}(y)| \xrightarrow{a.s.} O(r_{1d}(n) \vee r_{1d}(m)), \quad d \geq 1, \quad (2.6)$$

and so $h(\alpha_{nm}(\cdot)) \xrightarrow{\mathcal{D}} h(B(\cdot))$ for every continuous functional h on the space of real valued functions on I^d endowed with the supremum topology, which, in turn, implies

$$\int_{I^d} \alpha_{nm}^2(y) dy = \int_{R^d} \beta_{nm}^2(x) \prod_{i=1}^d dF_i(x_i) \xrightarrow{\mathcal{D}} W_d^2, \quad d \geq 1, \quad (2.7)$$

where for W_d^2 we refer to (1.19).

Hence for $W_{nm,d}^2 = \int_{R^d} \beta_{nm}^2(x) dS_{n+m}(x)$ of (2.1) we should also have the same convergence in distribution statement, i.e., that

$$W_{nm,d}^2 \xrightarrow{\mathcal{D}} W_d^2, \quad d \geq 1. \quad (2.8)$$

Indeed, it follows from an appropriate analogue in the present case of Lemma in Kiefer (1959) that the statement of (2.8) is true. We are going to present this appropriate analogue of Kiefer's Lemma here via extending the proof of Corollary 5.6.4 of the forthcoming monograph of Csörgö and Révész (1980) to the present multivariate situation. We have

PROPOSITION 1. Given $F = G \in F_0$, (2.6) is true, i.e.,

$$\lim_{n,m \rightarrow \infty} P\left\{ \int_{\mathbb{R}^d} \beta_{nm}^2(x) dS_{n+m}(x) \leq u \right\} = P\left\{ \int_{\mathbb{I}^d} B^2(y) dy \leq u \right\}, \quad u \geq 0 \quad (d \geq 1). \quad (2.9)$$

The proof of Proposition 1 is based on either of the next two preliminary corollaries, which might be of some interest on their own.

Let $y_i = F_i^{-1}(x_i)$ be the i th marginal distribution function of F and let $F_i^{-1}(\cdot)$ be its inverse. Define (cf. (2.3) for $B_{nm}(y)$)

$$W_{d,n,m}^2 = \int_{\mathbb{I}^d} B_{nm}^2(y) dS_{n+m}(y), \quad d \geq 1, \quad (2.10)$$

where $S_{n+m}(y) = \{nF_n^{-1}(y_1), \dots, F_d^{-1}(y_d)\} + \{mG_m^{-1}(y_1), \dots, F_d^{-1}(y_d)\} / (n+m)$.

In terms of our present terminology, $W_{nm,d}^2$ of (2.1) (cf. also (2.4)) can be written as

$$W_{nm,d}^2 = \int_{\mathbb{I}^d} \alpha_{nm}^2(y) dS_{n+m}(y), \quad d \geq 1, \quad (2.11)$$

and we have the following corollary to Theorem A à la (1.20).

COROLLARY 1. Given $F = G \in F_0$

$$|W_{nm,d}^2 - W_{d,n,m}^2| \stackrel{a.s.}{\leq} \begin{cases} O[(r_{1d}(n) \vee r_{1d}(m)) (\log \log(n \vee m))^{1/2}] & \text{if } d \geq 3, \\ O[(\rho_2(n) \vee \rho_2(m)) (\log \log(n \vee m))^{1/2}] & \text{if } d = 2, \\ O[(\rho_1(n) \vee \rho_1(m)) (\log \log(n \vee m))^{1/2}] & \text{if } d = 1. \end{cases} \quad (2.12)$$

PROOF. We give details for the case of $d \geq 3$. The proof in the latter two cases goes similarly. We have

$$\begin{aligned} |W_{nm,d}^2 - W_{d,n,m}^2| &\leq \int_{\mathbb{I}^d} |\alpha_{nm}(y) - B_{nm}(y)| \cdot |\alpha_{nm}(y) + B_{nm}(y)| dS_{n+m}(y) \\ &\leq \left(\sup_{y \in \mathbb{I}^d} |\alpha_{nm}(y) - B_{nm}(y)| \right) \left(\sup_{y \in \mathbb{I}^d} |B_{nm}(y) - \alpha_{nm}(y)| + 2 \sup_{y \in \mathbb{I}^d} |\alpha_{nm}(y)| \right) \\ &\stackrel{a.s.}{\leq} O(r_{1d}(n) \vee r_{1d}(m)) O(r_{1d}(n) \vee r_{1d}(m)) + O((\log \log(n \vee m))^{1/2}) \\ &\stackrel{a.s.}{\leq} O((r_{1d}(n) \vee r_{1d}(m)) (\log \log(n \vee m))^{1/2}), \end{aligned}$$

where the line before the last one follows from applying (2.6) twice and by the law of iterated logarithm for $\sup \{|\alpha_{nm}(y)| : y \in \mathbb{I}^d\}$. The latter, in turn, is done via applying the law of iterated logarithm to $\alpha_n^{(1)}$ and $\alpha_m^{(2)}$ (cf. (2.4)) separately via (1.10) and Theorem 2 in Csörgő and Chan (1976).

The other preliminary corollary to Proposition 1 we have in mind is

COROLLARY 2. Given $F = G \in F_0$, and $W_{d,n,m}^2$ and $W_{nm,d}^2$ as in (2.10) and (2.11) respectively, then

$$\sup_{0 < u < \infty} |P\{W_{nm,d}^2 \leq u\} - P\{W_d^2(n,m) \leq u\}| = \begin{cases} O(r_{1d}(n) r_{1d}(m)) & \text{if } d \geq 2, \\ O(\rho_2(n) \rho_2(m)) & \text{if } d = 2, \\ O(\rho_1(n) \rho_1(m)) & \text{if } d = 1. \end{cases} \quad (2.13)$$

PROOF. For $d \geq 2$ we use the inequality of (2.5) (an analogue of (1.8)), for $d = 2$ we use a similar analogue of (1.15) and for $d = 1$ that of (1.11) and, mutatis mutandis, repeat the proof of Theorem 1.

PROOF OF PROPOSITION 1. By (2.12)

$$\lim_{n,m \rightarrow \infty} |W_{nm,d}^2 - W_d^2(n,m)| \xrightarrow{\text{a.s.}} 0, \quad d \geq 1.$$

Hence, or by (2.13), it suffices to show that for any given $\epsilon > 0$ and $0 < \delta < 1$ there exists an $n_0 = n_0(\epsilon, \delta)$ such that

$$P\left\{ \left| \int_{I^d} B_{nm}^2(y) dS_{n+m}(y) - \int_{I^d} B_{nm}^2(y) dy \right| > \epsilon \right\} < \delta, \quad \text{whenever } n \geq n_0. \quad (2.14)$$

Let R_1, \dots, R_{ℓ^d} ($d \geq 1$) be d -dimensional rectangles of equal sides of length $1/\ell$, which are obtained by subdividing each side of I^d into ℓ equal parts. Then

The left hand side of inequality (2.14)

$$\begin{aligned} & \leq P\left\{ \sum_{k=1}^{\ell^d} \left| \int_{R_k} B_{nm}^2(y) dS_{n+m}(y) - \int_{R_k} B_{nm}^2(y) dy \right| > \epsilon \right\} \\ & \leq P\left\{ \sum_{k=1}^{\ell^d} \left| \int_{R_k} B_{nm}^2(y) dS_{n+m}(y) \right| + \sum_{k=1}^{\ell^d} \left| \int_{R_k} B_{nm}^2(y) dy \right| > \epsilon \right\} \\ & \leq P\left\{ \ell^d \max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| \left(\int_{R_k} dS_{n+m}(y) \right) > \epsilon/2 \right\} + P\left\{ \ell^d \max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| \frac{1}{\ell^d} > \epsilon/2 \right\} \\ & \leq P\left\{ \ell^d \left(\max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| \right) \left(\max_{1 \leq k \leq \ell^d} S_{n+m}(R_k) \right) > \epsilon/2 \right\} + P\left\{ \max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| > \epsilon/2 \right\} \\ & \leq P\left\{ \ell^d \left(\max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| \right) \left(\max_{1 \leq k \leq \ell^d} |S_{n+m}(R_k) - \frac{1}{\ell^d}| + \frac{1}{\ell^d} \right) > \epsilon/2 \right\} \\ & \quad + P\left\{ \max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| > \epsilon/2 \right\} \\ & \leq P\left\{ \max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| > \epsilon/4 \right\} + P\left\{ \max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| > \sqrt{\epsilon}/2 \right\} \\ & \quad + P\left\{ \ell^d \max_{1 \leq k \leq \ell^d} |S_{n+m}(R_k) - \frac{1}{\ell^d}| > \sqrt{\epsilon}/2 \right\} + P\left\{ \max_{1 \leq k \leq \ell^d} \sup_{y \in R_k} |B_{nm}^2(y)| > \epsilon/2 \right\} \\ & = P_1 + P_2 + P_3 + P_4. \end{aligned} \quad (2.15)$$

Given $\varepsilon > 0$ and $0 < \delta < 1$, we can choose now ℓ so big that $P_1 < \delta/4$, $P_2 < \delta/4$ and $P_4 < \delta/4$, since for each n and m

$$\max_{1 \leq k \leq \ell} \sup_{y \in R_k} |B_{nm}^2(y)| \stackrel{D}{=} \max_{1 \leq k \leq \ell} \sup_{y \in R_k} |B^2(y)| \stackrel{a.s.}{=} O((\log \ell)/\ell^d),$$

by the continuity modulus of $B(y)$, $y \in I^d$, or by simply saying that the latter Gaussian process is uniformly continuous over I^d . For the already given ℓ , $\varepsilon > 0$ and $0 < \delta < 1$, next we choose n and m so big that $P_3 < \delta/4$ by the Glivenko-Cantelli theorem. This also completes the proof of Proposition 1.

REMARK 2.1. Given $F \in F_0$, β_n as in (1.7), the statement of Proposition 1 can, of course, be also stated for the latter as follows :

$$\lim_{n \rightarrow \infty} P\left\{ \int_{R^d} \beta_n^2(x) dF_n(x) \leq u \right\} = P\left\{ \int_{I^d} B^2(y) dy \leq u \right\}, \quad u \geq 0 \quad (d \geq 1). \quad (2.16)$$

We note also that the calculation of the statistic $\int_{R^d} \beta_n^2(x) dF_n(x)$ is easier than that of $w_{n,d}^2$ of (1.17) (cf. Durbin (1970)) and, in the light of (2.16), the former can be used instead of the latter in practical situations. The proof of (2.16) is, mutatis mutandis, identical to that of (2.9).

3. SOME APPLICATIONS OF THE $w_{n,d}^2$ STATISTIC.

The results of sections 1-2 are valid in terms of the statistical goodness-of-fit hypothesis

$$H_0 : F \in F_0, \quad (3.1)$$

and are directly applicable to test such a one, provided the marginal distribution functions F_i of $F = \prod_{i=1}^d F_i$ under H_0 are completely specified. This, of course, also limits the direct use of $w_{n,d}^2$ to such completely defined goodness-of-fit situations only. On the other hand, if $X = (X_1, \dots, X_d)$ is a random vector with distribution function $F(x) = F(x_1, \dots, x_d) \in F$, then making the transformation T $Y = (Y_1, \dots, Y_d) = TX = T(x_1, \dots, x_d)$, where T is given by

$$y_i = P\{X_i \leq x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}\} = F_i(x_i | x_{i-1}, \dots, x_1), \quad i = 1, \dots, d,$$

the random vector $Y = TX$ is uniformly distributed over I^d , i.e., Y_1, \dots, Y_d are uniformly and identically distributed on $[0,1]$ (cf. Rosenblatt (1952a)). Thus T transforms the hypothetical distribution F to the uniform distribution on I^d , provided $y = (y_1, \dots, y_d)$ is completely specified, for then $w_{n,d}^2$ is computable. The transformation T with further conditions on $F \in F$ was also used by Csörgő and Révész (1975b) in order to prove a Theorem A-type strong invariance principle for the original (i.e., F is not necessarily in F_0) empirical process $\beta_n(x)$ of (1.6). There is, of course, nothing wrong with working with the transformation

T in order to prove a theorem for the original empirical process $\beta_n(x)$ of (1.6) of the random sample $X_j = (X_{j1}, \dots, X_{jd})$, $j = 1, 2, \dots, n$. But if we are to apply T to X_j to begin with, in order to get a uniformly distributed random sample $Y_j = TX_j$ ($j = 1, \dots, n$), then there are $d!$ transformations T of the type described above corresponding to the $d!$ ways in which one can number the coordinates x_1, \dots, x_d . Thus if T is to be applied to the original random sample X_j ($j = 1, \dots, n$), then, from a statistical point of view, an element of arbitrariness is introduced. Anyhow, it is hoped that, just like in the case of statistics based on the univariate sample distribution function, our results and tables will have a wider range of applicability than the test of multivariate goodness-of-fit (for more remarks in this direction we refer to Durbin (1970, Introduction)) even though our discussion was conducted in goodness-of-fit terms only.

Indeed, it is more appropriate to test the null hypothesis $H_0: F \in F_0$ against the alternative $H_1: F \in F - F_0$ in terms of the Hoeffding, Blum, Rosenblatt empirical process (cf. Hoeffding (1948) and Blum, Kiefer and Rosenblatt (1961))

$$T_n(x) = T_n(x_1, \dots, x_d) = n^{1/2} \left[F_n(x) - \prod_{i=1}^d F_{ni}(x_i) \right], \quad d \geq 2, \quad (3.2)$$

where $F_{ni}(x_i)$ is the marginal empirical distribution function of the i th component of X_j , than in terms of $\beta_n(x)$ of (1.7), since T_n of (3.2) does not depend on the particular form of F . Given $F \in F_0$, strong approximations of the process T_n are discussed by Csörgő (1979), and in a forthcoming paper we are going to carry out the program of the two parts of the present paper for the asymptotic distribution of the statistics $\int_{\mathbb{R}^d} T_n^2(x) \prod_{i=1}^d dF_i(x_i)$ and $\int_{\mathbb{R}^d} T_n^2(x) dF_n(x)$, $d \geq 2$.

Here we are going to apply our results to a characterization based goodness-of-fit test for normality and to some two-sample tests of independence. Towards a test for normality, we first quote a characterization theorem for the univariate normal family $N(x; \mu, \sigma^2)$:

Theorem D (Bondesson (1974)). Let Y_1, \dots, Y_m be univariate independent rv with continuous distribution functions and such that Y_1 and Y_2 have the same distribution function. Let

$$Z_i = \left(\sum_{k=1}^i Y_k - i Y_{i+1} \right) / (i(i+1))^{1/2}, \quad i = 1, \dots, m-1, \quad (3.3)$$

and

$$X_i = i^{1/2} Z_{i+1} / \left(\sum_{k=1}^i Z_k^2 \right)^{1/2}, \quad i = 1, \dots, m-2. \quad (3.4)$$

Then, provided $m \geq 6$, X_1, X_2, \dots, X_{m-2} are independent Student rv with 1, 2, ..., $m-2$ degrees of freedom respectively if and only if the Y_i ($i = 1, 2, \dots, m$) are i.i.d. $N(x; \mu, \sigma^2)$ rv.

A weaker version of this theorem first appeared in Csörgő, Seshadri and Yalovsky (1973). It can clearly serve as a basis for a test of normality of the Y_i ($i = 1, \dots, m$), since it replaces the latter composite statistical hypothesis by the equivalent simple one that the X_i ($i = 1, \dots, m-2$) are independent Student rv with $1, 2, \dots, m-2$ degrees of freedom respectively. The latter randomization reduction of m variables to $m-2$ variables is very economical: two nuisance (from the composite goodness-of-fit point of view) parameters are eliminated and we are left with $m-2$ variables to base our simple goodness-of-fit test on. Unfortunately our reduced problem is in terms of independent but not identically distributed rv, and this, in turn, presents difficulties when trying to construct exact tests (cf. Csörgő, Seshadri and Yalovsky (1973)). We are going to tackle this problem asymptotically here. Towards this end we first give a multi-sample version of Theorem D.

PROPOSITION 2. Let $Y_j = (Y_{j1}, Y_{j2}, \dots, Y_{jm})$, m (fixed) ≥ 6 , $j = 1, 2, \dots, n$, $n = 1, 2, \dots$, be i.i.d. random m -vectors with independent univariate components for each j . Assume that the latter marginal distribution functions are continuous and such that Y_{j1} and Y_{j2} are identically distributed. Let

$$Z_{ji} = \left(\sum_{k=1}^i Y_{jk} - i Y_{ji+1} \right) / (i(i+1))^{1/2}, \quad i = 1, \dots, m-1, \quad j = 1, \dots, n, \quad (3.5)$$

and

$$X_{ji} = i^{1/2} Z_{ji+1} / \left(\sum_{k=1}^i Z_{jk}^2 \right)^{1/2}, \quad i = 1, \dots, m-2, \quad j = 1, \dots, n. \quad (3.6)$$

Then $X_j = (X_{j1}, X_{j2}, \dots, X_{jm-2})$, $j = 1, \dots, n$, are i.i.d. random $(m-2)$ -vectors by definition, and, for each j , the components $X_{j1}, X_{j2}, \dots, X_{jm-2}$ are independent Student rv with $1, 2, \dots, m-2$ degrees of freedom respectively if and only if Y_{ji} ($1 \leq j \leq n$, $1 \leq i \leq m$) are i.i.d. $N(x; \mu, \sigma^2)$ rv.

Proposition 2 is simply a repetition of Theorem D n times, resulting in $m-2$ independent Student random samples of size n , each with respective degrees of freedom $1, 2, \dots, m-2$.

Let $y_i = F_i(x_i)$ be Student distribution functions with i degrees of freedom ($i = 1, \dots, m-2$), and let

$$F_n(x) = F_n(x_1, \dots, x_{m-2}) = n^{-1} \sum_{j=1}^n \prod_{i=1}^{m-2} I_{(-\infty, x_i]}(x_{ji})$$

be the empirical distribution function (cf. (1.4)) of $X_j = (X_{j1}, X_{j2}, \dots, X_{jm-2})$ of Proposition 2. In terms of these ingredients, define $\beta_n(x) = \alpha_n(y)$ as in (1.7). Then, with $d = m-2$ ($m \geq 6$), Theorem D holds for $X_j = (X_{j1}, X_{j2}, \dots, X_{jm-2})$, $j = 1, 2, \dots, n$, $n = 1, 2, \dots$, of Proposition 2, and whence, again with $d = m-2$ ($m \geq 6$), the appropriate statements of (1.18), (1.19), (1.20) and (1.27) are also true for

these X_j . Hence, when testing the composite goodness-of-fit hypothesis

$$H_0 : Y_j = (Y_{j1}, Y_{j2}, \dots, Y_{jm}) , m(\text{fixed}) \geq 6, j = 1, \dots, n, \text{ are } n \text{ i.i.d. } m\text{-vectors} \\ \text{of } m \text{ i.i.d. } N(x; \mu, \sigma^2) \text{ rv} \quad (3.7)$$

via the equivalent simple goodness-of-fit hypothesis

$$H'_0 : X_j = (X_{j1}, X_{j2}, \dots, X_{jm-2}) \text{ of Proposition 2, } m(\text{fixed}) \geq 6, j = 1, \dots, n, \\ \text{are } n \text{ i.i.d. } (m-2)\text{-vectors of independent Student rv with } 1, 2, \dots, m-2 \\ \text{degrees of freedom respectively,} \quad (3.8)$$

we can base our critical region on large values of

$$W_{n,m-2}^2 = \int_{R^{m-2}} \beta_n^2(x) \prod_{i=1}^{m-2} dF_i(x_i) \\ = \int_{R^{m-2}} n [F_n(x_1, \dots, x_{m-2}) - \prod_{i=1}^{m-2} F_i(x_i)]^2 \prod_{i=1}^{m-2} dF_i(x_i), m(\text{fixed}) \geq 6, \quad (3.9)$$

where $F_n(x_1, \dots, x_{m-2})$ is the empirical distribution function of $X_j = (X_{j1}, \dots, X_{jm-2})$, $j = 1, \dots, n$, of Proposition 2, and F_i ($i = 1, \dots, m-2$) are Student distribution functions with $i = 1, \dots, m-2$ degrees of freedom respectively. By (1.19) $W_{n,m-2}^2 \xrightarrow{D} \int_{T^{m-2}} B^2(y) dy$, $m \geq 6$, and significance points of the latter random variables are tabulated in Tables 1 and 2 of Cotterill and Csörgö (1980).

The nature of the transformations (3.3) and (3.4) is similar to the briefly mentioned T transformation of the first paragraph of this section in that, and in Theorem D too, we are randomizing, since the original Y_i can be taken in any order to form the Z_i , and then the latter can again be taken in any order to form the X_i . This is clearly a drawback of our suggested test for normality in terms of $W_{n,m-2}^2$. On the other hand, if the very nature of our problem is such that it directly fits the formulation of H_0 of (3.7), i.e., if we have a large number of independent small samples of size $m \geq 6$, then $W_{n,m-2}^2$ of (3.9) appears to be quite an economical way of assessing normality in the presence of unknown parameters. We may also have one very large set of data which we may wish to summarize somehow for the sake of testing for normality. We can then subdivide the original set of data into a large number of small sets of size $m \geq 6$, and then test for normality of the original set via $W_{n,m-2}^2$ of (3.9). Thus, when our problem is that inherently we have a very large number of observations instead of that of having too few, now that we have Tables 1 and 2 of Cotterill and Csörgö (1980) available, our suggested $W_{n,m-2}^2$ procedure might come in handy at least as an omnibus, preliminary test for normality.

Now we turn to a multivariate two-sample problem. Let $X_j = (X_{j1}, \dots, X_{jd})$ ($j = 1, \dots, n$), $Y_j = (Y_{j1}, \dots, Y_{jd})$ ($j = 1, \dots, m$) be two independent random samples with respective distribution functions F and G in F . Let F_i and G_i be the marginal

distribution function of the i th component of X_j resp. that of the i th component of Y_j . We wish to test the composite null hypothesis

$$H_0 : F = G \text{ and } F(x) \in F_0, \text{ for all } x = (x_1, \dots, x_d), \quad (3.10)$$

without knowing what that common distribution function F is.

First we consider the statistic $W_{nm,d}^2$ of (2.1), for which Proposition 1 holds under H_0 of (3.10), and hence our Table 1 and 2 in Cotterill and Csörgő (1980) are again available. Burke (1977) showed that the test for H_0 which is based on large values of $W_{nm,d}^2$ is consistent against the following two alternatives :

$$H_1 : F(x_0) \neq G(x_0) \text{ for some } x_0 \in R^d, \text{ and } F(x) \in F_0 \text{ for all } x \in R^d, \quad (3.11)$$

$$H_2 : F(x_0) \neq G(x_0) \text{ for some } x_0 \in R^d, \text{ and } F(x_0) \neq \prod_{i=1}^d F_i(x_{0i}) \text{ for some } x_0 \in R^d. \quad (3.12)$$

However, in general, the $W_{nm,d}^2$ test is not consistent against the alternative:

$$H_3 : F = G \text{ and } F(x_0) \neq \prod_{i=1}^d F_i(x_{0i}) \text{ for some } x_0 \in R^d. \quad (3.13)$$

H_1 , H_2 and H_3 exhaust all the possible alternatives to H_0 of (3.10). In order to handle H_3 , Burke (1977) proposed to consider the following two-sample empirical process:

$$Z_{nm}(x) = [nm/(n+m)]^{1/2} [F_n(x) + G_m(x) - 2 \prod_{i=1}^d F_{ni}(x_i)], \quad x = (x_1, \dots, x_d), \quad (3.14)$$

where $F_{ni}(x_i)$ is the marginal empirical distribution function of the i th component of X_j ($j = 1, \dots, n$). We have (cf. Theorem 2 in Burke (1977) and Theorem 5 in Csörgő (1979)) the following (2.5) type version of (1.8):

Given H_0 of (3.10), there exists a sequence of Brownian bridges B_{nm} so that for any $\mu > 0$ there is a $C > 0$ such that for each n and m

$$P\left\{ \sup_{x \in R^d} |Z_{nm}(x) - B_{nm}(F_1(x_1), \dots, F_d(x_d))| > C(r_{1d}(n) \vee r_{1d}(m)) \right\} \leq (n^{-\mu} \vee m^{-\mu}), \quad (3.15)$$

where $r_{1d}(\cdot)$ is as in (1.22). Whence

$$\sup_{x \in R^d} |Z_{nm}(x) - B_{nm}(F_1(x_1), \dots, F_d(x_d))| \stackrel{a.s.}{=} O(r_{1d}(n) \vee r_{1d}(m)), \quad d \geq 1, \quad (3.16)$$

and so

$$\int_{R^d} Z_{nm}^2(x) \prod_{i=1}^d dF_i(x_i) \xrightarrow{P} W_d^2, \quad d \geq 1, \quad (3.17)$$

where for W_d^2 we refer to (1.19).

We note also that (3.15) and (3.16) type versions of (1.11) and (1.13) and those of (1.15) and (1.16) can be written down very easily for $Z_{nm}(x)$.

Let

$$M_{nm,d}^2 = \int_{R^d} Z_{nm}^2(x) dS_{n+m}(x) \quad (3.18)$$

with $S_{n+m}(x) = (nF_n(x) + mG_m(x)) / (n+m)$, and let $W_d^2(n,m)$ of (2.10) be defined in terms of the sequence of Brownian bridges of (3.15). Given H_0 of (3.10), then Corollaries 1 and 2 as well as Proposition 1 hold for $M_{nm,d}^2$, and hence we have (cf. (2.9))

$$\lim_{n,m \rightarrow \infty} P\{M_{nm,d}^2 \leq u\} = P\left\{\int_{I^d} B^2(y) dy \leq u\right\} = V_d(u), \quad u \geq 0 \quad (d \geq 1). \quad (3.19)$$

Thus a test of H_0 of (3.10) can also be based on large values of $M_{nm,d}^2$ and Tables 1 and 2 of Cotterill and Csörgő (198)) can again be used. Burke (1977) showed that the test for H_0 of (3.10) is consistent against the alternative of (3.13).

When proving the consistency and asymptotic power properties of the $W_{nm,d}^2$ (cf. (2.1)) test against the alternatives H_1 and H_2 and that of the $M_{nm,d}^2$ test against the alternative H_3 , Burke (1977) used a theorem A type strong approximation theorem of Csörgő and Révész (1975b) for the empirical process $\beta_n(x)$ of (1.6). The latter strong invariance principle for $\beta_n(x)$ of (1.6) with $F \in \mathcal{F}$, i.e., with F not necessarily in \mathcal{F}_0 , was only one available at that time and, consequently, Burke (1977) in his proofs assumed the extra conditions on $F \in \mathcal{F}$ of Theorem 1 in Csörgő and Révész (1975b). These extra conditions on $F \in \mathcal{F}$ were proved to be superfluous by Philipp and Pinzur (1979), who also dropped the condition that $F \in \mathcal{F}$.

They proved

Theorem E (Philipp and Pinzur (1979)). Let X_1, \dots, X_n ($n = 1, 2, \dots$) be independent random vectors in R^d with an arbitrary common distribution function F . Then without changing the distribution of the empirical process $\{\beta_n(x); x \in R^d, n \geq 1\}$ of X_1, \dots, X_n ; $n \geq 1$ defined as in (1.6), we can redefine $\beta_n(x)$ on a richer probability space on which there exists an F -Kiefer process $\{K_F(x, t); x \in R^d, t \geq 0\}$ with mean zero and covariance function

$$(t_1 \wedge t_2) (F(x_1 \wedge x_2) - F(x_1)F(x_2)) \quad (3.20)$$

such that

$$\sup_{1 \leq k \leq n} \sup_{x \in R^d} |k^{1/2} \beta_k(x) - K_F(x, k)| = o_p(n^{1/2 - \delta}) \quad (3.21)$$

for any $\delta < (18 + 16d)^{-1}$ ($d \geq 1$), where $x_1 \wedge x_2 = (x_{11} \wedge x_{21}, \dots, x_{1d} \wedge x_{2d})$.

As to the F-Kiefer process $K_F(\cdot, \cdot)$ of (3.21) compared to the Kiefer process of D3. of Section 1, we should observe, of course, that the former lives on $\mathbb{R}^d \times [0, \infty)$ instead of $\mathbb{I}^d \times [0, \infty)$ and that instead of the Lebesgue measure $\lambda(\cdot)$ in the covariance function of $K(\cdot, \cdot)$ of D3. we now have the $F(\cdot)$ measure playing that role for $K_F(\cdot, \cdot)$.

We are going to close this section by mentioning a possible application of Proposition 2 in the field of statistical climatology.

Suppose that we wish to test the hypothesis that, say, the temperature readings of a central location, L_0 , sufficiently represent the temperature readings for the neighboring locations L_1, L_2, \dots, L_m . Let T_{j0} ($j = 1, 2, \dots, n$) be n independent temperature readings for each one of the locations L_i ($i = 1, 2, \dots, m$). Let

$$Y_j = (Y_{j1}, \dots, Y_{jm}), \text{ with } Y_{ji} = T_{ji} - T_{j0}, \quad (j = 1, \dots, n; i = 1, \dots, m). \quad (3.22)$$

Then, if it were to be true that readings for the central location L_0 sufficiently represented also the readings at the neighboring locations L_1, \dots, L_m , then Y_{ji} of (3.22) should essentially represent only independent errors of measurement rather than real differences in temperature. Whence, on the latter hypothesis it would be then reasonable to assume that $Y_j = (Y_{j1}, \dots, Y_{jm})$ of (3.22) are n i.i.d. m -vectors m i.i.d. $N(x; \mu, \sigma^2)$ rv with some unknown mean μ and variance $\sigma^2 > 0$. Thus the above climatological assumptions may be summarized via stating the composite goodness-of-fit hypothesis of (3.7), which, in turn, can then be tested via the equivalent simple goodness-of-fit hypothesis of (3.8), provided of course that $m \geq 6$, i.e., that we have at least six neighboring locations L_i to be compared to the central one L_0 .

The critical region for the latter simple hypothesis can be based on large values of $W_{n,m-2}^2$ of (3.9) or, equivalently, on large values of its modified version

$$\tilde{W}_{n,m-2}^2 = \int_{\mathbb{R}^{m-2}} n[F_n(x_1, \dots, x_{m-2}) - \prod_{i=1}^{m-2} F_i(x_i)]^2 dF_n(x_1, \dots, x_{m-2}), \quad m(\text{fixed}) \geq 6, \quad (3.23)$$

where everything is defined as for (3.9). By Proposition 1 we have, just like for $W_{n,m-2}^2$ of (3.9), that $\tilde{W}_{n,m-2}^2 \xrightarrow{D} \int_{\mathbb{I}^{m-2}} B^2(y) dy$, $m \geq 6$, and, as mentioned already, significance points of the latter random variables are tabulated in Tables 1 and 2 in Cotterill and Csörgő (1980).

Non-rejection of H'_0 of (3.8) in the present case would mean that temperature readings at locations L_1, \dots, L_m and temperature forecasts for the latter may reasonably well be made accordingly. Rejection of H'_0 of (3.8) would of course mean that at least one on the m locations might differ significantly from L_0 .

ACKNOWLEDGEMENT

This research was supported by a NSERC Canada Grant and a Canada Council Killam

Senior Research Scholarship at Carleton University, Ottawa.

REFERENCES

- Anderson, T.W. and Darling, D.A., 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic process. *Ann. Math. Statist.* 32:193-212.
- Blum, J.R., Kiefer, J. and Rosenblatt, M., 1961. Distribution free tests of independence based on the sample distribution function. *Ann. Math. Statist.* 32:485-498.
- Bondesson, L., 1974. Characterizations of probability laws through constant regression. *Z. Wahrs. Verw. Geb.* 30:93-115.
- Burke, M.D., 1977. On the multivariate two-sample problem using strong approximations of the EDF. *J. Mult. Anal.* 7:491-511.
- Cotterill, D.S. and Csörgő, M., 1980. On the limiting distribution of and critical values for the multivariate Cramér-von Mises statistic. To appear in *Ann. Statist.*
- Csörgő, M., 1979. Strong approximations of the Hoeffding, Blum, Kiefer, Rosenblatt multivariate empirical process. *J. Mult. Anal.* 9:84-100.
- Csörgő, M. and Chan, A.H.C., 1976. On the Erdős-Rényi increments and the P. Lévy modulus of continuity of a Kiefer process. In: P. Gaenssler and P. Révész (ed.), *Empirical Distributions and Processes. Lecture Notes in Math.* 566:1-16, Springer-Verlag.
- Csörgő, M. and Révész, P., 1975a. A new method to prove Strassen type laws of invariance principle, II. *Z. Wahrs. Verw. Geb.* 31:261-269.
- Csörgő, M. and Révész, P., 1975b. A strong approximation of the multivariate empirical process. *Studia Sci. Math. Hungar.* 10:427-434.
- Csörgő, M. and Révész, P., 1980. Strong Approximations in Probability and Statistics. Book manuscript in progress. Academic P., New York.
- Csörgő, M., Seshadri, V. and Yalovsky, M., 1973. Some exact tests for normality in the presence of unknown parameters. *J. Royal Statist. Soc. B.* 35:507-522.
- Csörgő, S., 1976. On an asymptotic expansion for the von Mises ω^2 statistic. *Acta Sci. Math. (Szeged.)* 38:45-67.
- Csörgő, S. and Stachó, L., 1979. A step toward an asymptotic expansion for the Cramér-von Mises statistics. To appear in *Coll. Math. J. Bolyai*.
- Dugue, D., 1969. Characteristic functions of random variables connected with Brownian motion and of the von Mises multidimensional ω_n^2 . In: P.R. Krishnaiah (ed.), *Multivariate Analysis, Vol. 2*: 289-301. Academic P., New York.
- Durbin, J., 1970. Asymptotic distributions of some statistics based on the bivariate sample distribution function. In: M.L. Puri (ed.), *Nonparametric Techniques in Statistical Inference*: 435-451. Camb. Univ. P.
- Götze, F., 1979. Asymptotic expansions for bivariate von Mises functionals. Reprints in *Statist.* 49. Univ. of Cologne.
- Hoeffding, W., 1948. A nonparametric test of independence. *Ann. Math. Statist.* 19:546-557.
- Kiefer, J., 1959. K-sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises tests. *Ann. Math. Statist.* 30:420-447.
- Komlós, J., Major, P. and Tusnády, G., 1975. An approximation of partial sums of independent R.V.'s and the sample D.F. I. *Z. Wahrs. Verw. Geb.* 32:111-131.
- Krivyakova, E.N., Martynov, G.V. and Tyurin, Yu.N., 1977. On the distribution of the ω^2 statistics in the multidimensional case. *Theory Prob. Appl.* 22:406-410.
- Philipp, W. and Pinzur, L., 1979. Almost sure approximation theorems for the multivariate empirical process. To appear.
- Rosenblatt, M., 1952a. Remarks on a multivariate transformation. *Ann. Math. Statist.* 23:470-472.
- Rosenblatt, M., 1952b. Limit theorems associated with variants of the von Mises statistics. *Ann. Math. Statist.* 23:617-623.
- Tusnády, G., 1977a. A study of statistical hypotheses (in Hungarian). *Cand. Dissert., Hungarian Acad. of Sci.*
- Tusnády, G., 1977. A remark on the approximation of the sample DF in the multidimensional case. *Periodica Math. Hungar.* 8:53-55.
- Tusnády, G., 1977b. Strong invariance principles. In: J.R. Barra et al. (ed.), *Recent Development in Statistics*. 289-300., North Holland Publ. C.

THE BEHAVIOUR OF BAYES DECISION FOR NORMAL MEAN UNDER NONSTANDARD PRIOR: UNKNOWN
PRECISION

A. K. BANSAL

Dept. Math. Stat., Univ. of Delhi, Delhi (India)

ABSTRACT

Bansal, A.K., The behaviour of Bayes decision for normal mean under nonstandard prior.
Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

In Bayesian inference, formulation and assessment of the prior distribution of unknown parameters has been considered since the time of Bayes. In the absence of the 'true' prior, the author advocated investigation of inference robustness of a Bayes decision with respect to the prior distribution.

In this paper, the Edgeworth-Gamma prior is employed to study the effect of non-normality on marginal posterior densities and Bayes estimators for the unknown mean and precision of a normal population. The zone of sensitivity to non-normality in the prior is also obtained for the Bayes analogue of the test of significance for the unknown mean. Some interesting cases and numerical illustrations are also discussed to bring out their behaviour when the 'true' prior of the unknown parameters is not the conventional normal-gamma distribution.

The derived expressions are directly applicable to an investigation of the robustness of the Bayes forecast to non-normality in the assumed prior of the unknown mean when the time series is adequately described by a constant process model in which 'noise' is normal with mean zero but variance unknown.

1. MOTIVATION

In normal theory Bayes decision problems, a random sample is assumed to be drawn from a normal population and then the conjugate prior is chosen for mathematical convenience to obtain appropriate decision rules. Following Box and Tiao (1962), researchers employed the class of symmetric exponential power distributions to investigate inference robustness to non-normality of Bayes decisions concerning the unknown mean. Edwards et. al. (1963) considered robustness with respect to the prior when the prior belonged to the family of natural conjugates. Recently, Rubin (1977) studied robustness of the Bayes estimator for the mean of a normal population (variance known) with normal, double exponential, logistic, and cauchy priors.

Quite often, situations arise when the investigator cannot assume exact normality and, therefore, he must choose a distribution belonging to a family of moderately non-normal distributions. Critical examination by Wallace (1958) revealed the fact that the Edgeworth population approach is conceptually more relevant to robustness problems than the asymptotic expansion in the sample size. It has enabled a fairly

accurate estimation of the error involved in use of the normal theory procedure for moderately non-normal variates. Further, when non-normality parameters $\lambda_3 (= \sqrt{\beta_1})$ and $\lambda_4 (= \beta_2 - 3)$ lie within the Barton Dennis (1952) region (BDR), the theoretical specification of the population by an Edgeworth series distribution (ESD) covers a variety of moderately non-normal populations. Singh (1967) showed that the ESD with $\lambda_3^2 = 0.5$ can be regarded as an effectively unimodal and proper density and that in some cases the errors in numerical terms are not serious even if λ_3^2 exceeds 0.5. Thus the representation may cover a wider class of non-normal populations. The author (see Bansal (1978a,b) and (1979)) utilized the Edgeworth approach to study the effects of skewness and kurtosis in the parent population and the conventional prior for the unknown mean on the Bayes estimator and on tests of significance in Jeffreys' framework. The precision of the parent normal population was assumed to be known in these papers.

2. MAIN ASSUMPTIONS AND POSTERIOR DISTRIBUTIONS

Let $\underline{X} = (X_1, X_2, \dots, X_n)$ be a random sample drawn from a normal population with unknown mean M and precision R . The value of the likelihood function when $M = m$, $R = r$, and $X_i = x_i$ ($i = 1, 2, \dots, n$) is given by:

$$f_n(\underline{x}|m, r) = (r/2\pi)^{n/2} \exp \left[-\frac{r}{2} \sum_{i=1}^n (x_i - m)^2 \right].$$

Assume that the prior joint distribution of M and R is as follows:

(i) The conditional density of M when $R = r$ (> 0) is represented by the first four terms of the ESD, that is:

$$\begin{aligned} \xi(m|R=r) = & \left[1 + \frac{1}{6} \lambda_3 H_3\{\sqrt{\tau r} (m - \mu)\} + \frac{1}{24} \lambda_4 H_4\{\sqrt{\tau r} (m - \mu)\} \right. \\ & \left. + \frac{1}{72} \lambda_3^2 H_6\{\sqrt{\tau r} (m - \mu)\} \right] (\tau r/2\pi)^{1/2} \exp \left[-\frac{1}{2} \tau r (m - \mu)^2 \right], \end{aligned} \quad (1)$$

where $H_k(\cdot)$ denotes the Hermite polynomial of degree k , $\lambda_3^2 = \beta_1$, $\lambda_4 = \beta_2 - 3$ such that $\mu \in (-\infty, \infty)$, $\tau > 0$, and (λ_3, λ_4) lie within the BDR.

(ii) The marginal distribution of R is a gamma with parameters

$$\xi(r) = \beta^\alpha r^{\alpha-1} e^{-\beta r} / \Gamma(\alpha); \quad \alpha > 0, \beta > 0. \quad (2)$$

The prior joint density of M and R is

$$\xi(m, r) = \xi(m|R=r) \xi(r). \quad (3)$$

Writing

$$A = n(\bar{x} - \mu)(\tau/\beta')^{1/2}/\tau'; \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i;$$

$$\beta' = \beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \tau n(\bar{x} - \mu)^2/2\tau';$$

$$\alpha_1 = (2\alpha + n)/2, \quad \tau' = \tau + n, \quad \tau_1 = \tau/\tau' - 1, \quad \mu' = (\tau\mu + n\bar{x})/\tau',$$

and

$$\begin{aligned} T = 1 + \frac{1}{6} \lambda_3 A \Gamma(\alpha_1 + 0.5) \{ (\alpha_1 + 0.5) A^2 + 3\tau_1 \} / \Gamma(\alpha_1) \\ + \frac{1}{24} \lambda_4 \{ \alpha_1 (\alpha_1 + 1) A^4 + 6\alpha_1 \tau_1 A^2 + 3\tau_1^2 \} \\ + \frac{1}{72} \lambda_3^2 \{ \alpha_1 (\alpha_1 + 1) (\alpha_1 + 2) A^6 - 15\alpha_1 (\alpha_1 + 1) \tau_1 A^4 \\ + 45\alpha_1 \tau_1^2 A^2 + 15\tau_1^3 \}, \end{aligned} \quad (4)$$

the marginal density of X with respect to the prior given in (3) is given by:

$$\begin{aligned} f_{\xi}(\underline{x}) = \int_{-\infty}^{\infty} \int_0^{\infty} (2\pi/r)^{1/2} \exp \left[-\frac{r}{2} \sum_{i=1}^n (x_i - m)^2 \right] \xi(m, r) \, dm \, dr \\ = (2\pi\beta')^{-n/2} (\beta/\beta')^{\alpha} (\tau/\tau')^{1/2} T \Gamma(\alpha_1) / \Gamma(\alpha), \end{aligned} \quad (5)$$

and the joint posterior density of M and R can be shown to be:

$$\begin{aligned} \xi_{\underline{x}}(m, r) = (r\tau'/2\pi)^{1/2} (\beta')^{\alpha_1} r^{\alpha_1-1} \exp [-r\{\tau'(m - \mu')^2/2 + \beta'\}] \times \\ [1 + \frac{1}{6} \lambda_3 H_3 \{\sqrt{\tau r} (m - \mu)\} + \frac{1}{24} \lambda_4 H_4 \{\sqrt{\tau r} (m - \mu)\} \\ + \frac{1}{72} \lambda_3^2 H_6 \{\sqrt{\tau r} (m - \mu)\}] / T. \end{aligned} \quad (6)$$

Let the three-parameter t -density have location parameter μ' , have precision $\tau'(\alpha' + k)/2\beta'$, and have $(\alpha' + k)$ degrees of freedom denoted by g_k ($k = 0, 1, \dots, 6$). The marginal posterior density (MPD) of the unknown M is then obtained by integrating out r from the expression given in (6). After simplification, the MPD of M may be written as:

$$\xi_{\underline{x}}(m) = [g_0 + \frac{1}{6} \lambda_3 \sqrt{2} B\{B^2(\alpha' + 1)g_3 - 3g_1\}] \Gamma(\alpha_1 + 0.5) / \Gamma(\alpha_1)$$

$$\begin{aligned}
& + \frac{1}{24} \lambda_4 \{ \alpha' (\alpha' + 2) B^4 g_4 - 6 \alpha' B^2 g_2 + 3 g_0 \} \\
& + \frac{1}{72} \lambda_3^2 \{ \alpha' (\alpha' + 2) (\alpha' + 4) B^6 g_6 - 15 \alpha' (\alpha' + 2) B^4 g_4 \\
& + 45 \alpha' B^2 g_2 - 15 g_0 \}] / T;
\end{aligned} \tag{7}$$

where $\alpha' = 2\alpha + n$, $B = (m - \mu) / (\tau/2\beta')^{1/2}$. Similarly, the MPD of the unknown precision R is obtained by integrating (6) with respect to m over the range $-\infty$ to ∞ . It can be shown to be:

$$\begin{aligned}
\xi_{\underline{x}}(r) = & [h_0 + \frac{1}{6} \lambda_3 A \{ (\alpha_1 + 0.5) A^2 h_{1.5} + 3 \tau_1 h_{0.5} \} \Gamma(\alpha_1 + 0.5) / \Gamma(\alpha_1) \\
& + \frac{1}{24} \lambda_4 \{ \alpha_1 (\alpha_1 + 0.5) A^4 h_2 + 6 \alpha_1 \tau_1 A^2 h_1 + 3 \tau_1^2 h_0 \} \\
& + \frac{1}{72} \lambda_3^2 \{ \alpha_1 (\alpha_1 + 0.5) (\alpha_1 + 1) A^6 h_3 + 15 \alpha_1 (\alpha_1 + 0.5) \tau_1 A^4 h_2 \\
& + 45 \alpha_1 \tau_1^2 A^2 h_1 + 15 \tau_1^3 h_0 \}] / T,
\end{aligned} \tag{8}$$

where h_k ($k = 0, 1/2, 1, \dots, 3$) denotes the gamma density with $(\alpha' + k)$ degrees of freedom and scale parameter β' .

3. BAYES ESTIMATOR FOR M AND R

Let the estimated values of the unknown parameters M and R be denoted by m and r , respectively. Consider the quadratic loss function:

$$L(m, r) = (m - \hat{m})^2 + (r - \hat{r})^2.$$

It is well known (see De Groot, 1970) that the Bayes estimator of M and R are the means of their respective marginal posterior densities. Thus the Bayes estimator for the unknown mean M when the precision R is unknown is:

$$\begin{aligned}
\delta_M^*(x) = & \int_{-\infty}^{\infty} m \xi_{\underline{x}}(m) dm \\
= & \mu' + \left[\frac{1}{2} \lambda_3 (\tau \beta')^{1/2} \Gamma(\alpha_1 - 0.5) \{ (\alpha_1 - 0.5) A^2 - \tau_1 \} / \{ \tau \Gamma(\alpha_1) \} \right. \\
& + \frac{1}{12} \lambda_4 (\tau_1 + 1) (\mu' - \mu) (2 \alpha_1 A^2 + 6 \tau_1 + 3) \\
& \left. + \frac{1}{24} \lambda_3^2 (\tau_1 + 1) (\mu' - \mu) \{ 2 \alpha_1 (\alpha_1 + 1) A^4 + 20 \alpha_1 \tau_1 A^2 + 30 \tau_1^2 - 15 \} \right] / T,
\end{aligned} \tag{9}$$

and that, for the unknown precision R , is given by:

$$\begin{aligned}\delta_R^*(\underline{x}) &= \int_0^\infty r \xi_{\underline{x}}(r) dr \\ &= \frac{\alpha_1}{\beta'} \left[1 + \frac{1}{24} \lambda_4 \{ (\alpha_1 + 1) (\alpha_1 + 2) A^4 + 6 \tau_1 (\alpha_1 + 1) A^2 \right. \\ &\quad \left. + 3 \tau_1^2 \} + \frac{1}{72} \lambda_3^2 \{ (\alpha_1 + 1) (\alpha_1 + 3) A^6 \right. \\ &\quad \left. + 15 \tau_1 (\alpha_1 + 1) (\alpha_1 + 2) A^4 + 45 \tau_1^2 (\alpha_1 + 1) A^2 + 15 \tau_1^3 \} \right] / T \\ &\quad + \frac{1}{6} \lambda_3 A \{ (\alpha_1 + 1.5) A^2 + 3 \tau_1 \} \Gamma(\alpha_1 + 1.5) / \{ \beta' T \Gamma(\alpha_1) \}.\end{aligned}\quad (10)$$

However, the Bayes risk ρ^* of the unknown parameter vector (M, R) against the ESD - Gamma type prior ξ is the sum of the variances of M and R of their respective MPD when the sample \underline{x} has been observed by the decision maker. These variances are as follows:

$$\begin{aligned}\text{Var}(M|\underline{x}) &= \int_{-\infty}^\infty \{m - \delta_M^*(\underline{x})\}^2 \xi_{\underline{x}}(m) dm \\ &= 2\beta' \left\{ 1 + \frac{1}{6} \lambda_3 A \{ A^2 (\alpha_1 - 0.5) + 9 \tau_1 + 6 \} \Gamma(\alpha_1 - 0.5) / \right. \\ &\quad \Gamma(\alpha_1 - 1) + \frac{1}{24} \lambda_4 \{ \alpha_1 (\alpha_1 - 1) A^4 + 6 (\alpha_1 - 1) (3 \tau_1 + 2) A^2 \\ &\quad \left. + 3 \tau_1 (5 \tau_1 + 4) \} + \frac{1}{72} \lambda_3^2 \{ \alpha_1 (\alpha_1^2 - 1) A^6 \right. \\ &\quad \left. + 15 \alpha_1 (\alpha_1 - 1) (3 \tau_1 + 2) A^4 + 45 \tau_1 (5 \tau_1 + 4) (\alpha_1 - 1) A^2 \right. \\ &\quad \left. + 15 \tau_1^2 (7 \tau_1 + 6) \} \right\} / \{ \tau' (\alpha' - 1) T \} - \{ \delta_M^*(\underline{x}) - \mu' \}^2,\end{aligned}\quad (11)$$

and

$$\begin{aligned}\text{Var}(R|\underline{x}) &= \int_0^\infty \{r - \delta_R^*(\underline{x})\}^2 \xi_{\underline{x}}(r) dr \\ &= \alpha_1 (\alpha_1 + 1) \left[1 + \frac{1}{24} \lambda_4 \tau_1 \{ (\alpha_1 + 2) (\alpha_1 + 3) A^4 + 6 (\alpha_1 + 2) A^2 + 3 \tau_1 \} \right. \\ &\quad \left. + \frac{1}{72} \lambda_3^2 \{ (\alpha_1 + 2) (\alpha_1 + 3) (\alpha_1 + 4) A^6 + 15 \tau_1 (\alpha_1 + 2) (\alpha_1 + 3) A^4 \right. \\ &\quad \left. + 45 \tau_1^2 (\alpha_1 + 2) A^2 + 15 \tau_1^3 \} \right] / \{ \beta'^2 T \} + \frac{1}{6} \lambda_3 A \{ (\alpha_1 + 2.5) A^2 \\ &\quad \left. + 3 \tau_1 \} \Gamma(\alpha_1 + 2.5) / \{ T \beta'^2 \Gamma(\alpha_1) \} - \delta_R^*(\underline{x})^2.\end{aligned}\quad (12)$$

4. BEHAVIOUR OF THE BAYES ESTIMATOR

The normal distribution is a member of ESD family of moderately non-normal distributions. It can be easily checked that for $\lambda_3 = \lambda_4 = 0$ the derived expressions for the joint and the marginal posterior distributions in Section 2, and also the Bayes estimator for the unknown mean M and precision R along with their associated risks against the prior when the sample of size n has been observed in Section 3, collapse to the corresponding normal theory expressions. Moreover, if the assumed conditional ESD prior is vague ($\tau \rightarrow 0$) it is once again, as expected, found that the MPD for each of M and R tend to normal theory marginal posterior densities.

The discrepancy in the prior information and the observed sample may be assigned to: misconception in the decision maker's prior attitude, inadequacy in the probability model, or bias in the data itself. Strong inconsistency between data and prior information may even lead to misleading decisions. In order to avoid such unpleasant situations, empirical bayesians are known to employ the sample to obtain suitable values for the parameters of their prior distribution. An interesting situation may arise when the investigator decides to revise the values of the prior mean in the light of the observed sample and takes $\mu = \bar{x}$. In this case $\mu' (= (\tau\mu + n\bar{x})/\tau')$ reduces to the sample mean \bar{x} . If one employs the notations,

$$\beta'_0 = \beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \text{and} \quad T_0 = 1 + \frac{1}{8} \tau_1^2 \lambda_4 + \frac{5}{24} \tau_1^3 \lambda_3,$$

the Bayes estimator for M , given in (9), simplifies to

$$\bar{x} - \frac{1}{2} \lambda_3 (\beta'_0 \tau)^{1/2} \tau_1 \Gamma(\alpha_1 - 0.5) / \{\tau' T_0 \Gamma(\alpha_1)\}. \quad (13)$$

Therefore, as in the known precision case (see Bansal, 1978a), a decision to revise one's opinion about the prior mean will not make the Bayes estimator free from the non-normality in the prior for any finite value of n . However, the Bayes estimator for R reduces to the normal theory estimator:

$$(2\alpha + n)/\beta'_0.$$

The Bayes risk of M , given in (11), reduces to:

$$2\beta'_0 \left[1 + \frac{1}{8} \lambda_4 \tau_1 (5\tau_1 + 4) + \frac{5}{24} \lambda_3 \tau_1^2 (7\tau_1 + 6) \right] / [\tau' T_0 (\alpha' - 1)] \\ - \frac{1}{4} \lambda_3^2 \beta'_0 \tau_1^2 \Gamma^2(\alpha_1 - 0.5) / [\tau' \Gamma(\alpha_1) T_0]^2, \quad (14)$$

which tends to

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 3)$$

as $\tau \rightarrow 0$. As before, the Bayes risk of R remains unaffected by the non-normality and we obtain the normal theory value $\alpha_1/\beta_0'^2$.

In order to illustrate the effect of non-normality in the 'true' prior distribution, the GAUSS subroutine was employed to draw samples of size 5 and 10 from standard normal populations. A number of ESD priors with $\mu = 0$ and $\tau = 1$, but with varying amounts of skewness and kurtosis, along with four gamma priors as marginals for R were employed to obtain Bayes estimates for M and R .

The Bayes estimate of M (see Table I) is seen to be affected by the non-normality as well as by the shape of the gamma prior of R . It is interesting to observe that the effect due to skewness counterbalanced, to some extent, that due to kurtosis in the conditional prior of M . Increasing the sample size to ten did not show any reduction in the effect. Further, the actual observed sample is seen to affect the Bayes estimate of M . In contrast to this, the Bayes estimate of the unknown precision R does not seem to be affected by non-normality of the conditional prior of M . But it is certainly affected by the shape of the chosen gamma prior to R . As expected, the Bayes estimate of the unknown precision of the normal population is significantly high (see Table II) for vague marginal prior of R . As β increases from 2.0 to 3.0, Bayes estimates of R tend to be larger for samples of size 10 priors, the estimate may behave in an unreasonable way.

TABLE I

Bayes estimate of the mean based on samples of size 5 and 10 (underlined) drawn from $N(0,1)$ with ESD-Gamma prior having non-normality parameter (λ_3, λ_4) , $\mu = 0$, $\tau = 1$, and Gamma with parameter (α, β) .

Gamma Parameter	Non-normality Parameter			
	(0, 0)	(0.3, 0.5)	(0.4, 1.5)	(0.5, 2.4)
(-0.5, 0)	0.0548	0.0555	0.0436	0.0349
	<u>0.1728</u>	<u>0.1718</u>	<u>0.1500</u>	<u>0.1342</u>
(1.0, 1.0)	0.0548	0.0629	0.0530	0.0466
	<u>0.1728</u>	<u>0.1728</u>	<u>0.1516</u>	<u>0.1364</u>
(1.0, 2.0)	0.0548	0.0677	0.0425	0.0538
	<u>0.1728</u>	<u>0.1743</u>	<u>0.1538</u>	<u>0.1394</u>
(1.0, 5.0)	0.0548	0.0776	0.0714	0.0683
	<u>0.1728</u>	<u>0.1779</u>	<u>0.1585</u>	<u>0.1453</u>

TABLE II

Bayes estimate of the precision based on samples of size 5 and 10 (underlined) drawn from $N(0,1)$ with ESD-Gamma prior having non-normality parameter (λ_3, λ_4) , $\mu = 0$, $\tau = 1$, and parameters (α, β) of marginal prior Gamma of R.

Gamma Parameter	Non-normality Parameter			
	(0, 0)	(0.3, 0.5)	(0.4, 1.5)	(0.5, 2.4)
(-0.5, 0)	<u>23.9245</u> <u>4.0631</u>	<u>23.8474</u> <u>4.0771</u>	<u>24.0606</u> <u>4.1596</u>	<u>24.2091</u> <u>4.2201</u>
(1.0, 1.0)	3.2300 <u>2.8469</u>	3.2265 <u>2.8544</u>	3.2298 <u>2.8943</u>	3.2320 <u>2.9293</u>
(1.0, 2.0)	1.6798 <u>1.9308</u>	1.6783 <u>1.9332</u>	30.2943 <u>1.9510</u>	1.6795 <u>1.9637</u>
(1.0, 5.0)	0.6885 <u>0.9824</u>	0.6881 <u>0.9825</u>	0.6881 <u>0.9868</u>	0.6882 <u>0.9899</u>

The Bayes risk associated with the computed estimate of the unknown mean M is presented in Table III. For a fixed conditional ESD type prior of M , the risk is seen to increase with an increase in the value of the shape parameter of the gamma prior. The effect of non-normality, in general, is not significant on the Bayes risk. An exceptional case occurs for $n = 5$ when $(\lambda_3, \lambda_4) = (0.4, 1.5)$ and $(\alpha, \beta) = (1.0, 2.0)$.

TABLE III

Bayes risk associated with the Bayes estimate of the mean with respect to ESD-Gamma prior having non-normality parameter (λ_3, λ_4) , parameters of Gamma prior (α, β) , and $\mu = 0$, $\tau = 1$ for samples of size 5 and 10 (underlined) from $N(0,1)$.

Gamma Parameter	Non-normality Parameter			
	(0, 0)	(0.3, 0.5)	(0.4, 1.5)	(0.5, 2.4)
(-0.5, 0)	0.0093 <u>0.0252</u>	0.0093 <u>0.0253</u>	0.0088 <u>0.0242</u>	0.0082 <u>0.0228</u>
(1.0, 1.0)	0.0602 <u>0.0348</u>	0.0592 <u>0.0348</u>	0.0562 <u>0.0335</u>	0.0539 <u>0.0320</u>
(1.0, 2.0)	0.1158 <u>0.0514</u>	0.1136 <u>0.0512</u>	0.0065 <u>0.0495</u>	0.1036 <u>0.0478</u>
(1.0, 5.0)	0.2824 <u>0.1010</u>	0.2767 <u>0.1004</u>	0.2624 <u>0.0974</u>	0.2519 <u>0.0948</u>

5. TEST OF SIGNIFICANCE FOR THE UNKNOWN MEAN

Consider the problem of testing a null hypothesis $H_0: m = m_0$ against the alternative $H_1: m \neq m_0$ as a binary decision problem in which decision d_i amounts to acceptance of the hypothesis H_i ($i = 0, 1$). Let $L_i(m, r)$ denote the loss incurred in taking decision d_i when $M = m$ and $R = r$. $L_i(m, r)$ is assumed such that:

$$L_1(m, r) = a r(m - m_0)^2; \quad m \in (-\infty, \infty), \quad a > 0,$$

and

$$L_2(m, r) = \begin{cases} 0 & \text{for } m \neq m_0, \\ b & \text{otherwise; } b > 0. \end{cases} \quad (15)$$

Following Jeffreys (1961), let the joint prior distribution of M and R be specified in two parts. Assume that a prior discrete probability p (> 0) is located at $M = m_0$ and the rest of the probability $(1 - p)$ is distributed in the remaining space of $M \times R$ such that the conditional prior densities of R are gamma with the same parameters α and β irrespective of whether or not the null hypothesis is correct. Furthermore, as in Section 2, assume that the conditional distribution of M , when $R = r$, is ESD given in (1). Thus, in this case, the conditional prior joint density of M and R , when $M \neq m_0$, is Edgeworth-Gamma.

The Bayes risk in taking decision d_0 is given by:

$$\begin{aligned} \rho_0 &= a(1 - p) E[R E\{(M - m_0)^2 | R = r\} | M \neq m_0] \\ &= a(1 - p) \int_0^\infty \int_{-\infty}^\infty r(m - m_0)^2 \xi_{\underline{x}}(m | R = r) \xi_{\underline{x}}(r) dm dr \\ &= a(1 - p) \int_0^\infty \int_{-\infty}^\infty r(m - m_0)^2 \xi_{\underline{x}}(m, r) dm dr \\ &= a(1 - p) [\alpha_1 (\mu' - m_0)^2 / \beta' + \frac{1}{\tau}, (\mu' - m_0) [\frac{1}{2} \lambda_3 (\tau | \beta')^{1/2} (\alpha_1 A^2 + \tau_1) \Gamma(\alpha_1 + 0.5) / \Gamma(\alpha_1) \\ &\quad + \frac{1}{6} \lambda_4 \tau (\mu' - \mu) \{\alpha_1 (\alpha_1 + 1) A^2 + 3\tau_1\} + \frac{1}{12} \lambda_3^2 \alpha_1 \tau (\mu' - \mu) \{(\alpha_1 + 1)(\alpha_1 + 2) A^4 \\ &\quad + 10(\alpha_1 + 1)\tau_1 A^2 + 15\tau_1^2\} / \beta'] + [1 + \frac{1}{6} \lambda_3 A \{(\alpha_1 + 0.5) A^2 + 3(3\tau_1 + 2)\} \Gamma(\alpha_1 + 0.5) / \\ &\quad \Gamma(\alpha_1) + \frac{1}{24} \lambda_4 \{\alpha_1 (\alpha_1 + 1) A^4 + 6\alpha_1 (3\tau_1 + 2) A^2 + 3\tau_1 (5\tau_1 + 4)\} + \frac{1}{72} \lambda_3^2 \{\alpha_1 (\alpha_1 + 1) \\ &\quad \times (\alpha_1 + 2) A^6 + 15\alpha_1 (\alpha_1 + 1) (3\tau_1 + 2) A^4 + 45\alpha_1 \tau_1 (5\tau_1 + 4) A^2 \\ &\quad + 15(7\tau_1^3 + 6\tau_1^2 - 6\tau_1 - 6)\} / \tau'] / \tau \end{aligned}$$

$$= C_1 + C_2 p \quad (16)$$

Since the likelihood function when the null hypothesis is true is:

$$\begin{aligned} f_n(\underline{x}|M = m_0) &= \int_0^\infty \xi(r|M = m_0) f_n(\underline{x}|m_0, r) dr \\ &= \beta^\alpha \Gamma(\alpha_1) / [(2\pi)^{n/2} \Gamma(\alpha) \{\beta + \frac{1}{2} \sum_{i=1}^n (x_i - m_0)^2\}^{\alpha_1}], \end{aligned} \quad (17)$$

the Bayes risk in taking decision d_1 is seen to be

$$\begin{aligned} \rho_1 &= b p f_n(\underline{x}|M = m_0) / [p f_n(\underline{x}|M = m_0) + (1 - p) f_\xi(\underline{x})] \\ &= b p / [p + (1 - p) C_3], \end{aligned} \quad (18)$$

where

$$\begin{aligned} C_3 &= T(\tau/\tau')^{1/2} [\{2\beta + \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - m_0)^2\} / \\ &\quad \{2\beta + \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2/\tau'\}]^{\alpha_1}. \end{aligned}$$

The Bayes decision function (BDF) for the binary decision problem will suggest to the investigator that he accept the null hypothesis H_0 whenever ρ_0 is less than ρ_1 .

The risk curve in the p -plane, as in the case of known precision (see Bansal, 1979), for decision d_0 is a line segment with $(1,0)$ as the right-hand end-point, whereas, that for decision d_1 is a segment of a rectangular hyperbola with end-points $(0,0)$ and $(1,b)$. These two curves intersect (see Fig.) for $p = p^*$, where p^* is the positive root (< 1) of the quadratic equation:

$$C_2(C_3 - 1)p^2 + (b + C_2 - 2C_2C_3)p + C_2C_3 = 0 \quad (19)$$

which is given by:

$$p^* = [(2C_2C_3 - b - C_2) - \sqrt{(b + C_2)^2 - 4bC_2C_3}] / [2C_2(C_3 - 1)].$$

Now the BDF for the hypothesis testing problem may be modified as follows:

$$D(x) = \begin{cases} d_0 & \text{if } p > p^* \\ d_1 & \text{otherwise.} \end{cases} \quad (20)$$

Let p_0^* denote the corresponding normal theory critical value of the probability p and write:

$$p_1 = \min(p_0^*, p^*) \text{ and } p_2 = \max(p_0^*, p^*). \quad (21)$$

The effect of non-normality on the BDF for the testing problem may be clearly observed in the shift of the intersection point of the risk curves for decisions d_0 and d_1 . The author (see Bansal, 1979) called the interval (p_1, p_2) , defined by the shift in intersection point, "Zone of Sensitivity to non-normality". The decision maker may err and accept a wrong hypothesis if his chosen prior probability p happens to fall in the zone of sensitivity. This incorrect decision will be due to the wrong assumption that the true conditional prior distribution of M ($\neq m_0$) is normal.

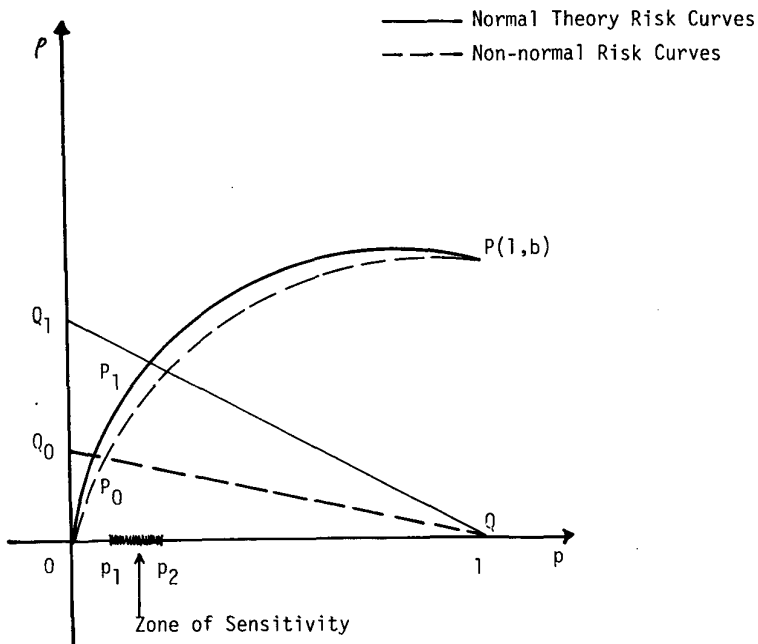


Fig. ZONE OF SENSITIVITY

As before, in order to illustrate the behaviour of the BDF to non-normality in the prior, consider the samples and priors of Section 4 with $a = b = 1$, $m_0 = 0$. Table IV, for example, indicates that if the true prior was defined by $(\alpha, \beta) =$

$(1, 2)$ and $(\lambda_3, \lambda_4) = (0.4, 1.5)$ and the sample of size 5 from the standard normal population happened to be $\underline{x} = (1.9619, -1.6159, 2.1711, -1.4922, 2.2688)$ then the decision maker must choose his discrete probability p in favour of $H_0: m = 0$ such that it does not fall in the zone of sensitivity $(0.0728, 0.1239)$. Further, his decision will not be affected by the non-normality in the prior if he decides to choose p in the region given by:

$$\{p: p \in (0, 0.0728) \cup (0.1239, 1)\}.$$

In Jeffrey's frame-work, any change in the extent of the zone of sensitivity reflects the effect of non-normality in the same way as in Criterion robustness of the t-test where one is interested in the effect of non-normality on the size of the test. Thus a decision maker should compute the zone of sensitivity for the worst departure from normality, i.e., choose extreme values of λ_3 and λ_4 along with those of α and β , and then choose the probability p which lies outside this interval.

TABLE IV

Critical value of the probability p in favour of the null hypothesis $H_0: m = 0$ against the alternative $H_1: m \neq 0$, when $a = b = 1$, non-normality parameter (λ_3, λ_4) , marginal Gamma prior parameter (α, β) , $\mu = 0$, $\tau = 1$ and samples of size 5 and 10 (underlined) from $N(0, 1)$.

Gamma Parameter	Non-normality Parameter			
	$(0, 0)$	$(0.3, 0.5)$	$(0.4, 1.5)$	$(0.5, 2.4)$
$(-0.5, 0)$	0.1178	0.1128	0.1107	0.1091
	<u>0.1150</u>	<u>0.1165</u>	<u>0.1192</u>	<u>0.1233</u>
$(1.0, 1.0)$	0.0757	0.0746	0.0747	0.0751
	<u>0.0834</u>	<u>0.0868</u>	<u>0.0910</u>	<u>0.0968</u>
$(1.0, 2.0)$	0.0728	0.0723	0.1239	0.0732
	<u>0.0623</u>	<u>0.0658</u>	<u>0.0699</u>	<u>0.0754</u>
$(1.0, 5.0)$	0.0709	0.0709	0.0715	0.0722
	<u>0.0439</u>	<u>0.0476</u>	<u>0.0516</u>	<u>0.0567</u>

6. CONCLUDING REMARKS

In many forecasting problems, subjective considerations are often required to make an initial forecast which is to be later revised in the light of the observed sample. Bayesian procedures are now in vogue to estimate parameters of the forecasting model. Most of the time the climatologist's interest lies with future rainfall, humidity, etc. If the observations are random samples from the same population and

if its mean does not change with time, then one may use a constant process model

$x_t = m + u_t$, where

x_t = rainfall in period t ;

m = the unknown process mean or mean of rainfall;

u_t = the random component or 'noise' in the process which is $N(0, 1/r)$.

Thus the rainfall x_t , in any period t , is normally distributed with mean m and precision r . Bansal (1978b) investigated the effect of non-normality on the Bayes forecast when (i) the noise is moderately non-normal and (ii) the prior distribution of the unknown mean M is not the conventional normal distribution. In both of these problems, precision of the noise was assumed to be known. Derived expressions and conclusions in the earlier sections imply that the Bayes forecast will depend heavily on future observations and also be affected by the non-normality in the assumed conditional prior of the unknown mean.

Bayesian techniques have not yet found their due place in climatological studies. For better weather forecasts and understanding of other meteorological phenomena future statistical climatologists will not only use relevant prior information but also check robustness of their Bayes forecasts to the underlying statistical assumptions.

ACKNOWLEDGEMENTS

The author is thankful to Professor Sadao Ikeda and Dr. Ian B. MacNeill for useful suggestions to improve this paper. Thanks are also due to the Delhi University Computer Centre for providing the necessary facilities for computations. He is grateful to the organizers of the International Meeting on Statistical Climatology, Department of Science and Technology (New Delhi) and the University of Delhi for providing partial financial assistance to present this paper in the Meeting.

REFERENCES

- Bansal, A.K., 1978a. Robustness of Bayes estimator for the mean of a normal population with non-normal prior. *Comm. Statist.* A7(5): 453-460.
- Bansal, A.K., 1978b. Robustness of Bayes forecasts to non-normality. *Jour. Korean Statist. Soc.* 7: 11-16.
- Bansal, A.K., 1979. Effect of non-normality on Bayes decision function for testing normal mean. *Jour. Korean Statist. Soc.* 8: (to appear).
- Barton, D.E. and Dennis, K.E., 1952. The conditions under which Gram Charlier and Edgeworth curves are positive, definite, and unimodal. *Biometrika* 39: 425-427.
- Box, G.E.P. and Tiao, G.C., 1962. A further look at robustness via Bayes' theorem. *Biometrika* 49: 419-433.
- De Groot, M.H., 1970. *Optimal Statistical Decisions*. McGraw-Hill, New York, Chapters 9 and 11.
- Edwards, W., Lindman, H. and Savage, L.J., 1963. Bayesian statistical inference for psychological research. *Psych. Rev.* 70: 193-242.
- Jeffreys, H., 1961. *Theory of Probability*. Oxford University Press.

- Rubin, E., 1977. Robust Bayesian Estimation. Statistical decision theory and related topics. II. Edited by Gupta, S.S. and Moore, D.S., Academic Press: 351-356.
- Singh, C., 1967. On the extreme value and range of samples from non-normal populations. Biometrika 54: 541-550.
- Wallace, D.L. 1958. Asymptotic approximations to distributions. Ann. Math. Statist. 29: 635-654.

SOME RESULTS ON EXCHANGEABILITY AND MAJORIZATION IN STATISTICS

A.M.ABOUAMMOH

Dept. Stat., Univ. of Riyadh, Riyadh (Saudi Arabia)

ABSTRACT

Abouammoh, A.M., Some results on exchangeability and majorization in statistics.
Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov. 29-Dec. 1, 1979

There are some properties and characteristics of distributions in statistics which can be easily identified and have been shown to be very interesting. In this paper, we shall investigate two of such characteristics, exchangeability and majorization. The random variables X_1, \dots, X_n are exchangeable if all the $n!$ permutations of the X 's have the same multivariate distribution. Also, the definition of Schur-distribution functions are presented through the concept of majorization of random vectors.

In addition, we have introduced the concept of related exchangeability which is more general than exchangeability and is motivated by a practical viewpoint. Further, some results on the behaviour of classes of distributions having the above structures are established under most commonly occurring functional operations such as closure under convolution, passage to a limit weakly, reversal, mixing and convolutional mixing. It is realized that exchangeability, Schur-functions and other related sub-classes cover, surprisingly, large families of distributions.

Moreover, some examples and applications have been pointed out to elucidate the main results of this paper.

1. INTRODUCTION AND BASIC RESULTS

Exchangeability or symmetric dependence is observed as a natural generalization of the random sample concept, see Ahmad (1975) and references cited therein. In fact the simplest example of exchangeability may be considered as follows: Let us have n matched pairs (X_i, Y_i) , $i = 1, \dots, n$, with a bivariate distribution $F(x, y)$. If we consider X as control and Y as 'treatment' response, then $F(x, y) = F(y, x)$ is equivalent to the assumption that there is no treatment effect. The extension to the k response situation is straightforward. Moreover, a sequence of random variables (r.v.s.) is exchangeable if for any finite number k , say, there is one k -dimensional distribution for all the $k!$ possible permutations, that is, $F(x_1, \dots, x_k) = F(\pi(x_1, \dots, x_k))$, where π is an element of S_k , the permutation group of integers $\{1, \dots, k\}$. For more examples and characterizations of exchangeable processes one can see Feller (1971) and De Fenitti (1975). The sequence of r.v.s. $\{X_n, n \geq 1\}$

is said to be spherical exchangeable if there exists a function g on the positive real line such that for each finite set $\{i_1, \dots, i_k\}$ of natural numbers, the joint characteristic function (ch.f.) f of X_{i_1}, \dots, X_{i_k} satisfies

$$f(t_1, \dots, t_k) = E \exp(i \sum_{j=1}^k t_j X_{i_j}) = g(\sum_{j=1}^k t_j^2) \quad (1.1)$$

Clearly each spherical exchangeable process is exchangeable.

It was established by Hardy et al (1952, p.49), see also Berge (1953, p.184) that an n -dimensional vector \underline{x} is said to be majorized by an n -dimensional vector \underline{y} if by rearrangement of the components to obtain $x_1 \geq x_2 \geq \dots \geq x_n$, $y_1 \geq y_2 \geq \dots \geq y_n$ one has

$$\sum_{j=1}^k x_j \leq \sum_{j=1}^k y_j, \quad k = 1, 2, \dots, n-1, \quad \text{and} \quad \sum_{j=1}^n x_j = \sum_{j=1}^n y_j, \quad (1.2)$$

and denote this by $\underline{x} \prec^* \underline{y}$ if relation (1.2) is satisfied. A function for which $\underline{x} \prec^* \underline{y}$ implies $f(\underline{x}) \geq (\leq) f(\underline{y})$ is called Schur-concave (convex) function or simply Schur-function, and such function is permutation-symmetric, that is, invariant under permutations of its components. Therefore $f(\underline{x})$ is Schur-function implies that the r.v.s. X_1, \dots, X_n are exchangeable. Thus a differentiable function $f(\underline{x})$ of exchangeable r.v.s. is Schur-concave (convex) if and only if

$$\left(\frac{\partial f(\underline{x})}{\partial x_i} - \frac{\partial f(\underline{x})}{\partial x_j} \right) (x_i - x_j) \leq (\geq) 0, \quad \text{for } i \neq j, \quad (1.3)$$

see Schur (1923) and Ostrowski (1952). The case $\underline{x} \prec^* \underline{y}$ can be expressed by $\underline{x} = D\underline{y}$ for some doubly stochastic matrix D . A matrix is called doubly stochastic if it is square in shape, its elements are real nonnegative numbers and the sum of each row and of each column is equal to one, Mirsky (1963). It is pointed out, see Marshall and Olkin (1974), that for any vector $\underline{x} = (x_1, \dots, x_n)$

$$(\sum x_i / n) (1, \dots, 1) \prec^* (x_1, \dots, x_n),$$

and therefore whenever $\sum_{i=1}^n x_i$ is fixed, Schur-concave (convex) function attains a maximum (minimum) point when the components are equal. In addition, we can see that exchangeability (complete permutation-symmetry) may not be realistic assumption in some practical situations. For example in Ahmad and Peterson (1978) the distribution of the generic data point does not lead to a reasonable hypothesis class. Rather an assumption of partial permutation-symmetry would be natural and necessary in some practical situations. Therefore, we shall introduce here the concept of partial exchangeability. For example, one may consider an n -dimensional r.v. to

represent the situation of physical and mental responses of a patient under some experimental stress. In such case one would allow permutation to take place between components which represent a particular response.

Let X_1, \dots, X_k be k r.v.s. such that the random vector (r.vec.) $\underline{X} = (X_1, \dots, X_k)$ represents some experimental data of s different responses or characteristics. Such responses could have different measurement units or different attitude of evaluations. Let S_{n_i} be the permutation symmetric group of n_i r.v.s., we may write $\underline{X} = (X_1, \dots, X_{n_1}, X_{n_1+1}, \dots, X_{n_2}, \dots, X_{n_{s-1}}, \dots, X_{n_s})$ where $\sum_{i=1}^s n_i = k$, then the distribution F of \underline{X} is said to be partially exchangeable if and only if $F(\underline{x}) = F(\pi(\underline{x}))$ for all π in the direct group $S_{n_1} \times S_{n_2} \times \dots \times S_{n_s}$ such that $1 \leq s \leq k$. The above defined r.vec. is called partially exchangeable.

In the following, we give the definitions of the main functional operations which are to be investigated in the next section. A sequence of distribution functions (d.fs.) $F_m(x)$, say, is said to converge weakly to a limit d.f. $F(x)$ if $\lim_{m \rightarrow \infty} F_m(x) = F(x)$ at all continuity points of $F(x)$. If $F_1(x)$ and $F_2(x)$ are two d.fs., then $F(x) = \int_{-\infty}^{\infty} F_1(x-y) dF_2(y)$ is a distribution function and is called the convolution of F_1 and F_2 . Also we define $F(x) = \sum_{i=1}^n \alpha_i F(x, \theta_i)$ for some finite sequence $F(x, \theta_i)$ for which $\alpha_i > 0$ for all $i = 1, \dots, n$ and $\sum_{i=1}^n \alpha_i = 1$, as the mixture of the sequence of d.fs. $F(x, \theta_i)$. Another form of this is the continuous mixture that is $F(x) = \int F(x, \theta) dG(\theta)$, where $G(\theta)$ is some d.f. of θ . In a similar way one may define the convolutional mixing of sequence of r.v.s. $\{X_i\}$, $i = 1, \dots, n$, as the d.f. of the r.v. $X = \sum_{i=1}^n \alpha_i X_i$ where $\alpha_i > 0$ and $\sum_{i=1}^n \alpha_i = 1$.

2. THE MAIN RESULTS

In this section we shall establish some results concerning the behaviour of classes of distributions having the structure of exchangeability and Schur-functions under some functional operations. Further some related results are also discussed.

For exchangeable r.v.s. X_1, \dots, X_k with probability of the form

$$P\{(X_1 - \theta_1, \dots, X_k - \theta_k) \in A\} = P(\underline{X} \in A + \underline{\theta}) \quad (2.1)$$

$\underline{\theta} \in \underline{\Theta} \subset \mathbb{R}^n$, is a parameter vector, often exhibit a monotonicity property in values of partially ordered according to majorization. Now the next lemma which is due to Marshall and Olkin (1974) shows that $P(\underline{X} \in A + \underline{\theta})$ is Schur-concave function of $\underline{\theta}$ whenever A has a Schur-concave indicator function.

Lemma 2.1 Let $f(\underline{x})$ be a Schur-concave function and consider a Lebesgue measurable

set $A \subset \mathbb{R}^n$ such that

$$\underline{y} \in A \text{ and } \underline{x} \stackrel{*}{<} \underline{y} \text{ implies } \underline{x} \in A. \quad (2.2)$$

Then $P(\underline{X} \in A + \underline{\theta}) = \int_{A+\underline{\theta}} f(\underline{x}) d\underline{x}$ is Schur-concave function of $\underline{\theta}$ where $\underline{\theta}$ is some parametric vector.

In fact condition (2.2) can be satisfied for every convex set A of exchangeable r.v.s., since $\underline{x} \stackrel{*}{<} \underline{y}$ implies $\underline{x} = D\underline{y}$ for some doubly stochastic matrix ($n \times n$) D , and since the set of such doubly stochastic matrices is the convex hull of the set of $n \times n$ permutation matrices (Birkoff's theorem, Mirsky (1963)). However condition (2.2) implies neither measurability nor convexity of A . Moreover, Mudholkar (1966) established lemma 2.1 where he generalized a result of Anderson (1955), concerning the unimodality of functions, but with an additional requirement on the set A , that is, A and the set $\{\underline{y}: f(\underline{y}) < c\}$ is convex for some fixed number c , i.e., $f(\underline{y})$ is Anderson unimodal (AU) and $f(\underline{y})$ is exchangeable, then condition (2.2) holds. Now we give the following interesting result.

Theorem 2.1 The class of Schur-concave (convex) functions is closed under reversal, passage to a limit weakly, mixing and convolution.

Proof. We shall give the proof for Schur-concave functions, whereas a similar argumentation can be carried out for the Schur-convex case. Clearly, the reversal property is valid, that is if $f(\underline{x})$ is Schur-concave function then so is $f(-\underline{x})$.

Now let $\{f_k\}$ be a sequence of Schur-concave functions. Then we have for any set A satisfying lemma 2.1, that $f_k / \leq h$ for each k , where h is integrable function on $A + \underline{\theta}$. Next, let f_k converge weakly to a function f . Then by Lebesgue Dominated Convergence Theorem and lemma 2.1, one get f is Schur-concave.

The mixture of finite number of Schur-concave functions is Schur-concave, which can be easily seen by applying relation (1.3).

Finally, to show that the class of Schur-concave functions is closed under convolution, let f_1 and f_2 be two Schur-concave functions. Then $f(-\underline{x})$ is also Schur-concave and we need to show that

$$f(\underline{\theta}) = \int_{\mathbb{R}^n} f_1(\underline{x}) f_2(\underline{\theta} - \underline{x}) d\underline{x} \quad (2.3)$$

for some parameter $\underline{\theta}$ is Schur-concave in $\underline{\theta}$. But by lemma 2.1, $\int_{A+\underline{\theta}} f_2(-\underline{x}) d\underline{x} = \int_{\mathbb{R}^n} I_A(-\underline{x}) f_2(\underline{\theta} - \underline{x}) d\underline{x}$ is Schur-concave in $\underline{\theta}$. Now approximate $f_1(\underline{x})$ by an increasing sequence of simple functions $h_k = \sum_i \alpha_i I_{A_i}$, where $\sum \alpha_i = 1$ and the sets A_i satisfy lemma 2.1. Hence by using Lebesgue Monotone Convergence Theorem the required result follows.

Now, we consider the case when the underlying r.v.s. are independent and identically

distributed and each has a common density g , say. Thus the joint density of \underline{x} is $f(\underline{x}) = \prod_{i=1}^n g(x_i)$ and in this case f is Schur-concave (convex) if and only if $\log g$ is concave (convex). Therefore, in such circumstances the function $f(\underline{x}) = \prod_{i=1}^n g(x_i)$ is AU and g is log-concave.

Corollary 2.1 Let $f(\underline{x})$ and $\phi(\underline{y})$ be two Schur-concave (convex) functions and assume $g(\underline{\theta}) = \int f(\underline{x}) \phi(\underline{\theta}-\underline{x}) \prod_{i=1}^n d\mu(x_i)$ is well defined. Then g is Schur-concave (convex).

In the above corollary both f and ϕ are Borel measurable functions and μ denotes the Lebesgue (or counting) measure. Further, in the next result a generalization, in some sense, of corollary 2.1 is given by taking the function to be totally positive of order two (TP_2) and satisfying the semigroup property that is (i) for any $x_1 < x_2$ and $\lambda_1 < \lambda_2$, $\phi(\lambda_1, x_1)\phi(\lambda_2, x_2) - \phi(\lambda_1, x_2)\phi(\lambda_2, x_1) \geq 0$, and (ii) for any Lebesgue measure μ on $[0, \infty)$ or counting measure on $\{0, 1, 2, \dots\}$ and in this case $\lambda \in [0, \infty)$ or $\lambda \in \{0, 1, 2, \dots\}$, $x \in [0, \infty)$ one has

$$\phi(\lambda_1 + \lambda_2, x) = \int \phi(\lambda_1, x-y)\phi(\lambda_2, y) d\mu(y).$$

Thus $g(\underline{\theta})$ in corollary 2.1 may be considered as mixture of the density $\phi(\underline{\theta}, \underline{x})$.

The proof of the following theorem can be carried out in similar manner to a result in Proschan and Sethuxaman (1977).

Theorem 2.2 Let $f(\underline{x})$ be Schur-concave (convex) function and consider the TP_2 function $\phi(\underline{\theta}, \underline{x})$ which is defined on $R_+^n \times R_+^n$ and satisfies the semigroup property. Thus, if

$$g(\underline{\theta}) = \int_{R_+^n} f(\underline{x}) \phi(\underline{\theta}, \underline{x}) \prod_{i=1}^n d\mu(x_i)$$

exists, then $g(\underline{\theta})$ is Schur-concave (convex) function.

It can be realized that theorem 2.2 above is more general than lemma 2.1, because $f(\underline{x}) = \prod_{i=1}^n f(x_i)$ is Schur-concave if $f(x+y)$ is TP_2 . Further, one may generalize the nonnegative (Lebesgue measurable) Schur-concave function as follows: Let G be a closed subgroup of $n \times n$ orthogonal matrices taking values in R^n and G_G be a convex hull of the G -orbits, $\{g\underline{x}; g \in G\}$ of \underline{x} . Assume that the group G has partial ordering \leq_* on R^n such that $\underline{y} \leq_* \underline{x}$ if and only if $\underline{y} \in G_G(\underline{x})$. A real valued function $f: R^n \rightarrow R^n$ is called G -monotone if $\underline{y} \leq_* \underline{x}$ implies $f(\underline{y}) \geq f(\underline{x})$. A G -monotone function is g -invariant. Let T_G be the class of all G -monotone functions $f: R^n \rightarrow R^n$ which are Lebesgue integrable over R^n . It was shown (see Eaton and Perlman (1977)) that T_G is closed under convolution. If G is permutation group, then \leq_* is exactly the majorization ordering \leq^* , and T_G is the class of nonnegative Schur-concave functions. Obviously, Schur-convex functions have

similar structure and can be easily established.

In the next theorem we summarize the behaviour of the class of exchangeable probability distributions under some functional operations.

Theorem 2.3 The class of exchangeable d.f.s. is closed under reversal, passage to a limit weakly, convolution and mixing of density functions (den.f.s.), but is not closed under the convolutional mixing, that is the mixing of r.v.s. in the general case.

Proof. It is obvious that if $F(\underline{x})$ is exchangeable d.f., then the d.f. of the r.v. $-\underline{x}$ is also exchangeable.

Let $\underline{Y}_m = (X_{1,m}, X_{2,m}, \dots, X_{n,m})$ be r.vec. where $m = 1, 2, \dots$ and n is a fixed number greater than or equal to two. Then $\lim_{m \rightarrow \infty} F_m(\underline{y}) = F(\underline{y})$ is exchangeable if \underline{Y}_m is exchangeable for every m , where each $\{X_{i,m}\}$, $i = 1, \dots, n$ forms a convergent sequence of r.v.s., and this proves the closure under the weak convergence.

Let \underline{X} and \underline{Y} be two r.vecs. with d.f.s. $F_1(\underline{x})$ and $F_2(\underline{y})$, then the d.f. of the r.v. $\underline{Z} = \underline{X} + \underline{Y}$ which is $F(\underline{z}) = \int F_1(\underline{x}-\underline{y}) dF_2(\underline{y})$ is also exchangeable. Consider the finite sequence of exchangeable den.f.s. $\{f_i(\underline{x})\}$, $i = 1, \dots, n$, and some $\alpha_i > 0$ where $\sum_{i=1}^n \alpha_i = 1$, then the mixture of such sequence $f(\underline{x}) = \sum_{i=1}^n \alpha_i f_i(\underline{x})$ is also a den.f. and is exchangeable, because $f(\underline{x})$ does not vary for all possible permutations of the components of the underlying r.vec. \underline{X} . However, the case is different when we consider the convolutional mixing which is of the form $\underline{Y} = \sum_{i=1}^n \alpha_i X_i$. Here, the den.f. of the r.vec. \underline{Y} can not be exchangeable unless $\alpha_1 = \alpha_2 = \dots = \alpha_n = 1/n$, that is, if one gives equal weights for all r.vecs. which does not exist in most of the real life situations.

3. FURTHER RESULTS ON EXCHANGEABILITY

We have already introduced the definition of partial exchangeability at the end of section 1. Thus, one can easily see that the r.vec. $\underline{X} = (X_1, \dots, X_k)$ is exchangeable if $s = 1$, and non-exchangeable (non-permutational symmetric) if $s = k$.

In particular, nonexchangeability means there are no two components of the underlying r.vec. which are interchangeable. In addition, if we denote by Ω_{pi}, Ω_i ,

$\Omega_{s_k}, \Omega_{se}$ the classes of sequences of r.vecs. (or their probability distributions, or their densities if they exist) which are partially independent, stochastically independent, partially exchangeable and spherical exchangeable, respectively, where $1 = s_1 < \dots < s_i < \dots < s_k = k$, clearly one has

$$\Omega_i \subset \Omega_{s_1} \subset \dots \subset \Omega_{s_i} \subset \dots \subset \Omega_{s_k} ; \quad \Omega_i \subset \Omega_{pi} \subset \Omega_{s_i} ; \quad \Omega_{se} \subset \Omega_{s_1} .$$

This kind of structure and implications are motivated from a practical viewpoint, therefore we introduce some other structures which might appear as well.

In the above set up, consider the case when the density $f(\underline{x})$ of the r.vec. \underline{X} is unchanged when the same permutation is applied to each of the n_i 's components, $i = 1, 2, \dots, s$. However, $f(\underline{x})$ increases or decreases when a certain permutation is applied to some but not all of the n_i 's. The r.vec. \underline{X} is called related exchangeable (REX). It is also called increasing in exchangeability (IEX) if the defined REX leads to an increasing $f(\underline{x})$, and decreasing in exchangeability (DEX) if it leads to a decreasing $f(\underline{x})$.

Now let $\underline{\theta} = (\theta_1, \dots, \theta_k)$ and $\underline{X} = (X_1, \dots, X_k)$ be parameter vector and r.vec. respectively. A function $f(\underline{x}, \underline{\theta})$ is said to be decreasing in transposition (DT) if for $\theta_1 \leq \theta_2 \leq \dots \leq \theta_k$ (i) f is unchanged when the same permutation is applied to $\underline{\theta}$ and to \underline{x} , and (ii) $f(\underline{x}, \underline{\theta}) \geq f(\underline{x}', \underline{\theta})$ whenever \underline{x} and \underline{x}' differ in two coordinates only, say i and j , such that $(x_i - x_j)(i - j) \geq 0$ and $x_i' = x_j$ and $x_j' = x_i$. Clearly $f(\underline{x}, \underline{\theta})$ is defined from $R^n \times R^n$ to R^n . This latter case has been investigated by Hollander et al (1977).

Further, there is another related concept of exchangeability which has been brought up recently, that is the concept of positive dependence by mixture (PDM). The probability d.f. $F(\underline{x})$ is called PDM if it can be represented by a mixture of d.fs. of exchangeable and independent r.vs. (see Shaked (1977)). In other words, the r.vec. \underline{X} with d.f. $F(x_1, \dots, x_n) = \int \prod_{i=1}^n F(x_i, \theta) dG(\theta)$, where $\theta \in \Theta \subset R^n$, G is a d.f. on Θ and the r.vs. X_1, \dots, X_n are independent and identically distributed, is called PDM. Such mixture may be said to have conditionally independence structure.

It is expected from the structure of the above classes that the classes of distributions with IEX, DEX, DT or PDM form of structure are closed under most of the functional operations studied earlier and some other kernel-type transformations.

4. CONCLUDING REMARKS AND APPLICATIONS

It is aimed in this section to discuss briefly some of the main applications where the concepts of exchangeability and majorization have played very interesting roles.

(i) Exchangeability is used by Ahmad and Abouammoh (1979) to establish the classes of infinitely A-divisible distributions where A refers to different forms of symmetric dependence instead of the stochastic independence of the underlying r.vs. in the class of infinitely divisible distributions.

(ii) In branching processes, namely population genetics in biology and other growth phenomena, the individuals of nonoverlapping generations are not distinguishable and the effects of individuals in the whole process are considered to be

exchangeable. See Kingman (1978) for the developement of such problem.

(iii) In time series analysis of various aspects of climatic changes the r.v.s. involved are, in fact, r.vecs. of exchangeable variables which are approximated by the resultant of their effects for the simplicity of studied model.

(iv) The concepts of exchangeability and its related classes are used in choosing prior probabilities in Bayesian statistics and subjective probability : see Hamaker (1977). In Ahmad (1975) it was found that for exchangeable hypotheses all distribution-free statistics are based on permutation-rank statistics.

(v) It was shown by Hollander et al (1977) that many d.fs. such as the multinomial, negative multinomial, multivariate hypergeometric, Dirichlet, multivariate normal, multivariate F, multivariate logarithmic series and some other d.fs. are all DT.

(vi) In Marshall and Olkin (1974) it was pointed out that if $A = [a_{ij}]$ is a positive definite matrix with $a_{11} = \dots = a_{nn}$ and $a_{ij} = a_{ji}$ for $i \neq j$, then $f(\underline{x}) = g(\underline{x}A\underline{x}')$ is Schur-concave, where g is an increasing function. In addition, the multivariate beta, F and chi-square are all Schur-concave distributions.

(vii) The structure of the concepts of exchangeability, its related classes and Schur-concavity (convexity) remain unchanged for the survival probability (or reliability) $\bar{F}(\underline{x}) = 1 - F(\underline{x})$ as it is for the probability distribution $F(\underline{x})$.

In particular, let us consider the following example.

Let $X(t) = (Y(t), Z(t))$ be some bivariate variable of some climatic factors such as temperature, precipitation, relative humidity etc., where $Y(t)$ and $Z(t)$ are first order stationary autoregressive, AR(1), with transfer function of the linear filter $1 - \phi D$ and $1 + \phi D$ where D is the backward shift operator such that $DX(t) = X(t-1)$, then $X(t)$ is exchangeable bivariate r.v.. Further $X(t)$ is also exchangeable if one considers $Y(t)$ and $Z(t)$ to be first order stationary autoregressive moving average, ARMA(1,1) with transfer functions of the linear filter $1 - \phi D$, $1 - \phi D$ and $1 + \phi D$, $1 + \phi D$. We may investigate this problem in some detail in future publication, specifically, its generalization, relation with REX and some of its applications.

ACKNOWLEDGEMENT

I would like to thank Dr. R. Ahmad of Strathclyde University for his advice. Also I would appreciate the assistance given to me by Dr. M.S.Ali Khan and A. Abd-Alla in revising parts of this paper.

REFERENCES

- Ahmad, R., 1975. Some characterizations of exchangeable processes and distribution-free tests. In: G.P.Patil et al (eds.), Statistical Distributions in Scientific Work., Vol.3., Reidel Pub. Co., Dordrecht : 237-248.

- Ahmad, R. and Abouammoh, A.M., 1979. On classes of A-independent probability distributions and on classes of infinitely A-divisible distributions. In: Proc. 14th Ann. Conf. in Stat., Computer Sci., Oper. Res. and Math., Cairo Univ. P., Vol.2: 9-34.
- Ahmad, R. and Peterson, M.M., 1978. Restricted permutation symmetry and hypotheses-generating group in statistics. In: Inform. Theory, Stat. Dec. Funct., Random Proc., Vol.A., Academic P., Prague : 71-82.
- Anderson, T.W., 1975. The integral of symmetric unimodal functions over a symmetric convex set and some probability inequalities. Proc. Amer. Math. Soc. 6 : 170-176.
- Berge, S., 1963. Topological Spaces. MacMillan.
- De Fenitti, B., 1975. Theory of Probability. Vol.2., Wiley.
- Eaton, M.L. and Perlman, M.D., 1977. Reflection groups, generalization and the geometry of majorization. Ann. Prob. 6: 829-860.
- Feller, W., 1971. An Introduction to Probability Theory and Its Applications. Vol.2, Wiley.
- Hamaker, H.C., 1977. Subjective probabilities and exchangeability from an objective point of view. Inter. Stat. Rev. 45: 223-231.
- Hardy, G.H., Littlewood, J.E. and Polya, G., 1952. Inequalities. Camb. Univ. P..
- Hollander, M., Proschan, F. and Sethuraman, J., 1977. Functions decreasing in transposition and their applications in ranking problems. Ann. Statist. 6: 139-151.
- Kingman, J.F.C., 1978. Uses of exchangeability. Ann. Prob. 6: 183-197.
- Marshall, A.W. and Olkin, I., 1974. Majorization in multivariate distributions. Ann. Statist. 2: 1189-1200.
- Mirsky, L., 1963. Results and problems on the theory of doubly stochastic matrices. Zeit. Wahrsch. Verw. Geb. 1: 319-334.
- Mudholkar, G.S., 1966. The integral of an invariant unimodal function over an invariant convex set, an inequality and applications. Proc. Amer. Math. Soc. 17: 1327-1333.
- Ostrowski, A., 1952. Sur quelques applications de fonctions convexes et concaves au sans de I. Schur. J. Math. Pure Appl. 31 : 253-292.
- Proschan, F. and Sethuraman, J., 1977. Schur functions in statistics, I. The preservation theorem. Ann. Statist. 6: 256-262.
- Schur, I., 1923. Über eine klass von mittelbildungen mit anwendungen auf die determinantentheorie. Sitzber. Berl. Math. Ges. 29: 9-20.
- Shaked, M., 1977. A concept for positive dependence for bivariate distributions. J. Amer. Statist. Assoc. 72 : 642-650.

PREDICTION OF A FUTURE ORDERED OBSERVATION BASED ON A SAMPLE FROM THE EXPONENTIAL POPULATION

E.H.GAN, M.SAFIUL HAK and M.M.ALI

Math. Dept., Univ. of Western Ontario, Ontario (Canada)

ABSTRACT

Gan, E.H., Hak, M.S. and Ali, M.M., Prediction of a future ordered observation based on a sample from the exponential population. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

The paper derives the prediction distribution of the smallest, the i -th and a set of m future ordered responses based on a sample arising from an exponential model with location parameter μ (>0) and scale parameter σ (>0). For the derivations, the structural relations between the responses and the error variables associated with the responses have been utilized. Here, because of the restriction $\mu > 0$, Fraser's (1968) general result on the structural model could not be used.

1. INTRODUCTION

Given a set of data from an experiment the distribution of a future response from the same experiment or similar experiment is called the prediction distribution. The methods for deriving the prediction distribution involve either the distribution of a pivotal quantity based on the present and the future response or an integration over the parameter space of the joint density of the parameter and the future response given the parameter. The integration formula for the prediction density is

$$p(y|x) = \int p(\theta|x)f(y|\theta)d\theta \quad (1.1)$$

where $p(\theta|x)$ represents the structural or posterior density of θ given the data x and $f(y|\theta)$ represents the density of the future response y for given θ . The works of Aitchison and Dunsmore (1975), Fisher (1959), Fraser and Haq (1969,70), Geisser (1965), Hora and Buehler (1967), Lindley (1972), and Zellner and Chetty (1965) may be mentioned in this respect.

In this paper we have used the integration formula to derive the prediction distribution of data from an exponential life testing model with location parameter $\mu > 0$, and the scale parameter $\sigma > 0$. We have utilized the structural relations

between the observations and the errors associated with the observations to obtain the distribution of the parameters. Because of the restriction that $\mu > 0$, we could not use the Fraser's general results (1968) on the structural model. However, since structural relations have been used to obtain the distribution of parameters, we have denoted the distribution of the parameters as the structural distribution.

In section 2, we describe the model; in section 3, we derive the structural distribution of μ and σ ; in section 4 we derive the prediction distribution for the smallest, the i -th, and a set of m future responses; in section 5 we obtain the point prediction of a future response for $\mu > 0$ and for $-\infty < \mu < \infty$ using mean square error criterion; and in section 6, we give an example.

2. THE MODEL

Let X be a random variable representing the operating time to failure (life length) of a device with the following exponential probability density function:

$$f(x;\mu,\sigma) = \begin{cases} (1/\sigma) \exp[-(x-\mu)/\sigma] & , \quad x > \mu, \quad \mu > 0, \quad \sigma > 0; \\ 0 & , \quad \text{otherwise.} \end{cases} \quad (2.1)$$

Here μ , the location parameter is recognized as the guarantee time or the minimum life of the device and is, therefore, assumed to be positive; and σ is the scale parameter.

Let x_1, x_2, \dots, x_n be a set of observed responses from the above model. The responses could be expressed as

$$x_1 = \mu + \sigma e_1, \quad x_2 = \mu + \sigma e_2, \dots, \quad x_n = \mu + \sigma e_n \quad (2.2)$$

where e_1, e_2, \dots, e_n are the error variables associated with the responses x_1, x_2, \dots, x_n respectively. e_1, e_2, \dots, e_n are not known, but their joint probability distribution is given by

$$\exp \left[- \sum_{\alpha=1}^n e_{\alpha} \right] \prod_{\alpha=1}^n de_{\alpha} \quad , \quad e_{\alpha} > 0, \quad \alpha = 1, \dots, n. \quad (2.3)$$

The responses are obtained from the error variables by a location-scale transformation. But due to the restriction on μ , the transformations do not form a group; therefore, the general results of a structural model (Fraser, 1968, p.41) are not applicable for deriving the structural distribution of μ and σ . With such restriction on the location parameter, Hoq, Ali and Templeton (1974) derived the structural distribution of μ and σ for a generalized life testing model through a series of transformations. We have used a simple transformation to obtain the distribution

of μ and σ and utilized the structural distribution to obtain the prediction distribution.

Prediction problems for data from exponential models have been studied by various authors, namely, Bury and Bernholtz (1971), Dunsmore (1974), Faulkenberg (1973), Kaminsky (1977a, 1977b), Kaminsky and Rhodin (1978), Lawless (1977), Likes (1974), Lingappaiah (1978), Nelson (1970), and Schafer and Agnes (1977). It is to be noted that none of the studies considered the case when the location parameter is positive.

3. THE STRUCTURAL DISTRIBUTION OF μ AND σ

Consider the model described in section 2. Let $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ be the order statistics with $x_{(1)} < x_{(2)} < \dots < x_{(n)}$. Then from (2.2) we have

$$x_{(\alpha)} = \mu + \sigma e_{(\alpha)}, \quad \alpha = 1, 2, \dots, n. \quad (3.1)$$

Also the error distribution (2.3) changes to

$$n! \exp\left[-\sum_{\alpha=1}^n e_{(\alpha)}\right] \prod_{\alpha=1}^n de_{(\alpha)}, \quad e_{(1)} < e_{(2)} < \dots < e_{(n)}. \quad (3.2)$$

Consider the following transformation from $e_{(1)}, e_{(2)}, \dots, e_{(n)}$ to $e_{(1)}, s(\underline{e}), d_3, \dots, d_n$:

$$e_{(1)} = e_{(1)}, \quad s(\underline{e}) = \left[\frac{1}{n-1} \sum_{\alpha=2}^n (e_{(\alpha)} - e_{(1)})^2 \right]^{1/2}, \quad d_{\alpha} = \frac{e_{(\alpha)} - e_{(1)}}{s(\underline{e})}, \quad \alpha = 3, \dots, n. \quad (3.3)$$

Then

$$e_{(1)} = e_{(1)}, \quad e_{(2)} = e_{(1)} + \left[(n-1) - \sum_{\alpha=3}^n d_{\alpha}^2 \right]^{1/2} s(\underline{e}), \quad e_{(\alpha)} = e_{(1)} + d_{\alpha} s(\underline{e}), \quad \alpha = 3, \dots, n.$$

It is readily seen that the transformations (3.3) are one to one and the Jacobian of the transformation is

$$|J| = s(\underline{e})^{n-2} \left[(n-1) - \sum_{\alpha=3}^n d_{\alpha}^2 \right]^{1/2}.$$

Therefore the joint density function of $e_{(1)}, s(\underline{e}), d_3, \dots, d_n$ is given by

$$n! \exp\left[-ne_{(1)} - s(\underline{e}) \left\{ \left[(n-1) - \sum_{\alpha=3}^n d_{\alpha}^2 \right]^{1/2} + \sum_{\alpha=3}^n d_{\alpha} \right\}\right] s(\underline{e})^{n-2} \left[(n-1) - \sum_{\alpha=3}^n d_{\alpha}^2 \right]^{1/2} \quad (3.4)$$

It is to be noted that $d_{\alpha} = (e_{(\alpha)} - e_{(1)})/s(\underline{e}) = (x_{(\alpha)} - x_{(1)})/s(\underline{x})$, $\alpha = 3, \dots, n$, are known. Therefore we obtain the conditional probability density function of

$e_{(1)}$ and $s(\underline{e})$, for the given values of d_3, \dots, d_n as

$$f(e_{(1)}, s(\underline{e}) | d_3, \dots, d_n) = k(\underline{d}) \exp[-ne_{(1)} - s(\underline{e}) \{ [n-1 - \sum_{\alpha=3}^n d_{\alpha}^2]^{1/2} + \sum_{\alpha=3}^n d_{\alpha} \}] s(\underline{e})^{n-2} \quad (3.5)$$

where $k(\underline{d})$ is the normalizing constant. Now from (3.1) we have

$$x_{(1)} = \mu + \sigma e_{(1)}, \quad x_{(1)} \geq \mu > 0, \quad s(\underline{x}) = \sigma s(\underline{e}). \quad (3.6)$$

The relations (3.6) are obtained from the structural relations (2.2) between the observations and the error variables. For the given set of responses, these relations along with the conditional probability element of $e_{(1)}$ and $s(\underline{e})$ are used to make probability statements about μ and σ :

$$g(\mu, \sigma | \underline{x}) d\mu d\sigma = \Psi(\underline{x}) \exp[-\frac{1}{\sigma} \sum_{\alpha=1}^n (x_{(\alpha)} - \mu)] \sigma^{-(n+1)} d\mu d\sigma \quad (3.7)$$

where

$$\begin{aligned} \Psi^{-1}(\underline{x}) &= \int_0^{\infty} \int_0^{x_{(1)}} \exp[-\frac{1}{\sigma} \sum_{\alpha=1}^n (x_{(\alpha)} - \mu)] \sigma^{-(n+1)} d\mu d\sigma \\ &= \frac{\Gamma(n-1)}{n} \{ [\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)})]^{-(n-1)} - [\sum_{\alpha=1}^n x_{(\alpha)}]^{-(n-1)} \}, \end{aligned} \quad (3.8)$$

and $0 < \mu < x_{(1)}$, and $\sigma > 0$. This result differs from that of Fraser (1968, p.41) since the derivation takes into account the restriction on μ . Hoq, Ali and Templeton (1974) obtained the same result through a series of transformations. The present derivation, however, avoids the series of transformations.

For a given set of data the probability statements about μ and σ form the basis of inference about μ and σ . The marginal distribution of μ is obtained as

$$g_1(\mu | \underline{x}) d\mu = n(n-1) \frac{[\sum_{\alpha=1}^n (x_{(\alpha)} - \mu)]^{-n} d\mu}{[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)})]^{-(n-1)} - [\sum_{\alpha=1}^n x_{(\alpha)}]^{-(n-1)}} \quad (3.9)$$

for $0 < \mu < x_{(1)}$, and the marginal distribution of σ is obtained as

$$g_2(\sigma | \underline{x}) d\sigma = \frac{\sigma^{-n}}{\Gamma(n-1)} \{ \exp[-\frac{1}{\sigma} \sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)})] - \exp[-\frac{1}{\sigma} \sum_{\alpha=1}^n x_{(\alpha)}] \} d\sigma \quad (3.10)$$

for $0 < \sigma < \infty$.

4. THE PREDICTION DISTRIBUTION

4.1 Prediction Distribution of the Smallest Future Response

Let y_1, y_2, \dots, y_m be m future observations from the statistical model described in section 2. Let $y_{(1)}$ be the smallest future response. Then the probability density of $y_{(1)}$ for given μ and σ is

$$\frac{m}{\sigma} \exp\left[-\frac{m}{\sigma}(y_{(1)} - \mu)\right], \quad \mu < y_{(1)} < \infty. \quad (4.1)$$

Therefore, using the formula (1.1) and the expressions (3.8) and (4.1) the prediction distribution of $y_{(1)}$ is obtained as

$$p(y_{(1)} | \underline{x}) = \Psi(\underline{x}) \int_0^\infty \int_0^{\min(x_{(1)}, y_{(1)})} \frac{1}{m\sigma^{-(n+2)}} \exp\left[-\frac{1}{\sigma} \sum_{\alpha=1}^n (x_{(\alpha)} - \mu) - \frac{m}{\sigma}(y_{(1)} - \mu)\right] d\mu d\sigma$$

$$= \begin{cases} \Psi(\underline{x}) \frac{m\Gamma(n)}{m+n} \left\{ \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) + m(y_{(1)} - x_{(1)}) \right]^{-n} - \left[\sum_{\alpha=1}^n x_{(\alpha)} + my_{(1)} \right]^{-n} \right\}, & \text{if } y_{(1)} > x_{(1)}, \\ \Psi(\underline{x}) \frac{m\Gamma(n)}{m+n} \left\{ \left[\sum_{\alpha=1}^n x_{(\alpha)} - ny_{(1)} \right]^{-n} - \left[\sum_{\alpha=1}^n x_{(\alpha)} + my_{(1)} \right]^{-n} \right\}, & \text{if } y_{(1)} < x_{(1)}, \end{cases} \quad (4.2)$$

where $\Psi(\underline{x})$ is given by (3.8).

4.2 The Prediction Distribution of the i -th Future Response

Let $y_{(i)}$, $i = 1, 2, \dots, m$, be the i -th order statistic for a set of m future responses from the statistical model described in section 2. Then the probability density function of $y_{(i)}$ for given μ and σ is

$$i \binom{m}{i} \sum_{j=0}^{i-1} (-1)^{i-j-1} \binom{i-1}{j} \frac{1}{\sigma} \exp\left[-\frac{1}{\sigma}(m-j)(y_{(i)} - \mu)\right]. \quad (4.3)$$

By the integration formula (1.1), the prediction distribution of $y_{(i)}$ for the given data \underline{x} , is obtained as

$$p(y_{(i)} | \underline{x}) = \Psi(\underline{x}) \int_0^\infty \int_0^{\min(x_{(1)}, y_{(i)})} i \binom{m}{i} \sigma^{-(n+2)} \sum_{j=0}^{i-1} (-1)^{i-j-1} \binom{i-1}{j} \\ \times \exp\left[-\frac{1}{\sigma}(m-j)(y_{(i)} - \mu) - \frac{1}{\sigma} \sum_{\alpha=1}^n (x_{(\alpha)} - \mu)\right] d\mu d\sigma$$

$$= \begin{cases} \psi(\underline{x}) i_{(i)}^{(m)} \sum_{j=0}^{i-1} (-1)^{i-j-1} \binom{i-1}{j} \frac{\Gamma(n)}{m+n-j} \left\{ \left[\sum_{\alpha=1}^n x_{(\alpha)} - (n-j)y_{(i)} \right]^{-n} - \left[\sum_{\alpha=1}^n x_{(\alpha)} + (m-j)y_{(i)} \right]^{-n} \right\} \\ \quad \text{for } y_{(i)} < x_{(1)}, \\ \psi(\underline{x}) i_{(i)}^{(m)} \sum_{j=0}^{i-1} (-1)^{i-j-1} \binom{i-1}{j} \frac{\Gamma(n)}{m+n-j} \left\{ \left[\sum_{\alpha=1}^n x_{(\alpha)} + (m-j)y_{(i)} - (m+n-j)x_{(1)} \right]^{-n} \right. \\ \quad \left. - \left[\sum_{\alpha=1}^n x_{(\alpha)} + (m-j)y_{(i)} \right]^{-n} \right\}, \text{ for } y_{(i)} > x_{(1)}. \end{cases} \quad (4.4)$$

From (4.4) one could easily obtain the prediction distribution of $y_{(1)}$, the smallest future response, or $y_{(m)}$, the largest future response.

4.3 The Prediction of m Ordered Future Responses

The joint probability distribution of m ordered future responses (for given μ and σ) from the statistical model described in section 2 is

$$m! \sigma^{-m} \exp \left[-\frac{1}{\sigma} \sum_{j=1}^m (y_{(j)} - \mu) \right] dy_{(1)} \dots dy_{(m)}; \mu < y_{(1)} < y_{(2)} < \dots < y_{(m)} < \infty, \quad (4.5)$$

where $y_{(j)}$, $j = 1, 2, \dots, m$, is the j -th order statistic. Then as before we obtain the prediction density of $y_{(1)}, \dots, y_{(m)}$ as

$$p(y_{(1)}, \dots, y_{(m)} | \underline{x}) = \psi(\underline{x}) \int_0^\infty \int_0^{\min(x_{(1)}, y_{(1)})} \frac{1}{m! \sigma^{-(n+2)}} \exp \left[-\frac{1}{\sigma} \sum_{\alpha=1}^n (x_{(\alpha)} - \mu) \right. \\ \left. - \frac{1}{\sigma} \sum_{j=1}^m (y_{(j)} - \mu) \right] d\mu d\sigma \\ = \begin{cases} \psi(\underline{x}) \frac{m! \Gamma(m+n-1)}{m+n} \left\{ \left[\sum_{\alpha=1}^n x_{(\alpha)} + \sum_{j=1}^m y_{(j)} - (m+n)x_{(1)} \right]^{-(m+n-1)} \right. \\ \quad \left. - \left[\sum_{\alpha=1}^n x_{(\alpha)} + \sum_{j=1}^m y_{(j)} \right]^{-(m+n-1)} \right\}, \text{ for } y_{(1)} > x_{(1)}, \\ \psi(\underline{x}) \frac{m! \Gamma(m+n-1)}{m+n} \left\{ \left[\sum_{\alpha=1}^n x_{(\alpha)} + \sum_{j=1}^m y_{(j)} - (m+n)y_{(1)} \right]^{-(m+n-1)} \right. \\ \quad \left. - \left[\sum_{\alpha=1}^n x_{(\alpha)} + \sum_{j=1}^m y_{(j)} \right]^{-(m+n-1)} \right\}, \text{ for } y_{(1)} < x_{(1)}. \end{cases} \quad (4.6)$$

This result could easily be extended to the case when k ($< m$) out of m future responses are smaller than $x_{(1)}$.

5. POINT PREDICTION OF A FUTURE RESPONSE

The predicted value of a future response given the set of data is called the point predictor. Using the mean square error criterion, the point predictor of the smallest future response $\hat{Y}_{(1)}$ for a future sample of size m is obtained as follows:

$$\begin{aligned}\hat{Y}_{(1)} &= E(Y_{(1)} | \underline{x}) = \int Y_{(1)} p(Y_{(1)} | \underline{x}) dY_{(1)} \\ &= \frac{n}{n+m} \left\{ \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) \right]^{-(n-1)} - \left[\sum_{\alpha=1}^n x_{(\alpha)} \right]^{-(n-1)} \right\}^{-1} \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) \right]^{-(n-1)} \\ &\quad \times \left(x_{(1)} + \frac{\frac{n-m}{2} \sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)})^2}{n(n-2)} + \frac{m}{n^2} \sum_{\alpha=1}^n x_{(\alpha)} \right. \\ &\quad \left. - \frac{\frac{n-m}{2}}{n^2(n-2)} \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) \right]^{n-1} \left[\sum_{\alpha=1}^n x_{(\alpha)} \right]^{-(n-2)} \right\}. \quad (5.1)\end{aligned}$$

If one is interested in the predicted value of a future response \hat{Y} , it is obtained as follows :

$$\begin{aligned}\hat{Y} &= E(Y | \underline{x}) = \int Y p(Y | \underline{x}) dY \\ &= \frac{n}{n+1} \left\{ \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) \right]^{-(n-1)} - \left[\sum_{\alpha=1}^n x_{(\alpha)} \right]^{-(n-1)} \right\}^{-1} \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) \right]^{-(n-1)} \\ &\quad \times \left(x_{(1)} + \frac{\frac{n-1}{2} \sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)})^2}{n(n-2)} + \frac{1}{n^2} \sum_{\alpha=1}^n x_{(\alpha)} \right. \\ &\quad \left. - \frac{\frac{n-1}{2}}{n^2(n-2)} \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) \right]^{n-1} \left[\sum_{\alpha=1}^n x_{(\alpha)} \right]^{-(n-2)} \right\}. \quad (5.2)\end{aligned}$$

It is readily seen that (5.2) can easily be obtained from (5.1) by putting $m = 1$.

For the exponential model with unrestricted location parameter the structural distribution of μ and σ can be obtained by using Fraser's general result (1968, p.41). Then the prediction distribution of the smallest future response from a sample of size m can be obtained by the integration formula (1.1) and then the point predictor of the smallest future response $\hat{Y}_{(1)}^*$ can be obtained as follows :

$$\hat{Y}_{(1)}^* = E(Y_{(1)}^* | \underline{x}) = x_{(1)} + \frac{n-m}{nm(n-2)} \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) \right]. \quad (5.3)$$

Again the point predictor of a future response for the unrestricted location parameter case is obtained as follows :

$$\hat{Y}^* = E(Y^* | \underline{x}) = x_{(1)} + \frac{n-1}{n(n-2)} \left[\sum_{\alpha=1}^n (x_{(\alpha)} - x_{(1)}) \right] . \quad (5.4)$$

It should be noted that the point predictor is valid only when $n > 2$. It is further noted that if n is large relative to m , the predictors are approximately the same for both cases.

6. AN EXAMPLE

We use the data studied by Grubbs(1971) on mileages at which 19 military carriers failed :

162, 200, 271, 302, 393, 508, 539, 629, 706, 777, 884
1008, 1101, 1182, 1463, 1603, 1984, 2355, 2880 .

From data it can be seen that

$$n = 19, \quad x_{(1)} = 162, \quad \text{and} \quad \sum_{\alpha=1}^{19} (x_{(\alpha)} - x_{(1)}) = 15,869 .$$

For $\mu > 0$, the prediction distribution of a future response Y is

$$p(y | \underline{x}) = \begin{cases} 1.12379 \times 10^{-3} \left\{ \left[1 + \frac{y-162}{15869} \right]^{-19} - \left[\frac{18947+y}{15869} \right]^{-19} \right\}, & \text{for } y > 162 \\ 1.12379 \times 10^{-3} \left\{ \left[1 + \frac{19(162-y)}{15869} \right]^{-19} - \left[\frac{18947+y}{15869} \right]^{-19} \right\}, & \text{for } y < 162, \end{cases}$$

and the value of the point predictor is $\hat{Y} = 1045.93$. For $-\infty < \mu < \infty$, the prediction distribution of Y^* is :

$$p(y^* | \underline{x}) = \begin{cases} 1.07757 \times 10^{-3} \left[1 + \frac{y-162}{15869} \right]^{-19}, & \text{for } y > 162, \\ 1.07757 \times 10^{-3} \left[1 + \frac{19(162-y)}{15869} \right]^{-19}, & \text{for } y < 162, \end{cases}$$

and the value of the point predictor is $\hat{Y}^* = 1046.34$.

7. DISCUSSION

It can easily be shown that the probability that the smallest future response is greater than the present smallest response is $n/(n+m)$, which, for large n , is close to one. Therefore, for large n , we could consider only the case when $y_{(1)}$ is greater than $x_{(1)}$.

It is observed that if instead of the mean square error criterion, the mode of

the prediction distribution is used to obtain the point predictor, it would be $x_{(1)}$ for both cases.

The prediction distributions obtained here, are based on the exponential model and the assumption that μ is positive. These distributions could be used for obtaining suitable prediction intervals or regions for future responses.

REFERENCES

- Aitchison, J. and Dunsmore, I.R., 1975. Statistical Prediction Analysis. Camb. Univ. P. Bury, K.V. and Burnholtz, B., 1975. Life testing : structural inference on the exponential model. *Infor.* 9: 148-159.
- Dunsmore, I.R., 1974. The Bayesian predictive distribution in life testing model. *Technometrics* 16: 455-460.
- Faulkenberg, G.D., 1973. A method of obtaining prediction intervals. *J. Amer. Statist. Assoc.* 68: 433-435.
- Fisher, R.A., 1959. Statistical Methods and Scientific Inference. Edinburgh : Oliver and Boyd.
- Fraser, D.A.S., 1968. The Structural Inference. New York, John Wiley.
- Fraser, D.A.S. and Haq, M.Safiul, 1969. Structural probability and prediction for the multivariate model. *J.R.Statist.Soc. B* 31:317-331.
- Fraser, D.A.S. and Haq, M.Safiul, 1970. Inference and prediction for the multilinear model. *J.Statist.Res.* 4:93-109.
- Geisser, S., 1965. Bayesian estimate in multivariate analysis. *Ann.Math.Statist.* 36: 150-159.
- Grubbs, F.E., 1971. Approximate fiducial bounds on reliability for the two parameter negative exponential distribution, *Technometrics* 13: 873-876.
- Hoq, A.K.M.S., Ali, M.M. and Templeton, J.G.C., 1974. Estimation of parameters of a generalized life testing model. *J.Statist. Res.* 8:67-69.
- Hora, R.B. and Beuhler, R.J., 1967. Fiducial theory and invariant prediction. *Ann.Math. Statist.* 38: 795-801.
- Kaminsky, K.S., 1977a. Comparison of prediction interval for failure times when life is exponential. *Technometrics* 19:83-86.
- Kaminsky, K.S., 1977b. Best prediction of exponential failure times when item may be replaced. *The Australian J.Statist.* 19:61-62.
- Kaminsky, K.S. and Rhodin, L.S., 1978. The prediction in the latest failure. *J.Amer. Statist.Assoc.* 73:863-866.
- Lawless, J.F., 1977. Prediction intervals for the two parameter exponential distribution. *Technometrics* 19: 469-472.
- Likes, J., 1974. Prediction of S^k -ordered observations from the two parameter exponential distribution. *Technometrics* 6:241-244.
- Lindley, D.V., 1972. Bayesian Statistics, A Review. SIAM, Philadelphia.
- Lingappaiah, G.S., 1978. Bayesian approach to the prediction problem in the exponential population. *IEEE Trans. on Reliability R-27*. NO.3: 222-225.
- Nelson, W., 1970. A statistical prediction interval for availability. *IEEE Trans. on Reliability R-19* : 179-182.
- Schafer, R.E. and Agnes, J.E., 1977. Predicting the confidence of passing life tests. *IEEE Trans. on Reliability R-26*:141-143.
- Zellner, A. and Chetty, V.K., 1965. Prediction and decision problem in regression models from Bayesian point of view. *J.Amer.Statist. Assoc.* 60:608-616.

TESTING HOMOGENEITY OF VARIANCES OF A SERIES OF LINEAR MODELS

Y.P. CHAUBEY

Math. Dept., Concordia Univ., Montreal, Quebec (Canada)

ABSTRACT

Chaubey, Y.P., Testing homogeneity of variances of a series of linear models. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

Several test procedures have been developed, recently, to be able to detect heteroscedasticity of observations in a linear model. All these procedures divide the observations into two groups and the resulting test statistic is equivalent to testing the equality of variances of two homoscedastic linear models. In practice, however, several observations may be lumped together to have the same variance. Hence, we should be able to get some alternative tests which will have more power than the above procedures if the alternative is more informative, namely, that some of the observations have the same variance. This, in turn, reduces to the testing of the equality of a series of regression models when the regression parameter is the same. This problem has been investigated in this paper.

1. INTRODUCTION

Consider a series of linear models given by

$$Y_i = X_i \beta + \epsilon_i, \quad i = 1, \dots, k, \quad (1.1)$$

$E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma_i^2 I_{n_i}$, $\text{cov}(\epsilon_i, \epsilon_j) = 0$, $i \neq j$, where Y_i is an n_i -vector of observations, X_i is a known $(p \times 1)$ matrix and ϵ_i is random disturbance for the i -th model of order n_i . Testing the null hypothesis $H_0: \sigma_1^2 = \dots = \sigma_k^2$ has been considered by many authors (see Chaubey (1978) and Harrison and McCabe (1979) and references therein). Almost all the papers considered the alternative hypothesis in some specified form, however, in some generality (e.g. multiplicative heteroscedasticity or increasing variances) so that the tests may be applicable to a wide variety of situations. No test, however, considered a general composite hypothesis. In this paper we propose a test based on ordinary least square residuals (which form a set of maximal invariant under translation change) similar to a likelihood ratio test. The present test criterion is simpler than the likelihood ratio test because it does not require computation of maximum likelihood estimates of $\sigma_1^2, \dots, \sigma_k^2$ which may be a formidable task (see Harville (1977)). A two moment approximation

in terms of a chi square variable is also suggested, which facilitates the use of the present test. Extensive numerical studies are planned for the power comparisons of the present test with various other tests. The present test is expected to be more powerful in comparison to tests of Goldfeld and Quandt (1965), Theil's BLUS test (1971, pp. 214-218) and the test of Harvey & Phillips (1974) because it takes into account the additional structure that n_i -observations have the same variance and does not require deletion of any residuals.

The test is described in section 2 and the approximation to its null distribution is given in section 3. Section 4 gives the details of computations for a simple model considered in Chaubey (1978).

2. THE TEST CRITERION

The form of the likelihood ratio test in the present context is given by

$$\Lambda = \prod_{i=1}^k (\hat{\sigma}_i^2)^{n_i/2} / (s^2)^{n/2} \quad (2.1)$$

where Λ is the likelihood ratio, $\hat{\sigma}_i^2$ is the maximum likelihood estimate of σ_i^2 and s^2 is the maximum likelihood estimate of σ^2 , the common value of σ_i^2 under H_0 . Since $\hat{\sigma}_i^2$'s are difficult to compute we will substitute

$$\hat{\sigma}_i^2 = \frac{1}{n_i} e_i' e_i \quad (2.2)$$

(see Rao and Chaubey (1978) for a justification) where $e_i = Y_i - X_i \hat{\beta}$ and $\hat{\beta}$ is the ordinary least squares estimator of β by considering all the k models as

$$Y = X\beta + \epsilon \quad (2.3)$$

where $Y = [Y_1' \dots Y_k']'$, $X = [X_1' \dots X_k']'$ and $\epsilon = [\epsilon_1' \dots \epsilon_k']'$, namely,

$$\hat{\beta} = (X'X)^{-1}X'Y. \quad (2.4)$$

It is to be noted that the above test is analogous to Bartlett's test of homogeneity of variances when all the means are equal and it depends on a set of maximal invariants $(e_1'e_1/e'e, \dots, e_k'e_k/e'e)$ where $e = Y - X\hat{\beta}$ under location change and orthogonal transformations. If $\hat{\sigma}_i^2$ could be assumed to be close to the corresponding maximum likelihood estimator, then, $-2 \log \Lambda$ could be approximated by a chi-square random variable. However, much care is needed in the present case.

Note that $s^2 = n^{-1}e'e$ and since

$$-2 \ln \Lambda = - \sum n_i \ln(e'_i e_i / e' e) + \sum n_i \ln(n_i / n) \quad (2.5)$$

we consider the test statistic

$$T = - \sum n_i \ln(e'_i e_i / e' e) \quad (2.6)$$

Large values of T lead to rejection of H_0 . A 2-moment chi-square approximation is given in the next section.

3. APPROXIMATION FOR THE DISTRIBUTION OF T

We may approximate T by a multiple of chi-square variable, because $e'_i e_i / e' e$ behaves like a beta variable and the negative logarithm of a beta variable behaves like a chi-square variable in the light of the observation made by Wise (1950).

Thus letting $T \sim a \chi_v^2$ we get

$$a = V(T)/2E(T), \quad v = 2[E(T)]^2/V(T). \quad (3.1)$$

To find $E(T)$ and $V(T)$ we consider the joint raw moments μ'_{st} of $Z_i = \ln(r_i/m_i)$ $Z_j = \ln(r_j/m_j)$ where $r_i = e'_i e_i / e' e$ and $m_i = E[r_i]$. We approximately have

$$\mu'_{st} = \sum_{r=0}^{\infty} \sum_{k=0}^{\infty} a_r(s) a_k(t) m_1^{-(2r+s)} m_2^{-(2k+t)} E(e_1^{2r+s} e_2^{2k+t}) \quad (3.2)$$

where $a_r(s) = s(s-1) \cdots (s-r+1)/r!$ and $a_0(s) = 1$ for all s . From (3.2) we obtain

$$\begin{aligned} \mu'_{10} &= m_i^{-3} m_{30}^{ij} + \dots, & \mu'_{01} &= m_j^{-3} m_{03}^{ij} + \dots, \\ \mu'_{20} &= m_i^{-2} m_{20}^{ij} + \dots, & \mu'_{02} &= m_j^{-2} m_{02}^{ij} + \dots, \\ \mu'_{11} &= m_i^{-1} m_j^{-1} m_{11}^{ij} + \dots \end{aligned} \quad (3.3)$$

where $m_{st}^{ij} = E[(r_i - m_i)^s (r_j - m_j)^t]$. From (3.3) and known relations between raw and central moments we get

$$\begin{aligned} E[\ln r_i] &= \ln m_i + m_i^{-3} m_{30}^{ij} + \dots \\ V[\ln r_i] &= m_i^{-2} m_{20}^{ij} + \dots \\ \text{Cov}(\ln r_i, \ln r_j) &= m_i^{-1} m_j^{-1} m_{11}^{ij} + \dots \end{aligned} \quad (3.4)$$

Thus, approximately,

$$E(T) = - \sum_i n_i \ln m_i$$

$$V(T) = \sum_i n_i^2 m_i^{-2} m_{20}^{ii} + \sum_{i \neq j} n_i n_j m_i^{-1} m_j^{-1} m_{11}^{ij} \quad (3.5)$$

Hence, a and v in (3.1) can be approximately obtained, once we know m_i , m_{20}^{ii} and m_{11}^{ij} . These, under the null hypothesis can be obtained easily by known results about the mean, variance and covariance of quadratic forms under normal variables and Geary's (1933) result that r_i is independent of its denominator. We obtain

$$m_i = \text{tr } Q_{ii} / (n - p)$$

$$m_{20}^{ii} = 2[\text{tr } Q_{ij}^2 - (n - p)^{-1}(\text{tr } Q_{ii})^2] / (n - p)(n - p + 2) \quad (3.6)$$

$$m_{11}^{ij} = -2\text{tr } Q_{ii} \text{tr } Q_{jj} / (n - p)^2 (n - p + 2)$$

where $Q_{ii} = I_{n_i} - X_i (X'X)^{-1} X_i'$.

4. A PARTICULAR MODEL

Consider the case of comparing 2 normal populations with common mean. Suppose that independent samples of size n_1 and n_2 are obtained from $N(\mu, \sigma_1^2)$, $N(\mu, \sigma_2^2)$, then for testing $H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$, T takes form

$$T = - \sum_i n_i \ln \sum_j (Y_{ij} - \bar{Y})^2 + n \ln \sum_i \sum_j (Y_{ij} - \bar{Y})^2 \quad (4.1)$$

where Y_{ij} is the j -th observation from i -th population ($i = 1, 2$, $j = 1, \dots, n_i$) and \bar{Y} is the usual sample mean. Using the method of section 3 we obtain that

$$E(T) = - \sum_{i=1}^2 n_i \ln(n_i/n), \quad V(T) = 2(n - 3)n_1 n_2 / (n - 1)(n + 1) \quad (4.2)$$

The case of k normal populations can be similarly carried out. It would be interesting to see the behavior of T as compared to the Bartlett's test statistic.

5. SOME APPLICATIONS

A well known problem in which the present method could be applied is in assessing the accuracy of k instruments or assuming that all the instruments are equally

precise, in assessing the accuracy of k observers. The present paper could also be used in testing the effectiveness of various psychometric tests which are supposed to be standardized.

An important area, in which the test mentioned in this paper applies, is time series cross section data. This particular aspect may be utilized in climatological studies to compare the heterogeneity of various variables over a given set of strata.

ACKNOWLEDGEMENT

This research is partially supported by a grant from NSERC of CANADA (A3661)

REFERENCES

- Chaubey, Y.P., 1978. Testing the equality of variances of two linear models. Proc. Amer. Statist. Assoc., Bus. Econ. 413-418.
- Geary, R.C., 1933. A general expression for the moments of certain symmetrical functions of normal samples. Biometrika, 25: 184-186.
- Goldfeld, S.M. and Quandt, R.E., 1965. Some tests of homoscedasticity. J. Amer. Statist. Assoc., 60: 539-547.
- Harrison, M.J. and McCabe, B.P.M., 1979. A test for heteroscedasticity based on ordinary least square residuals. J. Amer. Statist. Assoc., 74: 494-499.
- Harvey, A.C. and Phillips, G.D.A., 1974. A comparison of the power of some tests for heteroscedasticity in the general linear model. J. Econometrics, 2: 307-316.
- Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Statist. Assoc., 72: 320-338.
- Rao, P.S.R.S. and Chaubey, Y.P., 1978. Three modifications of the principle of the MINQUE. Comm. Stat. Theor. Meth., A7(8): 767-778.
- Theil, H., 1971. Principles of Econometrics. North Holland Publishing Co: Amsterdam.
- Wise, M.E., 1950. The incomplete beta function as a contour integral and a quickly converging series for its inverse. Biometrika, 37: 208-218.

SOME PROPERTIES OF GENERALIZED RIDGE ESTIMATORS IN LINEAR REGRESSION MODELS

T.D.DWIVEDI¹, J.M.LOWERRE² and V.K.SRIVASTAVA³

1. Dept. Math., Concordia Univ., Montreal, Quebec (Canada)

2. Clarkson Coll. Tech., Potsdam, New York

3. Dept. Math., Banaras Hindu Univ. (India)

ABSTRACT

Dwivedi, T.D., Lowerre, J.M. and Srivastava, V.K. Some properties of generalized ridge estimators in linear regression models. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

A number of properties and criticisms of biased estimators are detailed in this paper. The ridge estimators are compared with other biased estimators which have been proposed to counter the effects of an ill-conditioned $X'X$ matrix.

In addition some new estimators are proposed and their properties are studied in detail.

1. INTRODUCTION

Consider the linear regression model

$$Y = X\beta + u \quad (1.1)$$

where Y is a $n \times 1$ vector of observations on the variable to be explained, X is a $n \times p$ matrix, with full column rank, of n observations on p explanatory variables, β is a $p \times 1$ vector of p unknown coefficients associated with them and u is a $n \times 1$ vector of n unobservable disturbances with $E(u) = 0$ and $E(uu') = \sigma^2 I_n$.

It is well known that the best linear unbiased estimator of β is

$$b = (X'X)^{-1}X'Y \quad (1.2)$$

and

$$E(b-\beta)'(b-\beta) = \sigma^2 \text{tr}(X'X)^{-1} = \sigma^2 \sum_{i=1}^p 1/\lambda_i \quad (1.3)$$

where $\lambda_1, \lambda_2, \dots, \lambda_p$ are the eigen values of $X'X$.

Obviously, when some of the λ_i 's are small, the quantity defined by (1.3) is large - a situation with unpleasant prospects if the effect of an erroneous estimate is costly. To tackle this problem, Hoerl and Kennard (1970) advocated the method of ridge regression estimation. Adopting their philosophy, Dwivedi (1974) and Goldstein and Smith (1974) proposed a class of estimators characterized by a pair of scalars. The objective of this paper is to analyze the properties of this class

of estimators. The expressions for bias and total mean squared error are derived and studied. The concept of primary component is introduced and its relevance is examined. Then the results of a simple Monte Carlo experiment are reported and finally some remarks are placed.

2. ESTIMATORS AND THEIR PROPERTIES

For estimating β in (1.1), Dwivedi (1974) and Goldstein and Smith (1974) proposed a class of biased estimators:

$$\hat{\beta}_q = [kI_p + (X'X)^q]^{-1} (X'X)^{q-1} X'Y \quad (2.1)$$

where k and q are scalars characterizing the estimator.

When $q = 1$, we obtain the estimator suggested by Hoerl and Kennard (1970).

We can express $\hat{\beta}_q$ as

$$\hat{\beta}_q = C b \quad (2.2)$$

where

$$C = [I_p + k(X'X)^{-q}]^{-1}. \quad (2.3)$$

It is easy to see that

$$E(\hat{\beta}_q - \beta) = (C - I_p)\beta. \quad (2.4)$$

Similarly, the total mean squared error (TMSE) of $\hat{\beta}_q$ is

$$\begin{aligned} \text{TMSE}(\hat{\beta}_q) &= E(\hat{\beta}_q - \beta)'(\hat{\beta}_q - \beta) \\ &= E(b - \beta)'C'C(b - \beta) + \beta'(C - I_p)'(C - I_p)\beta \\ &= \sigma^2 \text{tr}(X'X)^{-1}C^2 + \beta'(C - I_p)^2\beta \\ &= \sigma^2 \text{tr}(X'X)^{-1} [I_p + k(X'X)^{-q}]^{-2} + k^2\beta'[kI_p + (X'X)^q]^{-2}\beta \end{aligned} \quad (2.5)$$

where we have used the relation

$$C = I_p - k[kI_p + (X'X)^q]^{-1}. \quad (2.6)$$

Let P be an orthogonal matrix which diagonalizes $X'X$. Since C is symmetric and commutes with $X'X$, the same orthogonal matrix, P , will diagonalize $(X'X)^{-1}C^2$. Thus from (2.5) we find

$$\text{TMSE}(\hat{\beta}_q) = \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i (1 + \frac{k}{\lambda_i^q})^2} + k^2 \sum_{i=1}^p \frac{\alpha_i^2}{(\lambda_i^q + k)^2} \quad (2.7)$$

$$= \sigma^2 \sum_{i=1}^p \frac{\lambda_i^{q-1}}{(\lambda_i^q + k)} + k \sum_{i=1}^p \frac{(k\alpha_i^2 - \sigma^2 \lambda_i^{q-1})}{(\lambda_i^q + k)^2}$$

where $\alpha = P\beta = (\alpha_1, \alpha_2, \dots, \alpha_p)'$. Using the relation

$$\frac{1}{\lambda_i} - \frac{\lambda_i^{q-1}}{(\lambda_i^q + k)} = \frac{k}{\lambda_i (\lambda_i^q + k)} > 0 \quad \text{for } k > 0 \quad (2.8)$$

it follows from (2.7) that

$$TMSE(\hat{\beta}_q) < \sigma^2 \sum_{i=1}^p \frac{1}{\lambda_i} + k \sum_{i=1}^p \frac{(k\alpha_i^2 - \sigma^2 \lambda_i^{q-1})}{(\lambda_i^q + k)^2} \quad (2.9)$$

Hence, from (1.3) we get

$$TMSE(\hat{\beta}_q) < E(b - \beta)'(b - \beta) = TMSE(b) \quad (2.10)$$

if $k > 0$ and

$$\sum_{i=1}^p \frac{(k\alpha_i^2 - \sigma^2 \lambda_i^{q-1})}{(\lambda_i^q + k)^2} < 0, \quad (q \geq 1). \quad (2.11)$$

The inequality (2.11) holds if

$$k \alpha_i^2 < \sigma^2 \lambda_i^{q-1}, \quad (q \geq 1; i = 1, 2, \dots, p) \quad (2.12)$$

which will be satisfied as long as

$$k < \sigma^2 \frac{\lambda_{\min}^{q-1}}{\alpha_{\max}^2}, \quad (q \geq 1) \quad (2.13)$$

where

$$\lambda_{\min} = \text{minimum}(\lambda_1, \lambda_2, \dots, \lambda_p), \quad \alpha_{\max} = \text{maximum}(\alpha_1, \alpha_2, \dots, \alpha_p). \quad (2.14)$$

Thus we have the following result:

Theorem I. For any given k such that

$$0 < k < \frac{\sigma^2 \lambda_{\min}^{q-1}}{\alpha_{\max}^2} \quad (q \geq 1), \quad (2.15)$$

the estimator $\hat{\beta}_q$ given by (2.1) is better than the least square estimator b according to total mean squares error criterion.

Setting $q = 1$ in (2.15) yields the result obtained by Hoerl and Kennard (1970).

It is interesting to note that as q increases the range of k specified by (2.15) narrows. It may be pointed out that our derivation of the condition (2.15) does not require $X'X$ to be necessarily in the form of a correlation matrix as assumed by Hoerl and Kennard.

From (1.3) and (2.7), we observe that the TMSE of each of the estimators b and

$\hat{\beta}_q$ is primarily affected by smallness of eigenvalues of $X'X$. It, therefore, appears to be reasonable to concentrate on that part of the TMSE, while comparing two estimators, which is most affected. We do it by choosing a cut-off number depending upon the magnitude of the eigenvalues. Let it be 1. Although we recognize that this choice is arbitrary and one could very well take a number less than 1, yet, such a choice includes, with certainty, all the terms that are likely to have a larger impact on TMSE owing to the smallness of λ_i 's. For this purpose we define the primary component of an estimator as the part of the mean squared error that involves terms in which the eigenvalues are equal to or less than 1. Thus in order to compare two biased estimators, we may confine our attention to the comparison of their primary components.

Let us assume now that the eigen values of $X'X$ have been ordered such that

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_t \leq 1 \leq \lambda_{t+1} \leq \dots \leq \lambda_p. \quad (2.16)$$

Thus from (2.7) the primary component (PC) of $\hat{\beta}_q$ is given by

$$PC(\hat{\beta}_q) = \sigma^2 \sum_{i=1}^t \frac{\lambda_i^{2q-1}}{(\lambda_i^q + k)^2} + k^2 \sum_{i=1}^t \frac{\alpha_i^2}{(\lambda_i^q + k)^2}. \quad (2.17)$$

For two estimators $\hat{\beta}_q$ and $\hat{\beta}_{q+1}$, it is observed that

$$PC(\hat{\beta}_{q+1}) < PC(\hat{\beta}_q) \quad (2.18)$$

when

$$\sigma^2 \sum_{i=1}^t \left[\frac{\lambda_i^{2q+1}}{(\lambda_i^{q+1} + k)^2} - \frac{\lambda_i^{2q-1}}{(\lambda_i^q + k)^2} \right] + k^2 \sum_{i=1}^t \alpha_i^2 \left[\frac{1}{(\lambda_i^{q+1} + k)^2} - \frac{1}{(\lambda_i^q + k)^2} \right] < 0. \quad (2.19)$$

Consider the following quantity

$$\begin{aligned} & \sigma^2 \left[\frac{\lambda_i^{2q+1}}{(\lambda_i^{q+1} + k)^2} - \frac{\lambda_i^{2q-1}}{(\lambda_i^q + k)^2} \right] + k^2 \alpha_i^2 \left[\frac{1}{(\lambda_i^{q+1} + k)^2} - \frac{1}{(\lambda_i^q + k)^2} \right] \\ &= \frac{\lambda_i^q (1 - \lambda_i) k}{(\lambda_i^q + k)^2 (\lambda_i^{q+1} + k)^2} \left[2\alpha_i^2 k^2 + \lambda_i^q (1 + \lambda_i) \left(\alpha_i^2 - \frac{\sigma^2}{\lambda_i} \right) k - 2\lambda_i^{2q} \sigma^2 \right] \end{aligned} \quad (2.20)$$

which is negative for $1 - \lambda_i > 0$ and $k > 0$ when

$$k < g_i \quad (2.21)$$

where

$$g_i = - \frac{\lambda_i^{q-1} (1 + \lambda_i) (\lambda_i \alpha_i - \sigma^2)}{4\alpha_i^2} + \left[\frac{4\lambda_i^{2q} \alpha_i^2 \sigma^2 + \lambda_i^{2q-2} (1 + \lambda_i)^2 (\lambda_i \alpha_i^2 - \sigma^2)^2}{16\alpha_i^4} \right]^{1/2} \quad (2.22)$$

Thus the inequality (2.19) will be satisfied when

$$0 < k < g_i \quad \text{for all } i \ (i = 1, 2, \dots, t) . \quad (2.23)$$

Writing $g_{\min} = \text{minimum } (g_1, g_2, \dots, g_t)$, the conditions (2.23) will hold so long as

$$0 < k < g_{\min} . \quad (2.24)$$

This leads to the following result:

Theorem II. The contribution of primary component of $\hat{\beta}_{q+1}$ will be smaller than that of $\hat{\beta}_q$ ($q \geq 1$) when k satisfies the condition (2.24).

From Theorem II, it follows that larger the value of q , smaller the contribution of the primary component to TMSE provided k satisfies the condition (2.24). Further, the value of g_{\min} tends to zero as q increases. In other words, the value of k will be very close to zero and then the estimator $\hat{\beta}_q$ will tend to $b = (X'X)^{-1}X'Y$, the least squares estimator.

3. A MONTE CARLO STUDY

From the previous section, it may be observed that the choice of k and q depends upon unknown parameters β and σ^2 . We have therefore carried out a Monte Carlo experiment in order to examine the behaviour of $\hat{\beta}_q$. For simplicity sake, only two cases $\hat{\beta}_1$ and $\hat{\beta}_2$ are considered.

Let us write

$$\begin{aligned} SV(\hat{\beta}_q) &= E[\hat{\beta}_q - E(\hat{\beta}_q)]' [\hat{\beta}_q - E(\hat{\beta}_q)] \\ SSB(\hat{\beta}_q) &= [E(\hat{\beta}_q) - \beta]' [E(\hat{\beta}_q) - \beta] \end{aligned} \quad (3.1)$$

so that

$$TMSE(\hat{\beta}_q) = SV(\hat{\beta}_q) + SSB(\hat{\beta}_q) . \quad (3.2)$$

From (2.8), we see that $SV(\hat{\beta}_q) < SV(b)$ for $k > 0$ whatever be the values of β and σ^2 . Since b is unbiased, $SSB(b) = 0$ and hence $TMSE(\hat{\beta}_q)$ could be larger than $TMSE(b) \equiv SV(b)$ for some β . In fact, $TMSE(\hat{\beta}_q)$ will be largest for that β which maximizes $SSB(\hat{\beta}_q)$. Considering the extreme case for simulation purpose, we choose β that maximizes $SSB(\hat{\beta}_q)$ subject to $\beta'\beta = 1$.

Now, we have from Theorem I

$$TMSE(\hat{\beta}_1) < TMSE(b) \quad \text{if } 0 < k < \sigma^2 / \alpha_{\max}^2 = K_1 \text{ (say)} \quad (3.3)$$

$$TMSE(\hat{\beta}_2) < TMSE(b) \quad \text{if } 0 < k < \sigma^2 \lambda_{\min} / \alpha_{\max}^2 = K_2 \text{ (say)} . \quad (3.4)$$

Lowerre (1974) demonstrated that the mean squared error of $\hat{\beta}_1$ will be component-wisely less than that of b when

$$0 < k < \min_{(i,m)} \left\{ \frac{\sigma^2 p_{im}}{\sum_{j=1}^p p_{jm} \beta_j} ; (p_{im})^2 \neq 0, \sum_{j=1}^p p_{jm} \beta_j \neq 0 \right\} = K_3 \text{ (say)} \quad (3.5)$$

where p_{im} is the (i,m) -element of the orthogonal matrix P that diagonalizes $X'X$.

Further, from Theorem II we observe that $PC(\hat{\beta}_2) < PC(\hat{\beta}_1)$ if $0 < k < K_4$ where K_4 is given by

$$K_4 = \min_i \left\{ - \frac{(1+\lambda_i)(\lambda_i \alpha_i^2 - \sigma^2)}{4\alpha_i^2} + \left[\frac{4\lambda_i^2 \alpha_i^2 \sigma^2 + (1+\lambda_i)^2 (\lambda_i \alpha_i^2 - \sigma^2)^2}{16\alpha_i^4} \right]^{1/2} \right\} \quad (3.6)$$

In practice, K_i 's cannot be calculated since they require the knowledge of $\alpha = P\beta$ and σ^2 . As an approximation, one may replace these quantities by their unbiased estimators $\hat{\alpha} = Pb$ and $\hat{\sigma}^2$ based on least squares estimator b of β and obtain \hat{K}_i 's. It is possible that the probability of \hat{K}_i exceeding K_i may be quite significant in some cases. A somewhat conservative method for estimating K_i 's could be as follows. Since $\alpha_{\max}^2 \leq \alpha'\alpha = \beta'\beta$ so that $1/\alpha_{\max}^2 \geq 1/\beta'\beta$, we may take $\hat{K}_1 = \hat{\sigma}^2 / \hat{\beta}'\hat{\beta}$ and $\hat{K}_2 = \hat{\sigma}^2 \lambda_{\min} / \hat{\beta}'\hat{\beta}$ which will have a fairly high probability of not exceeding K_1 and K_2 respectively, and at the same time are easily computable from the data. Clearly, $\hat{\sigma}^2 / \hat{\beta}'\hat{\beta}$ provides an estimate of the ratio $\sigma^2 / \beta'\beta$ which determines the extent of the bias that the system can tolerate without increasing the mean squared error. Similarly, we propose to replace α_i^2 in K_4 by $\hat{\beta}'\hat{\beta}$ to obtain \hat{K}_4 since K_4 is a decreasing function of α_i^2 . However, such a treatment cannot be given to K_3 owing to the structure of its formula. It is, therefore, suggested to replace unknowns in K_3 by their estimates using the normal equations in order to obtain \hat{K}_3 .

We have chosen two set of numerical values for the matrix X with $n = 5$ and $p = 3$ such that for one set $X'X$ has two eigen values less than 1 while for the other set it has all the three eigenvalues less than 1. The method of singular valued decomposition has been employed to construct X which comprises expressing X as $X = U\Lambda D$ where the columns of $n \times p$ matrix U are orthogonal, Λ is a $p \times p$ diagonal matrix and D is a $p \times p$ orthogonal matrix. The eigenvalues of $X'X$ are thus the square of the elements on the diagonal of Λ . We can then select three non-collinear vectors and orthogonalize them to form the matrix U . The matrix Λ can be appropriately selected to yield the desired eigen values.

The following four models are considered (for detailed specification see Table 1, I - IV):

	No. of eigen values less than 1	value of σ^2
Model I	2	4.00
Model II	3	4.00
Model III	2	0.04
Model IV	3	0.04

For each X , we draw a sample of size 100 from a multivariate normal population $N(X\beta, \sigma^2 I_n)$ for each specific value of σ^2 where β is determined such that $X'X\beta = \lambda_{\min} \beta$. Next, K_i 's are computed from (3.3), (3.4), (3.5) and (3.6). For each model, the average total mean squared error (ATMSE) and component-wise average mean squared error (CAMSE), the average being taken over 100 data vectors, of $\hat{\beta}_1$ and $\hat{\beta}_2$ are computed for various values of k ranging between 0 and K_i ($i = 1, 2, 3, 4$). A part of the results are shown in Table 2, I-IV. Then we set $k = \hat{K}_i$ in $\hat{\beta}_q$ ($q = 1, 2$) and calculate ATMSE and CAMSE for each model. These values have also been put on the lower line of each plot in the corresponding tables. The values of ATMSE and CAMSE for b are given in Table 1, I-IV.

An observation emerging from a study of the results is that ATMSE for both $\hat{\beta}_1$ and $\hat{\beta}_2$ is smaller than that of b over the range $0 \leq k \leq K_i$ ($i = 1, 2, 3, 4$). The same remains true even when K_i 's are estimated from the data. Similar is the case when a component-wise comparison is made of $\hat{\beta}_1$ and $\hat{\beta}_2$ with b . Further, it is observed that when all the eigen values of $X'X$ are less than 1, ATMSE of $\hat{\beta}_2$ is smaller than that of $\hat{\beta}_1$ over the admissible range of k . This is expected because TMSE is then identical with PC. A component-wise comparison also reveals that $\hat{\beta}_2$ is better than $\hat{\beta}_1$. However when only two eigen values are less than 1, ATMSE of $\hat{\beta}_2$ is smaller than that of $\hat{\beta}_1$ only if k is close to zero rather than to its upper bound. Substantial reduction is observed in all cases when the eigen values of $X'X$ are near to zero.

TABLE 1 , I (Model I)

$$X = \begin{pmatrix} .26546871 & .20709179 & .09294127 \\ .08699376 & -.38552773 & .46523600 \\ .31881716 & .18759224 & .16563288 \\ -.27648435 & -.29604647 & -.28391225 \\ .14186256 & .30173927 & -.25698986 \end{pmatrix}$$

Eigen values of $X'X$: (.57122921, .46219351, .01166019)

$$\beta' = (.67998685, -.53883530, -.48726693)$$

$$K_1 = 5.47716, K_2 = .06386, K_3 = .00374, K_4 = 3.74455$$

b	CAMSE	
First component	158.58493484	
Second component	96.51692759	ATMSE(b) = 343.03600337
Third component	87.93414093	

TABLE 1 , II (Model II)

$$X = \begin{pmatrix} .54472612 & .18926693 & .14683747 \\ .75709678 & -.15259512 & .08371406 \\ .53424734 & -.22413191 & .42268861 \\ .65087249 & -.44727052 & .31428611 \\ .52563128 & .22005858 & .41257239 \end{pmatrix}$$

Eigen values of $X'X$: (2.27291581, .30016923, .11622626)

$$\beta' = (-.42067037, -.05008295, .90583008)$$

$$K_1 = 6.93132, K_2 = .79498, K_3 = .02424, K_4 = 5.33420$$

<u>b</u>	<u>CAMSE</u>	
First component	8.43040920	
Second component	16.19485845	ATMSE(b) = 55.39227965
Third component	30.76701200	

TABLE 1, III (Model III)

$$X = \begin{pmatrix} .24891380 & -.24301846 & -.04039797 \\ .04540124 & .03156843 & .02590564 \\ .11213618 & -.10427252 & .01877457 \\ -.00895034 & .15426231 & -.02009906 \\ -.13908174 & .14568880 & .12652889 \end{pmatrix}$$

Eigen values of $X'X$: (.20513791, .01487921, .01102023)

$$\beta' = (.59804198, .36309580, .71449789)$$

$$K_1 = .04869, K_2 = .00054, K_3 = .01016, K_4 = .01731$$

<u>b</u>	<u>CAMSE</u>	
First component	1.92408687	
Second component	1.44546292	ATMSE(b) = 6.72598076
Third component	3.35640960	

TABLE 1, IV (Model IV)

$$X = \begin{pmatrix} .34864813 & -.30884291 & .10150767 \\ .50455832 & -.82061450 & -.63925233 \\ .37677604 & -.14154010 & .03074313 \\ .35251767 & -.56615090 & .13622661 \\ .32720439 & -.43233421 & .40706782 \end{pmatrix}$$

Eigen values of $X'X$: (2.02143737, .58988911, .05854955)

$$\beta' = (.78579218, .60556401, -.12578904)$$

$$K_1 = .07268, K_2 = .00426, K_3 = .00057, K_4 = .01247$$

<u>b</u>	<u>CAMSE</u>	
First component	.50219166	
Second component	.27130605	ATMSE(b) = .84678850
Third component	.07329076	

TABLE 2, I (Model I)

			$K_1 = 5.47716494$ (\hat{K}_1)	$K_2 = .06386478$ (\hat{K}_2)	$K_3 = .00364031$ (\hat{K}_3)	$K_4 = 3.74454611$ (\hat{K}_4)
ATMSE	$\hat{\beta}_1$		1.12507957 (202.83289433)	22.25003168 (338.27084401)	203.58135974 (342.70376957)	1.25247583 (277.18249100)
	$\hat{\beta}_2$		1.04046600 (57.66869892)	12.52573313 (204.64599871)	18.61359031 (331.13841616)	1.08275258 (92.99364875)
CAMSE	$\hat{\beta}_1$	1st	.48293412 (53.75170440)	8.10960925 (156.25272976)	93.34650021 (158.56647805)	.51297601 (126.803283)
		comp.	.37252900 (19.30884587)	7.05485667 (95.22779159)	57.11754356 (96.27751121)	.43626463 (78.53659289)
		2nd	.26960747 (91.41249101)	7.08556577 (86.79032265)	53.12730798 (87.85978030)	.30323519 (71.84261502)
		3rd				
		comp.				
	$\hat{\beta}_2$	1st.	.46449035 (5.49654940)	3.59010085 (92.01487526)	5.52104852 (154.09040397)	.47198581 (38.31262694)
		comp.	.32783245 (6.93780132)	4.88990263 (58.20168148)	6.59183694 (92.12401033)	.35570136 (26.06664074)
		2nd	.24814320 (6.87449515)	4.04562965 (54.42944197)	6.50070485 (84.92400186)	.25506541 (28.61438109)
		3rd				
		comp.				

TABLE 2, II (Model II)

			$K_1 = 6.83131924$ (\hat{K}_1)	$K_2 = .79397870$ (\hat{K}_2)	$K_3 = .02423727$ (\hat{K}_3)	$K_4 = 5.33419626$ (\hat{K}_4)
ATMSE	$\hat{\beta}_1$		1.12247945 (39.31714106)	3.61098149 (51.77196672)	41.30518691 (55.25744975)	1.18706403 (48.74156370)
	$\hat{\beta}_2$		1.33498227 (19.8829523)	2.50944747 (41.5613962)	17.12596213 (54.5065404)	1.43834198 (39.6558570)
CAMSE	$\hat{\beta}_1$	1st	.23170805 (6.13600249)	.97014942 (7.90088887)	6.23108849 (8.40269320)	.26438087 (7.41130945)
		comp.	.03990220 (11.67527136)	1.30423146 (15.24787101)	13.89199369 (16.15755545)	.06072269 (14.12752927)
		2nd	.85086820 (21.50586721)	1.33660060 (28.62320684)	21.18210473 (30.69720110)	.86196047 (27.20272949)
		3rd				
		comp.				
	$\hat{\beta}_2$	1st	.38705160 (3.20079369)	1.13934081 (6.26221833)	2.39570072 (8.23101463)	.46067084 (6.04552801)
		comp.	.01905048 (7.91272041)	.24686063 (13.75332149)	10.14086614 (16.06886835)	.02507132 (12.21896890)
		2nd	.92888188 (8.76941112)	1.1232460 (21.54585640)	4.58939526 (30.20665746)	.95259982 (21.39136015)
		3rd				
		comp.				

TABLE 2, III (Model III)

			$K_1 = .04869449$ (\hat{K}_1)	$K_2 = .00053662$ (\hat{K}_2)	$K_3 = .01015889$ (\hat{K}_3)	$K_4 = .01731486$ (\hat{K}_4)
ATMSE	$\hat{\beta}_1$		1.14009929 (4.55645767)	6.21895382 (6.66878235)	2.52337218 (1.99530175)	1.81137911 (6.22491024)
	$\hat{\beta}_2$		1.03869239 (1.46702207)	1.29361439 (4.73091244)	1.11167468 (1.06957139)	1.09801626 (3.86639216)
CAMSE	$\hat{\beta}_1$	1st	.37497688 (1.38045513)	1.77799825 (1.91058115)	.74115616 (.59063595)	.54978491 (1.92176302)
		comp.	.24613978 (1.01472415)	1.34747835 (1.43423951)	.58859981 (.40347954)	.42517389 (1.33010360)
		2nd	.51898271 (2.16127839)	3.09347733 (3.32396169)	1.19361621 (1.00119626)	.83642031 (3.07304361)
		3rd				
		comp.				

TABLE 2, III (cont'd)

	$\hat{\beta}_2$	1st comp. (.49732713)	.35694737 (.49732713)	.41764947 (1.41424850)	.37724036 (.36932507)	.37022506 (1.09008069)
		2nd comp. (.28762345)	.16616535 (.28762345)	.31205135 (1.07384657)	.21920089 (.18742376)	.20199587 (.82106256)
		3rd comp. (.68207129)	.51557967 (.68207129)	.56391356 (2.24281738)	.5152334 (.5128225)	.5167943 (1.9552489)

TABLE 2, IV (Model IV)

			$K_1 = .07207756$ (K_1)	$K_2 = .00425524$ (K_2)	$K_3 = .00065733$ (K_3)	$K_4 = .01247174$ (K_4)
ATMSE	$\hat{\beta}_1$.57957099 (.82959764)	.76370943 (.83094092)	.84316009 (.78689404)	.66062924 (.79508745)
	$\hat{\beta}_2$.98629497 (.76649574)	.59404971 (.77931162)	.69118144 (.76001745)	.76773622 (.78408929)
		1st comp. (.46180563)	.34123943 (.46180563)	.95098703 (.49319188)	.49489009 (.47829969)	.38743250 (.47546367)
CAMSE	$\hat{\beta}_1$	2nd comp. (.24411326)	.17601998 (.24411326)	.24086543 (.26577177)	.26667591 (.24440164)	.20321432 (.25527115)
		3rd comp. (.06067685)	.06231158 (.06067685)	.07194697 (.07197727)	.07309409 (.06419272)	.06998241 (.06485263)
		1st comp. (.50487290)	.59561077 (.50487290)	.34288262 (.46232007)	.40513744 (.46723988)	.45038870 (.47131889)
		2nd comp. (.2641137)	.32546166 (.2641137)	.17621887 (.24433088)	.21455004 (.22804595)	.23980572 (.24829852)
	$\hat{\beta}_2$	3rd comp. (.06061157)	.06522255 (.06061157)	.07494882 (.07266068)	.07249396 (.06473162)	.07754180 (.06447188)

4. SOME REMARKS

We have considered a class of biased estimators for β characterized by a pair of scalars k and q . It is demonstrated that for $q \geq 1$ and $k > 0$, we can always pick up an estimator from this class which will be better than the least squares estimator according to total mean squared error (TMSE) criterion. It is seen that the range of k over which $\hat{\beta}_q$ will be better, narrows as q increases. The concept of the primary component (PC) of an estimator is introduced which actually refers to that part of TMSE that is attributable to small eigen values of $X'X$. Effect of change in q on the PC is examined and it is found that as q grows large, PC of the corresponding estimator decreases over a specific range of k . Again, it is observed that the upper bound of k moves to zero as q increases. However, a very value of q may not be desirable from computation point of view, for then the computer may be out of its bits. These considerations pose an interesting problem pertaining to the optimum choice of k and q . One simple solution could be to choose that value of k and q for which the quantity $(\hat{\beta}_q - b)'(\hat{\beta}_q - b)$ is maximum. Few other proposals can be forwarded but their performance is yet to be investigated.

We have carried out a simple Monte Carlo experiment. A more extensive study is required with other values of n and p , and better methods need be evolved to estimate K_i 's - the bounds. The scalar q should be assigned values greater than 2. Investigations incorporating these suggestions will probably throw more light on the properties of the biased estimators. This is an area which requires a further careful expolation.

ACKNOWLEDGEMENT

This work was partially supported by NRC Grant No. A3989.

REFERENCES

- Dwivedi, T.D., 1974. Biased estimation in regression models. Presented at the Western Regional Meeting, Edmonton, Alta., August 1974. Abstract appeared in IMS Bull. 3: 158.
- Goldstein, M. and Smith, A.F.M., 1974. Ridge-type estimators for regression analysis. JRSS, B 36:284-291.
- Hoerl, A.E. and Kennard, R.W., 1970. Ridge-regression: Biased estimation for non-orthogonal problems. Technometrics 12:55-67.
- Lowerre, J.M., 1974. On the mean square error of parameter estimates for some biased estimators. Technometrics 16:461-464.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

SELECTION OF THE NUMBER OF REGRESSION PARAMETERS IN SMALL SAMPLE CASES

R. SHIBATA

Dept. Math., Tokyo Inst. of Technology, Tokyo (Japan)

ABSTRACT

Shibata, R., Selection of the number of regression parameters in small sample cases.
Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

The author proposes an asymptotically optimal selection of the number of parameters in autoregressive process and in multiple linear regression. The method is asymptotically equivalent to Akaike's Information Criterion and Mallows' C_p method. In this paper, we report how the asymptotics hold true in relatively small sample cases. Many results by computer simulation are shown.

INTRODUCTION

As a first step to statistical model fitting, the selection of the number of parameters plays an important role in many practical situations. If the structure of the population is well known a priori and it remains only to know the value of some parameters, there might be no need of the selection. Otherwise, the selection of the number of parameters may improve the efficiency of an estimation or a prediction. Various methods of the selection have been proposed and investigated (Allen, 1971; Akaike, 1973; Mallows, 1973; Hocking, 1976; Box & Jenkins, 1976; Bhansali & Downham, 1977; Schwarz, 1978). Also for the meteorological or climatological data, some of them have been applied to the case of fitting a Markov chain or a time series model or a multiple regression (Gates & Tong, 1976; Box & Jenkins, 1976; Ozaki, 1977). But it seems that the statistical optimality has not been so clear.

In the recent paper the author, while introducing the concept of efficient selection, proposed an asymptotically efficient selection for the case of an autoregressive model fitting in time series analysis (Shibata, 1980). The result also applies to the multiple regression analysis (Shibata, 1979). Furthermore it may also apply to the Markov chain. The purpose of the present paper is to show how the results hold true in small sample cases and to suggest some finite corrections in practical applications. As the results are very analogous, to simplify discussions we will restrict our attention to the selection in multiple regression. We first sketch the asymptotics, and will see the behavior of small sample cases from the results of computer simulations.

Let $f(x)$ be a regression function, then the observational equation is

$$y = f(x) + \varepsilon,$$

where ε is a sampling error or measurement error. Assume that $f(x)$ can be written as

$$f(x) = (\underline{x}, \underline{\beta})$$

where $f(x)$ is determined by a vector \underline{x} at the sampling point x and by the vector of parameters $\underline{\beta}$, both in the Hilbert space \mathcal{L}_2 of sequences of real numbers with the inner product (\cdot, \cdot) and the norm $\|\cdot\|$. The above representation can be justified if one considers, for example, a polynomial regression or a finite Fourier approximation to a smooth function $f(x)$. Given samples y_1, \dots, y_n at n sampling points x_1, \dots, x_n , we can obtain the least squares estimates $\hat{\underline{\beta}}(k) = (\hat{\beta}_1(k), \dots, \hat{\beta}_k(k), 0, \dots)'$ and $\hat{\sigma}^2(k)$ by fitting a model with k regression parameters,

$$y_\alpha = (\underline{x}_\alpha, \underline{\beta}(k)) + \varepsilon_\alpha, \quad \varepsilon_\alpha \sim N(0, \sigma^2), \quad \alpha = 1, \dots, n,$$

where $\underline{\beta}(k)$ belongs to a subspace $V(k) = \{(\beta_1(k), \dots, \beta_k(k), 0, \dots)'\}$ and $\underline{x}_\alpha = (x_{\alpha 1}, x_{\alpha 2}, \dots)'$. We will see what selection is optimal for estimating regression parameters or predicting future observations.

ASYMPTOTICS

Loss function

Consider the prediction of future observations $\underline{y}^* = (y_1^*, \dots, y_n^*)'$ at the same sampling points where y_1, \dots, y_n were taken. An estimated predictor is then given by

$$\hat{\underline{y}}(k) = \underline{x} \hat{\underline{\beta}}(k)$$

where

$$\underline{x} = \begin{pmatrix} \underline{x}'_1 \\ \vdots \\ \underline{x}'_n \end{pmatrix}$$

is a $n \times \infty$ matrix, which is also considered as a bounded linear operator on \mathcal{L}_2 .

If expectation is taken with respect to future observations, the mean squared error of prediction is

$$E^* \| \underline{\hat{y}}(k) - \underline{y}^* \|^2 = \| \underline{x\beta} - \underline{x\hat{\beta}}(k) \|^2 + n\sigma^2.$$

Because the last term of the right hand side is independent of the number k , we adopt as a loss function

$$L_n(\underline{\beta}, \underline{\hat{\beta}}(k)) = \| \underline{x\beta} - \underline{x\hat{\beta}}(k) \|^2.$$

This loss function is also the squared error of the estimated mean vector. Letting $\underline{\beta}^{(n)}(k)$ be the vector which minimizes $\| \underline{x\beta} - \underline{x\beta}(k) \|$ in $V(k)$, we can rewrite the loss function as

$$L_n(\underline{\beta}, \underline{\hat{\beta}}(k)) = \| \underline{x\beta} - \underline{x\beta}^{(n)}(k) \|^2 + \| \underline{x\beta}^{(n)}(k) - \underline{x\hat{\beta}}(k) \|^2.$$

The first term on the right hand side signifies the bias of the parameter estimation and the last represents the dispersion. For fixed n , the first term is a decreasing function of k , and the last is an increasing function. Therefore, the best selection is the one which balances those two terms. We note here that our main concern is not about the number k itself but about whether the loss $L_n(\underline{\beta}, \underline{\hat{\beta}}(k))$ is a minimum or not (See Schwarz, 1978).

Assumptions

Let a range of selection $1 \leq k \leq K_n$ be given a priori. We need the following assumptions for K_n and $L_n(k) = E(L_n(\underline{\beta}, \underline{\hat{\beta}}(k)))$.

(A1) For any k in $1 \leq k \leq K_n$, the principal submatrix $M(k)$ of the information matrix $M = X'X$ has full rank and K_n/n goes to zero as $n \rightarrow \infty$.

(A2) For any $0 < \delta < 1$,

$$\lim_{n \rightarrow \infty} \sum_{k=1}^{K_n} \delta^{L_n(k)} = 0.$$

The assumption (A1) is the one usually taken for assuring the uniqueness and the efficiency of the least squares estimate $\underline{\hat{\beta}}(k)$. The assumption (A2) is satisfied when $\{K_n\}$ is a divergent sequence and

$$\| \underline{x\beta} - \underline{x\beta}^{(n)}(k) \|^2$$

diverges to infinity for any fixed $k > 0$ as $n \rightarrow \infty$. In many cases these conditions hold true if $\underline{\beta}$ has infinitely many nonzero elements. For a detailed result see Shibata, 1979.

A lower bound for the loss

We will present the following Proposition and Theorem without proof (see Shibata, 1979).

Proposition

Under Assumptions (A1) and (A2),

$$\text{p-lim}_{n \rightarrow \infty} \frac{L_n(\underline{\beta}, \hat{\underline{\beta}}(k))}{L_n(k)} = 1,$$

uniformly in $1 \leq k \leq K_n$.

This proposition shows that the behavior of $L_n(\underline{\beta}, \hat{\underline{\beta}}(k))$ is asymptotically equivalent to that of $L_n(k)$. Therefore, putting k_n^* the number which minimizes $L_n(k)$ in $1 \leq k \leq K_n$, we obtain a lower bound for the loss $L_n(\underline{\beta}, \hat{\underline{\beta}}(\tilde{k}))$.

Theorem

Under Assumptions (A1) and (A2), for any selection $1 \leq \tilde{k} \leq K_n$, possibly depending on the given samples y_1, \dots, y_n , and for any $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr\left(\frac{L_n(\underline{\beta}, \hat{\underline{\beta}}(\tilde{k}))}{L_n(k_n^*)} > 1 - \varepsilon\right) = 1.$$

Accordingly, k_n^* is asymptotically optimal as the number of parameters. Because it depends on the value of parameters, it is necessary to find the other selection which behaves as k_n^* but does not depend on the value of parameters. Define the pathwise efficiency of a selection \tilde{k} by

$$\text{p.eff.} = \frac{L_n(k_n^*)}{L_n(\underline{\beta}, \hat{\underline{\beta}}(\tilde{k}))}.$$

We call a selection \tilde{k} asymptotically pathwise efficient if the above efficiency converges to 1 in probability as $n \rightarrow \infty$.

Methods

We will compare the following six methods. Each selection \tilde{k} is defined as that

k which minimizes the corresponding statistic in $1 \leq k \leq K_n$.

$$S_1 = \hat{\sigma}^2(k) \exp(k/n) \quad (\text{mean squared error}) \quad (\text{i})$$

$$C_p = \hat{\sigma}^2(k) \left(1 + \frac{2k}{n}\right) \quad (\text{Mallows, 1973; Shibata, 1979}) \quad (\text{ii})$$

$$\text{AIC} (= S_2) = \hat{\sigma}^2(k) \exp\left(\frac{2k}{n}\right) \quad (\text{Akaike, 1973}) \quad (\text{iii})$$

$$S_3 = \hat{\sigma}^2(k) \exp\left(\frac{3k}{n}\right) \quad (\text{see Bhansali \& Downham, 1977}) \quad (\text{iv})$$

$$S_4 = \hat{\sigma}^2(k) \exp\left(\frac{4k}{n}\right) \quad (\text{v})$$

$$\text{BIC} = \hat{\sigma}^2(k) \exp\left(\frac{(\log n)k}{n}\right) \quad (\text{Schwarz, 1978; Akaike, 1978}) \quad (\text{vi})$$

The result in Shibata, 1979, shows that among these six methods, (ii) and (iii) are asymptotically pathwise efficient under Assumptions (A1) and (A2). The BIC behaves as S_3 when $n = 20$, and as S_4 when $n = 50$.

TABLE I

n	$\log n$
20	2.9950
30	3.4012
50	3.9120
100	4.6517
200	5.2983
400	5.9915

COMPARISONS IN SMALL SAMPLE CASES

We will examine how the optimality derived in the preceding section holds true in relatively small samples. As a typical example of a regression function, we take

$$f(x) = -\log(1-x), \quad 0 \leq x < 1$$

which is a rather rapidly increasing function, or a unimodal function

$$f(x) = \exp(-10(x-0.5)^2), \quad 0 \leq x < 1.$$

Each of them has an infinite Taylor series expansion or a Fourier series expansion.

We generate n samples y_1, \dots, y_n at

$$x_\alpha = \left(\frac{\alpha - 1}{n} \right) \delta, \quad \alpha = 1, \dots, n,$$

and we fit one of the following models and select the number k by the methods (i) to (vi).

a) Polynomial regression

$$y_\alpha = \sum_{l=0}^{k-1} x_\alpha^l \beta_{l+1} + e_\alpha, \quad E(e_\alpha) = 0, \quad \alpha = 1, \dots, n.$$

b) Finite Fourier series regression

$$y_\alpha = \sum_{l=0}^{k-1} \frac{\cos \pi l (x_\alpha / \delta)}{l+1} \beta_{l+1} + e_\alpha, \quad E(e_\alpha) = 0, \quad \alpha = 1, \dots, n.$$

Here, Assumptions (A1) and (A2) are satisfied if $\delta < 1$ and K_n diverges to infinity so as $K_n/n \rightarrow 0$ (see Theorem 3.1 in Shibata, 1979).

Choice of K_n

The lower bound $L_n(k_n^*)$ depends on the choice of K_n , because it is the minimum in $1 \leq k \leq K_n$ of $L_n(k)$. It becomes high as the lower K_n is chosen. So, K_n should be chosen large enough for $L_n(k_n^*)$ to be the lowest bound. On the other hand, for a good asymptotic approximation, it is required that K_n increase as slowly as possible, and that k_n^* be moderately large. Therefore, we may choose K_n so that these requirements are balanced. By some experiments, we have a choice $K_n = n^{0.85}$. If K_n is chosen larger than this, the asymptotic behavior of $\hat{\sigma}^2(k)$ is often unsatisfactory. On the contrary, if K_n is chosen smaller, for example, as $K_n = n^{0.65}$, each efficiency of the methods is improved at the cost of an increase of the value of the loss itself, but the order of the efficiencies of the six methods changes very little.

Other conditions

We put $\delta = 0.99$ for boundedness of $f(x) = -\log(1-x)$. We take $\sigma = 0.01$ or 0.02 for avoiding the case $k_n^* = K_n$. Otherwise the difference of the methods does not become so clear. Here $\sigma^2 = E(\epsilon^2)$ is the variance of the error variable ϵ .

Computations

We can obtain the least squares estimates $\hat{\beta}(k)$ and $\hat{\sigma}^2(k)$ and can estimate the value of the loss $L_n(\beta, \hat{\beta}(k))$ at a time for all k by applying the Householder transform (Kitagawa & Akaike, 1978). We have performed 200 repetitions for $n = 20$ to 200, and 100 repetitions for $n = 400$, generating sequences of pseudo normal random numbers.

Pathwise Efficiency

From the result of asymptotic approximation, the pathwise efficiency does not exceed 1 for sufficiently large n . However, it can be seen in Table IV or V, that $n \leq 400$ is not entirely sufficient for such asymptotics in our case. Therefore, we will compare methods by calculating the mean efficiency instead of the pathwise efficiency.

Mean Efficiency

Mean efficiency of a selection \tilde{k} is defined by

$$\text{m. eff.} = \frac{L_n(k_n^*)}{E(L_n(\beta, \hat{\beta}(\tilde{k})))}$$

Using a proof similar to the proof of pathwise efficiency, we can show that the mean efficiency does not exceed 1 for sufficiently large n , and it converges to 1 for (ii) or (iii). In the following tables the efficiency of k_n^* is shown for reference, although k_n^* does not apply in practice. The underline indicates the maximum of the six efficiencies.

TABLE II

Mean Efficiency (Polynomial Regression)

$f(x)$	σ	n	K_n	k_n^*	$L_n(k_n^*)$	S_1	C_p	AIC	S_3	S_4	BIC	k_n^*
-log(1-x)	0.01	20	13	9	.00095	.768	.767	.808	.836	<u>0.843</u>	0.836	1.003
		30	18	10	.00114	.690	.691	.753	.830	<u>0.882</u>	0.864	1.039
		50	28	13	.00137	.506	.515	.646	.786	<u>0.831</u>	0.833	0.986
		100	50	16	.00170	.364	.396	.699	<u>.888</u>	<u>0.888</u>	0.865	1.043
		200	90	18	.00200	.227	.283	.792	<u>.890</u>	0.858	0.815	0.995
		400	163	21	.00226	.140	.288	.830	<u>.885</u>	0.841	0.764	0.968
	0.02	20	13	8	.00343	.696	.695	.745	.793	<u>0.805</u>	0.793	1.012
		30	18	9	.00406	.617	.619	.695	.799	<u>0.848</u>	0.840	1.059
		50	28	11	.00488	.453	.463	.607	.783	0.819	<u>0.821</u>	1.003

TABLE II (Continued)

$f(x)$	σ	n	K_n	k_n^*	$L_n(k_n^*)$	S_1	C_p	AIC	S_3	S_4	BIC	k_n^*
$\exp(-10(x-0.5)^2)$	0.01	100	50	14	.00605	.325	.357	.678	<u>.880</u>	0.870	0.842	1.052
		200	90	16	.00710	.201	.264	.760	<u>.871</u>	0.844	0.777	1.002
		400	163	21	.00806	.124	.277	.827	<u>.897</u>	0.866	0.756	0.983
		20	13	7	.00084	.685	.685	.746	<u>.813</u>	<u>0.858</u>	0.813	1.012
		30	18	9	.00090	.557	.558	.662	<u>.833</u>	<u>0.926</u>	0.878	1.067
		50	28	9	.00090	.339	.350	.520	<u>.746</u>	<u>0.845</u>	0.842	1.017
		100	50	9	.00091	.196	.236	.538	<u>.828</u>	0.818	0.787	1.066
		200	90	9	.00092	.104	.158	.647	<u>.858</u>	0.823	0.730	0.997
	0.02	400	163	9	.00094	.061	.215	.833	.991	<u>1.043</u>	0.934	1.098
		20	13	7	.00294	.604	.605	.662	.731	<u>0.746</u>	0.731	1.014
		30	18	7	.00300	.464	.469	.581	.756	<u>0.887</u>	0.816	1.074
		50	28	7	.00313	.294	.306	.476	.783	<u>0.917</u>	0.915	1.014
		100	50	7	.00343	.187	.232	.548	.992	<u>0.988</u>	<u>1.008</u>	1.052
		200	90	9	.00362	.102	.158	.651	.843	<u>0.856</u>	0.997	
		400	163	9	.00364	.058	.207	.732	<u>.804</u>	0.767	0.726	1.038

The methods S_1 and C_p are inferior to all others. The efficiency of AIC becomes high as n increases. In most cases, the highest efficiency is attained by S_3 when $n \leq 100$, and by S_4 when $n \leq 50$. The efficiency of BIC is the highest in the three cases, but there is not so much difference from that of S_4 . An interesting result is the case $f(x) = -\log(1-x)$, $\sigma = 0.02$ and $n = 50$. As seen in Table I the multiplier of BIC stands between those of S_3 and S_4 when $n = 50$. The high efficiency of BIC indicates that the optimal multiplier is between 3 to 4.

TABLE III

Mean Efficiency (Finite Fourier Series Regression)

$f(x)$	σ	n	K_n	k_n^*	$L_n(k_n^*)$	S_1	C_p	AIC	S_3	S_4	BIC	k_n^*
$-\log(1-x)$	0.01	20	13	13	.00289	<u>1.019</u>	<u>1.019</u>	0.985	0.968	0.944	0.968	1.006
		30	18	18	.00365	<u>1.021</u>	<u>1.021</u>	<u>1.021</u>	1.009	1.000	1.006	1.021
		50	28	28	.00493	<u>0.993</u>	<u>0.993</u>	0.989	0.974	0.946	0.948	0.995
		100	50	50	.00729	<u>1.001</u>	<u>1.001</u>	0.995	0.977	0.938	0.905	1.005
		200	90	90	.01069	<u>0.997</u>	<u>0.997</u>	0.990	0.964	0.899	0.765	1.000
		400	163	130	.01583	0.906	0.907	0.920	<u>0.929</u>	0.862	0.695	0.989
	0.02	20	13	13	.00697	1.003	<u>1.004</u>	0.986	0.967	0.932	0.967	1.012
		30	18	18	.00905	1.026	<u>1.028</u>	1.009	0.978	0.937	0.966	1.035
		50	28	28	.01333	<u>0.987</u>	<u>0.987</u>	0.975	0.952	0.909	0.911	0.993
		100	50	48	.02215	<u>0.998</u>	<u>0.998</u>	0.986	0.955	0.891	0.822	1.009
		200	90	70	.03417	0.910	0.910	<u>0.924</u>	0.919	0.844	0.734	1.007
		400	163	96	.04839	0.727	0.736	<u>0.847</u>	<u>0.916</u>	0.831	0.663	0.988
$\exp(-10(x-0.5)^2)$	0.01	20	13	9	.00122	<u>0.874</u>	0.873	0.848	0.764	0.647	0.746	1.032
		30	18	9	.00157	0.742	0.740	<u>0.780</u>	0.758	0.609	0.697	1.028
		50	28	11	.00147	0.531	0.542	0.633	<u>0.744</u>	0.720	0.725	0.997
		100	50	13	.00173	0.362	0.380	0.660	<u>0.840</u>	0.801	0.768	1.041
		200	90	15	.00206	0.237	0.308	0.758	<u>0.823</u>	0.768	0.716	1.005
		400	163	19	.00246	0.152	0.239	0.851	<u>0.867</u>	0.821	0.706	1.018
	0.02	20	13	3	.00299	0.580	0.583	0.660	0.765	<u>0.864</u>	0.763	1.011

TABLE III (Continued)

$f(x)$	σ	n	K_n	k_n^*	$L_n(k_n^*)$	S_1	C_p	AIC	S_3	S_4	BIC	k_n^*
		30	18	7	.00360	0.533	0.540	0.624	0.788	<u>0.879</u>	0.843	1.057
		50	28	7	.00406	0.372	0.393	0.528	0.704	<u>0.719</u>	0.411	1.008
		100	50	9	.00483	0.256	0.284	0.592	<u>0.750</u>	<u>0.653</u>	0.605	1.038
		200	90	11	.00578	0.167	0.246	0.760	<u>0.841</u>	0.770	0.623	0.995
		400	163	13	.00691	0.108	0.213	0.809	<u>0.832</u>	0.744	0.653	1.039

When $K_n = k_n^*$ or nearly so, S_1 or C_p is most efficient, but the differences from the other AIC, S_3 or S_4 are not so significant. Interesting is the case $f(x) = \exp(-10(x - 0.5)^2)$, $\sigma = 0.01$ and $n = 20$. Although k_n^* is different from K_n , S_1 is the most efficient. We can interpret this result by the graph of $L_n(k)$ in Figure 1. The $L_n(k)$ does not increase so much with $k > k_n^*$, and is similar to the case $K_n = k_n^*$. It automatically follows that the selection of a value near K_n has high efficiency when the increase of $L_n(k)$ in $k > k_n^*$ is very small.

Otherwise the results are similar to those of Table II.

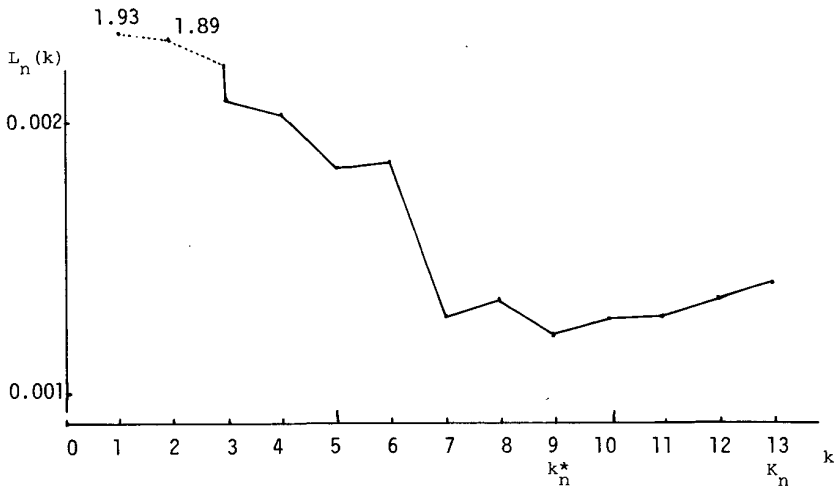


Fig. 1.

Standard Deviations

The standard deviations of $L_n(\hat{\beta}, \hat{\beta}(k))/L_n(k_n^*)$ are shown in Tables IV and V. The mean efficiency is the reciprocal of the expectation of this variable. The underline indicates where the standard deviation is a minimum. The place is almost the same where the maximum efficiency is attained. The value for k_n^* shows that standard deviation of the normalized prediction error decreases as n increases when the

TABLE IV

Standard Deviation of $L_n(\underline{\beta}, \hat{\underline{\beta}}^{\vee}(k))/L_n(k^*)$ (Polynomial Regression)											
$f(x)$	σ	n	K_n	k_n^*	S_1	C_p	AIC	S_3	S_4	BIC	k_n^*
$-\log(1-x)$	0.01	20	13	9	<u>0.517</u>	<u>0.517</u>	0.526	0.540	0.536	0.540	0.459
		30	18	10	<u>0.499</u>	<u>0.500</u>	0.517	0.481	<u>0.408</u>	0.454	0.360
		50	28	13	0.561	0.566	0.631	0.468	0.411	<u>0.409</u>	0.342
		100	50	16	0.681	0.922	0.834	0.407	<u>0.355</u>	<u>0.388</u>	0.317
		200	90	18	0.888	1.751	0.717	<u>0.308</u>	0.315	0.315	0.279
	0.02	400	163	21	0.870	3.120	0.398	<u>0.317</u>	0.323	0.383	0.298
		20	13	8	0.579	0.578	0.586	<u>0.570</u>	0.571	<u>0.570</u>	0.489
		30	18	9	0.562	0.567	0.607	<u>0.531</u>	<u>0.466</u>	0.479	0.369
		50	28	11	0.641	0.681	0.751	0.529	<u>0.455</u>	0.457	0.326
		100	50	14	0.799	1.082	0.936	0.382	<u>0.371</u>	0.394	0.323
		200	90	16	1.006	2.129	0.833	<u>0.357</u>	<u>0.372</u>	0.422	0.293
		400	163	21	0.981	3.551	0.433	0.308	<u>0.301</u>	0.379	0.309
	0.01	20	13	7	0.597	0.596	0.606	0.601	<u>0.580</u>	0.601	0.435
		30	18	9	0.651	0.657	0.735	0.565	<u>0.494</u>	0.516	0.414
		50	28	9	0.899	0.979	1.096	0.696	<u>0.445</u>	0.448	0.400
		100	50	9	1.380	2.200	1.593	0.509	<u>0.462</u>	<u>0.440</u>	0.424
		200	90	9	1.978	4.721	1.575	<u>0.594</u>	0.634	0.672	0.421
		400	163	9	3.125	7.050	0.683	<u>0.493</u>	<u>0.488</u>	0.721	0.470
	0.02	20	13	7	<u>0.692</u>	0.693	0.720	0.770	0.818	0.770	0.495
		30	18	7	<u>0.791</u>	0.812	0.933	0.754	<u>0.659</u>	0.697	0.426
		50	28	7	1.054	1.184	1.338	0.788	<u>0.522</u>	<u>0.522</u>	0.433
		100	50	7	1.543	2.457	1.683	0.500	<u>0.428</u>	<u>0.428</u>	0.385
		200	90	9	2.008	4.823	1.529	0.471	0.425	<u>0.390</u>	0.428
		400	163	9	2.395	7.089	0.747	<u>0.487</u>	0.488	<u>0.435</u>	0.398

TABLE V

Standard Deviation of $L_n(\underline{\beta}, \hat{\underline{\beta}}^{\vee}(k))/L_n(k^*)$ (Finite Fourier Series Regression)											
$f(x)$	σ	n	K_n	k_n^*	S_1	C_p	AIC	S_3	S_4	BIC	k_n^*
$-\log(1-x)$	0.01	20	13	13	<u>0.177</u>	<u>0.177</u>	0.191	0.199	0.206	0.199	0.170
		30	18	18	<u>0.147</u>	<u>0.147</u>	<u>0.147</u>	0.160	0.168	0.160	0.146
		50	28	28	<u>0.148</u>	<u>0.148</u>	<u>0.148</u>	0.153	0.176	0.172	0.147
		100	50	50	<u>0.141</u>	<u>0.141</u>	<u>0.142</u>	0.152	0.176	0.184	0.139
		200	90	90	<u>0.125</u>	<u>0.125</u>	<u>0.123</u>	0.125	0.176	0.253	0.126
	0.02	400	163	130	0.129	0.129	0.131	<u>0.119</u>	0.150	0.221	0.110
		20	13	13	<u>0.293</u>	<u>0.293</u>	0.298	<u>0.306</u>	0.333	0.306	0.290
		30	18	18	<u>0.236</u>	<u>0.236</u>	0.238	0.256	0.282	0.262	0.235
		50	28	28	<u>0.217</u>	<u>0.217</u>	0.219	0.227	0.246	0.240	0.217
		100	50	48	0.181	0.181	0.180	<u>0.175</u>	0.216	0.277	0.175
		200	90	70	0.158	0.158	0.159	<u>0.154</u>	0.210	0.273	0.143
		400	163	96	0.173	0.190	0.207	<u>0.133</u>	0.169	0.213	0.136
	0.01	20	13	9	0.402	0.402	0.403	0.435	<u>0.393</u>	0.435	0.352
		30	18	9	<u>0.409</u>	0.410	0.425	0.481	<u>0.554</u>	0.528	0.293
		50	28	11	0.535	0.555	0.602	<u>0.499</u>	0.574	0.565	0.295
		100	50	13	0.719	0.855	0.819	0.363	<u>0.345</u>	0.361	0.278
		200	90	15	0.855	1.694	0.709	0.308	<u>0.300</u>	0.380	0.250
		400	163	19	1.161	3.039	0.336	0.269	<u>0.263</u>	0.344	0.225
$\exp(-10(x-0.5)^2)$	0.02	20	13	3	0.671	0.675	0.705	0.668	<u>0.507</u>	0.671	0.317
		30	18	7	0.620	0.633	0.664	0.505	<u>0.328</u>	0.399	0.349

TABLE V (Continued)

$f(x)$	σ	n	K_n	k_n^*	S_1	C_p	AIC	S_3	S_4	BIC	k_n^*
		50	28	7	0.800	0.912	1.004	0.560	<u>0.318</u>	0.318	0.328
		100	50	9	1.117	1.491	1.193	<u>0.526</u>	0.533	0.530	0.331
		200	90	11	1.307	2.718	0.870	<u>0.325</u>	0.418	0.708	0.276
		400	163	13	1.857	4.471	0.423	<u>0.361</u>	0.441	0.426	0.244

optimal number k_n^* is applied.

The mean of the selected number \tilde{k}

The selected number itself is not our main concern, but is included for reference. We show a part of the results when $n = 400$.

TABLE VI

The mean of the selected number \tilde{k} (Polynomial regression)

$f(x)$	σ	K_n	k_n^*	S_1	C_p	AIC	S_3	S_4	BIC
$-\log(1-x)$	0.01	163	21	156.84	68.36	21.43	19.55	18.66	17.68
	0.02	163	18	156.84	62.41	19.07	17.02	16.27	15.27
$\exp(-10(x-0.5)^2)$	0.01	163	9	152.71	38.89	10.14	9.26	9.10	8.94
	0.02	163	9	154.38	36.96	9.89	8.47	7.95	7.41

TABLE VII

The mean of the selected number \tilde{k} (Finite Fourier series regression)

$f(x)$	σ	K_n	k_n^*	S_1	C_p	AIC	S_3	S_4	BIC
$-\log(1-x)$	0.01	163	130	160.87	160.35	147.34	123.32	105.78	89.64
	0.02	163	96	158.67	154.87	119.74	86.83	74.50	62.06
$\exp(-10(x-0.5)^2)$	0.01	163	19	154.82	89.81	19.89	15.60	13.93	12.39
	0.02	163	12	153.22	69.46	14.18	11.18	9.51	8.24

It can be seen that the selection whose mean is very near to k_n^* is not always the most efficient one. The best selection is rather biased towards the lower number.

A Conclusion

Although the AIC or C_p method is asymptotically efficient, it does not behave so well if the sample size is less than or equal to 400. Especially C_p is inferior in our cases, because the difference between their multipliers significantly affects the behavior, although which is negligible in large samples.

It is necessary to apply some finite corrections. If the sample size is from 100 to 400, the use of S_3 is recommended, and if it is 50, S_4 . If it is less than or equal to 30, in many cases, S_4 is superior, but in certain circumstances AIC is superior.

ACKNOWLEDGEMENT

The computations were done by the HITAC M-180 system of the Information Processing Center, Tokyo Institute of Technology. The author wishes to thank his wife for her help in preparing the manuscript.

REFERENCES

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: B.N. Petrov and F. Csáki (eds.), 2nd International Symposium on Information Theory. Budapest: Akadémia Kiado, pp. 267-81.
- Akaike, H., 1978. A Bayesian analysis of the minimum AIC procedure. Ann. Inst. Statist. Math., 30A: 9-14.
- Allen, D.M., 1971. Mean square error of prediction as a criterion for selecting variables. Technometrics, 13: 469-81.
- Bhansali, R.J. and Downham, D.Y., 1977. Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion. Biometrika, 64: 547-51.
- Box, E.P. and Jenkins, M., 1976. Time Series Analysis; forecasting and control. Holden-day, San Francisco.
- Gates, P. and Tong, H., 1976. On Markov chain modeling to some weather data. J. Appl. Met., 15: 1145-51.
- Hocking, R.R., 1976. The analysis and selection of variables in linear regression. Biometrics, 62: 1-49.
- Kitagawa, G. and Akaike, H., 1978. A procedure for the modeling of nonstationary time series. Ann. Inst. Statist. Math. 30B: 351-63.
- Mallows, C.L., 1973. Some comments on C_p . Technometrics, 15: 661-75.
- Ozaki, T., 1977. On the order determination of ARMA models. J. R. Statist. Soc., C 26: 290-301.
- Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist., 6: 461-4.
- Shibata, R., 1979. An optimal selection of regression variables. submitted to Biometrika.
- Shibata, R., 1980. Asymptotically efficient selection of the order of the model for estimating parameters of a linear process. Ann. Statist., 8: 147-64.

MODELLING WEATHER DATA AS A MARKOV CHAIN

L.BILLARD¹ and M.R.MESHKANI²

1 Dept. Stat., Florida State Univ., Florida

2 Coll. of Stat. and Informatics (Iran)

ABSTRACT

Billard, L. and Meshkani, M.R., Modelling weather data as a Markov chain. Proc. 1-st Intern. Conf. on Stat. and Climat., held in Tokyo, Nov.29-Dec.1, 1979

Weather data is taken and by classifying the various degrees of wetness and dryness, it can be modelled as a Markov chain. Data are recorded over several years. Estimates for the transition probabilities are obtained which utilize the experience (data) of previous years giving so-called empirical Bayes estimates. These contrast with the maximum likelihood estimates which use the data for the year under discussion alone.

1. INTRODUCTION

Some time ago Gabriel and Neuman (1962) showed how weather could be modelled as a Markov chain. They modelled days as either dry or wet (corresponding to chain with states 0 or 1). Obviously it is possible to subclassify the wet days into (s-1) classes designating the degree of wetness as measured by the amounts of rainfall actually received. This then gives a Markov chain with s states. As is well known once the appropriate transition probabilities of the chain are known together with the initial distribution everything is completely known about the Markov process. In this case then we are able to ascertain the transition probabilities of moving from one weather classification type to another in any given number of days. We can determine the long run probabilities of certain weather types and so forth.

In the Gabriel and Neuman work, estimates for the transition probabilities were determined via maximum likelihood techniques. However, it would be advantageous to be able to obtain estimates which incorporate the available data from previous years as well. Thus, in the following sections we first outline a theoretical approach which allows a researcher to utilize past experience represented by previous data when estimating these transition probabilities yielding so called empirical Bayes estimates. This approach is quite general and therefore it is presented within a general framework.

Then, in Sections 5-7 we apply these techniques to some weather data from Tallahassee, Florida. For each of computation we take the classes of the Markov chain

to be simply "dry" and "wet". Exactly the same techniques apply to the cases of a larger number of classes (that is, $s > 2$) as well as to data from broader regions such as other cities, states, localities, or whatever. It should be remarked that some underlying assumptions are made about the process such as stationarity of the chain, independence of the transition probabilities from year to year, etc. Quite clearly, the techniques discussed here can only be used for data sets for which such assumptions hold true. With weather data there will be some regions for which these assumptions are invalid. Thus, in Section 6, verification of these assumptions for the Tallahassee data is carried out.

2. THE PROBABILITY MODEL

2.1. Preliminaries

For the sake of brevity, we shall not repeat the well-known results pertaining to single Markov chains. These are found in, e.g., Feller (1968).

Suppose $\{X_t, t \in T_0\}$ is a Markov chain with values in the finite state space $S = \{1, \dots, s\}$ where $T = \{1, \dots, T\}$ and $T_0 = \{0\} \cup T$. We assume the chain is simple, i.e., its order of dependency is 1. Furthermore, it is stationary and has an irreducible t.p.m. $\underline{\Lambda}$ with elements Λ_{jk} , $j, k \in S$. Let the initial distribution be $\underline{\theta}$ with element θ_j , $j \in S$.

The data are outcomes of $(n+1)$ repeated experiments. In such experiment, we observe and record the states visited by the chain during a fixed period of time, $T > 1$. The outcomes of the first n experiments will be referred to as the "past data". Let a realization of an experiment be $\underline{x}_T = (x_0, x_1, \dots, x_T)$, where the subscripts refer to the order in which the observations were taken and not to their values.

Definition 2.1 Let \underline{F} be an $s \times s$ matrix whose (j, k) -th element F_{jk} is the number of times that the state k has followed the state j in a sequence of states visited by a Markov chain $\{X_t, t \in T_0\}$. That is, F_{jk} is the number of times the event $\{X_{t-1} = j, X_t = k; t \in T\}$ has occurred. For each fixed $T > 1$, \underline{F} is called the frequency count matrix (f.c.m.) of the chain up to time T .

Then, the probability of observing a particular ordered sequence of states is

$$P(X_0 = u, X_1 = x_1, \dots, X_T = x_T) = P(X_0 = u) \prod_{t \in T} P(X_t = x_t | X_{t-1} = x_{t-1}) = \theta_u \prod_{j, k \in S} \Lambda_{jk}^{F_{jk}} \quad (2.1)$$

where $\theta_u \in \Theta$ with $\Theta = \{\underline{\theta} : \theta_j > 0, j \in S, \sum_{j \in S} \theta_j = 1\}$, and $\Lambda_{jk} \in \Omega_s$ with $\Omega_s = \{\underline{\Lambda} : \Lambda_{jk} \geq 0, j, k = 1, \dots, s, \sum_{k \in S} \Lambda_{jk} = 1, j \in S\}$, and where $x_0 = u$ is the initial state of the chain.

It is clear that \underline{F} is a sufficient statistic for $\underline{\Lambda}$ and $\underline{\theta}$. In the sequel, we deal mainly with \underline{F} .

Before observing the outcome x_T , the integer x_0 and the matrix F are random quantities. The conditional distribution of F given the initial state is u and the t.p.m. is $\underline{\Lambda}$, was first derived by Whittle (1955). This and some other related distributions have been discussed in detail by Martin (1967).

2.2. Conditional distributions.

We are interested in the unconditional distribution of F given $x_0 = u$, and in the posterior distribution of $\underline{\Lambda}$ given F . We shall derive these distributions utilizing Martin's results on conditional distributions.

Let $x_0 = u$, $x_T = v$. Then by the definition of F ,

$$F_{j+} - F_{+j} = \delta_{ju} - \delta_{jv}, \quad j \in S, \quad (2.2)$$

where

$$F_{j+} = \sum_{k \in S} F_{jk}, \quad F_{+k} = \sum_{j \in S} F_{jk}.$$

For a given F and a fixed u , the equations (2.2) uniquely determine v , and vice versa. The restriction on F is essentially the defining characteristic of the space of values of F .

Let M be the set of positive integers and $M_0 = M \cup \{0\}$. For fixed u , $u \in S$, $\underline{\Lambda} \in \Omega_S$ and $T \in M$, we define the following sets:

$$\Phi_S(u, v, T, \underline{\Lambda}) = \{F: F_{jk} \in T_0, \mathbf{1}' F \mathbf{1} = T, F_{j+} - F_{+j} = \delta_{ju} - \delta_{jv}, F_{jk} = 0 \text{ if } \Lambda_{jk} = 0, j, k \in S\}, \quad (2.3)$$

$$\Phi_S(u, T, \underline{\Lambda}) = \bigcup_{v \in S} \Phi_S(u, v, T, \underline{\Lambda}), \quad u \in S, T \in M, \underline{\Lambda} \in \Omega_S, \quad (2.4)$$

$$\Phi_S^*(T, \underline{\Lambda}) = \bigcup_{u \in S} \Phi_S(u, T, \underline{\Lambda}), \quad T \in M, \underline{\Lambda} \in \Omega_S, \quad (2.5)$$

$$\Phi_{S1}^*(T, \underline{\Lambda}) = \{F: F \in \Phi_S^*(T, \underline{\Lambda}), F_{j+} = F_{+j}, j \in S\}, \quad (2.6)$$

and

$$\Phi_{S2}^*(T, \underline{\Lambda}) = \Phi_S^*(T, \underline{\Lambda}) - \Phi_{S1}^*(T, \underline{\Lambda}). \quad (2.7)$$

For each f.c.m. F , we define $F^* = (F_{jk}^*)$ where, for $j, k \in S$,

$$F_{jk}^* = \begin{cases} \delta_{jk} - F_{jk}/F_{j+}, & F_{j+} > 0, \\ \delta_{jk}, & F_{j+} = 0. \end{cases} \quad (2.8)$$

The (v, u) -th cofactor of F^* will be denoted by $F_{(vu)}^*$.

The conditional p.m.f. of F given u and $\underline{\Lambda}$, known as the Whittle distribution, is

$$P(F|u, \underline{\Lambda}) \equiv P^{(S)}(F|u, T, \underline{\Lambda}) = F_{(vu)}^* A(F) \prod_{j, k \in S} \Lambda_{jk}^{F_{jk}}, \quad F \in \Phi_S(u, T, \underline{\Lambda}), \quad (2.9)$$

where v is the unique solution of (2.2) and

$$A(\underline{F}) = \prod_{j \in S} (F_{j+}! / \prod_{k \in S} F_{jk}!).$$

Here and elsewhere, the convention $0^0 = 1$ will be observed.

The joint distribution of \underline{F} and X_0 which is called the Whittle-1 distribution, is

$$P_1(\underline{F}, u | \underline{A}, \underline{\theta}) \equiv P_1^{(s)}(\underline{F}, u | \underline{A}, \underline{\theta}) = \theta_u P(\underline{F} | u, \underline{A}), \quad u \in S, \underline{F} \in \Phi_S(u, T, \underline{A}). \quad (2.10)$$

The marginal distribution of U for a given probability vector $\underline{\theta} = (\theta_1, \dots, \theta_s)$ is a multinomial distribution, $M_s(1, \underline{\theta})$. The marginal distribution of \underline{F} for a given \underline{A} is given as follows.

There are exactly s pairs of integers $(x, y) = (u, u)$, $u \in S$, which satisfy the equations

$$F_{j+} - F_{+j} = \delta_{jx} - \delta_{jy}, \quad j \in S, \quad (2.11)$$

if $\underline{F} \in \Phi_{s1}^*(T, \underline{A})$. There is a unique solution $(x, y) = (u, v)$, $u \neq v$, to these equations if $\underline{F} \in \Phi_{s2}^*(T, \underline{A})$, see Martin (1967, Lemma 6.1.5). Then, the marginal distribution of \underline{F} for a given t.p.m. known as the Whittle-2 distribution, is

$$P_2(\underline{F} | \underline{A}, \underline{\theta}) \equiv P_2^{(s)}(\underline{F} | T, \underline{A}, \underline{\theta}) = \begin{cases} A(\underline{F}) \left(\sum_{j \in S} \theta_j F_{jj}^* \right) \prod_{j, k \in S} \Lambda_{jk}^{F_{jk}}, & \underline{F} \in \Phi_{s1}^*(T, \underline{A}), \\ A(\underline{F}) \theta_u F_{vu}^* \prod_{j, k \in S} \Lambda_{jk}^{F_{jk}}, & \underline{F} \in \Phi_{s2}^*(T, \underline{A}), \end{cases} \quad (2.12)$$

where (u, v) is the unique solution to (2.11) when $\underline{F} \in \Phi_{s2}^*(T, \underline{A})$.

2.3. Unconditional distributions.

We shall assume the "natural conjugate priors" for $\underline{\theta}$ and \underline{A} to be independent of each other. The "natural conjugate prior" for $\underline{\theta}$ is a Dirichlet distribution and for \underline{A} is a matrix beta distribution. We denote these distributions by $D(\underline{\alpha})$ and $MB(\underline{\rho})$, respectively. The resultant unconditional distribution will be named the Beta-Whittle distribution.

To specify the space of values of \underline{F} , we define the following sets:

$$\Phi_S(u, v, T) = \bigcup_{\underline{A} \in \Omega_S} \Phi_S(u, v, T, \underline{A}), \quad (2.13)$$

$$\Phi_S(u, T) = \bigcup_{\underline{A} \in \Omega_S} \Phi_S(u, T, \underline{A}), \quad (2.14)$$

$$\Phi_S^*(T) = \bigcup_{\underline{A} \in \Omega_S} \Phi_S^*(T, \underline{A}), \quad (2.15)$$

$$\Phi_{s1}^*(T) = \bigcup_{\underline{\Lambda} \in \Omega_s} \Phi_{s1}^*(T, \underline{\Lambda}), \quad (2.16)$$

and

$$\Phi_{s2}^*(T) = \Phi_s^*(T) - \Phi_{s1}^*(T). \quad (2.17)$$

Now, we derive the unconditional distributions by integrating the conditional ones w.r.t. $q_1(\underline{\theta})$ and $q(\underline{\Lambda})$, the prior distributions of $\underline{\theta}$ and $\underline{\Lambda}$, respectively. That is,

$$q_1(\underline{\theta}) = g(\underline{\alpha}) \prod_{j \in S} \theta_j^{\alpha_j - 1}, \quad \underline{\theta} \in \Theta,$$

with the parameter $\underline{\alpha} = (\alpha_j)$, $\alpha_j > 0$, $j \in S$, and

$$g(\underline{\alpha}) = \Gamma(\alpha_+) / \prod_{j \in S} \Gamma(\alpha_j),$$

and $\alpha_+ = \sum_{j \in S} \alpha_j$; and

$$q(\underline{\Lambda}) = C(\underline{\rho}) \prod_{j, k \in S} \Lambda_{jk}^{\rho_{jk} - 1}, \quad \underline{\Lambda} \in \Omega,$$

where the parameter $\underline{\rho} = (\rho_{jk})$, $\rho_{jk} > 0$, $j, k \in S$, and

$$C(\underline{\rho}) = \prod_{j \in S} \{ \Gamma(\rho_{j+}) / \prod_{k \in S} \Gamma(\rho_{jk}) \},$$

and $\rho_{j+} = \sum_{k \in S} \rho_{jk}$, $j \in S$. Thus, from (2.9), the Beta-Whittle distribution for a MB($\underline{\rho}$) prior and known u is

$$P(\underline{F}|u) = F_{(vu)}^* A(\underline{F}) \int_{\Omega_s} \left(\prod_{j, k \in S} \Lambda_{jk}^{F_{jk}} \right) q(\underline{\Lambda}) d(\underline{\Lambda}) = F_{(vu)}^* \cdot A(\underline{F}) B(\underline{\rho}, \underline{F}), \quad \underline{F} \in \Phi_s(u, T), \quad (2.18)$$

where v is the unique solution to (2.2) and where

$$B(\underline{\rho}, \underline{F}) = \prod_{j \in S} \{ [\Gamma(\rho_{j+} + F_{j+})] / \prod_{k \in S} [\Gamma(\rho_{jk} + F_{jk}) / \Gamma(\rho_{jk})] \}.$$

Similarly, from (2.10) when assuming a D($\underline{\alpha}$) prior for $\underline{\theta}$, we obtain the Beta-Whittle-1 distribution,

$$\begin{aligned} P_1(\underline{F}, u) &= \int_{\Theta} \int_{\Omega_s} \theta_u q_1(\underline{\theta}) P(\underline{F}|u, \underline{\Lambda}) q(\underline{\Lambda}) d(\underline{\Lambda}) d(\underline{\theta}) \\ &= A(\underline{F}) \cdot B(\underline{\rho}, \underline{F}) \cdot C(F_{(vu)}^*, \underline{\alpha}), \quad u \in S, \quad \underline{F} \in \Phi_s(u, T), \end{aligned} \quad (2.19)$$

where

$$C(F_{(vu)}^*, \underline{\alpha}) = [\Gamma(\alpha_+) / \prod_{j \in S} \Gamma(\alpha_j)] \cdot [F_{(vu)}^* \cdot \Gamma(\alpha_u + 1) \prod_{\substack{k \in S \\ k \neq u}} \Gamma(\alpha_k) / \Gamma(\alpha_+ + 1)].$$

Finally, the Beta-Whittle-2 distribution is derived from (2.12). Thus,

$$P_2(\underline{F}) = \int_{\Theta} \int_{\Omega_S} P_2(\underline{F} | \underline{\Lambda}, \underline{\theta}) q_1(\underline{\theta}) q(\underline{\Lambda}) d(\underline{\Lambda}) d(\underline{\theta})$$

$$= \begin{cases} A(\underline{F}) \int_{\Theta} \left(\prod_{j \in S} \theta_j F_{jj}^* \right) q_1(\underline{\theta}) d(\underline{\theta}) \int_{\Omega_S} \left(\prod_{s \in S} \prod_{k \in S} \Lambda_{jk}^{F_{jk}} \right) q(\underline{\Lambda}) d(\underline{\Lambda}), & \underline{F} \in \Phi_{s1}^*(T), \\ F_{(vu)}^* \cdot A(\underline{F}) \int_{\Theta} \theta_u q(\underline{\theta}) d(\underline{\theta}) \cdot \int_{\Omega_S} \left(\prod_{s \in S} \prod_{k \in S} \Lambda_{jk}^{F_{jk}} \right) q(\underline{\Lambda}) d(\underline{\Lambda}), & \underline{F} \in \Phi_{s2}^*(T), \end{cases}$$

where (u, v) is the unique solution to (2.11) when $\underline{F} \in \Phi_{s2}^*(T)$. Therefore, the distribution is

$$P_2(\underline{F}) = \begin{cases} A(\underline{F}) \cdot B(\underline{\rho}, \underline{F}) \cdot C(\underline{F}^*, \underline{\alpha}), & \underline{F} \in \Phi_{s1}^*(T), \\ A(\underline{F}) \cdot B(\underline{\rho}, \underline{F}) \cdot C(\underline{F}_{(vu)}^*, \underline{\alpha}), & \underline{F} \in \Phi_{s2}^*(T), \end{cases} \quad (2.20)$$

where $A(\underline{F})$, $B(\underline{\rho}, \underline{F})$ and $C(\underline{F}^*, \underline{\alpha})$ have been defined in (2.9), (2.18) and (2.19), respectively, and where

$$C(\underline{F}^*, \underline{\alpha}) = [\Gamma(\alpha_+) / \prod_{j \in S} \Gamma(\alpha_j)] \cdot \left[\prod_{j \in S} F_{jj}^* \Gamma(\alpha_j + 1) \prod_{\substack{k \in S \\ k \neq j}} \Gamma(\alpha_k) / \Gamma(\alpha_+ + 1) \right].$$

When u is known, $P(X_0 = u | \underline{\theta}) = \theta_u = 1$. Then, (2.19) reduces to (2.18). In the sequel, we shall consider both cases and treat them simultaneously.

3. BAYES ESTIMATE OF $\underline{\Lambda}$

3.1. Posterior distribution of $\underline{\Lambda}$.

We assume squared error loss. Hence, the loss function associated with the estimation of $\underline{\Lambda}$ by $\underline{d} = (d_{jk})$, $j, k \in S$, is given by, from DeGroot (1970),

$$L(\underline{\Lambda}, \underline{d}) = \sum_{j, k \in S} (d_{jk} - \Lambda_{jk})^2.$$

It can be easily shown that the minimum risk is achieved when each Λ_{jk} , $j, k \in S$, has least possible risk. Thus, the Bayes estimate of $\underline{\Lambda}$ is found by finding the Bayes estimate for each Λ_{jk} , $j, k \in S$. This in turn is given by the posterior mean of $\underline{\Lambda}$ for given \underline{F} .

Theorem 3.1. Let \underline{F} be the f.c.m. of a single stationary Markov chain up to time T . Let $\underline{\Lambda}$ be the t.p.m. of the chain. Assume $\underline{\Lambda}$ has a $MB(\underline{\rho})$ prior distribution. Then, the posterior distribution of $\underline{\Lambda}$ given \underline{F} is a $MB(\underline{\rho} + \underline{F})$. Furthermore, the conclusion is true whether the initial state $X_0 = u$ is known or unknown.

Proof. First, we suppose u is not known. Then, from (2.10) and (2.20), we have

$$q^*(\underline{\Lambda}, \underline{\theta}) = K(\underline{F}, \underline{\alpha}, \underline{\rho}) \theta_u^{\alpha_u} \prod_{\substack{k \in S \\ k \neq u}} \theta_k^{\alpha_k - 1} \prod_{j, k \in S} \Lambda_{jk}^{F_{jk} + \rho_{jk} - 1}, \quad \theta \in \Theta, \underline{\Lambda} \in \Omega_S,$$

where $K(\cdot)$ is free of θ_j and Λ_{jk} , $j, k \in S$. Then,

$$q^*(\underline{\Lambda}) = \int_{\Omega} q^*(\underline{\Lambda}, \underline{\theta}) d\underline{\theta} \propto \prod_{j,k \in S} \Lambda_{jk}^{F_{jk} + \rho_{jk} - 1}, \underline{\Lambda} \in \Omega_S. \quad (3.1)$$

It is obvious that (3.1) is a $MB(\underline{\rho} + \underline{F})$.

When u is known, we have $\theta_u = 1$, $u \in S$, and the above derivation more easily gives (3.1). This proves the theorem.

Theorem 3.2. Let \underline{F} be the f.c.m. of a single stationary Markov chain up to time T . Let $\underline{\Lambda}$ have a prior distribution $MB(\underline{\rho})$. Then, the Bayes estimate of $\underline{\Lambda}$ relative to the squared error loss function, whether the initial state $X_0 = u$ is known or unknown, is

$$\underline{\Lambda}_B = \underline{\Lambda}_B(\underline{F}, \underline{\rho}) = (\Lambda_{B;jk}) \quad (3.2)$$

where

$$\Lambda_{B;jk} = (F_{jk} + \rho_{jk}) / (F_{j+} + \rho_{j+}), \quad j, k \in S.$$

Proof. It is enough to find the Bayes estimate of Λ_{jk} . For the squared error loss function, the posterior mean is the Bayes estimate. Thus,

$$\Lambda_{B;jk} = \int_{\Omega_S} \Lambda_{jk} q^*(\underline{\Lambda}) d(\underline{\Lambda}) = (F_{jk} + \rho_{jk}) / (F_{j+} + \rho_{j+}).$$

This completes the proof.

The maximum likelihood estimate (MLE) of $\underline{\Lambda}$ based on \underline{F} , which will be denoted by

$$\underline{\Lambda}_{ML} = (\Lambda_{ML;jk}), \text{ is}$$

$$\Lambda_{ML;jk} = F_{jk} / F_{j+}, \quad j, k \in S.$$

(See Bartlett (1951) or Billingsley (1961)). Note that $\Lambda_{B;jk}$ is a convex combination of $\Lambda_{ML;jk}$ and $E(\Lambda_{B;jk}) = \rho_{jk} / \rho_{j+}$.

4. EMPIRICAL BAYES ESTIMATE OF $\underline{\Lambda}$

4.1. Preliminaries.

In this section, we shall estimate ρ_{jk} from the "past data". Then, we shall substitute these values in (3.2). The resultant value will be called an EB estimate of $\underline{\Lambda}$.

Let $N = \{1, \dots, n\}$. Here, the "past data" refers to the set \underline{F}_i , $i \in N$ which are independent of $\underline{F} \equiv \underline{F}_{n+1}$ which represents the "current data", but they are identically distributed as \underline{F} .

We have seen in (2.19) that the pair (\underline{F}, u) is distributed according to a Beta-Whittle-1 distribution. The marginal distribution of U is identical to a Dirichlet-Multinomial distribution. The EB procedure for estimation of parameters of this

distribution has been considered in Billard and Meshkani (1978).

Now, we address ourselves to the estimation of ρ_{jk} from $\{F_i, i \in N\}$. The marginal distribution of F was given in (2.20) which contains $s(s+1)$ parameters \underline{a} and $\underline{\rho}$. We can readily estimate s parameters \underline{a} by methods proposed in Billard and Meshkani (1978). Therefore, in the rest of this section, we concentrate only on the estimation of $\underline{\rho}$.

4.2. Method of moments estimate of $\underline{\rho}$.

Exact formulae for moments of F are too complicated to be useful in estimating $\underline{\rho}$. Using some results of Martin (1967), we have

$$E(F_{jk}) = E_2[E_1(F_{jk})] = \sum_{t=0}^{T-1} E_2(\Lambda_{uj}^{[t]} \Lambda_{jk}), \quad j, k \in S,$$

where the subscript 1 (2) indicates the expectations have been taken for a given $\underline{\Lambda}$ (w.r.t the distribution of $\underline{\Lambda}$). We also have

$$E(F_{jk} F_{gh}) = \begin{cases} \delta_{jg} \delta_{kh} E(F_{jk}), & T = 1, \\ \delta_{jg} \delta_{kh} E(F_{jk}) + \sum_{t=1}^{T-1} E_2\{\Lambda_{uj}^{[T-1-t]} \Lambda_{jk} \sum_{m=0}^{t-1} \Lambda_{kg}^{[m]} \Lambda_{gh} \\ + \Lambda_{ug}^{[T-1-t]} \Lambda_{gh} \sum_{m=0}^{t-1} \Lambda_{hj}^{[m]} \Lambda_{jk}\}, & T \geq 2, \quad j, k, g, h \in S. \end{cases}$$

Evaluation of the expectations in the above equations will lead to polynomials of degree $(T-1)$ in ρ_{jk} , $j, k \in S$. When $T \geq 3$, the resultant equations will be almost intractable. Since for single chains, T is usually far greater than 3, setting $T \leq 3$ above to obtain solvable equations, would be a waste of available information. Moreover, the estimates would not be very efficient.

We shall seek some functions of F which render simpler expressions for their moments. One of these functions is

$$M_{jk} = F_{jk}/F_{j+}, \quad j, k \in S. \quad (4.1)$$

Since $\underline{\Lambda}$ is assumed to be irreducible, $\Lambda_{j+} \neq 0$, for all $j \in S$. Thus, from the condition (2.2), for T large enough, $F_{j+} > 0$, $j \in S$. We assume $F_{j+} > 0$, $j \in S$, so that we can use M_{jk} to estimate ρ_{jk} .

Whittle (1955), under the assumption that $F_{j+} > 0$, $j \in S$, gave

$$E_1(M_{jk}|u) = \Lambda_{jk}(T + a_{jk})/T + O(T^{-3/2}), \quad (4.2)$$

and

$$\text{Cov}_1(M_{jk}, M_{gh}|u) = \delta_{jg}(\delta_{kh} - \Lambda_{jk}\Lambda_{gh}) \cdot E_1(F_{j+}^{-1}|u) + O(T^{-3/2}), \quad (4.3)$$

where a_{jk} is the (j, k) -th element of the matrix of right eigenvectors. By appropriate normalization of \underline{a} , we can make $0 \leq a_{jk} \leq 1$.

Now, using (4.2) and (4.3) and $a_{jk} = 1$, we shall find the unconditional expectations and covariances relative to the $MB(\rho)$ prior for \underline{A} . In the sequel, we shall assume T is large enough so that we can ignore $O(T^{-3/2})$. Thus

$$E(M_{jk}) = [(T+1)/T] \rho_{jk}/\rho_{j+}, \quad j, k \in S, \quad (4.4)$$

and

$$\text{Cov}(M_{jk}, M_{gh}) = \omega_j \delta_{jg} \rho_{jk} (\delta_{kh} \rho_{j+} - \rho_{jh}) / \rho_{j+}^2, \quad j, k, g, h \in S, \quad (4.5)$$

where

$$\omega_j = \{\rho_{j+} E[F_{j+}^{-1}] + [(T+1)/T]^2\} / (\rho_{j+} + 1), \quad j \in S. \quad (4.6)$$

The result (4.5) indicates that different rows of the matrix $\underline{M} = (M_{jk})$ are uncorrelated. Since $\underline{M} \underline{1} = \underline{1}$, we shall delete its last column to avoid singularity in the covariance matrix of \underline{M} . The covariance matrix of the first $(s-1)$ columns of \underline{M} will be denoted by $\underline{\Sigma}^*$. Then, $\underline{\Sigma}^*$ is a block diagonal matrix of order $s(s-1)$. That is, $\underline{\Sigma}^* = \text{Diag}\{\underline{\Sigma}_{jj}^*\}$, where the elements $\sigma_{jk,jh}^* = \text{Cov}(M_{jk}, M_{jh})$ of $\underline{\Sigma}_{jj}^*$ are defined in (4.5).

We observe that for each $j \in S$, the relations (4.4) give $(s-1)$ linearly independent equations in s unknowns ρ_{jk} , $k \in S$. We need one more equation. This is established as follows.

From (4.5), we may write

$$\text{Cov}(M_{jk}, M_{jh}) = \omega_j E(M_{jk}) [\delta_{kh} - E(M_{jh})], \quad k, h \in S.$$

In matrix form, we have

$$\underline{\Sigma}_{jj}^* = \omega_j \underline{\Sigma}_{jj}, \quad j \in S, \quad (4.7)$$

where we define the elements of $\underline{\Sigma}_{jj}$ to be $\sigma_{jk,jh} = E(M_{jk}) [\delta_{kh} - E(M_{jh})]$, $k, h \in S$. We can solve (4.7) for ω_j to obtain

$$\omega_j = \{|\underline{\Sigma}_{jj}^*| / |\underline{\Sigma}_{jj}|\}^{1/(s-1)}, \quad j \in S.$$

Therefore, substituting for ω_j in (4.6) and solving for ρ_{j+} , we have

$$\rho_{j+} = \{[(T+1)/T]^2 - \omega_j\} / [\omega_j - E(F_{j+}^{-1})], \quad j \in S. \quad (4.8)$$

This, together with (4.4) which is rearranged into

$$\rho_{jk} = T \rho_{j+} E(M_{jk}) / (T+1), \quad j, k \in S,$$

allows us to solve for ρ_{jk} , $j, k \in S$.

The equations (4.4) and (4.8) give the parameters in terms of the moments of M_{jk} and F_{j+}^{-1} . Now, we substitute the sample moments obtained from the "past data" in (4.4) and (4.8) to obtain the method of moments estimates of ρ_{jk} , $j, k \in S$.

These estimates will be denoted by r_{jk} , $j, k \in S$.

For each $j \in S$ and $k, h \in S$, let us define the sample means $\bar{M} = (\bar{M}_{jk})$ and $\bar{G} = (\bar{G}_j)$, and sample covariances $\hat{\Sigma}_{jj}^* = (\hat{\sigma}_{jk,jh}^*)$ and $\hat{\Sigma}_{jj} = (\hat{\sigma}_{jk,jh})$ where the elements are respectively defined by

$$\bar{M}_{jk} = n^{-1} \sum_{i \in S} (F_{i,jk} / F_{i,j+}) , \quad (4.9)$$

$$\bar{G}_j = n^{-1} \sum_{i \in S} F_{i,j+} , \quad (4.10)$$

$$\hat{\sigma}_{jk,jh}^* = (n-1)^{-1} \sum_{i \in S} (M_{i,jk} - \bar{M}_{jk})(M_{i,jh} - \bar{M}_{jh}) , \quad (4.11)$$

and

$$\hat{\sigma}_{jk,jh} = \bar{M}_{jk} (\delta_{kh} - \bar{M}_{jh}) . \quad (4.12)$$

Then, the estimates of ω_j , ρ_{j+} and ρ_{jk} respectively are

$$c_j = \{ |\hat{\Sigma}_{jj}^*| / |\hat{\Sigma}_{jj}| \}^{1/(s-1)} , \quad j \in S , \quad (4.13)$$

$$r_{j+} = \{ [(T+1)/T]^2 - c_j \} / [c_j - \bar{G}_j] , \quad j \in S , \quad (4.14)$$

and

$$r_{jk} = \text{Tr}_{j+} \bar{M}_{jk} / (T+1) , \quad j, k \in S . \quad (4.15)$$

Consequently,

$$r_{js} = r_{j+} - \sum_{k \in S} r_{jk} = r_{j+} (T \bar{M}_{js} + 1) / (T+1) , \quad j \in S .$$

Therefore, from (3.2), the EB estimate of $\underline{\Lambda}$, denoted by $\underline{\Lambda}_{EB}$, is obtained by replacing ρ_{jk} by r_{jk} .

Definition 4.1. The EB estimate of $\underline{\Lambda}$ obtained by the method of moments is the matrix $\underline{\Lambda}_{EB}$ whose elements $\Lambda_{EB;jk}$ are given by

$$\Lambda_{EB;jk} = (F_{jk} + r_{jk}) / (F_{j+} + r_{j+}) , \quad j, k \in S .$$

5. OUR PROBLEM

Our interest is to apply the theoretical results of the previous section to some rainfall data compiled from the government publication Local Climatographical Data (1961-1977). Summer days (June through August) with a measurable amount of precipitation (that is, at least 0.01 inches) at Tallahassee were counted for the years 1961 through 1977 inclusive. A day with measurable precipitation is called a wet day. The sequence of wet and dry days is assumed to form a two-state stationary simple Markov chain. The frequency count matrix for each year is given in Table 1.

We assume there is independence between the different years.

TABLE 1.

Frequency count matrix of summer days precipitation at Tallahassee 1961-1977.

1961				1962				1963			
$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}
1	28	19	47	1	28	17	45	1	28	18	46
2	18	26	44	2	18	28	46	2	17	28	45
F_{+j}	46	45	91	F_{+j}	46	45	91	F_{+j}	45	46	91

1964				1965				1966			
$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}
1	27	13	40	1	30	16	46	1	41	17	58
2	14	37	51	2	15	30	45	2	15	18	33
F_{+j}	41	50	91	F_{+j}	45	46	91	F_{+j}	56	35	91

1967				1968				1969			
$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}
1	22	19	41	1	33	18	51	1	30	20	50
2	20	30	50	2	18	22	40	2	19	22	49
F_{+j}	42	49	91	F_{+j}	51	40	91	F_{+j}	49	42	91

1970				1971				1972			
$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}
1	28	17	45	1	17	17	34	1	41	16	57
2	18	28	46	2	16	41	57	2	15	19	34
F_{+j}	46	45	91	F_{+j}	33	58	91	F_{+j}	56	35	91

1973				1974				1975			
$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}
1	26	19	45	1	31	20	51	1	17	20	37
2	19	27	46	2	19	21	40	2	20	34	54
F_{+j}	45	46	91	F_{+j}	50	41	91	F_{+j}	37	54	91

1976				1977			
$t/t+1$	1	2	F_{j+}	$t/t+1$	1	2	F_{j+}
1	37	17	54	1	30	18	48
2	17	20	37	2	18	25	43
F_{+j}	54	37	91	F_{+j}	48	43	91

Since we have a two-state Markov chain, it is completely specified by just two parameters Λ_1 and Λ_2 , say, where

$$\Lambda_1 \equiv \Lambda_{11} = P\{X_t=1|X_{t-1}=1\} \text{ and } \Lambda_2 \equiv \Lambda_{21} = P\{X_t=1|X_{t-1}=2\}.$$

Then,

$$\underline{\Lambda} = \begin{pmatrix} \Lambda_1 & 1 - \Lambda_1 \\ \Lambda_2 & 1 - \Lambda_2 \end{pmatrix}.$$

Our objective is to estimate Λ_1 and Λ_2 for the year 1977, say, using the past data (years 1961-1976) and the current data (year 1977) according to the empirical Bayes procedure.

6. VERIFICATION OF THE ASSUMPTIONS

A crucial assumption is that the $\underline{\Lambda}_i$, $i = 1, \dots, n+1$, are independent and identically distributed, where in our case $n = 16$. In this example, it may appear that this assumption does not hold. However in personal communications, meteorologists Gleeson and Stuart at the Florida State University believe that due to shower activity in summer in the Tallahassee area, the degree of dependence between the $\underline{\Lambda}_i$, $i = 1, \dots, n+1$, from year to year is very small, if any.

To substantiate this belief of independence we applied a Runs Test. If there is not a pattern of variation among the $\underline{\Lambda}_i$, $i = 1, \dots, n+1$, they should fluctuate around their mean or median in a random manner. We first consider the signs of $\Lambda_1(i) - \bar{\Lambda}_1$. They are - + - +++ - + - + - + --- ++. Thus the sample size is $n = 17$, the number of runs is $u = 12$, the number of minuses is $n_1 = 8$, and the number of pluses is $n_2 = 9$. The null hypothesis H_0 is that the signs are randomly arranged. For these observations $P(U \leq 12 | H_0) = 0.939$ and $P(U \geq 9 | H_0) = 0.157$. Therefore, H_0 can not be rejected. Similarly, for $\Lambda_2(i) - \bar{\Lambda}_2$, we have + ---- + - ++ -- +++ - ++. In this case, $n = 17$, $u = 9$, $n_1 = 9$, $n_2 = 8$ and hence $P(U \leq 9 | H_0) = 0.50$. Thus again, H_0 cannot be rejected. Therefore, $\underline{\Lambda}_i$, $i = 1, \dots, n+1$, can be regarded as independent random variables.

In addition to the Runs Test we determined the maximum likelihood estimate $\hat{\Lambda}_1(i)$ and $\hat{\Lambda}_2(i)$ for each year $i = 1, \dots, n+1$, and plotted $\hat{\Lambda}_1(i-1)$ against $\hat{\Lambda}_1(i)$, as well as $\hat{\Lambda}_2(i-1)$ against $\hat{\Lambda}_2(i)$. These are shown in Fig. 1. There does not appear to be any detectable relationship between consecutive $\underline{\Lambda}_i$. Thus, these plots also suggest the assumption of independence of $\underline{\Lambda}_i$ for different years is valid.

Another assumption was that the Markov chain was stationary. To avoid lengthy computations we chose one year at random, 1967, to test for stationarity. The chain contained 92 observations and was broken into 6 consecutive pieces each of 15 days with the final 2 observations being ignored. The resulting frequency count matrices were obtained. The null hypothesis H_0 is that the chain is stationary, that is, $\Lambda(t) = \Lambda$, $t = 1, \dots, 6$. From Anderson and Goodman (1957), if H_0 is true the test statistic

$$Q = \sum_{t=1}^6 \sum_{j,k=1}^2 F_{jk}(t) \ln \hat{\Lambda}_{jk} / \hat{\Lambda}_{jk}(t)$$

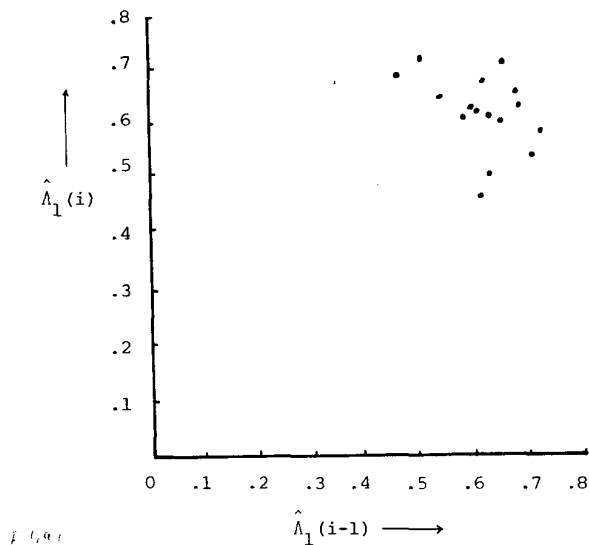


Fig.1(a). Plot of $\hat{\Lambda}_1(i)$ against $\hat{\Lambda}_1(i-1)$.

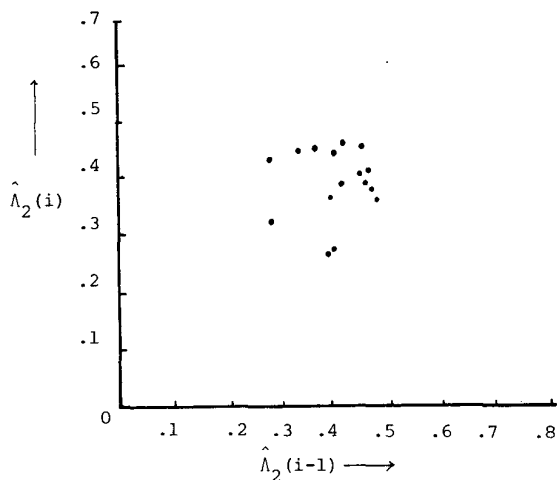


Fig.1(b). Plot of $\hat{\Lambda}_2(i)$ against $\hat{\Lambda}_2(i-1)$.

is asymptotically chi-square distributed with $d = (T-1)s(s-1) = 5 \times 2 \times (2-1) = 10$ degrees of freedom. Our observed value of Q upon substituting becomes 9.8934 while the tabulated value at the 5% level of significance is 18.307. Thus, H_0 can not be rejected and the chain is stationary.

Finally we test the assumption that the Markov chain is simple, that is, the order of dependency is one. We first test the null hypothesis H_0 that the chain is of order zero, that is, $\{X_t\}$, $t = 1, \dots, T$, are independent and identically distributed binomial variables against the alternative that the order of the chain is not zero.

A chi-square statistic

$$Q = \sum_{i=1}^{17} \chi_i^2$$

is used where χ_i^2 is the chi-square value within the i -th year and where Q has a chi-square distribution with 17 degrees of freedom. The observed value for Q is 78.77 while the tabulated value for the 5% level of significance is 27159. Thus, H_0 is rejected and the chain has an order of dependence greater than or equal to one.

We now test the null hypothesis H_0 that the order of the chain is one against the alternative hypothesis that it is two. The transition probability matrix for the second order chain can be expressed as

$$\begin{array}{c} \text{States at time } t-2, t-1 \\ \begin{array}{cc} 11 & 12 \\ 12 & 21 \\ 21 & 22 \end{array} \end{array} \begin{array}{c} \text{States at time } t-1, t \\ \begin{array}{cc} 11 & 12 \\ 21 & 22 \end{array} \end{array}$$

$$\begin{array}{cc} \begin{array}{cc} 11 & 12 \\ 12 & 21 \\ 21 & 22 \end{array} & \begin{array}{cc} \begin{array}{cc} \Lambda_{111} & \Lambda_{112} \\ 0 & 0 \end{array} & \begin{array}{cc} 0 & \Lambda_{121} \\ \Lambda_{211} & \Lambda_{212} \\ 0 & 0 \end{array} \\ \begin{array}{cc} \Lambda_{121} & \Lambda_{122} \\ 0 & 0 \end{array} & \begin{array}{cc} \Lambda_{221} & \Lambda_{222} \end{array} \end{array} \end{array}$$

The test statistic based on the likelihood ratio is

$$Q = \sum_{i=1}^{17} \sum_{j,k,l=1}^2 F_{i;jkl} \ln \hat{\Lambda}_{i;k} / \hat{\Lambda}_{i;jkl}$$

where Q is chi-square distributed with $17s(s-1)^2 = 34$ degrees of freedom. The observed value of Q is 34.55 while the tabulated value at the 5% level of significance is 48.57. Thus, H_0 cannot be rejected, that is, we can assume safely we have a first order Markov chain.

7. THE ESTIMATES

Once the basic assumptions have been verified, it is a simple matter to substitute the data values into the formulae of section 4. Thus, we obtain

$$\begin{aligned} \hat{\sigma}_{11}^* &= 474794 \times 10^{-8}, & \hat{\sigma}_{21}^* &= 353397 \times 10^{-8} \\ \hat{\sigma}_{11} &= 122905 \times 10^{-8}, & \hat{\sigma}_{21} &= 553235 \times 10^{-8}, \\ \bar{G}_1 &= 0.02186, & \bar{G}_2 &= 0.02307 \\ c_1 &= 3.8635, & c_2 &= 0.6388, \end{aligned}$$

and

$$r_{11} = 8.9633, \quad r_{12} = 0.$$

Therefore, $\Lambda_{EB;1} = 0.622$, $\Lambda_{EB;2} = 0.419$. That is,

$$\Lambda_{EB} = \begin{bmatrix} .622 & .378 \\ .419 & .581 \end{bmatrix}.$$

We could compare this empirical Bayes estimate with the maximum likelihood estimate for 1977 which is $\Lambda_{ML;1} = 30/48 = 0.625$, $\Lambda_{ML;2} = 0.419$, that is

$$\Lambda_{ML} = \begin{bmatrix} .625 & .375 \\ .419 & .581 \end{bmatrix}.$$

The advantage of the empirical Bayes estimate is that all the data from previous years is being used, that is, we are benefiting from the past experience whereas the maximum likelihood estimate is found by using the data of 1977 alone.

8. CONCLUSION

Once $\hat{\Lambda}$, the estimate of Λ , has been obtained there are many applications and quantities of interest that can be further estimated. One such example relates to the work of Gabriel and Neuman (1962) in which they used $\hat{\Lambda}$ to determine the distribution of weather cycles. For this purpose a wet(dry) spell of k days is defined as a sequence of k wet(dry) days preceded and followed by a dry(wet) day. Let W denote the length of a wet spell. Then,

$$P(W = k) = \hat{\Lambda}_2(1 - \hat{\Lambda}_2)^{k-1}, \quad k = 1, 2, \dots$$

A weather cycle is defined as combinations of a wet spell and an adjacent dry spell. Let C denote the length of a weather cycle. Then

$$P(C = m) = m\hat{\Lambda}_2(1 - \hat{\Lambda}_2)/(1 - \hat{\Lambda}_1 - \hat{\Lambda}_2)(1 - \hat{\Lambda}_2)^{m-1} - \hat{\Lambda}_1^{m-1}, \quad m = 1, 2, \dots$$

We refer the reader to the paper cited for more applications of this type.

REFERENCES

- Anderson, T.W. and Goodman, L.A., 1957. Statistical inference about chains. AMS 28: 89-110.
- Bartlett, M.S., 1951. The frequency goodness of fit test for probability chains. Proc. Cambridge Phil. Soc. 47:86-95.
- Billard, L. and Meshkani, M.R., 1978. Empirical Bayes estimation for the multinomial distribution. Florida State Univ., Dept. Stat. Tech. Rep. M475.
- Billingsley, P., 1961. Statistical Inference for Markov Processes. Univ. Chicago P., Chicago.
- DeGroot, M.H., 1970. Optimal Statistical Decisions. McGraw-Hill, New York.
- Feller, W., 1968. An Introduction to Probability Theory and Its Applications. Vol. 1, 3rd ed., John Wiley, New York.
- Gabriel, K.R. and Neuman, J., 1962. A Markov chain model for daily rainfall occurrence at Tel Aviv. Quart. J. Roy. Met. Soc. 88:90-95.
- Martin, J.J., 1967. Bayesian Decision Problems and Markov Chains. John Wiley, New York.
- Whittle, P., 1955. Some distribution and moment formulae for Markov chain. JRSSB17:235-42.

ON RED NOISE AND QUASI-PERIODICITY IN THE TIME SERIES OF ATMOSPHERIC TEMPERATURE

O.M.ESSENWANGER

DRSMI-TRA(R&D),Tech.Lab.MICOM, and Univ. of Alabama in Huntsville, Alabama

ABSTRACT

Essenwanger, O.M., On red noise and quasi-periodicity in time series of atmospheric temperature. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

The author has developed a methodology which permits the separation of the time series into three components: cycles, quasi-cycles, and red (or white) noise. The method utilizes power spectrum and Fourier analysis which is economically feasible for large amount of data if one uses the algorithm of the Fast Fourier Transform. The process which is based on the utilization of statistical significance criteria for testing the amplitudes of the Fourier series is described in detail.

After separation into the three components, the red noise is based on a modified first lag correlation. The method is illustrated with six-hourly observations of the temperature for five consecutive years at four stations from typical climatic regimes. While the red noise share is largest in the tropics, the quasi-cycles contribute more to the variance in the subtropical and temperature zones than in the polar or tropical regions.

1. INTRODUCTION

Various tools are available for the statistical modelling of atmospheric time series. E.g., the derivation of autoregressive, moving average of mixed models (if applicable) is one alternative for an analytical representation. In atmospheric science spectral analysis is in widespread use. In the latter, however, one is soon confronted with the reality of meteorological quasi-cycles which is an intricate and complex problem. The problem has not been solved in the past (e.g., Bartels (1943), Brier et al. (1964), Craddock (1965), Shapiro (1975) ; etc.) and may elude an easy answer, at least in part, in the near future.

The author has attacked the problem in spectral analysis by examining the noise. We may consider the atmospheric time series as consisting of three components: cycles, quasi-cycles, and noise (white or red). It is not difficult to recognize true cycles in statistical significance testing. If we were able to deduce the noise we could separate the three components because the quasi-cycles would emerge as the remainder after cycles and noise have been identified. Thus, this article serves two purposes.

First, a methodology is developed to separate cycles, quasi-cycles and noise based on statistical significance testing. After elimination of cycles and quasi-cycles the residual variance is noise (red or white) which is predictable only in statistical terms. Quasi-cycles also may not be predictable in a strict sense but

the knowledge of their existence may be useful.

Second, the methodology is illustrated with an analysis of the temperature time series in four typical climatic zones. Some quasi-cycles in the periodogram withstand rigorous statistical testing and can be explained as being associated with synoptic scale phenomena. Other high amplitudes of statistical significance may be related to the asymmetry of meteorological cycles, the seasonal variation of the amplitude of cycles, or side lobes.

2. SPECTRUM ANALYSIS

Usually different statistical techniques are available for spectral analysis : Fourier series, power spectrum and periodogram. Cycles are expressed in terms of sine waves although they may not be precisely sine curves. Deviations will lead to more than one sine wave for one cycle.

The Fourier series for a set of data y_t , $t = 1, \dots, N$, is defined by

$$y_t - \bar{y} = \sum A_j \sin(j\alpha_t + \psi_j) \quad (1a)$$

where usually $j = 1, \dots, N/2$ and

$$\alpha_t = 2\pi t/L \quad (1b)$$

where L designates the basic period, generally comprising all data N so that $L \equiv N$. $\theta = 2\pi j/L$ is the angular frequency, $\lambda = 2\pi/\theta$ the wave length, A_j the amplitude, and ψ_j the phase angle. While the Fourier series and the power spectrum are usually plotted against the angular frequency θ or the wave number j , the periodogram is based on the wave length λ . The ordinates reflect the variability of A or the squared amplitude A^2 . These ordinates may be utilized either in their original or in standardized form by division with $2\sigma^2$ (power spectrum).

Although in the past Fourier series and the periodogram required elaborate calculations, the computational effort has been reduced today by the Fast Fourier Transform (see e.g., Cooley and Tukey (1965), Bloomfield (1976), etc.) and the availability of electronic data processing systems. More details can be found in the literature, such as Kendall and Stuart (1966), Taubenheim (1969), Kendall (1973), Bloomfield (1976), etc. In the subsequent method of separating temperature data into three components, power spectra and Fourier analyses are utilized.

In the spectral analysis of atmospheric cycles the daily or annual cycles can be spotted easily in most cases although the precise mathematical formulation in terms of the Fourier series is not always simple. Consideration must be given to "leakage", side lobes, asymmetry and the modulation of the amplitude during the seasons. E.g., let us assume a daily cycle (subscript d):

$$y_t = A_j \sin(j_d \alpha_t + \psi_d) \quad (2a)$$

with a seasonal fluctuation. Then the amplitude varies, e.g.,

$$A_j = B_j + D_j \sin(j_s \alpha_t + \psi_s) \quad (2b)$$

with α_t (eqn.(1b)), $t = 1, \dots, N$. This type of "modulation" of the amplitude is reflected in the spectrum. We can combine (2a) and (2b):

$$y_t = B_j \sin(j_d \alpha_t + \psi_d) + D_j \sin(j_d \alpha_t + \psi_d) \sin(j_s \alpha_t + \psi_s) . \quad (3)$$

While the first term is a sine wave with periodicity $p = 2 j_d / N$, the second term in eqn. (3) resembles:

$$2 \sin(A + B) \cos(A - B) = \sin(2A) \pm \sin(2B) . \quad (4)$$

Consequently, at wave number j_d only the amplitude B (or B^2) will appear in the spectrum. Depending on j_s the amplitude $D^2/4$ will occur at $j_d \pm j_s$ (see e.g., Table 1). Under these circumstances the total contribution of the daily cycle to the variance would be $B^2 + D^2/2$ but at the exact wavelength of the daily cycle only the fraction

$$A_d^2 = B^2 / (B^2 + D^2/2) \quad (5)$$

can be found. Thus, a modulation of the amplitude (e.g., during the seasons) is simulated in the Fourier series by amplitudes at several waves. It is especially important to remember these peculiarities of the mathematical tools for the quasi-cycles, which may be classified as cycles that last for a few repetitions and/or show modulated amplitudes and then disappear. When they reoccur later with a different phase angle ψ the amplitude is reduced, too. Since the power spectrum is independent of the phase angle, quasi-cycles may be reflected better in the power spectrum than in the Fourier analysis or periodogram unless the data series is broken into subparts. Furthermore, meteorological cycles may not produce a plain sine wave but may display asymmetry which adds to the complexity of the analysis and separation problems because several terms of the Fourier series are needed to provide a mathematical approximation (representation) of the cycle.

TABLE 1.

Example of spectral amplitude for daily cycle of the form of equation (1) and (2), $j_d = 360$, $j_s = 2$, $N = 1440$, $B = 3$, $D = 3$.

j	A_j^2	%
358	2.25	16.7
359	0.00	00.0
360	9.00	66.7
361	0.00	00.0
362	2.25	16.7
E	13.50	

3. RESIDUAL ERROR

Under the assumption that cycles (C_j) and/or quasi-cycles (Q_j) are present in atmospheric time series we can formulate the data series y_t , $t = 1, \dots, N$ by the mathematical expression

$$y_t - \bar{y} = \sum_{j=1}^{n_1} C_j(\theta) + \sum_{j=n_1+1}^{n_2} Q_j(\theta) + \varepsilon_t. \quad (6)$$

Usually the number of cycles (n_1) and quasi-cycles ($n_2 - n_1$), $n_1 \leq n_2$, is not known a priori and ε_t is a product of random errors which may or may not include random instrumental errors. For modelling purposes this error ε_t may be negligible in many cases, but its consideration is non-negligible in atmospheric time series although the error may prove statistically significant. Furthermore, it is assumed that the time series is stationary.

As mentioned before, $C_j(\theta)$ and $Q_j(\theta)$ can be expressed in terms of a Fourier series while ε_t would be formulated either as white or red noise. White noise requires that all amplitudes in the spectrum be of equal size and the phase angles follow a rectangular distribution. Red noise shows some typical pattern in the distribution of the amplitudes (see, e.g., later eqn. (7)). Thus, the residual error ε_t would disclose one of these patterns.

The most common red noise in meteorology follows a plain exponential sequence in the autocorrelation series:

$$\rho_t = \exp(-bt) \quad (7)$$

with $t \geq 0$, $b > 0$. This exponential series is identical with the first Markov chain (see e.g., Box and Jenkins (1970), etc.):

$$\rho_k = \rho^k \quad (8)$$

and $\rho_1 = \rho$, where ρ_1 is the first lag correlation. Thus, $b = \ln \rho = \ln \rho_1$, $\rho_1 > 0$.

The power spectrum for the exponential red noise can be written (e.g., see Gilman et al. (1963)):

$$L_1 = [(1-\rho)/(1+\rho^2-2\cos k\pi/m)]/m \quad (9)$$

where m is the number of lags. In turn,

$$\rho_t = \sum_{k=1}^m L_k \cos(tk\pi/m). \quad (10)$$

Since atmospheric time series may comprise a mixture of cycles, quasi-cycles and a remaining error, the autocorrelation function can be considered to be a composite of several terms such as

$$\rho_t = \sum_{j=1}^{n-1} w_j \rho_{jt} + w_n \rho_{\varepsilon t} \quad (11)$$

where the weights $\Sigma w_j = 1$, and $\rho_{\epsilon t}$ is the correlation of the residual error component. The summation term in eqn. (11) is composed of the cycles or quasi-cycles in accordance with the Fourier transform (e.g., see Kendall (1973)) of the correlation coefficient

$$\rho_t = \Sigma C_j \exp(-itj). \quad (12)$$

with $i = \sqrt{-1}$. The validity of eqn. (11) can be demonstrated (see Essenwanger (1977,1979)).

4. TESTING FOR STATISTICAL SIGNIFICANCE

If the reality of cycles or quasi-cycles were known a priori no test for statistical significance would be necessary. Unfortunately, random processes in meteorological time series can also produce large amplitudes in the spectrum without physical reality. E.g., Kendall and Stuart (1966) assume that the distribution of the squared amplitudes follows an exponential law. Consequently, the probability that a threshold A_{th}^2 is exceeded is:

$$P(A_{th}^2 \geq 4\sigma^2 k/N) = \exp(-k). \quad (13)$$

In many cases, especially when the number of independent amplitudes is large and the probability is small, Walker (1914) found the probability from eqn. (13) to be incorrect. The probability that one of n independent amplitudes exceeds the threshold A_{th}^2 according to Walker is

$$P(A_{th}^2 \geq 4\sigma^2 k/N) = 1 - (1 - e^{-k})^n. \quad (14)$$

Bloomfield (1976) expresses the probability that the largest squared amplitude I_{max} among n independent amplitudes exceeds the threshold $k_2 + \ell n n$ as

$$P(I_{max} \geq k_2 + \ell n n) = 1 - \exp(-e^{-k_2}). \quad (15a)$$

This test is related to Fisher's statistic G_n :

$$P(nG_n \geq k_2 + \ell n n) \approx \exp(-e^{-k_2}) \quad (16)$$

with

$$G_n = A_{max}^2 / \sum_{j=1}^n A_j^2 = (k_2 + \ell n n) / n. \quad (17)$$

Consequently:

$$P(A_{max}^2 \geq 4\sigma^2 h^2/N) \approx 1 - \exp(-e^{-k_2}) \quad (15b)$$

where $h^2 = k_2 + \ell n n$.

Brooks and Carruthers (1953) recommended testing by

$$A_{th}^2 \geq \sigma^2 h_A^2 / N \quad (18)$$

where $h_A^2 = 4 \ln(n/P)$. this leads to the same threshold as eqn. (15b) because $h^2 \sim \ln(n/P)$ for $P \leq 0.05$. For $P = 0.05$, $N = 1440$, $n = 720$ we obtain the same test threshold $G_n = 1.33\%$ from (14), (15b), (16) and (18). $A_j^2 \geq G_n$ would be significant. Only (13) provides a smaller value: $A_{th}^2 = 0.83\%$. This value stems from the exponential distribution. A 5% exceedance means that 5% of the (squared) amplitudes, i.e., 36, can be higher than A_{th}^2 . This is a different test basis and no contradiction to eqn. (17) or (15a).

Another test procedure which is based on the principle of the analysis of variance was developed by Hartley (1949). The test statistic is the F-test, but the calculations for the test are more elaborate than for eqn. (13)-(18) because the individual amplitudes A_j^2 enter. The reader is referred to the quoted reference.

The criteria which were discussed above do not take into consideration persistence. Stumpff (1937) has shown that the threshold A_{th}^2 should be revised to A_P^2 , incorporating persistence by including the (smoothed) autocorrelations:

$$A_P^2 = A_{th}^2 \left(1 + 2 \sum_{i=1}^{n-1} r_i w_i \cos i\theta \right) \quad (19)$$

where the weights are $w_i = (1-i)/N$, and $\theta = 2\pi j/N$ designates the angular frequency. The author (1950) has shown that the contribution of the term in parenthesis for daily pressure values of Europe may amount to a factor of four to five. In our case this adjustment is not necessary because we can include persistence in the testing of the power spectrum.

It can be shown that $2\pi A^2/\sigma^2$ has a χ^2 distribution with 2 degrees of freedom (see e.g., Kendall and Stuart (1966)). In the power spectrum usually we test:

$$R_a = L_j/L_E = \chi^2/\nu \quad (20)$$

Blackman and Tukey (1958) deduce that $\nu = 2N/m - 2/3$, where m is the number of lags. L_E is the expected value of the power spectrum. For white noise we would set $L_E = \bar{L}$. More frequently L_E is assumed to be the smooth spectrum. Several smoothing formulae have been suggested (see e.g., Kendall (1973), p.110ff).

In our case L_E is replaced by the red noise spectrum. Because L_j/L_E can be either greater or smaller than unity both tailends must be tested.

Sneyers (1975,1976) has recently suggested testing the statistical significance of the residual errors ϵ_t after selecting the first, second, etc. harmonic components of the Fourier analysis. The Fourier series is discontinued after the residual errors ϵ_t fall below the predetermined level of significance. This selection of subsequent steps in order of the harmonics is rigorously valid if the size of the amplitudes decreases with increasing order of the harmonics such as in the example

by Sneyers for the annual cycle in Uccle (Belgium). Otherwise, the harmonics must be picked in sequence of the size of the amplitude. If the latter selection procedure is chosen the result should resemble the author's method by which an iterative testing by size of the amplitudes is performed (see details in Section 5). Because Sneyers methodology requires the recalculation of ϵ_t for every added harmonic, the computational efforts become quite elaborate for larger data samples N and number of harmonics $N/2$. Thus, the author preferred testing of the individual amplitudes rather than the residual errors ϵ_t as the noise in the time series implies that these residuals are statistically insignificant.

5. SEPARATION OF NOISE AND CYCLES

The author's suggested method of separating cycles and noise requires two phases: first, testing of the amplitudes of the Fourier series; second, testing of the power spectrum.

The testing of the amplitudes of the Fourier series is based on eqn. (17). As recommended by Brooks and Carruthers (1953), an iterative process is used by which the statistically significant maximum amplitude is eliminated. The maximum of the remaining amplitudes is then retested until no significant amplitude is left. This process is somewhat cumbersome for a large number of significant amplitudes and can be abbreviated by extracting more than one amplitude during one step. Let us denote the iterative threshold by $G_{n\tau}$. Then:

$$G_n \approx G_{n1} [s^2 - \sum_{k=1}^{\tau-1} A_k^2 / 2] / s^2 . \quad (21)$$

Because G_n increases only very slowly for large N , $G_n \leq G_{n\tau-1}$. Consequently, replacing $G_{n\tau}$ by $G_{n\tau-h}$, $\tau > h \geq 1$, does not add components which would not be selected by the detailed process.

The procedure is illustrated in Table 2. The first threshold $G_{n1} = 1.33\%$ of $2\sigma_y^2$ (see Section 4). Three amplitudes in the Fourier series, $n = 720$, exceeded 1.33% . These were the amplitudes of the wave number $j = 2, 360$ and 720 , or periods of 180, 1 and 0.5 days. The significant waves in this first round are considered as the true cycles and proved to be the yearly, daily and semidaily cycle. They amount to 76.1% of the variance (see Table 2). Consequently, $G_{n4} \approx G_{n1} \cdot 23.9\% = 0.319\%$ of $2\sigma_y^2$. This threshold appears very low, and at first one may expect at this low level that $A_k^2 > G_{n4}$ for many amplitudes. This is not the case, however, as the following calculations demonstrate. For $n = 720$ the average (squared) amplitude would amount to $1/720 = 0.14\%$ of $2\sigma_y^2$ but 717 amplitudes contribute only 23.9% . Thus, the average reduces to $23.9/717 = 0.033\%$ which is about one-tenth of G_{n4} . This is precisely what can be observed in the spectrum of the 717 remaining amplitudes. In fact, 69% of the amplitudes stay under 0.033% although under

the assumption of an exponential distribution (see Essenwanger (1976, p.113)) we would expect only 60% .

By lowering the threshold to 0.319% only six more amplitudes were added, $j = 1, 14, 25, 358, 359$ and 361 , with a total of 3.51% . This leads to $G_{n10} = 0.274\%$. Four more waves ($j = 3, 5, 53, 362$) were found to exceed this third threshold . The test criterion reduces to $G_{n14} = 0.255\%$, but no other amplitude was larger than G_{n14} . It should be noticed (Table 2) that the fourteen amplitudes comprise the annual, daily and semidaily cycles, $j = 4$ was added for continuity, and the remaining three amplitudes may imply the existence of quasi-cycles.

TABLE 2.

Example of the procedure to separate cycles and noise.

Example of the procedure to separate cycles and noise:											
Cycles j	Power spectrum L_j L_j		j_F	Selection of significant cycles				L'_j	Adj. to 100% L_j	Red Noise L_{Rj}	Ratio L_j/L_{Rj}
				First step		Third step					
	j_F	per cent		$G_{n14}=0.274\%$ j_F	%						
0	5.9	5.9%	0 to 20	2	2.47%	1 - 5	3.84	1.7%	8.9%	3.7	2.4
1	2.7	8.6	21 - 60			14	0.38	1.9	9.9	7.3	1.4
						25	0.46				
						43	0.31				
2	1.4	10.0	61 - 100					1.4	7.4	7.2	1.0
3	1.1	11.1	101 - 140					1.6	5.7	7.0	0.8
4	1.0	12.1	141 - 180					1.0	5.2	6.7	0.8
5	0.6	12.7	181 - 220					0.6	3.1	6.5	0.5
6	0.8	13.5	221 - 260					0.8	4.2	6.2	0.7
7	0.9	14.4	261 - 300					0.9	4.7	5.9	0.8
8	0.9	15.3	301 - 340					0.9	4.7	5.6	0.8
9	67.9	83.2	341 - 380	360	64.51	358 - 362	66.72	1.2	6.2	5.4	1.1
10	1.4	84.6	381 - 420					1.4	7.4	5.1	1.5
11	1.2	85.8	421 - 460					1.2	6.2	4.9	1.3
12	0.7	86.5	461 - 500					0.7	3.6	4.7	0.8
13	0.6	87.1	501 - 540					0.6	3.1	4.5	0.7
14	0.7	87.8	541 - 580					0.7	3.6	4.4	0.8
15	0.7	88.5	581 - 620					0.7	3.6	4.3	0.8
16	1.0	89.5	621 - 660					1.0	5.2	4.3	1.2
17	0.9	90.4	661 - 700					0.9	4.7	4.2	1.1
18	9.6	100.0	701 - 720	720	9.08	720	9.08	0.5	2.6	2.1	1.2
Total					76.06%	-	80.79%	19.2%	-	-	-

j_F is used for the Fourier series in order to distinguish between power spectrum and Fourier series. χ^2 test criterion for L_j/L_{Rj} and 160 degrees of freedom at 0.5% and 99.5% is 0.67 to 1.41. The cycle length = $1/j$ days or $360/j_F$.

After completion of the selection of amplitudes from the Fourier series we subtract the cycles (quasi-cycles) from the y_t series to obtain ϵ_t . Then the power spectrum for ϵ_t is calculated. This power spectrum should now resemble white or red noise. The power spectrum (L_j) of ϵ_t is displayed in Table 2 in the third column from the right.

The testing of the power spectrum may now take place with a more stringent significance level, e.g., $P = 0.99$. We find the χ^2 boundaries for $\nu = 160$ as $0.67 \leq R_a \leq 1.41$. Let us first consider the assumption of white noise. $\bar{L} = 1/18 = 5.56\%$,

and $3.72 \leq L_j \leq 7.74\%$. Several L_j value fall outside the boundaries. Thus, we may reject the hypothesis that the spectrum of ϵ_t is white noise.

To confirm the decision we may test r_1 , the first lag correlation from ϵ_t , y_t after elimination of the cycles. We test whether r_1 is significantly different from $\rho_1 = 0$. The original first lag correlation of y_t was $r_1 = -0.01$ due to the strong influence of the daily cycle (see eqn. (11)). After elimination of the cycles (third step, Table 2) the first lag correlation coefficient for ϵ_t changed to $r_1 = 0.14$. This correlation is significantly different from zero. (The upper significance boundary is $r_{0.025} = 0.052$, $r_1 > r_{0.052}$.) Consequently, the existence of a red noise spectrum is more likely than white noise.

The red noise spectrum L_{Rj} based on r_1 for ϵ_t and the ratio $R_a = L_j/L_{Rj}$ are listed in the last two columns of Table 2. We find three classes outside the boundaries, i.e., $j = 0, 5, 10$. Several alternatives for a correlation may apply.

First, we may consider that the data series of ϵ_t is only a fraction of the observations. Thus, we should test only a fraction of N . The original number of degrees of freedom $v = 160$ may be too high. E.g., instead of N , we may use 19.2% of N . This reduces v to 30 and expands the boundaries to $0.36 \leq R_a \leq 2.07$. Only the spectral value $j = 0$ lies outside these boundaries. The procedure leads to the re-examination of white noise because the boundaries for L_j/\bar{L} would be equally expanded. However, the first lag correlation $r_1 = 0.14$ is still significant even at the reduced level of N , ($r_1 > 0.12$).

A second alternative is a substitution of L_j by the smooth spectrum \tilde{L}_j . Under utilization of the "Hamming window" (see Taubenheim (1969, p.289) or Blackman and Tukey (1958)), the new ratios L_j/L_{Rj} for $j = 0, 5, 10$ are now 2.5, 0.6, 1.3. This brings the ratios within the boundaries except for $j = 0$. Since L_{10} is not significantly above \tilde{L}_{10} we may be satisfied. Similar arguments are valid for L_5 . This leaves only L_0 .

The third option is a correction by an addition or dropping of amplitudes which fall within the spectral classes. We would augment the 14 amplitudes selected in phase one by adding the largest, second largest, etc., amplitude of the Fourier series within the spectral class. E.g., in our case for $j = 0$ the class comprises j_F from 1 to 20. The largest remaining amplitude was 0.22% for $j = 6$. We need to reduce $L_0 = 1.7\%$ by almost 0.7%. We continue with the next highest amplitudes which were $j = 16$ and 18. They provide another 0.36%. Finally, we add $j = 7$ with 0.15%. It should be noticed that in this case we have not created additional "cycles" but have only expanded existing cycles. The annual wave has been augmented to $j_F = 7$; $j_F = 1$ to 7 is now 4.21%. The quasi-cycle $j_F = 14$ comprises now $j_F = 14$ to 18 with 0.83%.

If we had decided that L_{10} were significantly above L_{R10} we would have added waves in the class $j = 11$ with $j_F = 380$ to 420. The maximum amplitudes would occur at $j_F = 385$ with 0.20%. This addition to the cycles would bring the test ratio

within the boundaries. Although the amplitude 0.20% below the last G_n criteria justification for selection is the peak in the power spectrum. Because the power spectrum is independent of the phase angle quasi-periodicity is better reflected by this tool.

If the cycle or quasi-cycle at $j_F = 385$ is real it should reappear in a subdivision of the data. In both half-years (15 July 55 - 11 Jan 56 and 12 Jan 56 - 10 July 56) a maximum around 192-193 can be found in the spectrum (0.35% at $j_F = 192$, first 180 days, and 0.31% at $j_F = 194$, second 180 days). These amplitudes stay below the last iterative significance threshold ($G_n = 0.61\%$ or 0.41%) of the half-year selection although they place peak amplitudes. We would need to search for significance by further subdivision or we consider the maximum as a side lobe of the daily cycle. For the subsequent tabulations the author has decided on the second choice, and only the selections as given in Table 2, step 3, were adopted.

In order to correct a ratio which is too low we have two options. First, we may drop some amplitudes within the spectral class. If this is not possible (such as in our case) we may add amplitudes in other classes or augment cycles, etc.. This supplementation in other spectral classes will reduce the share of the total noise contribution to the variance and will raise the low L_j in the new spectrum. Usually after the waves have been selected in the first phase, the first lag correlation of ϵ_t changes only slightly in the adjustment process during the second phase, and the red noise spectrum remains approximately the same.

More sophisticated procedures could be engaged, such as a prorating of waves into shares of cycles and red noise, but the computational efforts for this sophistication are disproportionate to the improvements or changes of L_j and L_{R_j} .

One may object to the correction procedure and the augmentation of the originally selected number of waves because one could also continue this procedure until the residual error is white noise. This argument is factually correct, and in the case of the example illustrated in Table 2 the selection of a few more waves would leave a white noise spectrum. The drawback is, however, that we may add too many unreal "quasi-cycles" which are only produced by random processes and persistence. The latter is taken into account with red noise, however.

In the example of Albroom the difference between white and red noise is very small. This was one of the reasons for choosing this example for illustration. The difference is much larger in other climates as displayed by the second example for 6-hourly data at Huntsville, 15 July 1959 to 10 July 1960 (Figure 1).

The observant reader may notice that the annual and daily waves which are listed as being eliminated in the center part of Figure 1 amount to 73.2%. In Table 3 the share of these cycles adds up to 80.0%, however. This apparent discrepancy can be explained. In the construction of Figure 1 (center part) only the amplitudes for $j = 1$ and $j = 360$ were eliminated, amounting to 73.2%. In the final analysis (Table 3) the annual wave comprises the amplitudes for $j = 1-6$, and the daily cycle

$j = 359-361$ which sum up to 80.0%.

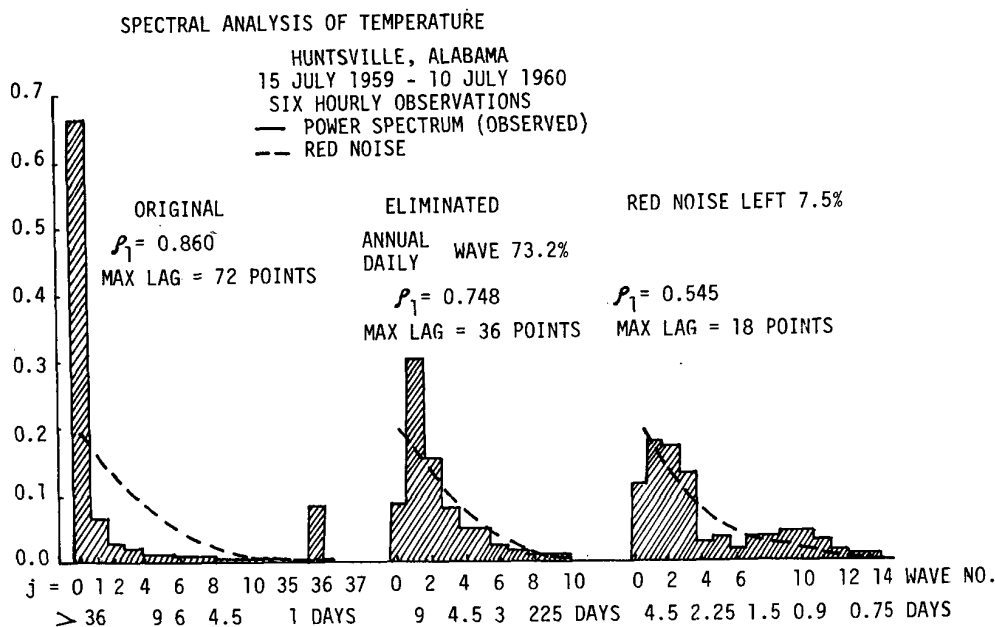


Fig. 1. Spectral analysis of temperature

6. CYCLES, QUASI-CYCLES AND NOISE

The procedure of selecting significant amplitudes has been applied to 6-hourly temperature data of five consecutive years from four typical climates. For the simplicity of the computer program 360 days were chosen as the basic period (Fast Fourier Transform) but results from the exact 365 days disclosed little change. Table 3 summarizes the contributions to the variance by the yearly, daily and semi-daily cycles, the quasi-cycles and the noise.

Although it is common knowledge that the daily cycle dominates in the tropics (Albrook) and the annual cycle in the other climatic regions, the amount of the contribution to the total variance of the data may not be readily available. In the tropical area the high share of red noise may be a surprise but quasi-cycles are of little practical consequence. As expected, a strong semidiurnal cycle exists in the tropics. Its existence proves to be statistically significant in the subtropical and temperature zones, albeit with a much smaller contribution. The highest percentage of quasi-cycles can be found in the subtropical (Huntsville) and the temperate (Frankfurt) zones, but the strong showing in the polar region in some years may be somewhat of a surprise.

TABLE 3.

Fractional share of cycles, quasi-cycles and red noise.

Albrook (Canal Zone)						
Year	1955	1956	1957	1958	1959	
Mean	26.1	26.7	27.1	26.7	26.4	degree C (degree C) ²
Variance	8.0	9.5	8.6	8.4	8.2	
Annual Cycle	3.8	2.4	2.6	1.4	1.5 %	
Daily Cycle	66.7	69.9	67.0	67.8	67.4 %	
Semi-Daily Cycle	9.1	9.8	11.6	10.5	8.9 %	
Quasi Cycle	1.2	0.6	1.3	1.6	0.5 %	
Red Noise	19.2	17.3	17.5	18.7	21.7 %	

Huntsville (Alabama)					
Year	1959	1960	1961	1962	1963
Mean	15.6	15.4	16.1	15.7	17.3
Variance	107.4	92.8	96.9	113.2	104.9
Annual Cycle	71.4	66.8	66.1	68.8	70.7
Daily Cycle	8.6	11.4	10.3	9.2	10.6
Semi-Daily Cycle	0.4	0.6	0.5	0.5	0.5
Quasi-Cycle	12.1	11.9	14.6	11.6	9.9
Red Noise	7.5	9.3	8.5	9.9	8.3

Frankfurt (F.R.Germany)					
Year	1958	1959	1960	1961	1962
Mean	10.2	10.3	9.8	8.9	7.5
Variance	63.8	65.6	48.1	63.3	95.0
Annual Cycle	70.6	65.9	64.6	63.7	77.3
Daily Cycle	11.5	13.4	12.4	9.6	8.5
Semi-Daily Cycle	0.3	0.2	0.2	0.2	0.2
Quasi-Cycle	10.4	13.6	11.8	18.4	7.9
Red Noise	7.2	6.9	11.0	8.1	6.1

Barrow (Alaska)					
Year	1958	1959	1960	1961	1962
Mean	-11.9	-13.4	-14.1	-11.4	-11.1
Variance	174.5	175.6	164.8	143.6	169.0
Annual Cycle	91.9	87.7	91.9	83.0	86.2
Daily Cycle	0.6	0.4	0.5	0.4	0.3
Quasi-Cycle	3.4	8.3	5.0	10.9	8.8
Red Noise	4.1	3.6	2.6	5.7	4.7

The statistical criteria utilized in the selection of significant waves are found in Table 4. The first row for every station after the designation of the year provides the last threshold of G_n separating significant and insignificant amplitudes. Although this value may appear to be very low, especially at Barrow, the threshold must be brought into perspective with its association of the majority of amplitudes. First we notice that everywhere the (lowest) test criterion G_n is about ten times as high as the average of the remaining amplitudes (third row, Table 4). Let us examine the lowest threshold, $G_n = 0.05\%$, at Barrow (1960). A frequency distribution of the (squared) amplitude revealed that the median (squared) amplitude

TABLE 4.

Quasi-cycle selection.

Albrook						Huntsville				
Year	1955	1956	1957	1958	1959	1959	1960	1961	1962	1963
$A_{th}^2/2\sigma^2$	0.26%	0.24%	0.24%	0.25%	0.29%	0.12%	0.13%	0.13%	0.14%	0.12%
n_A	13	11	12	11	8	53	45	53	41	40
n_C	9	8	7	7	7	10	9	10	8	9
$\bar{A}_R^2/2\sigma^2$	0.027%	0.024%	0.025%	0.026%	0.031%	0.011%	0.014%	0.013%	0.015%	0.012%
r_1	-0.01	-0.03	-0.04	-0.03	-0.02	0.86	0.82	0.84	0.85	0.84
r_{RN}	0.14	0.14	0.13	0.16	0.09	0.54	0.55	0.54	0.58	0.58

Frankfurt						Barrow				
Year	1958	1959	1960	1961	1962	1958	1959	1960	1961	1962
$A_{th}^2/2\sigma^2$	0.11%	0.11%	0.16%	0.12%	0.09%	0.08%	0.07%	0.05%	0.08%	0.07%
n_A	40	34	38	48	35	42	46	56	41	53
n_C	10	12	10	9	12	19	10	19	16	15
$\bar{A}_R^2/2\sigma^2$	0.011%	0.010%	0.016%	0.012%	0.009%	0.006%	0.005%	0.004%	0.008%	0.007%
r_1	0.83	0.82	0.80	0.85	0.88	0.98	0.98	0.98	0.97	0.98
r_{RN}	0.51	0.51	0.59	0.54	0.48	0.78	0.84	0.83	0.81	0.86

 A_{th}^2 = lowest significant threshold n_A = number of amplitudes $> A_{th}^2$ \bar{A}_R = average of remaining amplitudes n_C = number of amplitudes for annual, daily, semi-daily cycle

TABLE 5.

Examples of quasi-cycles.

Albrook 1958		Huntsville 1961		Frankfurt 1961		Barrow 1961	
Cycle (days)	PV	Cycle (days)	PV	Cycle (days)	PV	Cycle (days)	PV
45-51	1.0%	36-50	1.5%	33-60	5.1%	14-21	5.8%
6-8	0.6%	24-30	2.1%	17-26	4.5%	12-13	1.5%
		13-18	3.9%	13-16	3.0%	9-10	2.2%
		9-11	3.5%	11	0.6%	6-7	1.1%
		6-7	1.2%	7-10	4.3%	5-5.5	0.3%
		4-5	2.4%	4-6	0.5%	4	0.1%
				3-2	0.4%		

PV = percentage of variance

was $\leq 0.001\%$. This would lead to a standard deviation of 0.0014% in an exponential distribution. Under the postulation of an exponential distribution and $\sigma = 0.002$, i.e., a standard deviation even higher than the one deduced from the median value, only 6% of the amplitudes are expected to exceed 0.005% which is one-tenth of the

threshold. The criterion of 0.05% corresponds to an expectation of less than 1 value in 10^8 data points in an exponential distribution. For 715 amplitudes the moments fit leads to $\sigma = 0.013\%$ which would be one-fourth of this lowest $G_n = 0.05\%$. Even then only 1.5% of the amplitudes, or 10, are normally expected to exceed this threshold. Thus, the significance threshold is not so low as it appears from its numerical value. Besides, the numerical value of the amplitude for 0.05% is $A = 0.4^\circ\text{C}$ which is twice as large as the selection criterion $G_n = 0.25\%$, i.e., $A = 0.2^\circ\text{C}$, in the tropics.

The derived quasi-cycles cannot be totally listed in the frame of this article but one example for the year with the highest contribution is given for every station (Table 5). Although the individual amplitudes must fulfill the individual test criterion they can later be lumped together whenever they show consecutive wave numbers thus forming a "window" or resembling a "filter band". This combination is justified because the seasonal variation of the amplitudes and/or the change in the length of periodicity are reflected in the spectrum by a spreading over several Fourier terms which are connecting wave numbers in the majority of cases. In order to take conditions into account which are similar to the illustration in Table 1, occasionally the bandwidth needs to be expanded by connecting isolated waves or adding too separate bands with two or three adjacent amplitudes. Therefore, the number of selected amplitudes is not identical with the number of separate quasi-cycles. In the present analysis the number of quasi-cycles was held to seven or less for every one of the analyzed hourly sets of data.

We learn from the given examples that the combined amplitudes of one quasi-cycle would exceed the first selection criterion $G_{n1} = 1.33\%$ in most cases. We could adopt a rule that only quasi-cycles which meet this additional criterion should be considered statistically significant. This rule would lead us back, however, to the problem of significant peaks in the power spectrum of ϵ_t whenever these selected amplitudes are included in the noise.

The physical reality of these quasi-cycles cannot be determined without a detailed analysis of the atmospheric fluctuations, their causes and their reflections on the particular meteorological element under study. E.g., the six to seven day cycle at Huntsville appears to be associated with cyclonic movements (cyclone families) while a quasi-cycle of nine to eleven days seems related to repetitive synoptic situations during cold air outbreaks in winter (see Essenwanger (1977,1979)). Detailed analysis of the quasi-cycles could also include the study of the phase angle ψ and the fluctuation of the amplitude in time.

Finally, it may be noted that the method to separate the data series into three components resembles, in principle, Craddock's (1956) proposal of filtering meteorological time series. Two significant differences from Craddock's method exist, however. First, the cycles (quasi-cycles) are subtracted from the original data series which leaves the noise as a defined statistical quantity. Second, a relatively

simple mathematical formulation of the components is found in terms of the Fourier series and red (or white) noise. The differences are inherent in the dissimilarity of the analysis in Craddock's and the author's methods.

One additional fact about quasi-periodicity will be called to the author's attention. Quasi-cycles comprise usually a limited number of cycles whose total length is shorter than the entire data series. They disappear and may reappear but the phase angles of the individual parts may not be aligned. This produces a reduction of the amplitude in the Fourier terms of the total data sample. Because the power spectrum is independent of the phase angle a statistically significant cycle may be indicated in this tool while testing against the red noise background.

In such a case the only method to determine its Fourier representation for elimination from the data in order to separate cycles and noise is a subdivision of the data series and a subsequent harmonic (or periodogram) analysis of these subsamples. This research for quasi-cycles and their reality may sometimes be somewhat cumbersome but no simple tool exists by which quasi-periodicity can be easily identified. The filter analysis could prove more advantageous in such cases but statistical testing is more difficult in filter analysis. The reader should keep in mind that quasi-periodicity is also produced by random processes. Therefore, statistical testing and a careful analysis are essential before hasty conclusions are drawn.

We may observe in Table 4 that in all cases the first lag correlation of the original data series is not identical with the first lag correlation of the red noise component. Where the daily cycle is dominant this original correlation coefficient is lower but where the annual wave controls most of the variance the red noise lag correlation is lower than the original correlation.

7. CONCLUSIONS

The author has attempted to define quasi-cycles in atmospheric time series as the third component after extracting cycles and separating the red (or white) noise by testing the power spectrum. The Fourier series of the spectrum is utilized as an auxiliary tool to extract significant amplitudes by an iterative process and locate the quasi-cycles.

The representation by three components enables us to formulate a few mathematical terms for meteorological time series without a lengthy listing of the individual frequencies of the power spectrum, and without loss of detailed information.

Some reservations can be made to the use of the Fourier series. The fluctuations of the amplitude of quasi-cycles, the disappearance and reappearance with a different phase angle and the associated diminution of the amplitude may distort the recognition of quasi-cycles in harmonic analysis. The described process combining harmonic analysis and spectrum analysis decreases this deficiency. In turn, the step by step significance testing minimizes the probability that quasi-cycles are

selected which are produced only by chance without physical reality.

The first lag correlation of the regular time series is governed by the presence of periodicities (such as the annual or diurnal cycle) or quasi-periodicities . Therefore, this lag correlation is not well suited for testing of significance against red noise background in power spectra of the entire data series. After cycles and quasi-cycles have been removed, the presence of red (or white) noise can be tested by the first lag correlation of the residual error. One should reconsider this type of noise as the "real" or "residual" noise in meteorological time series because it is free of cycles and quasi-cycles and is predictable only in form of statistical characteristics.

After extraction of the quasi-cycles they must be further investigated as to their physical reality unless the physical background is obvious.

Quasi-periodicities may not always prove to be statistically significant in the power spectrum of the entire data series as tested against the "classical" red noise background but the entire window (filter band) may prove statistically significant. Caution must be exercised, however, in the interpretation of quasi-cycles because they can also imply side lobes of existing cycles or quasi-cycles. Thus, a detailed analysis of the atmospheric fluctuations, their causes and their reflections on the particular meteorological element is essential before a complete interpretation of these quasi-cycles can be made. The identification of the quasi-cycles can be helpful for the follow-up study and problem formulation.

ACKNOWLEDGEMENT

The author wishes to express his sincere thanks to Dr. Dorathy A. Stewart and Mrs. Helen M. Boyd for their critical review of the manuscript. The author appreciates the assistance of Mr. Jerry D. Smith in establishing the computer program and running the various phases of calculations on the CDC 6600 system. Mrs. Clara B. Brooks deserves credit for her diligent efforts of expeditiously typing and assembling the manuscript.

REFERENCES

- Bartels, J., 1943. Gesetz und Zufall in der Geophysik. *Naturw.* 31: 421-435.
 Blackman, R.B. and Tukey, J.W., 1958. *The Measurement of Power Spectra*. Dover, New York, p.190.
 Bloomfield, P., 1976. *Fourier Analysis of Time Series : An Introduction*. Wiley & Sons, New York, p.258.
 Box, E.P. and Jenkins, G.M., 1970. *Time Series Analysis, Forecasting and Control*. Holden Day, San Francisco, p.553.
 Brier, G.W., Shapiro, R. and MacDonald, N.J., 1964. A test for the period of 18 cycles per year in rainfall data. *J. Appl. Meteor.* 3:53-57.
 Brooks, C.E.P. and Carruthers, N., 1953. *Handbook of Statistical Methods in Meteorology*. Her Majesty's Stationery Office, London, p.412.
 Cooley, J.W. and Tukey, J.W., 1965. An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* 19:297-301.

- Craddock, J.M., 1965. The analysis of meteorological time series for use in forecasting. *The Statist.* 15:167-190.
- Essenwanger, O.M., 1950. Wahre Expectanz und Erhaltungsneigung des Luftdrucks. *Meteor. Rundsch.* 3:62-65.
- Essenwanger, O.M., 1976. *Applied Statistics in Atmospheric Science, Part A: Frequencies and Curve Fitting.* Elsevier, Amsterdam, p.412.
- Essenwanger, O.M., 1977. Red noise analysis in autocorrelogram and power spectrum of atmospheric temperature. Preprint Vol., 5th Conf. Prob. Stat., Nov.15-18, 1977, publ. by Amer. Meteor. Soc., 1977.
- Essenwanger, O.M., 1979. Red noise in the power spectrum of atmospheric temperature data. In: ARO 79-2, Proc. 24th Conf. on Design of Experiments in Army Research, Development and Testing : 51-61.
- Gilman, D.L., Fuglister, F.J. and Mitchell, J.M. Jr., 1963. *J. Atmosph. Scie.* 20:182-184.
- Hartley, H.O., 1949. Test of significance in harmonic analysis. *Biometrika* 36:194-201.
- Kendall, M.G., 1973. *Time Series.* Griffin, London, p.197.
- Kendall, M.G. and Stuart, A., 1966. *The Advanced Theory of Statistics, Vol.3, Design and Analysis, and Time Series.* Hafner, New York, p.552.
- Shapiro, R., 1975. The variance spectrum of monthly mean central England temperatures. *Qu. J. Roy. Meteor. Soc.* 101:679-681.
- Sneyers, R., 1975. Sur l'analyse statistique des series d'observations. *Note Techn.* 143, OMM 415, p.192.
- Sneyers, R., 1976. Application of least squares to the search of periodicities. *J. Appl. Meteor.* 15:387-393.
- Stumpff, K., 1937. *Grundlagen und Methoden der Periodenforschung.* Springer, Berlin, p.332.
- Taubenheim, J., 1969. *Statistische Auswertung Geophysikalischer und Meteorologischer Daten.* Akadem. Verlagsgesellschaft, Leipzig, p.386.
- Walker, G.T., 1914. Correlation in seasonal variation of weather, III. On the criterion for the reality of relationships of periodicities. *Ind. Meteor. Dept. Mem. (jimla)* 21, p.22.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

DETECTION OF CHANGES IN THE PARAMETERS OF PERIODIC OR PSEUDO-PERIODIC SYSTEMS WHEN THE CHANGE TIMES ARE UNKNOWN

I.B.MACNEILL

Stat. and Actua. Sci. Group, Dept. Math., Univ. of Western Ontario, Ontario (Canada)

ABSTRACT

MacNeill, I.B., Detection of changes in the parameters of periodic or pseudo-periodic systems when the change times are unknown. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

A method of detecting changes in regression when change times are unknown has been discussed by the author (Annals of Statistics, 1978). An alternative to this method, which has based on cumulative sums of raw regression residuals, is presented in this paper. The new method, which is based on a likelihood-ratio type test, is applied to various climatological data sets, including tree-ring data, temperature data and sunspot series.

1. INTRODUCTION

Most climatological and meteorological phenomena are, at least in part, probabilistic in nature and, as such, are properly described by stochastic models. These models may be characterized by certain parameters. As an example, consider the daily maximum temperature at a station. A reasonable model might be of a regression type with a year cycle plus constant forming the mean value function, and with a random element superimposed upon this function. More specifically, if $T(t)$ represents the daily maximum temperature for the t -th day in the series and $\epsilon(t)$ represents the corresponding random element, then, for appropriate parameters μ , A , ω and ϕ , the stochastic model might take the form:

$$T(t) = \mu + A \sin(\omega t + \phi) + \epsilon(t), \quad t = 1, 2, \dots$$

Usually, one would expect a temperature series to possess the property of stationarity, and hence one would find the statistical problem associated with such a model to be one of the estimation of parameters. However, in some instances, it may be that the stationarity assumption is not valid and that the parameter values change with time. If the change times are known, then the statistical problem becomes a two-sample test of hypothesis, which in this case possesses a reasonably standard solution. However, if the change times are unknown, the problem of detecting change in μ

and/or A becomes non-standard. The problem of testing for change of regression at unknown times was first considered by Quandt (1958, 1960) who proposed a test for no change versus one change based upon the likelihood ratio. Hinkley (1969) and Feder (1975) also explored the likelihood ratio test approach. Brown, Durbin and Evans (1975) proposed tests based upon recursively generated residuals and the associated sequences of partial sums of these residuals. MacNeill (1978a, 1978b) examined the properties of sequences of partial sums of raw regression residuals and proposed a Cramér-von Mises type statistic for use in testing for change of regression at unknown time. In the sequel, a related test statistic is proposed. This statistic is derived using an approach of Chernoff and Sacks (1964), Gardner (1969), and MacNeill (1974) for the detection of parameter changes at unknown times in a sequence of independent and identically distributed random variables.

Another series, much speculated upon by climatologists and meteorologists, is the Wolfer sunspot series. Models of the autoregressive type, which can be used to characterize pseudo-periodic phenomena, have been shown to fit this series better than those with strict periodic components. Also, it has been suggested that the series is not stationary, and that the parameters of the fitted autoregressive models are different in different parts of the series. The question then naturally arises as to where the change points are. This again is a problem of detection of change of (auto) regression at unknown time. The test proposed below can be considered for such problems although distribution theory is not the same as in the previous example. In the examples considered below, the sunspot series is analysed as a periodic phenomenon and autoregressions are fitted to several other series.

2. DERIVATION OF THE TEST STATISTIC

The regression model considered below is characterized by a set of regressor functions $\{f_k(t), t = [0,1], k = 0, 1, 2, \dots, p\}$ and a set of independent and identically distributed error terms $\{\epsilon_j, j \geq 1\}$ each normally distributed with zero mean and variance $\sigma^2 > 0$. Without loss of generality, the time parameter is scaled into the unit interval. The dependent variables $\{Y_j, j = 1, \dots, n\}$ are then defined by:

$$Y_j = \beta' \underline{f}(t_j) + \epsilon_j, \quad j = 1, 2, \dots, n,$$

where

$$\underline{f}'(t_j) = \{f_0(t_j), f_1(t_j), \dots, f_p(t_j)\}$$

and

$$\underline{\beta}' = \{\beta_0, \beta_1, \dots, \beta_p\}$$

is the vector of regression coefficients. In the standard matrix formulation, this may be written as:

$$\underline{y}_n = \underline{x}_n' \underline{\beta} + \varepsilon_n,$$

where \underline{X} is the design matrix whose ij -th component is $f_j(t_i)$. The Gauss-Markov estimator for $\underline{\beta}_p$, denoted by $\hat{\underline{\beta}}_p$, is:

$$\hat{\underline{\beta}}_p = (\underline{X}'\underline{X})^{-1} \underline{X}'\underline{y}_n.$$

The subscripts on the vectors and matrices are omitted where no confusion results. The vector of regression residuals is defined to be $\underline{y} - \hat{\underline{y}}$ where the i -th component of $\hat{\underline{y}}$ is:

$$\hat{y}_i = \hat{\underline{\beta}}' \underline{f}(t_i).$$

The null hypothesis that is to be tested is:

$$H_0: E(y_i) = \underline{\beta}_0' \underline{f}(t_i), \quad i = 1, 2, \dots, n$$

where $\underline{\beta}_0$ is fixed but unknown. The alternative hypothesis requires changes in $\underline{\beta}_0$ at unknown times. To specify alternatives, we let

$$\underline{\delta}_i' = \{\delta_{0i}, \delta_{1i}, \dots, \delta_{pi}\}$$

represent the changes in the vector of regression coefficients effected between the i -th and the $(i + 1)$ -th observation. That is, if $\underline{\beta}_i$ is the vector of regression coefficients for the i -th observation, then

$$\underline{\beta}_{i+1} = \underline{\beta}_i + \underline{\delta}_i.$$

So that the Bayes-type argument introduced by Chernoff and Sacks (1964) may be used to eliminate nuisance parameters, we assume that $\underline{\delta}$ has a multivariate normal distribution with zero mean and covariance matrix $\tau^2 \underline{I}$ where $\tau^2 > 0$. We then let a particular change sequence be defined by:

$$\underline{\omega}' = \{\omega_1, \omega_2, \dots, \omega_{n-1}\}$$

where ω_i is 1 if a change in β occurs between the i -th and $(i + 1)$ -th observation and is zero otherwise. Thus, a single change through the series of observations would require one component of ω to be 1 and the rest zero. The assignment of a prior distribution to the collection of all possible change sequences, ω , then makes it possible to formulate the problem in a way introduced by Gardner (1969). The nuisance parameters, δ_i , can then be integrated out and, with τ^2 small, the likelihood ratio statistic for testing H_0 against change sequences ω , with a uniform prior can be shown to be approximately proportional to:

$$Q_n = \frac{1}{\sigma^2} \sum_{k=1}^{n-1} \mathbf{y}' (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{X}_k^k \mathbf{X}_k^{k'} (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{y},$$

where \mathbf{X}_k^k is \mathbf{X} with the first k rows identically equal to zero. The approximation becomes exact as τ^2 vanishes. Note that:

$$Q_n = \frac{1}{\sigma^2} \sum_{k=1}^{n-1} \left\| \varepsilon' (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{X}_k^k \right\|^2 = \frac{1}{\sigma^2} \sum_{k=1}^{n-1} \left\| (\mathbf{y}' - \mathbf{y}') \mathbf{X}_k^k \right\|^2$$

where, if

$$\mathbf{z}' = \{z_1, z_2, \dots, z_k\},$$

$$\|\mathbf{z}\|^2 = z_1^2 + z_2^2 + \dots + z_k^2.$$

Associated with the sequence of partial sums of regression residuals is a generalized Brownian Bridge (see MacNeill 1978b) which we shall denote by $\{\beta_f(t), t \in [0, 1]\}$. The stochastic integral

$$\int_0^1 \beta_f^2(t) dt$$

is then related to a Cramér-von Mises type statistic defined upon the sequence of partial sum of regression residuals; some examples are considered by MacNeill (1978a). Let μ_f and σ_f^2 denote the mean and variance of the stochastic integral. Then it may be shown that:

$$E(Q) \approx \sigma_f^2 \mu_f \sum_{i=1}^n (i-1) (\mathbf{x}_i \cdot \mathbf{x}_i)$$

and

$$\text{Var}(Q) \approx 2\sigma_f^4 \sum_{i=2}^n \sum_{j=2}^n [\min\{(i-1), (j-1)\}]^2 (\mathbf{x}_i \cdot \mathbf{x}_j)^2$$

where X_i is the i -th row of the design matrix X , and

$$(X_i \cdot X_j) = \sum_{\ell=0}^P X_{i\ell} X_{j\ell} = \sum_{\ell=0}^P f_{\ell}(t_i) f_{\ell}(t_j).$$

Distributions of the statistic Q are skewed as can be seen from Table 2 of MacNeill (1978a), where it can be noted also that the upper 5% point is approximately two standard deviations above the mean.

3. APPLICATION OF THE TEST STATISTIC

Three examples are considered below.

Example 1: Average annual riverflow of the Nile at Aswan for the period 1870-1945.

The 75 observations in this series are plotted in Figure 1. Inspection of the data suggests a decline in average flow over the course of the series. An autoregressive model of order 1 fitted to the mean corrected data yields the following:

$$(Z(t) - 2947.7) = 0.6773\{Z(t-1) - 2947.7\} + \epsilon(t).$$

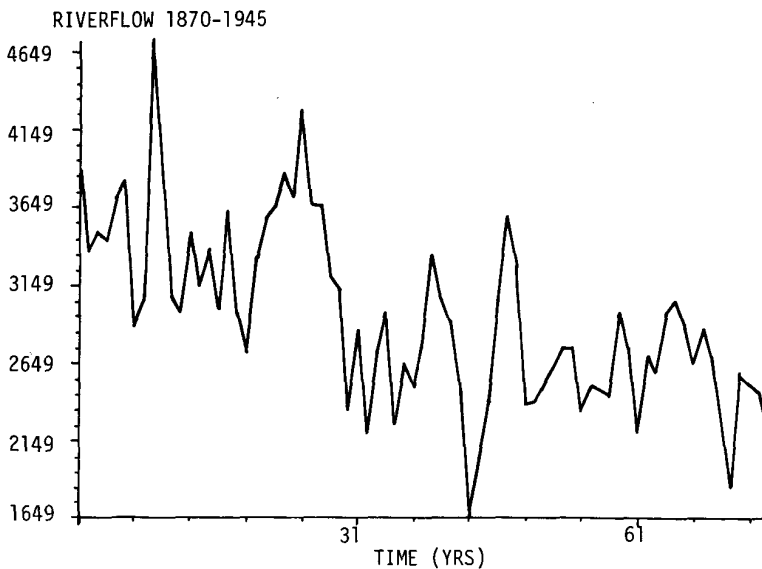


Fig. 1. Average annual riverflow of the Nile at Aswan for 1870-1945.

The computations for test of change of the parameters of the model at unknown time yield the following:

$$Q = 93.8 \times 10^6,$$

$$E(Q) = 48.4 \times 10^6,$$

and

$$\sqrt{\text{Var}(Q)} = 1.17 \times 10^6.$$

Although the mean and variance represent approximations, it is evident that the test has detected what is an obvious change of parameter.

Example 2: Tree ring data for Jeffery Pine from the Tioga Pass, California, for the period 1384-1964.

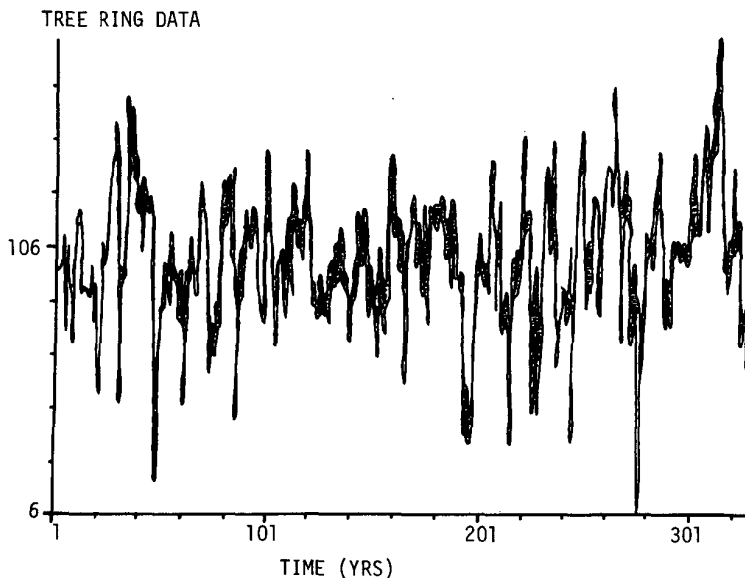


Fig. 2. Tree ring data for Jeffery Pine from Tioga Pass, California, 1384-1964.

The 661 observations in this series are plotted in Figure 2. An autoregression of order 1 was fitted to the mean corrected data and the statistics for testing for change of parameter at unknown time were calculated as follows:

$$Q = 21.1 \times 10^6,$$

$$E(Q) \approx 20.1 \times 10^6,$$

and

$$\sqrt{\text{Var}(Q)} \approx 1.45 \times 10^6.$$

Thus this computation does not appear to support the hypothesis of change of parameter.

Example 3: Wolfer sunspot series for the period 1700-1960.

The 261 observations in this series are plotted in Figure 3. These data are often regarded as periodic with a period of approximately 11.2 years and corresponding frequency of 0.56 radians per year; this conclusion is suggested by a spectral analysis. However, it has been found that finite parameter schemes, such as the autoregression of order 2, do a better job of fitting the data than does a model with a periodic component. Part of the explanation for this could be that changes in mean level, amplitude, or phase make it difficult for a periodic model with fixed parameters to fit the data.

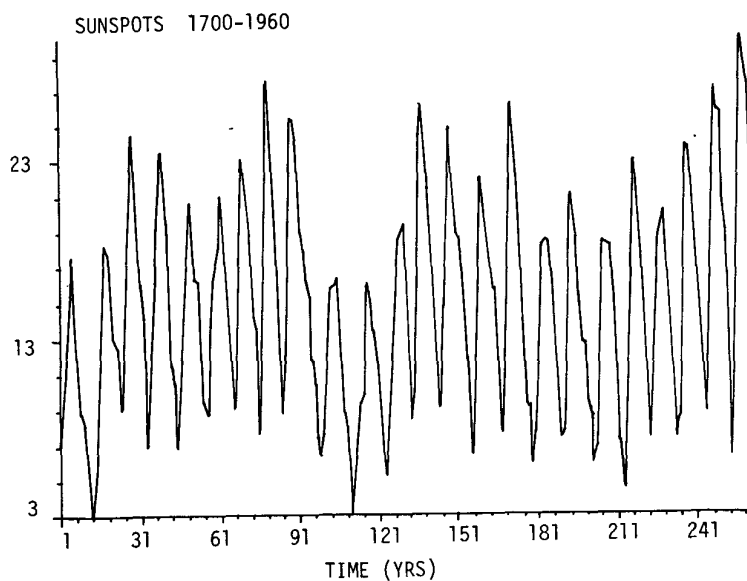


Fig. 3. Wolfer sunspot series, 1700-1960.

The following model is fitted to the sunspot series:

$$Z(t) = \mu + A \sin(\omega t + \phi) + \varepsilon(t)$$

where μ , A and ϕ are estimated by least squares. This yields, as estimated model, the following:

$$\hat{Z}(t) = 14.47 + 4.66 \sin(0.56t - 1.38).$$

This fitted model is superimposed over the original data in Figure 4. The fit is rather poor; several probable reasons are: first, the mean level of series fluctuates; and, second, the fitted series becomes out of phase with the actual observations at certain times.

The statistics of the test for change of parameters at unknown time are as follows:

$$Q = 27.09 \times 10^5,$$

$$E(Q) \approx 2.05 \times 10^5,$$

and

$$\sqrt{\text{Var}(Q)} \approx 0.75 \times 10^5.$$

The test strongly suggests that a change of parameters occurs within the time period of the observations.

As suggested above, the mean level of the series appears to fluctuate with time. To investigate the series further, a mean value function was fitted empirically to the data. This function, which is a rough estimate of the actual mean value function appears as the broken line graph superimposed on the original series in Figure 5. The difference between the original series and the estimated mean value function appears in Figure 6. Again, a sinusoid with frequency 0.56 was fitted by least squares to the mean corrected data resulting in only a modest improvement of the fit; the fitted sinusoid is superimposed on the mean-corrected series in Figure 7. One problem that is immediately apparent is the serious phase difference between the two series in the time period running approximately from 1770 to 1800. If this period is deleted from the series, and a sinusoid of frequency 0.56 is fitted to the data, a much better fit occurs. This fit is graphed in Figure 8. If separate fits are made to the pre 1770 data and to the post 1800 data, still better fits occur; Figure 9 and 10 show the respective fits.

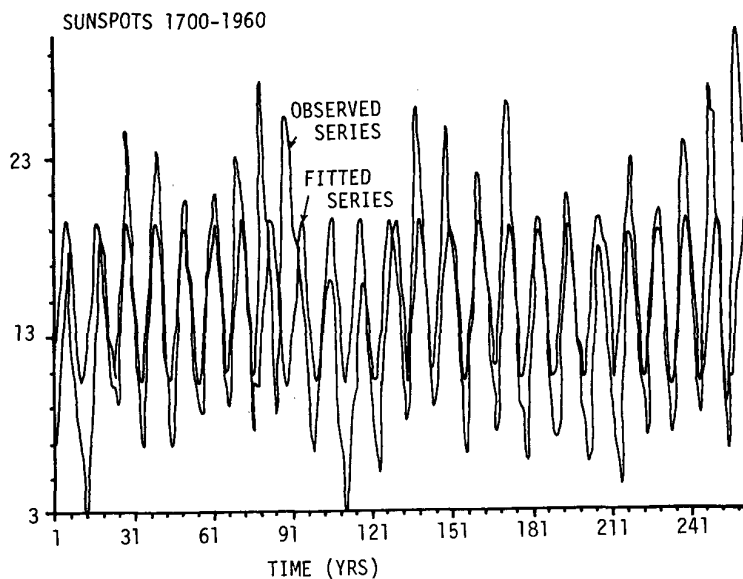


Fig. 4. Wolfer sunspot series and fitted sinusoid, 1700-1960.

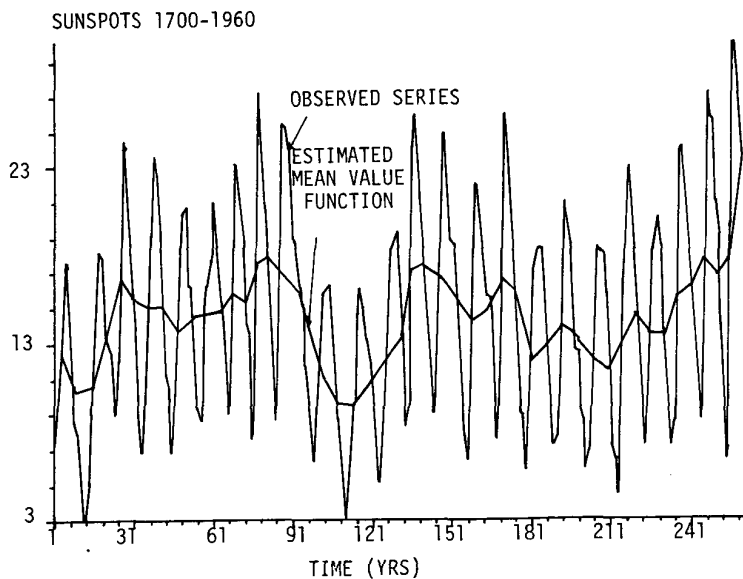


Fig. 5. Wolfer sunspot series and estimated mean value function.

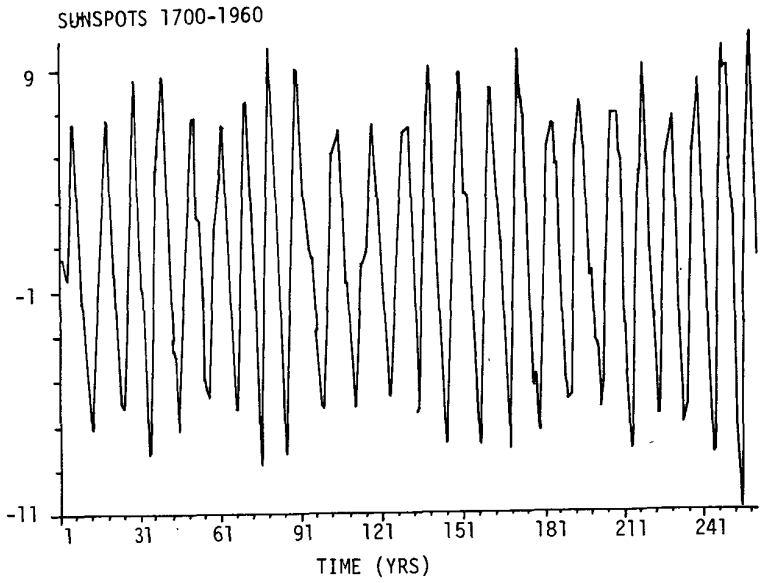


Fig. 6. Mean corrected Wolfer sunspot series.

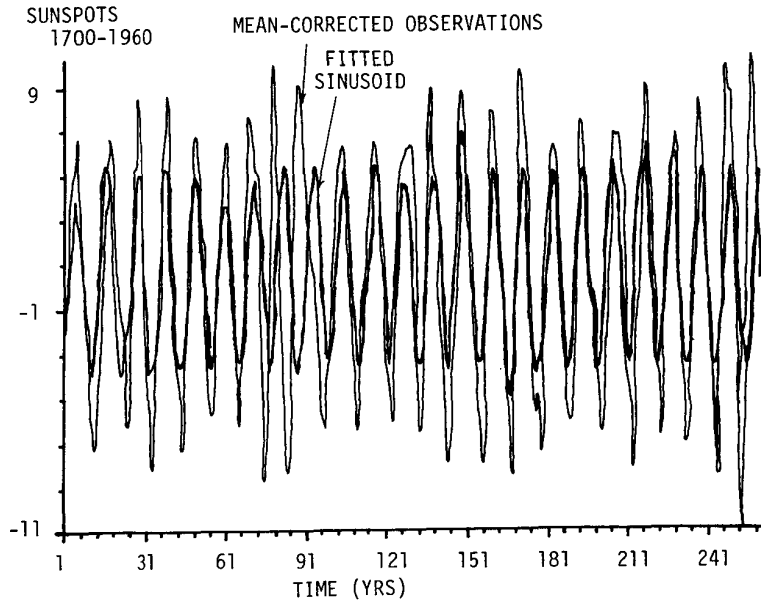


Fig. 7. Mean-corrected Wolfer sunspot series and fitted sinusoid.

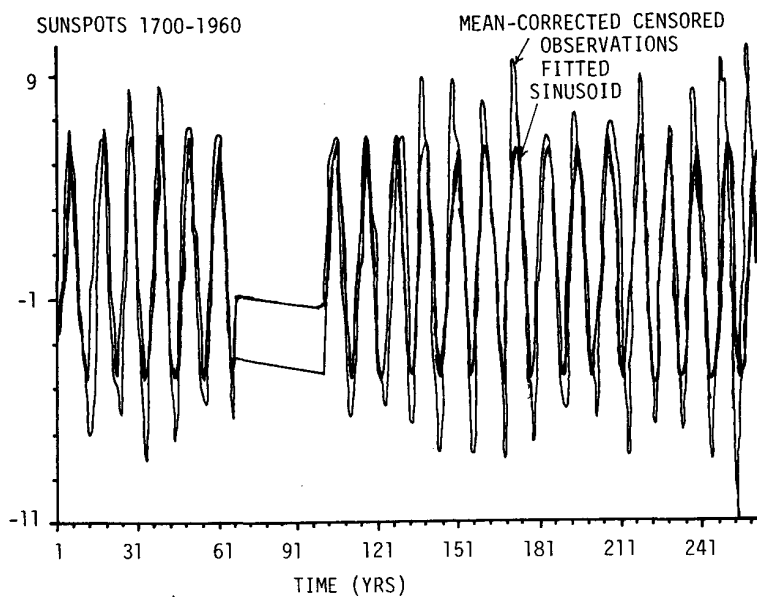


Fig. 8. Mean-corrected censored Wolfer sunspot series and fitted sinusoid.

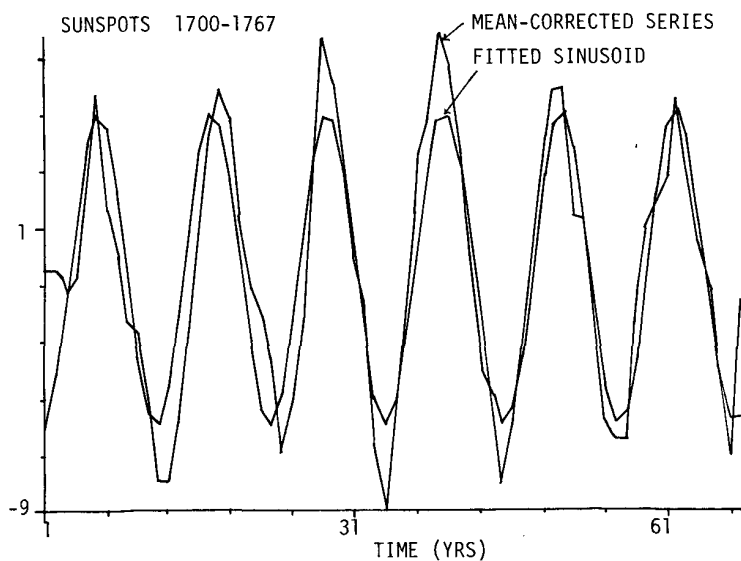


Fig. 9. Mean-corrected Wolfer sunspot series and fitted sinusoid, 1700-1767.

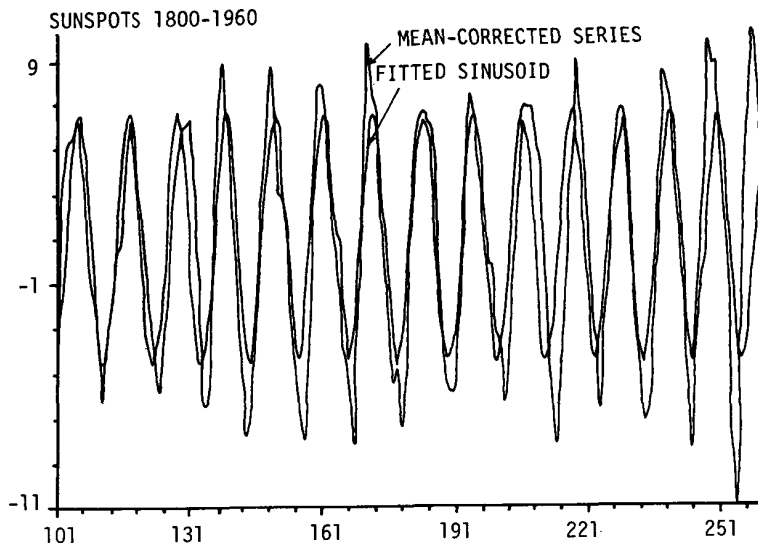


Fig. 10. Mean-corrected Wolfer sunspot series and fitted sinusoid, 1800-1960.

4. ACKNOWLEDGEMENTS

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

The author would like to thank Dr. A.I. McLeod and Mr. C.I. Young for valuable contributions to the data analysis of this paper.

REFERENCES

- Brown, R.L., Durbin, J. and Evans, J.M., 1975. Techniques for testing the constancy of regression relationships over time. *J. Roy. Statist. Soc. Ser. B* 37: 149-192.
- Chernoff, H. and Zacks, S., 1964. Estimating the current mean of a normal distribution which is subject to changes in time. *Ann. Math. Statist.*, 35: 990-1018.
- Feder, P.I., 1975. The log likelihood ratio in segmented regression. *Ann. Statist.*, 3: 84-97.
- Gardner, L.A., 1969. On detecting changes in the mean of normal variates. *Ann. Math. Statist.*, 40: 116-126.
- Hinkley, D.V., 1969. Inference about the intersection in two-phase regression. *Biometrika*, 56: 495-504.
- MacNeill, I.B., 1974. Tests for change of parameter at unknown time and distributions of some related functionals on Brownian motion. *Ann. Statist.*, 2: 950-962.
- MacNeill, I.B., 1978a. Properties of sequence of partial sum of polynomial regression residuals with applications to tests for change of regression at unknown times. *Ann. Statist.*, 6: 422-433.
- MacNeill, I.B., 1978b. Limit processes for sequences of partial sums of regression residuals. *Ann. Probab.*, 6: 695-698.

- Quandt, R.E., 1958. The estimation of parameters of a linear regression system obeying two separate regimes. J. Amer. Statist. Assoc., 53: 873-880.
- Quandt, R.E., 1960. Tests of the hypothesis that a linear regression system obeys two separate regimes. J. Amer. Statist. Assoc., 55: 324-330.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

PRECIPITATION SIMULATION PROCESS WITH MARKOV CHAIN MODELING

O.P.BISHNOI and K.K.SAXENA

Dept. Agronomy and Dept. Math. and Stat., Haryana Agric. Univ., Hissar (India)

ABSTRACT

Bishnoi, O.P. and Saxena, K.K. Precipitation simulation process with Markov chain modeling. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov. 29-Dec. 1, 1979

The daily precipitation data for two stations Hissar (Arid) and Ambala (Dry sub-humid) have been analysed with the help of Markov chain models representing the conditional dependence. This has been done by using a loss function composed of a log likelihood ratio term and a degree of freedom term is used as a decision criterion. It has been found that the third order Markov chain model represents better the daily precipitation occurrence during the S.W. monsoon season. The common practice of assuming first order Markov chain is unjustified. If the record length is less than 2000 days at Ambala and 4000 days at Hissar, there is a tendency for a low order chain to be misrepresented as the proper model. A specific example in which third order model is required to depict the precipitation occurrence has been given in detail. Therefore, the proper Markov order describing the daily precipitation occurrence process has to be determined and cannot be assumed a priori. The common practice of assuming the first order model is unjustified.

INTRODUCTION

The daily precipitation occurrence can be approximated by simple Markov chain model, but there have been instances in which the simple Markov chain is not able to depict the daily occurrence properly. On the basis of a hypothesis testing procedure, Chin (1976) presented definitive examples in which the simple first order chain should be rejected and a second order model was proposed for such occasions. In the present work this has been shown for the S.W. monsoon and N.E. monsoon precipitation. Instead of the conventional chi-square test, a decision procedure based on the extension of the maximum likelihood principle has been used. The primary objective is to determine the proper order of Markov chains that would be appropriate to represent conditional dependence of daily precipitation occurrence in the S.W. monsoon and N.E. monsoon seasons. A second objective is to ascertain to what extent, if any, the Markov order is effected by sample size variation.

MATERIAL AND METHOD

The daily precipitation records during S.W. monsoon season (June to September)

and N.E. monsoon season (December to February) for 30 years at Ambala (1941-70), and 55 years at Hissar (1915-70, except 1945) during the S.W. monsoon season and 40 years during N.E. monsoon season (1931-70) have been utilized in the present study.

A markov chain is a sequence of discrete random variables and is said to be of order k if k is the smallest positive integer, such that for all n the following equation relating the conditional probabilities is satisfied:

$$P\{x_n | x_{n-1}, x_{n-2}, \dots, x_{n-k}, x_{n-k-1}, \dots\} = P\{x_n | x_{n-1}, x_{n-2}, \dots, x_{n-k}\}.$$

An approach to the problem of model identification based on an extension of the maximum likelihood principle was proposed by Bartlett (1951), Akaike (1972, 1974), Tong (1975) and Gates and Tong (1976). Akaike (1974) defined the $AIC(\theta)$ in terms of k such that

$$AIC(\theta) = -2 \log(\text{maximum likelihood}) + 2k,$$

where k is the number of independently adjusted parameters within the model. θ is an unknown vector parameter in the probability density function. When there are several competing estimates in a model identification problem, the estimate that minimizes AIC is the appropriate choice.

Using a simplified notation, the transition probabilities of an ergodic Markov chain is denoted by $P_{i,i+1,i+2,\dots,j-1,j}$ and the frequency of these transitions by $N_{i,i+1,i+2,\dots,j-1,j}$. In dealing with large samples or relatively long records of observations the log likelihood function is given by

$$\begin{aligned} \log L &= \sum_{i,i+1,\dots,j-1,j} N_{i,i+1,\dots,j-1,j} \log P_{i,i+1,\dots,j-1,j} \\ &= \sum_{i,i+1,\dots,j-1,j} N_{i,i+1,\dots,j-1,j} \log (N_{i,i+1,\dots,j-1,j} / N_{i,i+1,\dots,j-1,j}) \end{aligned}$$

where $(N_{i,i+1,\dots,j-1,j} / N_{i,i+1,\dots,j-1,j})$ is the sample maximum likelihood estimate of the unknown transition probability of the population. Specifically we want to compare the relative validity, for example, between an m th-order and $(m-1)$ th-order model. Let $P_{i,i+1,\dots,j-1,j}$ represent the transition probabilities of an m th-order chain and S be the number of states; we want to find out if

$$P_{i,i+1,\dots,j-1,j} = P_{i+1,\dots,j-1,j} \quad \text{where } i = 1, 2, \dots, S.$$

A pertinent step is to form the log-likelihood ratio $\log \lambda_{m-1,m}$. Hoel (1954) has shown that $-2 \log \lambda_{m-1,m}$ for an ergodic chain is asymptotically a chi-squared variate with $\nabla^2 S^{m+1}$ degrees of freedom. Here ∇ is the difference operator on the subscript:

$$\nabla x^a = x^a - x^{a-1}, \quad \nabla^2 x^a = \nabla(\nabla x^a).$$

Rewriting $-2 \log \lambda_{m-1,m}$ as $_{m-1}H_m$, we have

$$_{m-1}H_m = 2 \sum_{i,i+1,\dots,j-1,j} N_{i,i+1,\dots,j-1,j} [\log \{N_{i,i+1,\dots,j-1,j} / N_{i,i+1,\dots,j-1}\} - \log \{N_{i+1,\dots,j-1,j} / N_{i+1,\dots,j-1}\}] .$$

The right hand side of the above equation is a measure of how well the observed sequence supports the hypothesis of an m th-order versus an $(m-1)$ th-order chain, where the count of the successive number of states observed in the sequence $i, i+1, \dots, j-1, j$ adds up to m . For $k < (m-1)$ it can be shown that

$$_kH_m = _kH_{k-1} + _{k+1}H_{k+2} + \dots + _{m-1}H_m .$$

We assume that the individual terms on the right hand side of this equation to be asymptotically independent. Then, given that the chain is of order k , $_kH_m$ has a chi-squared distribution with degrees of freedom given by $v = \sqrt{S^{m+1}} - \sqrt{S^{k+1}}$.

To formulate a decision procedure related to the problem of Markov order identification, the fundamental step is the choice of an appropriate loss function. On the basis of the AIC approach, Tong (1975) proposed the following loss function:

$$R(k) = _kH_m - 2(\sqrt{S^{m+1}} - \sqrt{S^{k+1}}),$$

where k is the order of the fitting model and m is the highest order under consideration. The loss function $R(k)$ thus defined is composed of two counter-acting terms representing the log-likelihood ratio and the degree of freedom, respectively. As one tries to fit the observed data with Markov chains of higher and higher orders, the log-likelihood ratio terms will most likely decrease progressively. But this reduction in variance is achieved at the price of increasing model complexity and is reflected as a graduated penalty by the increasing degree of freedom term. The selected Markov orders k is the one that minimizes the sum of these two terms. This is the minimum AIC estimated (MAICE).

A day has been described as a wet day (w) if at least 1.00 mm rainfall is recorded on that day, otherwise defined as a dry day (d). The past four days weather phenomenon has been presented in Table 1 for Hissar and Ambala.

The MAICE procedure indicated that the precipitation occurrence process in the monsoon months can best be described by a third order model. A comparison of this model with second and first order Markov chain models in their ability to represent monsoon precipitation characteristics would be instructive. Any model that can describe the daily precipitation occurrence process well should be able to represent the distribution of dry and rainy sequences as well. A dry sequence of length r is defined as a succession of r dry days preceded and followed by at least one rainy day. A wet sequence is defined correspondingly. The positive integer r may take a value of 1. The distribution of dry sequence based on the third order model is given by

$$P(d_r) = \begin{cases} p(w)p(d|w)p(w|d,w) & r = 1 \\ p(w)p(d|w)p(d|d,w)p(w|d,d,w) & r = 2 \\ p(w)p(d|w)p(d|d,w)p(d|d,d,w)p(d|d,d,d)^{r-3}p(w|d,d,d) & r \geq 3. \end{cases}$$

Here $P(d_r)$ is the marginal probability of getting a dry sequence of r days, and $p(w)$ is the marginal probability of day being wet. Calculations of second and first order distributions were carried out by using similar procedures.

TABLE 1.

Occurrence of wet and dry days alongwith their weather phenomenon on preceding days of Ambala and Hissar for (a) S.W.monsoon period (June-Sept.)

Preceding days					Hissar			Ambala		
t-4	t-3	t-2	t-1		Observation day t			Observation day t		
					w	d	Total	w	d	Total
					1236	5474	6710	914	2746	3660
			w		452	784	1236	434	480	914
			d		764	4710	5474	480	2266	2746
		w	w		189	263	452	230	204	434
		d	w		268	496	764	204	276	480
		w	d		161	623	784	116	364	480
		d	d		579	4131	4710	308	1958	2266
	w	w	w		96	93	189	128	102	230
	d	w	w		103	165	268	99	105	204
	w	d	w		60	101	161	42	74	116
	d	d	w		189	390	579	145	163	308
	d	d	d		475	3656	4131	244	1714	1958
	w	d	d		114	509	623	57	307	364
	d	w	d		110	386	496	68	208	276
	w	w	d		60	203	263	41	163	204
d	d	d	d		681	2975	3656	188	1526	1714
d	d	d	w		159	316	475	106	138	244
d	d	w	d		91	299	390	37	126	163
d	d	w	w		73	116	189	70	75	145
d	w	d	d		85	301	386	22	186	208
d	w	d	w		39	71	110	26	42	68
d	w	w	d		44	121	165	14	91	105
d	w	w	w		56	47	103	49	50	99
w	d	d	d		74	435	509	51	256	307
w	d	d	w		35	79	114	25	32	57
w	d	w	d		32	69	101	26	48	74
w	d	w	w		23	37	60	24	18	42
w	w	d	d		33	170	203	24	139	163
w	w	d	w		22	38	60	16	25	41
w	w	w	d		16	77	93	14	88	102
w	w	w	w		41	55	96	74	54	128
(b) N.E. monsoon season (Dec.-Feb.)										
					186	3513	3699	235	2472	3707
			w		41	145	186	82	153	235
			d		146	3367	3513	154	2318	2472
		w	w		9	32	41	25	57	82
		d	w		31	115	146	55	99	154
		w	d		8	137	145	8	145	153
		d	d		137	3230	3367	140	2178	2318
w	w	w	w		2	7	9	4	21	25

TABLE 1 (cont'd)

d	w	w	7	24	31	21	34	55
w	d	w	2	6	8	3	5	8
d	d	w	28	109	137	47	97	140
d	d	d	121	3109	3230	134	2044	2178
w	d	d	10	126	136	6	139	145
d	w	d	9	106	115	6	93	99
w	w	d	0	32	32	2	55	57
d	d	d	116	2996	3112	118	1926	2044
d	d	d	30	91	121	50	84	134
d	d	w	15	95	110	6	91	97
d	d	w	7	21	28	21	26	47
d	w	d	14	93	107	2	91	93
d	w	d	3	6	9	2	4	6
d	w	w	2	22	24	2	32	34
d	w	w	2	5	7	4	17	21
w	d	d	9	116	125	4	135	139
w	d	d	0	10	10	1	5	6
w	d	w	0	6	6	1	4	5
w	d	w	0	1	1	0	3	3
w	w	d	3	29	32	4	51	55
w	w	d	0	0	0	0	2	2
w	w	w	0	7	7	1	20	21
w	w	w	0	2	2	0	4	4

With the state space composed of d and w, there are 16 permutation in a 4-day sequence. In considering all possible 4-day sequences in a long record, repeated counting upto 4 times for each day, in making up the sequences is permitted. An example of observed and expected 4-day sequences without any restriction of preceding and following day are shown in Table 2.

The model expectations were computed by using marginal probabilities of each sequence. Some examples for the formulation of these probabilities for the third order model are as follows:

$$p(dddd) = p(d)p(d|d)p(d|d,d)p(d|d,d,d),$$

$$p(wwdd) = p(w)p(w|w)p(d|w,w)p(d|d,w,w) .$$

Since the contents of the 16 classes in Table 2 are not independent and are not mutually exclusive, the chi-square test for goodness of fit is not applicable. However, the relative magnitude of the quantity

$$\chi^2 = \sum_{i=1}^{16} \frac{(n_i - e_i)^2}{e_i} , \quad n_i \text{ and } e_i \text{ being the observed and expected frequencies,}$$

could still provide a qualitative indication of how well the models fit, in spite of the differences in degree of freedom. The χ^2 for M_3 is the smallest out of the three. Without making any assertions about the probability values, it seems likely that the third order model can simulate the distribution of 4-day sequences in this case with much more fidelity than either the second or the first order model.

TABLE 2.

Observed distribution of 4-day sequences compared with those computed with the first order (M1), second order (M2) and third order (M3) Markov chain models.

(S.W. Monsoon Season)							
(a) <u>Ambala</u>							
Sequence	Observed	M3	M2	M1	M3 - Obs.	M2 - Obs.	M1 - Obs.
dddd	1714	1712	1692	1543	-2	-22	-171
dddw	244	243	266	326	-1	22	82
ddwd	163	164	177	208	1	14	45
ddww	145	145	136	188	0	-9	43
dwd	208	209	209	208	1	1	0
dwdw	68	68	66	44	0	-2	-24
dwd	105	105	100	120	0	-5	15
dwww	99	99	108	108	0	9	9
wddd	307	306	315	327	-1	8	20
wddw	57	57	49	69	0	-8	12
wdwd	74	75	66	44	1	-8	-30
wdww	42	42	49	40	0	7	-2
wwdd	163	164	147	188	1	-16	25
wwdw	41	41	49	40	0	8	-1
wwwd	102	103	108	108	1	6	6
wwww	128	127	123	99	-1	-5	-29
χ^2 -Value					0.06	11.94	110.90

(b) Hissar

dddd	3656	3655	3622	3486	-1	-34	-170
dddw	475	477	509	565	2	34	90
ddwd	390	389	376	416	-1	-14	26
ddww	189	190	205	240	1	16	51
dwd	386	386	376	416	0	-10	30
dwdw	110	109	102	71	-1	-8	-39
dwd	165	165	155	176	0	-10	11
dwww	103	107	118	106	4	15	3
wddd	509	509	546	580	0	37	71
wddw	114	113	76	94	-1	-38	-20
wdwd	101	100	116	69	-1	15	-32
wdww	60	60	56	40	0	-4	-20
wwdd	203	202	209	246	-1	6	43
wwdw	60	60	54	40	0	-6	-20
wwwd	93	93	111	105	0	18	12
wwww	96	95	79	60	-1	-17	-36
χ^2 -Value					0.10	38.95	137.72

(N.E. Monsoon Season)

(a) Ambala

χ^2 -Value	0.10	4.62	6.42
-----------------	------	------	------

(b) Hissar

χ^2 -Value	0.15	7.34	8.67
-----------------	------	------	------

RESULTS AND DISCUSSION

The Markov chain order to represent the daily precipitation process was selected on the basis of minimizing the loss function. This choice represented the best

compromise between the two competing requirements for reducing the residual variance without incurring a higher than necessary cost in the fitting process.

We have shown as an example for Hissar and Ambala stations that the second or third order model is required to depict the precipitation occurrence process. The various values of the loss function $R(k)$ calculated for Hissar and Ambala are shown in Table 3.

TABLE 3.

Loss function $R(k)$ with $m = 1, 2, 3, 4$ for Hissar and Ambala.

	k	m	v	R(k) during S.W.Monsoon		R(k) during N.E.Monsoon	
				Hissar	Ambala	Hissar	Ambala
R(0)	0	1	1	602.56	394.85	793.19	439.06
	0	2	3	830.85	704.60	758.83	518.46
	0	3	7	848.77	844.50	492.95	559.56
	0	4	15	-568.56	1085.57	248.35	719.98
R(1)	1	2	2	228.28	309.75	-52.36	79.40
	1	3	6	449.65	449.65	-300.24	120.50
	1	4	14	-1169.03	718.72	-544.84	280.92
R(2)	2	3	4	17.97	139.89	-247.88	41.10
	2	4	12	-1399.42	380.96	-492.48	201.52
R(3)	3	4	8	-1417.39	241.07	-244.59	160.42
R(4)	4	4	0	0.00	0.00	0.00	0.00

The results of Table 2 clearly indicate that the M3 - Obs. have approached to very small values and thus the daily precipitation occurrence seems to be well defined by a third order Markov model for the stations under study. Therefore, for $m = 4$ it becomes evident that a second order or third order is required to describe better the simulation of daily rainfall properly in monsoon seasons particularly S.W. monsoon season. The considerable difference between second and third order loss functions further indicates the stable feature of the third order characteristics.

We have also shown that at least a third order model is needed in this case to reproduce the observed fluctuations in the distribution of dry day sequences. Since the decision procedure applied to the model identification was based on the asymptotic behaviour of the log likelihood ratio, it is inherently a large sample method. Experiments were carried out to determine the possible effects of sample size on the order of conditional dependence. Loss functions were computed initially from total span of data successively reduced by 5 years interval and the last by 1 year respectively, the 1970 (Table 4). As the sample size, n , decreases the Markov orders degenerate towards lower values of fluctuate. The minimum sample size that will yield a stable estimate of the proper order seems to depend on the climate and the season. An n value of 1830 days at Ambala (semi arid station) and 4270 days for Hissar (arid station) should generate stable estimate. This suggests that this apparent

third order characteristics should be considered spurious. When the sample size is less than these limits there is a tendency for a low order chain to be misrepresented as the correct model.

TABLE 4.

Effect of sample sizes on the variation of loss function.

Years	n	Loss function R(k)			
		k = 0	k = 1	k = 2	k = 3
Hissar, July - Sept.					
1915-70	6710	-568.56	-1169.13	-1399.42	-1417.39*
1921-70	6100	-607.19	-1183.34	-1386.55	-1602.89*
1926-70	5490	-618.55	-1121.37	-1314.25*	-1304.69
1931-70	4880	-655.49	-1092.50	-1258.52	-1268.09*
1936-70	4270	-860.14	-1256.89	-1394.42*	-1360.00
1941-70	3660	-630.64	-991.34	-1100.03	-1483.44*
1946-70	3050	-125.85	-479.37	-593.73*	-581.34
1951-70	2440	-545.24	-861.07	-589.64	-864.82*
1956-70	1830	246.79	51.46	28.51	17.41*
1961-70	1220	-159.60	274.11	-288.65	-320.20*
1966-70	610	109.10	43.40	4.42*	10.62
1970	122	-3.75	-13.58	-24.77*	-12.42
Ambala, July - Sept.					
1941-70	3660	1085.57	690.72	380.96	241.07*
1946-70	3050	656.21	246.02	133.29	183.62*
1951-70	2440	550.84	217.97	103.94	-134.54*
1956-70	1830	286.63	22.06	-51.14	-91.98*
1961-70	1220	280.75	106.64	66.65	4.89*
1966-70	610	223.23	114.28	58.67	-0.91*
1970	122	-7.67	-12.19	-21.75	-34.70*
Hissar, Dec. - Feb.					
1931-70	3610	248.35	-544.84*	-492.48	-244.59
1936-70	3159	241.76	-67.51	-92.09	-204.21*
1941-70	2707	225.18	-31.33	-50.91	-131.80*
1946-70	2256	161.26	-75.69	-81.22	-159.34*
1951-70	1805	44.24	-57.59	-69.80*	-1.06
1956-70	1354	93.79	6.93	10.93	-124.15*
1961-70	902	129.68	41.67	32.82	-54.62*
1966-70	451	123.74	79.47	70.34	-30.56*
1970	90	-16.12	-18.64	-12.06	-21.40*
Ambala, Dec. - Feb.					
1941-70	2707	719.98	280.92	201.52	160.42*
1946-70	2256	510.50	192.46	109.96	93.24*
1951-70	1805	415.66	186.30	112.16	76.74*
1956-70	1354	170.06	32.88	10.98*	67.68
1961-70	902	34.30	-13.86*	73.44	26.88
1966-70	451	-115.92	-214.40*	-135.82	-115.12
1970	90	-20.96	-63.56*	-48.54	-29.38

* lowest value in the row.

The predominance in Markov orders can be attributed to inherent physical causes responsible for the formation and movements of S.W. monsoon currents. These migratory systems will have a characteristic length scale and a characteristic life cycle

time in the monsoon depressions. At any station the precipitation occurrences associated with monsoon depression passage would most likely indicate a conditional dependence with Markov order higher than one. The likelihood occurrence of these S.W. monsoon depressions, however, is dependent on whether the synoptic meteorological fields provide a favourable environment. One combination of environmental factors favourable to the occurrence of air mass, thunderstorms could be a wind field leading to local moisture convergence, unstable thermal structure, and the passage of an upper air shortwave perturbations.

ACKNOWLEDGEMENT

The authors are highly grateful to the Chief Scientist, Dry Land Research Project, Haryana Agricultural University, Hissar, for kindly providing the facilities.

REFERENCES

- Akaike, H., 1972. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N. and Csaki, F. (ed.) The Second Intern. Symp. on Inform. Theo., 267-281. Akad. Kido, Budapest.
- Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. AC-19:716-723.
- Bartlett, M.S., 1951. The frequency goodness of fit test for probability chains. Proc. Camb. Phil. Soc. 47:86-95.
- Chin, E.H.A., 1976. A second order Markov chain model for daily rainfall occurrences. Presented at Conf. on Hydromet., Amer. Met. Soc., Fort Worth, Tex., Apr. 20-22.
- Chin, E.H.A., 1977. Modeling daily precipitation occurrence process with Markov chain. Water Resour. Res. 13-6.
- Gates, P. and Tong, H., 1976. On Markov chain modeling to some weather data. J. Appl. Met. 15:1145-1151.
- Hoel, P.G., 1954. A test for Markov chains. Biometrika 41:430-433.
- Tong, H., 1975. Determination of the order of a Markov chain by Akaike's information criterion. J. Appl. Prob. 12:488-497.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

STATISTICAL PREDICTION OF CLIMATOLOGICAL EXTREME VALUE AND RETURN PERIOD IN THE CASE OF SMALL SAMPLES

E.SUZUKI, M.MIYATA and S.HONGO

Inform. Sci. Res. Center, Aoyama-Gakuin Univ., Shibuya, Tokyo (Japan)

ABSTRACT

Suzuki, E., Miyata, M. and Hongo, S., Statistical prediction of climatological extreme value and return period in the case of small samples. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

Suppose X be a random variable having a continuous distribution function, and basing upon an ordered sample obtained from the observations on X , we shall identify and predict the extreme value and the return period of the population distribution.

Various methods were proposed for this problem so far in the case of large samples; in the present paper we propose some methods applicable in the case of small samples. For the prediction of extreme values we give a simple method based on an approximation of conditional expectation instead of calculating the definite integral involved there, and for the prediction of return period the discrete linear filtering method in the Kalman Filtering Theory is applied.

1. INTRODUCTION

By using the data of the annual maximum snowfall amounts or the maximum wind speeds obtained over a period of about twenty years, we would like to predict the possible extreme value and the corresponding return period.

As the first step of predicting such an extreme value, we give a method utilizing the conditional expectation, $E\{X_{n+1} | x_1, x_2, \dots, x_n\}$, obtained from the conditional probability density function of order statistics, as is stated in the following section. Section 3 is devoted to discuss the prediction problem of return period, where we propose a practical method derived by applying the Kalman filtering theory.

A numerical example is given in section 4, and finally some remarks are given in section 5.

2. PREDICTION OF THE EXTREME VALUE

Let $X_1 \leq X_2 \leq \dots \leq X_n \leq X_{n+1}$ be order statistics based on a random sample of size $n+1$ from a population having a continuous distribution function $F(x)$ and the probability density function $f(x)$. The joint probability density function of the order statistics is then given by

$$f(x_1, x_2, \dots, x_n, x_{n+1}) = (n+1)! \prod_{i=1}^{n+1} f(x_i), \quad (-\infty < x_1 < \dots < x_n < x_{n+1} < \infty). \quad (2.1)$$

Since the marginal distribution of $x_1 \leq x_2 \leq \dots \leq x_n$ is given by

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= \int_{x_n}^{\infty} f(x_1, x_2, \dots, x_n, x_{n+1}) dx_{n+1} \\ &= (n+1)! \prod_{i=1}^n f(x_i) \cdot \{1 - F(x_n)\}, \end{aligned} \quad (2.2)$$

the conditional probability density function of x_{n+1} given x_1, \dots, x_n turns out to be

$$f(x_{n+1} | x_1, \dots, x_n) = \begin{cases} f(x_{n+1}) / \{1 - F(x_n)\}, & x_n < x_{n+1} < \infty, \\ 0, & \text{elsewhere.} \end{cases} \quad (2.3)$$

The conditional expectation and the conditional variance of x_{n+1} are given respectively by

$$E\{x_{n+1} | x_1, \dots, x_n\} = \frac{1}{1 - F(x_n)} \int_{x_n}^{\infty} x f(x) dx, \quad (2.4)$$

and

$$\text{Var}\{x_{n+1} | x_1, \dots, x_n\} = \frac{1}{1 - F(x_n)} \int_{x_n}^{\infty} x^2 f(x) dx - [E\{x_{n+1} | x_1, \dots, x_n\}]^2. \quad (2.5)$$

These values may be used to predict x_{n+1} . In many cases, however, it is impossible to get an explicit formula for the value of the definite integrals involved in (2.4) and (2.5). For example, consider the Gamma probability density:

$$f(x) = \frac{\beta^v}{\Gamma(v)} e^{-\beta x} x^{v-1}, \quad (0 < x < \infty; \quad v > 0, \quad \beta > 0), \quad (2.6)$$

which is of frequent use in statistical climatology. In this case, the parameters are estimated easily from the sample mean \bar{x} and variance s^2 by the moment method,

$$v = \bar{x}^2 / s^2, \quad \beta = \bar{x} / s^2, \quad (2.7)$$

while the definite integrals in (2.4) and (2.5) are not obtained explicitly.

We shall now introduce an approximation procedure to calculate (2.4) and (2.5). Firstly, it is well known (Kendall and Stuart 1958) that $F(x_n)$ is well approximated by its expected value:

$$F(x_n) \approx n/(n+1) \quad (= E\{F(x_n)\}). \quad (2.8)$$

Secondly, we shall take the exponential smoothing procedure given by

$$f(x) \approx f(x_n) e^{-c(x-x_n)}, \quad c > 0, \quad x_n < x < \infty. \quad (2.9)$$

Then, since

$$y(X_i) = C(X_i)T(X_i) + \varepsilon(X_i) \quad (3.2)$$

where

X_1, X_2, \dots, X_n : given ordered sample,

$T(X_{i+1}), T(X_i)$: true value of the return period corresponding to X_{i+1} and X_i ,

$A(X_i)$: transformation coefficients for prediction

$\eta(X_i)$: error of the return period,

$y(X_i)$: value of the return period corresponding to the observed ordered sample,

$C(X_i)$: transformation coefficients for observation,

$\varepsilon(X_i)$: error of observation system.

The identification of return periods is given by the following process. First, $y(X_i)$ ($i = 1, 2, \dots, n$) are obtained by empirical extrapolation under the Thomas Plot rule. In order to identify $T(X_i)$ ($i = 1, 2, \dots, n$), we assume the following linear equations:

$$\hat{T}(X_i) = \alpha y(X_i) + \beta, \quad (i = 1, 2, \dots, n), \quad (3.3)$$

for which the coefficients α and β are obtained as the solutions of the equations:

$$\begin{aligned} \sum_{i=1}^n \{(\alpha-1)y(X_i) + \beta c + \varepsilon(X_i)\}cy(X_i) &= 0, \\ \sum_{i=1}^n \{(\alpha-1)y(X_i) + \beta c + \varepsilon(X_i)\}c &= 0, \end{aligned} \quad (3.4)$$

where C is the transformation coefficient of the Kalman filtering model, and in practical application, we replace $C(X_i)$ by the constant c as a simple model.

Solving the equations in (3.4), and assuming that $E\{\varepsilon(X_i)\} = 0$, we get the relation (see, Arimoto 1977):

$$\hat{T}(X_i) = \bar{T} + \frac{c\sigma_\varepsilon^2}{\sigma_T^2 + c^2\sigma_\varepsilon^2}(y(X_i) - c\bar{T}). \quad (3.5)$$

We make use of this result to give a procedure of predicting the return period.

In the Kalman filtering theory, the transformation coefficients $A(X_i)$ are given functions, but in our present case they are unknown because of their dependence on the observed order statistics. Hence we assume the following relations to obtain the values of $A(X_i)$ successively:

$$A(X_1) = \hat{T}(X_2)/\hat{T}(X_1), \quad A(X_2) = \hat{T}(X_3)/\hat{T}(X_2), \dots, \quad A(X_{n-1}) = \hat{T}(X_n)/\hat{T}(X_{n-1}), \quad (3.6)$$

and then $A(X_n)$ can be determined by extrapolation under the linear model assumption:

$$A(X_i) = aA(X_{i-1}) + b(X_i - X_{i-1}) + \xi(X_i), \quad (i = 1, 2, \dots, n). \quad (3.7)$$

The coefficients a and b are obtained by the normal equation:

$$\left\{ \sum_{i=2}^n A(X_{i-1})^2 \right\} a + \left\{ \sum_{i=2}^n A(X_{i-1})(X_i - X_{i-1}) \right\} b = \sum_{i=2}^n A(X_i)A(X_{i-1}), \quad (3.8)$$

$$\left\{ \sum_{i=2}^n A(X_{i-1})(X_i - X_{i-1}) \right\} a + \left\{ \sum_{i=2}^n (X_i - X_{i-1})^2 \right\} b = \sum_{i=2}^n A(X_i)(X_i - X_{i-1}).$$

Solving the equation to get the solution \hat{a} and \hat{b} , we then have

$$\tilde{A}(X_i) = \hat{a} A(X_{i-1}) + \hat{b} (X_i - X_{i-1}). \quad (3.9)$$

Then, we can predict $T(X_{n+1})$ by the following recurrence relation:

$$\hat{T}(X_{n+1}) = \tilde{A}(X_n) \hat{T}(X_n), \quad (3.10)$$

where X_i is obtained by the relations (2.13).

4. NUMERICAL EXAMPLES

We shall show two examples for identification and prediction of the annual maximum snowfall amounts based on 20 records in Nara Prefecture and Toyama Prefecture in Japan.

The algorithm of the calculation of $T(X_{i+1})$ is as follows:

$$\hat{T}(X_{i+1}) = \hat{T}(X_i) + p_i c \sigma_{\varepsilon(X_i)} \{y(X_i) - (cT^*(X_i) \bar{\varepsilon}(X_i))\}$$

where

$$T^*(X_i) = A_{i-1} \hat{T}(X_{i-1}) + \bar{\eta}(X_{i-1}), \quad p_i = (m_i^{-1} + c^2 \sigma_{\varepsilon(X_i)}^{-1})^{-1}, \quad m_i = A_{i-1}^2 p_{i-1} + \sigma_{\eta(X_{i-1})},$$

with initial conditions

$$T^*(X_0) = \bar{T}(X_0), \quad m_0 = \sigma_{T(X_0)}.$$

Here, we have assumed that $\bar{\varepsilon}(X_i)$, $\bar{\eta}(X_i)$, $\sigma_{\varepsilon(X_i)}$ and $\sigma_{\eta(X_i)}$ are known. The $T^*(X_i)$, m_i and p_i are defined by

$$T^*(X_i) = E\{T(X_i)\}, \quad m_i = E\{(T(X_i) - T^*(X_i))^2\}, \quad p_i = E\{(\hat{T}(X_i) - T(X_i))^2\}.$$

TABLE 1.

Numerical results obtained by applying the Kalman filter model (Nara Pref.)

n	x	T	\hat{T}	\tilde{A}	$\hat{\sigma}_{\xi}^2$	P	M	T*
1	0	1.05	1.03	1.00	—	0.60	1.50	1.00
2	2	1.11	1.08	1.00	—	0.61	1.60	1.03
3	2	1.17	1.13	1.00	—	0.62	1.61	1.03
4	2	1.24	1.20	1.06	—	0.62	1.62	1.13
5	3	1.31	1.30	1.10	0.0003	0.63	1.70	1.27
6	3	1.40	1.41	1.10	0.0002	0.64	1.76	1.42
7	4	1.50	1.52	1.08	0.0002	0.64	1.77	1.55
8	4	1.62	1.63	1.07	0.0002	0.64	1.75	1.65
9	5	1.75	1.75	1.08	0.0002	0.63	1.74	1.75
10	6	1.91	1.90	1.09	0.0001	0.64	1.74	1.89

TABLE 1 (cont'd)

11	6	2.10	2.09	1.11	0.0001	0.64	1.76	2.08
12	7	2.33	2.33	1.12	0.0001	0.64	1.78	2.31
13	8	2.63	2.62	1.13	0.0001	0.64	1.80	2.60
14	8	3.00	2.99	1.15	0.0001	0.65	1.82	2.96
15	10	3.50	3.48	1.17	0.0001	0.65	1.85	3.44
16	12	4.20	4.16	1.21	0.0001	0.65	1.89	4.07
17	13	5.25	5.13	1.26	0.0002	0.66	1.96	5.03
18	16	7.00	6.84	1.36	0.0004	0.67	2.05	6.25
19	17	10.50	10.13	1.51	0.0010	0.69	2.24	9.30
20	22	21.00	19.40	2.07	0.0102	0.72	2.57	15.29
\hat{x}								
21	24.11		40.15	2.18	0.0096	0.80	4.08	
22	27.17		87.48			0.83	4.81	

Note: P is the error variance for identification and prediction.

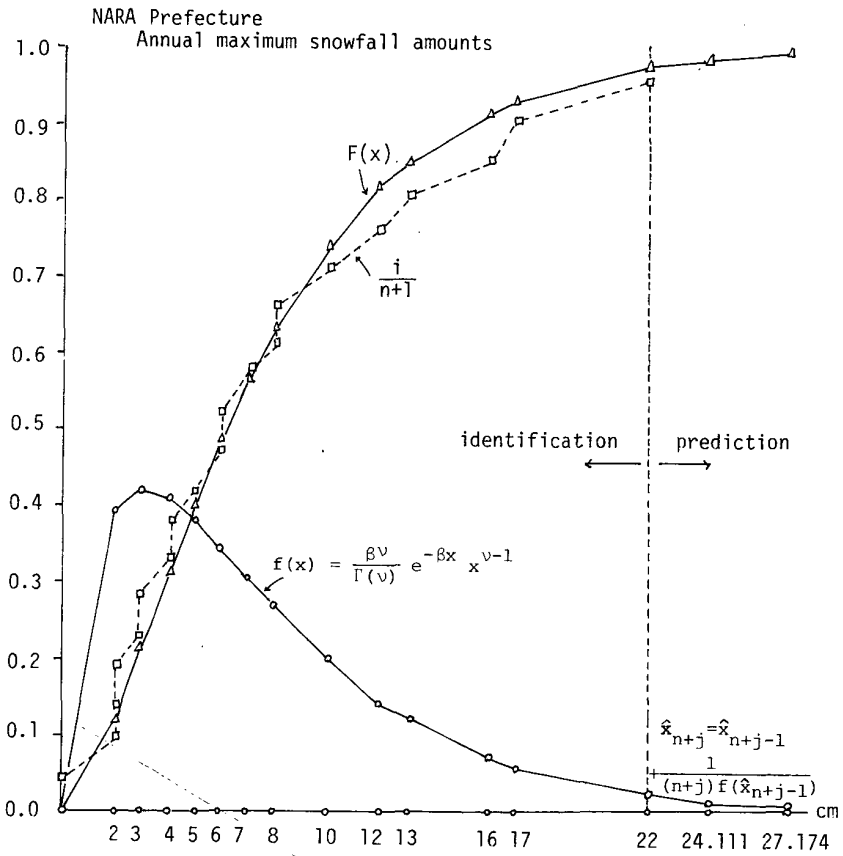


Fig. 1. Identification and prediction of extreme values (Nara Pref.).

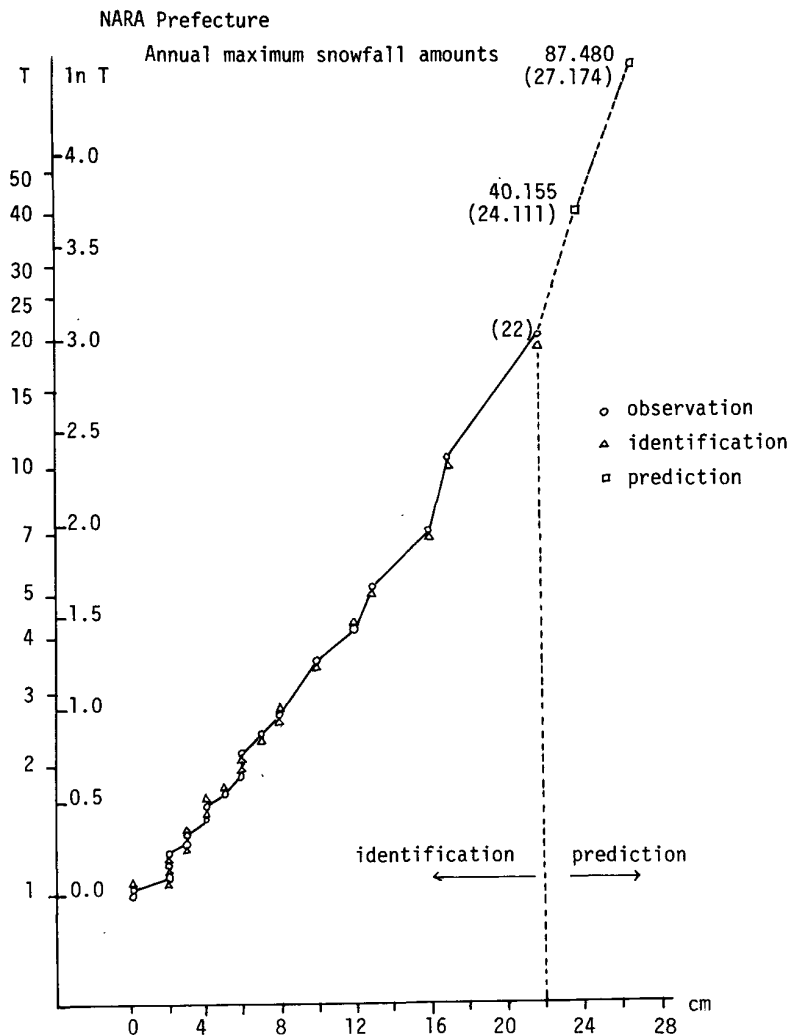


Fig. 2. Identification and prediction of return period (Nara Pref.)

TABLE 2.

Numerical results obtained by applying the Kalman filter model (Toyama Pref.)

n	x	T	\hat{T}	\tilde{A}	$\hat{\sigma}_{\xi}^2$	P	M	T^*
1	37	1.05	1.03	1.00	—	0.60	1.50	1.00
2	40	1.11	1.08	1.00	—	0.62	1.60	1.03
3	42	1.17	1.13	1.00	—	0.62	1.62	1.08
4	43	1.24	1.20	1.06	—	0.62	1.62	1.13
5	44	1.31	1.30	1.11	0.0001	0.63	1.69	1.27
6	46	1.40	1.41	1.10	0.0001	0.64	1.77	1.43
7	48	1.50	1.52	1.08	0.0002	0.64	1.77	1.55
8	59	1.62	1.63	1.01	0.0002	0.64	1.75	1.64
9	65	1.75	1.71	1.04	0.0002	0.62	1.64	1.64

TABLE 2 (cont'd).

10	66	1.91	1.86	1.10	0.0004	0.63	1.68	1.78
11	68	2.10	2.08	1.13	0.0004	0.64	1.76	2.05
12	70	2.33	2.34	1.13	0.0004	0.64	1.81	2.35
13	75	2.63	2.63	1.13	0.0003	0.65	1.83	2.65
14	77	3.00	2.99	1.15	0.0003	0.65	1.83	2.98
15	104	3.50	3.48	1.13	0.0003	0.65	1.85	3.43
16	107	4.20	4.10	1.19	0.0003	0.65	1.83	3.93
17	108	5.25	5.13	1.27	0.0005	0.66	1.92	4.89
18	110	7.00	6.83	1.36	0.0008	0.67	2.05	6.49
19	165	10.50	10.12	1.46	0.0016	0.69	2.24	9.27
20	208	21.00	19.22	2.16	0.0040	0.71	2.48	14.80
	\hat{x}							
21	213.38		41.55	2.21	0.0037	0.81	4.33	
22	219.46		91.92			0.83	4.98	

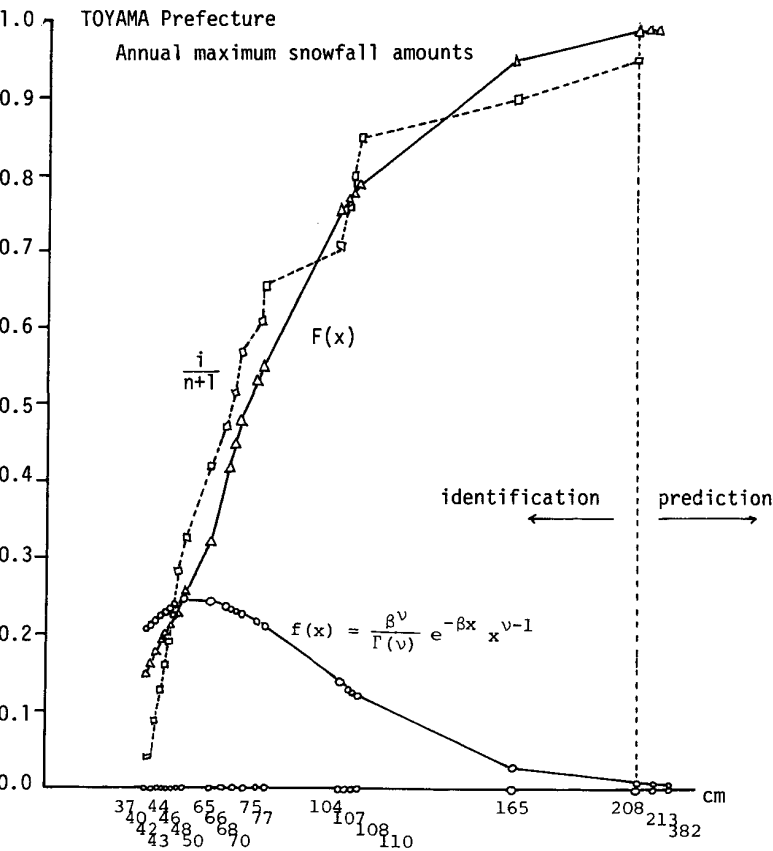


Fig. 3. Identification and prediction of extreme values (Toyama Pref.).

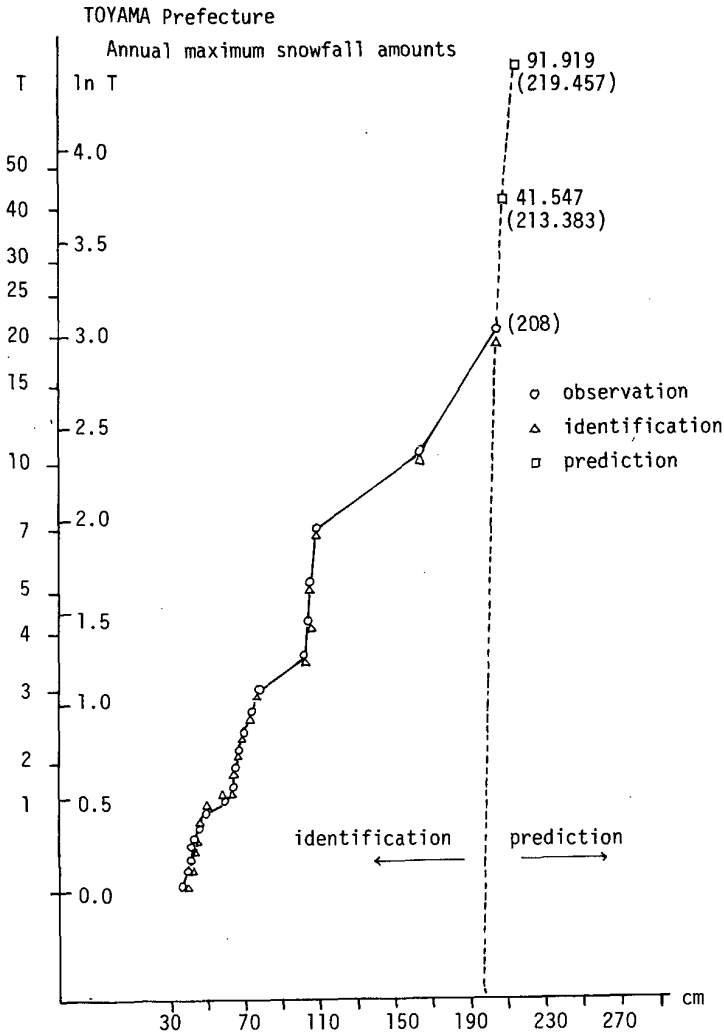


Fig. 4. Identification and prediction of return period (Toyama Pref.)

5. CONCLUDING REMARKS

We can identify the return periods $T(X_i)$, $i = 1, 2, \dots, n$, by (3.5), and the prediction system can be denoted by

$$\hat{T}(X_{n+1}) = \{a(X_{n-1}) + b(X_n - X_{n-1})\} \hat{T}(X_n) + \xi(X_n) \hat{T}(X_n) + \eta(X_n).$$

Evaluation of the prediction error, $\xi(X_n) \hat{T}(X_n) + \eta(X_n)$, has not been done yet, which

is left open to the study in the future.

In the last place, summarizing the theoretical and empirical knowledges up to the present, we shall list the ranges of sample size n and the methods which are suitable for n in each of the ranges in the identification and prediction problem of extreme values and return periods:

- 1) $n < 10$. In this case, there will be no method applicable in practice.
- 2) $10 < n < 30$. We can use the curve fitting method to the empirical data by making use of the Kalman filtering theory as studied above.
- 3) $30 < n < 70$. The same method as in 2) above will be applicable in this case, too.

The theoretical models, Gumbel type, Weibull type and others, have been shown to be applicable in practice.

- 4) $70 < n$. Theoretical models, $e^{-e^{-y}}$ -type, $e^{-y^{-\alpha}}$ -type, are applicable, but a correction of the difference between the theoretical model and the practical model would be necessary.

ACKNOWLEDGEMENT

The authors wish to express their thanks to Dr. L. Billard for her kind advice and comments.

REFERENCES

- Arimoto S., 1977. Kalman Filter. Saiensu-Sha, Tokyo (in Japanese).
 Gibbons, J.D., 1971. Nonparametric Statistical Inference. McGraw Hill.
 Jazwinski, A.H., 1970. Stochastic Process and Filtering Theory. Math. in Sci. and Engin. Vol. 64, Academic P.
 Kendall, M.G. and Stuart, T., 1958. The Advanced Theory of Statistics. Vol. 1. Charles Griffin, London.
 Morrison, G.W. and Pike, D.H., 1977. Kalman filtering applied to statistical forecasting. Manag. Sci. 23.
 Suzuki, E., 1968. Statistical Meteorology. Chijin Shokan, Tokyo (in Japanese).
 Suzuki, E., 1975. Statistical Methods in Environmental Science. Chijin Shokan, Tokyo (in Japanese).
 Takeuchi, K., 1963. Mathematical Statistics. Toyo Keizai Shinpo Sha, Tokyo (in Japanese).
 Japanese Standard Association (ed.), 1972. Statistical Tables and Formulae with Computer Applications (in Japanese).

ON THE USE OF EXPONENTIAL SMOOTHING FOR THE ESTIMATION OF CLIMATIC ELEMENTS

M. OGAWARA

Chiba Univ. of Commerce, Chiba (Japan)

ABSTRACT

Ogawara, M. On the use of exponential smoothing for the estimation of climatic elements. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

According to the agreement of the World Meteorological Organization (WMO), the latest 30 arithmetical mean values of meteorological elements are currently adopted as the climatic records and they are to be revised every ten years. This scheme is based on the consideration of climatic change.

In spite of the consideration, however, the 30 year arithmetical mean (AM) can not always represent the recent climatic situation owing to the climatic change. Here, by the word "recent" we mean the latest 10 year period, which may be a period of present-day human lives and activities.

On this subject, we shall show that the method of exponential smoothing (ES) is generally better than AM at least in diminishing the bias.

For routine work, however, the author proposes the use of ES with a common value of smoothing parameter at all meteorological stations in the world and for all climatic elements. The common value will have to be decided by the WMO on a basis of world-wide examinations.

1. A STOCHASTIC MODEL FOR CLIMATIC CHANGE

Let $X(t)$ be a monthly mean or total of a meteorological element for the year t . In almost all cases we may suppose that $X(t)$ is a stochastic process

$$X(t) = m(t) + Y(t) \quad , \quad (t = 0, \pm 1, \dots) \quad , \quad (1.1)$$

where $m(t) = EX(t)$ is a trend function and $Y(t)$ is a stationary process with $EY(t) = 0$. Here, we assume that

I. $m(t)$ is approximately linear in t over one or two decades T :

$$m(t + u) = a + bu \quad , \quad |u| < T \quad (t = 0, \pm 1, \dots) \quad , \quad (1.2)$$

where $a = m(t)$ and $b = b(t)$,

II. $Y(t)$ is an AR(1) process (so called red noise) with

$$EY(s)Y(t) = \sigma^2 \rho^{|s-t|} \quad , \quad |\rho| < 1 \quad , \quad (1.3)$$

where in most actual cases ρ is very small.

If we assume that $m(t)$ is identically constant, i.e., $X(t)$ is stationary, then

the autocorrelation of $Y(t)$ would have more complicated structure; On the other hand, if we assume that $Y(t)$ is an independent process, then $m(t)$ may be of a high order in t .

In relation to our stochastic model, we may define some climatological concepts as follows:

Definition 1. $m(t)$ is called the function of climatic change, the value of $m(t)$ at the t -th year is the climatic value of the element, and $m(t) - m(s)$, ($s < t$), is called the climatic change in the interval (s, t) .

Definition 2. Climatic record is an estimate (or a prediction) $\hat{m}(t + u)$ of $m(t + u)$ ($u \geq 0$) from the data up to the year t .

In the following, we consider two special cases, $u = 0$ and $u = 5.5$, where $\sum_{u=1}^{10} (a + bu) / 10 = a + 5.5 b$.

Now, the current method (AM) of estimating $m(t + u)$ is given by

$$\hat{m}_a(t + u) = \frac{1}{N} \sum_{j=0}^{N-1} X(t - j), \quad (N = 30), \quad (1.4)$$

while exponential smoothing (ES) is

$$\hat{m}_e(t + u) = \alpha \sum_{j=0}^{N-1} \beta^j X(t - j), \quad (N \leq \infty), \quad (1.5)$$

where $0 < \beta < 1$ and $\alpha = (1 - \beta)/(1 - \beta^N)$. A merit of ES is in the fact that the effect of a possibly large discrepancy of an approximate value $a - b_j$ from the true value $m(t - j)$ for a large j is diminished by the small weight β^j . The reason why we do not use higher order exponential smoothing is that we want to put larger weights on the data more up-to-date.

2. CLIMATIC CHANGE IN JAPAN

Let $m_1 = \sum_{j=0}^9 (a - bj)/10$, $m_2 = \sum_{j=10}^{19} (a - bj)/10$, then we have $b = (m_1 - m_2)/10$. If we put

$$\hat{b} = (\bar{x}_1 - \bar{x}_2) / 10, \quad (2.1)$$

where \bar{x}_1 and \bar{x}_2 are 10 year sample mean values corresponding to m_1 and m_2 , respectively, then $E\hat{b} = b$ and if the difference $\bar{x}_1 - \bar{x}_2$ is significant we can recognize a climatic change in the two decades. A test of significance can be done by assuming that the data are independent random samples from normal distributions with a common variance (if necessary the data should be suitably transformed) and an estimate of the variance is approximately given by

$$\hat{\sigma}^2 = \frac{1}{19} \sum_{j=0}^{19} [X(t-j) - \hat{a} + \hat{b}j]^2 / 20, \quad (\hat{a} = (\bar{x}_1 + \bar{x}_2)/2 + (19/2)\hat{b}). \quad (2.2)$$

$\hat{\sigma}$ may be also approximated by $\hat{\sigma} \approx (R_1 + R_2)/6$, R_1 and R_2 being the range in each decade.

By using this method, about 150 (stations) times 12 (months) cases are tested for the period 1951-1970 on the rough significance level 5 percent, and the following results are obtained:

<u>Temperature</u> is <u>decreasing</u> at <u>48</u> stations in <u>Feb.</u> or <u>Mar.</u>		
	<u>increasing</u>	<u>12</u>
<u>Precipitation</u> <u>decreasing</u> { <u>7</u>		
		<u>17</u>
	<u>increasing</u>	{ <u>2</u>
		<u>1</u>
		<u>May</u>
		<u>Feb.</u> or <u>Mar.</u>
		<u>Sep.</u> or <u>Oct.</u>
		<u>Jun.</u> or <u>Jul.</u>
		<u>Sep.</u>

Even if $\bar{x}_1 - \bar{x}_2$ is not significant on the level of 5 percent, there must be more or less climatic change.

Some examples are shown in Fig. 1.1 and Fig.1.2 in which horizontal dotted lines are 30 year arithmetical means and the horizontal full lines are \bar{x}_1 and \bar{x}_2 for which $\bar{x}_1 - \bar{x}_2$ is significant. In almost all cases, the discrepancy between the climatic record by WMO and the last 10 year mean is quite large.

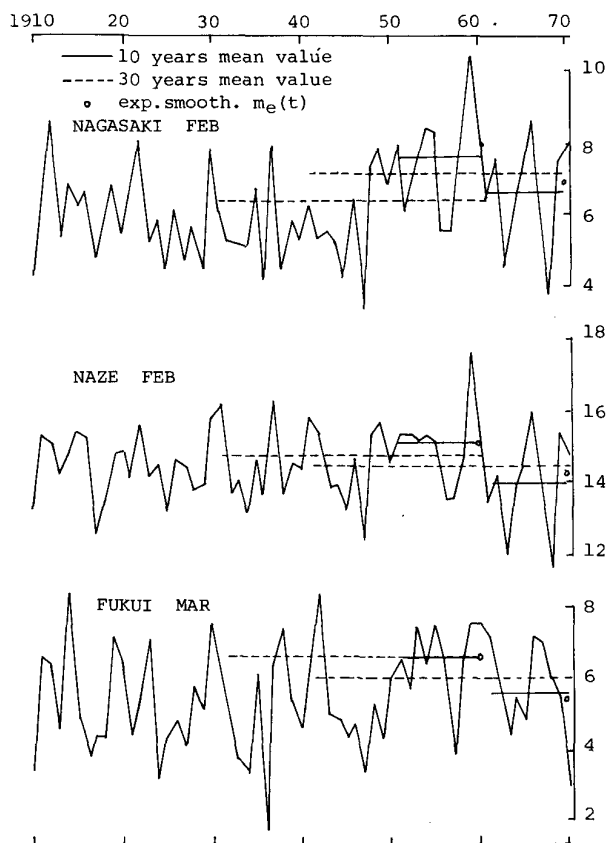


Fig. 1.1. Examples of climatic change in Japan - Temperature (°C).

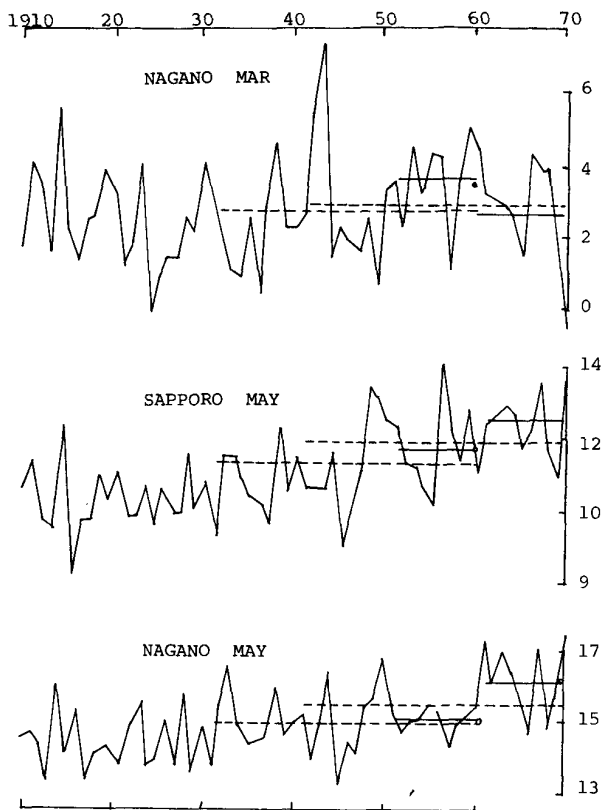


Fig. 1.1. (cont'd)

3. OPTIMUM VALUE OF SMOOTHING PARAMETER

The mean square error of $\hat{m}_e(t+u)$ is given by

$$\begin{aligned}
 \text{MSE(ES)} &= E[\hat{m}_e(t+u) - (a+bu)]^2 \\
 &= \sigma^2 \alpha^2 \sum_{j,k=0}^{\infty} \rho^{j+k} |j-k| + b^2 \left(\frac{\beta}{1-\beta} + u \right)^2 \\
 &= \sigma^2 \frac{(1-\beta)(1+\beta\rho)}{(1+\beta)(1-\beta\rho)} + b^2 \left(\frac{\beta}{1-\beta} + u \right)^2 = [\text{variance(ES)}] + [\text{bias(ES)}]^2.
 \end{aligned} \tag{3.1}$$

The optimum value of $\beta = 1 - \alpha$ can be obtained by minimizing MSE(ES) w.r.t. β , and the optimum value of $\alpha = 1 - \beta$ can be shown to be a root of the equation

$$\begin{aligned}
 &[\gamma^2 \rho - (\gamma^2 + u - 1) \rho^2] \alpha^5 - [2(\gamma^2 + u - 1) \rho - (2\gamma^2 + 6u - 7) \rho^2] \alpha^4 \\
 &+ [\gamma^2 - u + 1 + 2(5u - 6) \rho - (\gamma^2 + 13u - 19) \rho^2] \alpha^3 + [4u - 5 - 2(8u - 13) \rho + (12u - 25) \rho^2] \alpha^2 \\
 &- [4(u - 2) - 8(u - 3) \rho + 8(u - 3) \rho^2] \alpha - 4(1 - \rho)^2 = 0,
 \end{aligned} \tag{3.2}$$

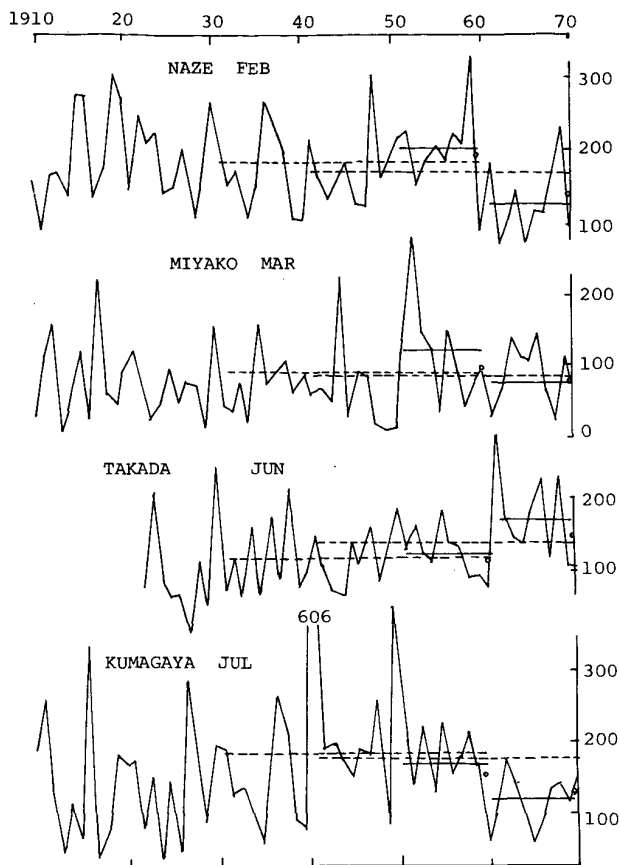


Fig. 1.2. Examples of climatic change in Japan — Precipitation (MM)

where $\gamma^2 = \sigma^2/b^2$. If $\rho = 0$, (3.2) reduces to the cubic equation

$$(\gamma^2 - u + 1)\alpha^3 + (4u - 5)\alpha^2 - 4(u - 2)\alpha - 4 = 0, \quad (3.3)$$

and if $u < \gamma^2$ the root exists uniquely in the interval $(0, 1)$. The optimum value of $\beta = 1 - \alpha$ for $u = 0, 1, 2$ and 5.5 can be found from the diagram in Fig. 2. For large values of γ^2 the variation of β is slight. For a finite N in (1.5) and a non-zero but small ρ , the value of β given by the diagram is approximately optimum.

On the other hand,

$$\begin{aligned} \text{MSE(AM)} &= E[\hat{m}_a(t+u) - (a+bu)]^2 \\ &= \frac{\sigma^2}{N} \left[\frac{1+\rho}{1-\rho} - \frac{2\rho(1-\rho^N)}{N(1-\rho)^2} \right] + b^2 \left[\frac{N-1}{2} + u \right]^2 \\ &= [\text{variance(AM)}] + [\text{bias(AM)}]^2. \end{aligned} \quad (3.4)$$

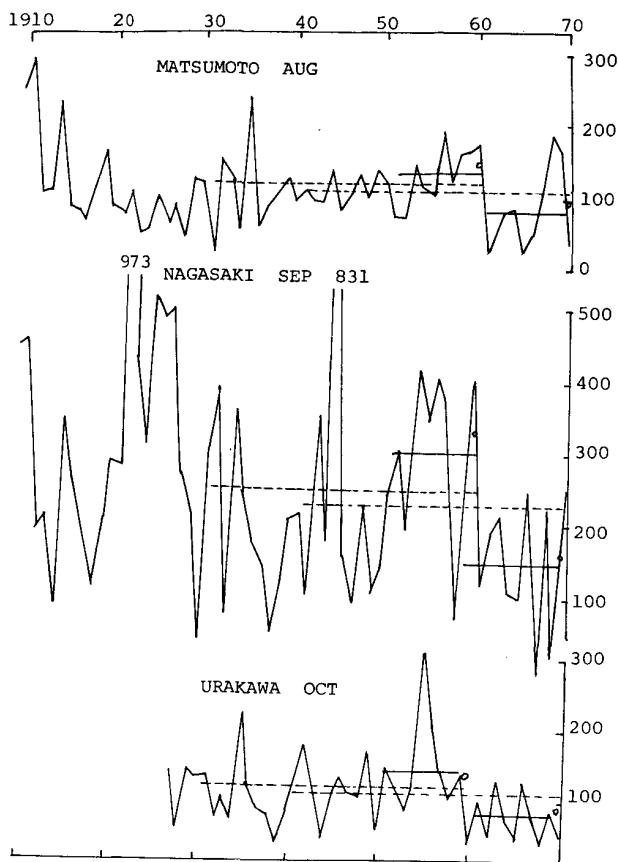


Fig. 1.2. (cont'd)

The optimum value of N is a root of the equation

$$N^4 + (2u - 1)N^3 - 2\gamma^2 \frac{1+\rho}{1-\rho} N + \frac{4\gamma^2 \rho}{(1-\rho)^2} = 0, \quad (3.5)$$

and if $\rho = 0$ the equation reduces to

$$N^3 + (2u - 1)N^2 - 2\gamma^2 = 0. \quad (3.6)$$

If we use the optimum β and the optimum N for each γ^2 , ES and AM are comparable in both variance and bias as we see in Table 1. When $\rho = 0$, inequalities $(1 - \beta)/(1 + \beta) \geq 1/N$ and $\beta/(1 - \beta) \leq (n - 1)/2$ are mutually equivalent, and hence the variance and the bias are complementary. Supposing however we attach more importance to the bias than to the variance, for any N and u , $\beta \leq (N-1)/(N+1)$ implies $\text{bias(ES)} \leq \text{bias(AM)}$. For instance, when $N = 30$ this is true if $\beta \leq 0.9355$, i.e.,

$$\gamma^2 \leq 10^4.$$

The influence of ρ is only on the variance, and if $\rho > 0$ and $2N > (1-\beta\rho)/(1-\beta)$ or if $\rho < 0$ and $2N < (1-\beta\rho)/(1-\beta)$ then $\text{var}(\text{ES}) < \text{var}(\text{AM})$.

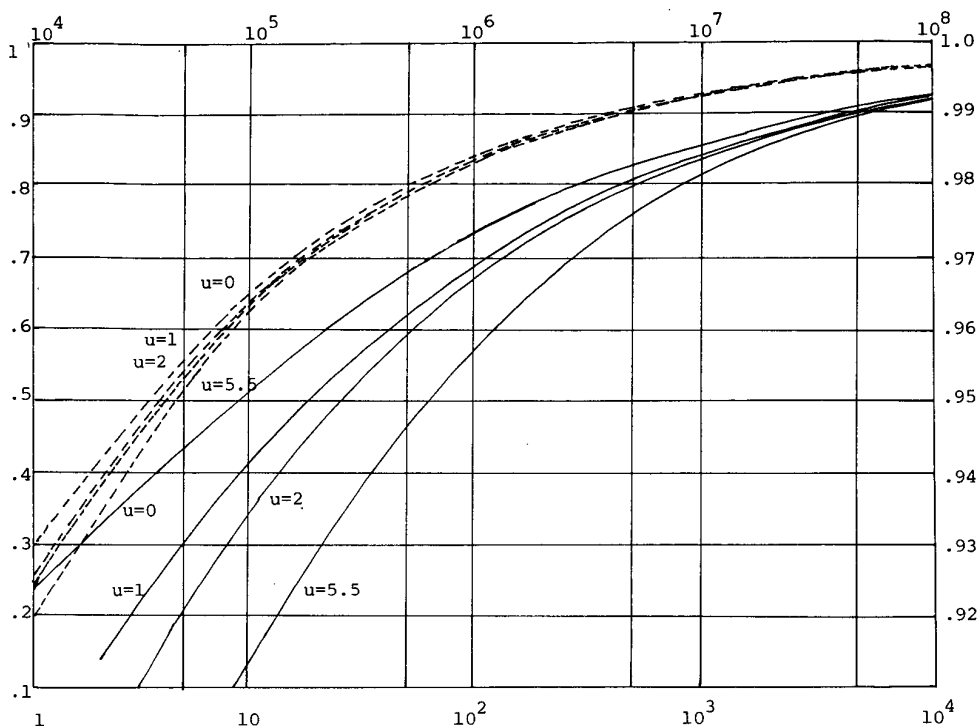


Fig. 2. Optimum value of β for $\gamma^2 = \sigma^2/b^2$, ($\rho = 0$). (Scales of dotted curves are up and right.)

The variance of unbiased estimate $\hat{a} + \hat{b}u$ Of the regression $a + bu$ by a sample $X(t-j)$ ($0 \leq j \leq N$), is given by

$$\text{var}(\hat{a} + \hat{b}u) = \frac{12\sigma^2}{N(N+1)} \left(2N-1+u+\frac{u^2}{N-1} \right)$$

which is fairly large, because we can not take N owing to the generally short span of the linearity of local trend. For instance, when $N = 20$ and $u = 0$, $\text{var}(\hat{a} + \hat{b}u) = 1.114\sigma^2$, which suggests that this method is inadequate.

4. ESTIMATION OF OPTIMUM ES PARAMETER AND SOME EXAMPLES

Most simple way of estimating optimum β is to use (2.1), (2.2) and the diagram

TABLE 1.

Variance and bias for optimum value of parameter ($\rho = 0$).

	γ^2 (σ^2/b^2)	Optimum values of parameter		(Variance) Coefficient of σ^2		(Bias) Coefficient of b	
		β	N	ES	AM	ES	AM
u = 0	10	0.512	3	0.3288	0.3333	1.0	1.0
	50	0.680	5	0.1905	0.2000	2.1	2.0
	10 ²	0.739	6	0.1501	0.1667	2.8	2.5
	10 ³	0.857	13	0.0770	0.0769	6.0	6.0
	10 ⁴	0.930	27	0.0363	0.0370	13.3	13.0
	1.3×10 ⁴	0.935	30	0.0336	0.0333	14.4	14.5
	10 ⁵	0.966	59	0.0172	0.0169	28.4	29.0
	10 ⁶	0.984	126	0.0081	0.0079	61.5	62.5
	10 ⁸	0.997	585	0.0015	0.0017	332.3	292.0
u = 5.5	10 ²	0.584	4	0.2626	0.2500	6.9	7.0
	10 ³	0.818	10	0.1001	0.1000	10.0	10.0
	10 ⁴	0.921	24	0.0411	0.0417	17.2	17.0
	1.8×10 ⁴	0.9355	30	0.0333	0.0333	20.0	20.0
	10 ⁵	0.964	55	0.0180	0.0182	33.1	32.5
	10 ⁶	0.984	122	0.0081	0.0082	67.0	66.0
	10 ⁸	0.997	582	0.0015	0.0017	337.8	296.0

(Fig.2) with $\gamma^2 = \hat{\sigma}^2 / \hat{b}^2$ and u. The error of the estimate \hat{b} may be inferred from the fact that $\text{var}(\hat{b}) = \sigma^2/500$ and the distribution of $t' = \sqrt{500\hat{b}}/\sqrt{20\hat{\sigma}^2/18} = 21.21/\hat{\gamma}$ is approximately non-central t with degrees of freedom $\nu = 18$ and non-centrality parameter $\Delta = b$. From the Pearson-Hartley (1972) Table 27, we can also find the approximate confidence limits of b. However, the sampling error of \hat{b} is fairly large owing to the small degrees of freedom, and so we should consider \hat{b} as a descriptive value for a special sample series.

Example 1. For Sapporo May Temperature (0.1°C), t = 1970, u = 5.5, we get $\hat{b} = 0.8$, $\hat{\sigma}^2 = 125$, $\gamma^2 = \hat{\sigma}^2/\hat{b}^2 = 195$. Thus we have a rough estimate $\hat{b} = 0.67$.

The confidence interval of b with confidence coefficient 90 percent is found to be (-0.09, 3.09) which illustrates the remark mentioned above.

For the same data, if we use the values of β ,

$$\beta = 0.60 \quad 0.67 \quad 0.70 \quad 0.80 \quad 0.90$$

then we get

$$\hat{m}_e(t + 5.5) = 125 \quad 125 \quad 125 \quad 124 \quad 122 \quad (0.1^\circ\text{C})$$

respectively. From such examples we can see the robustness of \hat{m} with respect to small change of β .

Some examples of ES with roughly estimated β are shown in Table 2 and Table 3 in comparison with AM. Table 3 suggests that the prediction of $m(t+5.5)$ by any method is very difficult owing mainly to the unexpected change of local trend. On the other hand, as we see in Table 2, Fig.1.1 and Fig.1.2, $\hat{m}_e(t)$ is much superior to 30 year AM, $\hat{m}_a(t)$.

TABLE 2.

Examples of estimation of $m(t)$.

Temperature (0.1°C)

Station	Month	ES		AM	\bar{x}_2	ES		AM	\bar{x}_1
		β	-1960 (\sqrt{MSE})	1931-60 (\sqrt{MSE})	1951 -60	β	-1970 (\sqrt{MSE})	1941-70 (\sqrt{MSE})	1961 -70
NAGASAKI	FEB	0.72	81 (7.9)	76 (26.3)	76	0.88	69 (5.1)	71 (6.5)	66
NAZE	FEB	0.79	151 (4.6)	147 (10.4)	151	0.76	143 (5.9)	145 (16.1)	140
FUKUI	MAR	0.76	67 (5.8)	66 (17.6)	66	0.78	54 (5.1)	60 (14.7)	56
NAGANO	MAR	0.84	35 (6.4)	27 (10.2)	36	0.79	23 (6.4)	29 (14.8)	26
SAPPORO	MAY	0.85	117 (4.1)	113 (6.2)	117	0.78	124 (5.2)	118 (11.8)	125
NAGANO	MAY	0.87	150 (2.4)	149 (3.1)	150	0.65	162 (2.1)	154 (16.0)	161

Precipitation (MM)

NAZE	FEB	0.84	200 (26)	184 (45)	205	0.72	137 (34)	170 (111)	129
MIYAKO	MAR	0.78	94 (35)	88 (86)	123	0.78	78 (27)	88 (68)	76
TAKADA	JUN	0.95*	119 (7)	121 (8)	125	0.74	146 (24)	140 (70)	173
KUMAGAYA	JUL	0.73	148 (34)	182 (110)	166	0.73	125 (24)	175 (79)	117
MATSUMOTO	AUG	0.65	159 (19)	116 (94)	131	0.73	97 (24)	107 (77)	78
NAGASAKI	SEP	0.85	339 (61)	253 (90)	303	0.68	163 (55)	233 (222)	151
URAKAWA	OCT	0.97**	121 (12)	121 (13)	143	0.74	76 (29)	114 (88)	83

* $N = 39$, $= 0.0570$; ** $N = 35$, $= 0.0456$

Remarks

1. In order to improve on the rough estimate $\hat{\beta}$, we may apply a successive approximation method as follows. Let a_1 , b_1 , σ_1^2 , γ_1^2 and $\beta_1 = 1 - \alpha_1$ be the first approximations. Then the second approximations a_2 , b_2 , σ_2^2 , γ_2^2 and β_2 can be obtained by minimizing $\alpha_1 \sum_{j=0}^{\infty} \beta_1^j [X(t-j) - a + bj]^2$ w.r.t. a and b , and they turn out to be

$$a_2 = (1 + \beta_1)s_1 - \alpha_1 s_2, \quad b_2 = (\alpha_1^2 / \beta_1)s_2 - \alpha_1 s_1,$$

$$\sigma_2^2 = \alpha_1 \sum_{j=0}^{\infty} \beta_1^j [X(t-j) - a_2 + b_2 j]^2 = s_0 - a_2 s_1 + b_2 s_2,$$

where

$$s_0 = \alpha_1 \sum_{j=0}^{\infty} \beta_1^j X(t-j)^2, \quad s_1 = \alpha_1 \sum_{j=0}^{\infty} \beta_1^j X(t-j) \quad \text{and} \quad s_2 = \alpha_1 \sum_{j=0}^{\infty} j \beta_1^j X(t-j).$$

However, this method may have a negative effect on the robustness of the estimate

TABLE 3.

Examples of prediction of $m(t + 5.5)$.Temperature (0.1°C)

Station	Month	ES		AM	\bar{x}_1	b/year		ES		AM
		β	-1960 ($\sqrt{\text{MSE}}$)	1931-60 ($\sqrt{\text{MSE}}$)	1961 -70	1941 -60	1951 -70	β	-1970 ($\sqrt{\text{MSE}}$)	1941-70 ($\sqrt{\text{MSE}}$)
NAGASAKI	FEB	0.55	85 (14.8)	63 (36.1)	66	1.8	-1.0	0.70	70 (13.2)	71 (20.2)
NAZE	FEB	0.70	153 (7.2)	147 (14.1)	140	0.7	-1.1	0.63	143 (10.0)	145 (22.1)
FUKUI	MAR	0.64	69 (11.1)	66 (24.1)	56	1.2	-1.0	0.66	50 (9.6)	60 (20.1)
NAGANO	MAR	0.79	37 (8.8)	27 (10.0)	26	0.7	-1.0	0.69	19 (10.1)	29 (20.2)
SAPPORO	MAY	0.80	118 (5.5)	113 (8.2)	125	0.4	0.8	0.67	125 (7.8)	118 (10.1)
NAGANO	MAY	0.85	151 (3.0)	149 (4.0)	161	0.06	1.10	0.42	166 (8.0)	154 (22.0)

Precipitation (MM)

NAZE	FEB	0.77	202 (41)	184 (60)	129	3.0	-7.6	0.58	133 (53)	170 (152)
MIYAKO	MAR	0.66	90 (50)	88 (118)	76	5.8	-4.6	0.66	127 (45)	88 (93)
TAKADA	JUN	0.89*	120 (10)	121 (8)	173	-0.2	4.8	0.60	135 (27)	140 (96)
KUMAGAYA	JUL	0.56	123 (63)	182 (151)	117	-7.5	-5.4	0.56	131 (38)	175 (109)
MATSUMOTO	AUG	0.35	170 (45)	116 (129)	78	6.4	-5.3	0.57	93 (45)	107 (106)
NAGASAKI	SEP	0.82	290 (81)	253 (120)	151	5.7	-15.2	0.46	181 (115)	233 (305)
URAKAWA	OCT	0.97**	121 (13)	121 (14)	83	0.3	-6.0	0.59	69 (52)	114 (121)

* $N = 39$, $\alpha = 0.1110$; ** $N = 35$, $\alpha = 0.0456$ \hat{m} and the proof of the convergence of this procedure may also be difficult.Example 2. For the data in Example 1, $a_1 = 121.0$, $b_1 = 0.8$, $\sigma_1^2 = 125$, $\gamma_1^2 = 195$, $\beta_1 = 0.67 = 0.70(\text{say})$:

then we have

 $a_2 = 126.8$, $b_2 = 0.78$, $\sigma_2^2 = 130$, $\gamma_2^2 = 214$, $\beta_2 = 0.68$.

2. Empirical minimization of $\sum_t [X(t+u) - \alpha \sum_j \beta^j X(t-j)]^2$ w.r.t. $\alpha = 1 - \beta$ may be inadequate, because the local trend varies with each short span of years and the long range mean value of α or β does not always adapt to the recent situation except in the stationary case.

5. GENERAL WEIGHTED MEAN AND ES

Let

$$M(w) = \sum_{j=0}^{\infty} w_j X(t-j), \quad \sum_{j=0}^{\infty} w_j = 1, \quad \sum_{j=0}^{\infty} j |w_j| < \infty \quad (5.1)$$

be a general weighted mean for the estimation of $m(t+u) = a+bu$; then the MSE is given by

$$\phi(w) = \sum_{j,k=0}^{\infty} (\sigma^2 \rho^{|j-k|} + b^2 jk) w_j w_k + 2b^2 u \sum_{j=0}^{\infty} j w_j + b^2 u^2. \quad (5.2)$$

If we put

$$w_j = \alpha \beta^j + \alpha v_j, \quad \sum_{j=0}^{\infty} v_j = 0, \quad \sum_{j=0}^{\infty} j |v_j| < \infty,$$

then we get

$$\Delta\phi = \phi(\alpha\beta^j + \alpha v_j) - \phi(\alpha\beta^j) = \sum_{j,k=0}^{\infty} A_{jk} v_j v_k + 2 \sum_{j=0}^{\infty} B_j v_j, \quad (5.3)$$

where

$$A_{jk} = \alpha^2 (\sigma^2 \rho^{|j-k|} + b^2 jk), \quad B_j = \frac{\alpha^2 \sigma^2 (1-\rho^2) \beta^{j+1} - \alpha (1-\rho\beta) \rho^{j+1}}{(\beta-\rho)(1-\rho\beta)} + b^2 (\alpha u + \beta) j, \quad (5.4)$$

($\beta \neq \rho$).

The first term on the right hand side of (5.3) is an infinite dimensional version of a positive definite quadratic form, while the sign of the second term depends on the series $\{v_j\}$.

Example. If $v_j = c$ for $j = 0, 1, \dots, n$; $= -d$ for $j = n+1, \dots, n+m$; $= 0$ for $j > n+m$, $nc = md$, then

$$\Delta\phi = \alpha^2 \sigma^2 c [n(c+d) + 2 - n(n+m)d] + b^2 cn(n+m) [cn(n+m)\alpha^2 - 4(u-1)\alpha - 4]/4.$$

Sufficient conditions for $\Delta\phi > 0$ are given by

$$c > 0, \quad \frac{n(c+d)+2}{n(n+m)} > \alpha > \frac{2(u-1)+2(u-1)^2+cn(n+m)}{cn(n+m)},$$

or

$$c < 0, \quad \frac{n(c+d)+2}{n(n+m)} < \alpha.$$

6. CONCLUDING REMARKS

Although our illustrations are limited to the data in Japan, we may conclude as follows:

1. Prediction of future climatic value $m(t+u)$ ($u > 0$) is difficult, but the estimate $\hat{m}_e(t)$ of the present (or very near future) situation by means of exponential smoothing is better in general than 30 year arithmetical mean $\hat{m}_a(t)$. To find the optimum value of exponential smoothing parameter, the curve with $u = 0$ in the diagram Fig.1

may be used.

2. However, it may be complicated to estimate and to use different smoothing coefficient for each station and for each climatic element. For routine work, therefore, it may be convenient and yet effective to use an appropriately decided common smoothing coefficient throughout all the stations in the world and for all the climatic elements.

3. The common value of $\beta = 1 - \alpha$ will have to be decided by the WMO on a basis of world-wide examinations. However, it should neither be too small nor too large, and it may be reasonable to select a value around 0.90. Then, in the presence of a climatic change the bias will be smaller than that for the 30 year AM, and in the stationary case the variance of the estimate will be comparable to that for the 30 year AM.

ACKNOWLEDGEMENT

The author thanks the reviewer, Professor T. Jayachandran, for his kind advices and comments.

REFERENCES

- Ogawara, M., 1969. On exponential smoothing. Tokyo Woman's Christ. Coll. Science Rep.: 7-11.
 Ogawara, M. and Ohkubo, E., 1972. On exponential smoothing. II. *ibid.* :19-23.
 Pearson, E.S. and Hartley, H.O. (ed.), 1972. Biometrika Table for Statisticians. 2. Cambridge Univ. P.
 The Japan Meteor. Agency, Climatic Records of Japan. -1954, 1951-60, 1961-70.

AN OPTIMUM LINEAR RESTRICTION IN THE ESTIMATION PROBLEM FOR A GENERALIZED LINEAR MODEL AND ITS APPLICATION TO CLIMATIC DATA

E.SUZUKI, T.OOHASHI and S.HONGO

Inf. Sci. Res. Center, Aoyama-Gakuin Univ., Tokyo (Japan)

ABSTRACT

Suzuki, E., Oohashi, T. and Hongo, S., An optimum linear restriction in the estimation problem for a generalized linear model and its application to climatic data. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.31, 1979

A method of imposed linear restriction for a generalized linear model has already been shown by some authors; however, in practical applications it is rather hard to determine a specific linear restriction because there can be many linear restrictions to obtain a conditional BLUE. So we impose one additional condition on the linear restriction in order to find the conditional BLUE having the minimum trace of its variance-covariance matrix. Further, the set of these estimators contains an estimator obtained by using the Moore-Penrose generalized inverse.

Since a method using the above conditional BLUE has not been applied so far to climatic data, we show an application of the method to a particular example, which is an attempt among possible approaches to the environmental impact assessment. Our method is applicable to various types of data and its algorithm is very simple.

1. INTRODUCTION

In this paper a useful method is proposed to obtain a minimum variance, linear, conditionally unbiased estimator ("conditional BLUE") for a generalized linear model, and an application of the method to certain climatological data is demonstrated.

A generalized linear model is defined as

$$y = X\beta + \epsilon, \quad (1.1)$$

where y is the $n \times 1$ vector of observations, X is an $n \times k$ ($n \geq k$) constant matrix of rank(X) = r ($< k$), β is a $k \times 1$ vector of unknown parameters, and ϵ is an $n \times 1$ vector of errors with zero expectation and variance matrix $E(\epsilon\epsilon') = \sigma^2 I$.

In ordinary regression analysis, the normal equation and its solution are given, respectively, by $X'X\hat{\beta} = X'y$ and $\hat{\beta} = (X'X)^{-1}X'y$ (assuming that $r = k$), and $\hat{\beta}$ is known to be a minimum variance, linear unbiased estimator. However, in a generalized linear model (1.1), $X'X$ has no inverse and the solution of the normal equation is not unique. Therefore, a linear restriction is imposed on the model (1.1) to obtain a unique solution. Such a method has been already considered by Chipman (1964) and summarized by Pringle and Rayner (1971), in which, however, the conditional BLUE is not uniquely determined. So we impose an additional condition on those

conditional BLUE's in order to obtain a useful conditional BLUE minimizing the trace of its variance-covariance matrix.

Now, we state an example to which our present method will be applied in later section.

A tunneling work caused a trouble of ground water withdrawal during the period 1973-1977 in an area at the foot of Mt. Haruna, Japan, composed of rice field, town field and mulberry field. This is a rectangular area with one side east to west of 6.5 km and the other side of 8.0 km, and the tunnel is running under the center line of this area (see Fig.1).

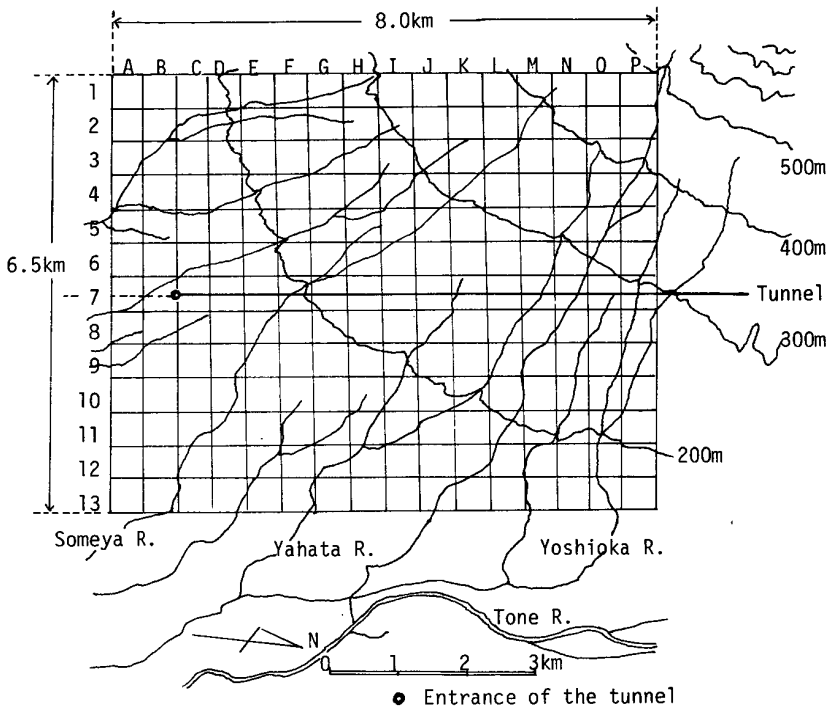


Fig. 1. Topographical map around the tunnel.

The data in Table 1 are obtained from the partitioned areas, each being a $0.5\text{km} \times 0.5\text{km}$ square. The observation vector y contains six discrete indices 0,1,2,3,4,5 corresponding to six states of ground water withdrawal due to tunneling work over the previous five years, the index of which is an intersection-multiplicity about regions of water loss obtained by observation of wells every year. The constant matrix X consists of four kinds of variables: The column $X(0)$, being composed of 1, corresponds to the constant term of the regression equation. The columns $X(i)$ ($i = 1,2,\dots,7$) are continuous variables indicating the topography and infiltration

TABLE 1.

Original data.

N	NO	y	X(0)	X(1)	X(2)	X(3)	X(4)	X(5)	X(6)	X(7)	Z(1)	Z(2)	Z(3)	Z(4)	Z(5)	Z(6)	Z(7)	Z(8)
1	7A	0	1	138	70	-0.2	0.8	0.0	0.0	1.0	0	0	1	0	1	0	0	0
2	8A	0	1	136	84	0.6	0.9	0.5	0.0	10.0	1	0	0	0	1	0	0	0
3	9A	0	1	133	87	0.2	1.0	1.0	0.0	10.0	0	1	0	0	1	0	0	0
4	7B	0	1	145	76	-0.2	1.0	0.0	0.0	1.0	0	0	1	0	0	0	1	0
5	8B	0	1	145	95	0.3	1.0	0.5	0.0	1.0	0	0	1	0	0	0	0	1
6	9B	0	1	141	95	0.5	0.9	1.0	0.0	10.0	0	0	1	0	1	0	0	0
7	5C	0	1	163	94	0.5	1.7	1.0	1.0	1.0	1	0	0	1	0	0	0	0
8	6C	1	1	159	91	0.4	1.5	0.5	1.0	1.0	0	1	0	0	0	0	1	0
9	7C	0	1	156	88	0.3	1.4	0.0	2.0	1.0	0	0	1	0	0	0	1	0
10	8C	1	1	154	101	0.2	1.0	0.5	1.0	1.0	0	0	1	0	0	0	0	1
11	9C	1	1	149	104	0.9	0.8	1.0	0.0	1.0	0	0	1	0	1	0	0	0
12	10C	0	1	141	101	0.7	0.5	1.5	0.0	10.0	0	0	1	0	1	0	0	0
13	11C	0	1	136	115	0.5	0.5	2.0	0.0	10.0	0	0	1	0	1	0	0	0
14	4D	0	1	185	118	0.5	2.1	1.5	1.0	1.0	0	0	1	0	0	1	0	0
15	5D	0	1	179	114	0.7	2.1	1.0	1.0	1.0	0	0	1	0	0	1	0	0
16	6D	1	1	174	110	0.6	1.8	0.5	2.0	1.0	0	0	1	0	0	0	1	0
17	7D	2	1	168	105	0.6	1.4	0.0	3.0	1.0	0	1	0	1	0	0	0	0
18	8D	4	1	163	111	0.6	1.1	0.5	1.0	1.0	0	0	1	0	0	0	0	1
19	9D	0	1	156	112	1.0	0.8	1.0	0.0	1.0	0	0	1	0	0	0	1	0
20	10D	0	1	147	109	1.0	0.7	1.5	0.0	10.0	0	0	1	0	1	0	0	0
21	3E	0	1	212	142	1.1	2.2	2.0	2.0	5.0	0	0	1	0	0	1	0	0
22	4E	0	1	203	134	1.1	2.1	1.5	2.0	1.0	0	0	1	0	0	1	0	0
23	5E	0	1	195	128	0.6	1.6	1.0	2.0	1.0	0	0	1	0	0	1	0	0
24	6E	0	1	189	123	0.9	1.7	0.5	4.0	1.0	0	0	1	1	0	0	0	0
25	7E	2	1	181	117	0.9	1.5	0.0	5.0	1.0	1	0	0	1	0	0	0	0
26	8E	1	1	172	118	1.2	1.0	0.5	2.0	1.0	0	0	1	0	0	0	1	0
27	9E	0	1	162	118	1.0	0.7	1.0	0.0	1.0	0	0	1	0	0	0	0	1
28	10E	0	1	153	119	1.0	0.7	1.5	0.0	1.0	1	0	0	0	1	0	0	0
29	11E	0	1	147	123	0.4	0.7	2.0	0.0	10.0	0	0	1	0	1	0	0	0
30	3F	0	1	232	161	1.3	2.3	2.0	2.0	5.0	0	0	1	0	0	1	0	0
31	4F	0	1	221	151	1.2	2.2	1.5	2.0	1.0	0	0	1	0	0	1	0	0
32	5F	0	1	212	141	0.9	2.3	1.0	3.0	1.0	0	0	1	0	0	1	0	0
33	6F	0	1	304	134	1.1	1.6	0.5	5.0	1.0	0	0	1	1	0	0	0	0
34	7F	1	1	194	127	1.0	1.4	0.0	8.0	1.0	0	0	1	0	0	0	1	0
35	8F	1	1	184	125	1.4	1.7	0.5	3.0	1.0	0	0	1	0	0	0	1	0
36	9F	1	1	171	120	1.5	1.3	1.0	1.0	1.0	0	0	1	0	0	0	1	0
37	10F	1	1	160	117	1.0	0.8	1.5	0.0	1.0	0	0	1	1	0	0	0	0
38	4G	0	1	240	165	1.4	1.9	1.5	3.0	5.0	0	0	1	0	0	1	0	0
39	5G	0	1	230	156	0.9	1.8	1.0	4.0	1.0	0	0	1	0	0	1	0	0
40	6G	1	1	219	149	1.5	2.1	0.5	6.0	1.0	0	0	1	0	0	1	0	0
41	7G	1	1	208	137	1.1	1.7	0.0	10.0	1.0	0	0	1	0	0	1	0	0
42	8G	2	1	196	134	1.5	1.1	0.5	4.0	1.0	0	0	1	0	0	1	0	0
43	9G	1	1	181	128	2.0	1.0	1.0	1.0	1.0	0	0	1	1	0	0	0	0
44	10G	1	1	167	121	1.2	0.9	1.5	0.0	1.0	0	1	0	1	0	0	0	0
45	4H	0	1	256	182	1.5	1.9	1.5	3.0	5.0	0	0	1	0	0	0	0	1
46	5H	0	1	246	173	0.8	1.9	1.0	5.0	1.0	0	0	1	0	0	1	0	0
47	6H	1	1	236	165	1.5	1.7	0.5	8.0	1.0	0	0	1	0	0	1	0	0
48	7H	1	1	222	154	1.7	1.7	0.0	12.0	1.0	0	0	1	0	0	1	0	0
49	8H	1	1	207	147	1.9	1.3	0.5	5.0	1.0	0	0	1	0	0	1	0	0
50	9H	0	1	190	134	1.9	1.1	1.0	1.0	1.0	0	0	1	1	0	0	0	0
51	3I	0	1	288	218	1.8	2.2	2.0	3.0	5.0	0	0	1	1	0	0	0	0
52	4I	0	1	274	203	1.3	2.2	1.5	4.0	5.0	1	0	0	0	0	0	0	1
53	5I	0	1	262	192	1.5	1.7	1.0	5.0	5.0	0	0	1	0	1	0	0	0
54	6I	2	1	249	179	1.5	1.3	0.5	9.0	1.0	0	0	1	0	1	0	0	0
55	7I	2	1	235	165	1.8	1.2	0.0	4.0	1.0	0	0	1	0	0	1	0	0
56	8I	1	1	216	152	2.5	0.9	0.5	5.0	1.0	0	0	1	0	0	1	0	0

(Table 1 cont'd)

57	9I	2	1	199	140	1.4	1.0	1.0	2.0	1.0	0	0	1	1	0	0	0	0
58	10I	1	1	184	132	2.2	1.0	1.5	0.0	1.0	0	1	0	1	0	0	0	0
59	2J	0	1	325	255	1.8	2.7	2.5	3.0	5.0	0	0	1	1	0	0	0	0
60	3J	0	1	309	239	1.7	2.6	2.0	4.0	5.0	1	0	0	1	0	0	0	0
61	4J	0	1	293	224	2.0	2.0	1.5	5.0	5.0	1	0	0	0	0	0	0	1
62	5J	0	1	276	209	1.9	1.4	1.0	6.0	5.0	0	0	1	0	1	0	0	0
63	6J	2	1	262	193	1.1	1.8	0.5	10.0	1.0	0	0	1	0	0	1	0	0
64	7J	1	1	246	175	2.6	1.4	0.0	16.0	1.0	0	0	1	0	0	1	0	0
65	8J	1	1	226	158	2.0	1.4	0.5	6.0	1.0	0	0	1	0	0	1	0	0
66	9J	2	1	211	149	1.6	1.6	1.0	2.0	1.0	0	0	1	1	0	0	0	0
67	10J	1	1	192	138	2.6	1.0	1.5	1.0	1.0	0	0	1	1	0	0	0	0
68	2K	0	1	348	277	1.9	2.6	2.5	4.0	5.0	0	0	1	1	0	0	0	0
69	3K	0	1	331	260	2.0	2.3	2.0	4.0	5.0	1	0	0	0	0	0	0	1
70	4K	0	1	309	240	3.1	1.6	1.5	5.0	5.0	1	0	0	0	0	0	0	1
71	5K	0	1	287	215	1.9	1.2	1.0	6.0	5.0	0	0	1	0	0	0	0	1
72	6K	1	1	272	197	1.4	0.5	0.5	11.0	5.0	0	0	1	0	0	0	0	1
73	7K	1	1	255	176	2.6	0.5	0.0	17.0	1.0	0	0	1	1	0	0	0	0
74	8K	1	1	237	186	1.4	1.1	0.5	7.0	1.0	0	0	1	1	0	0	0	0
75	9K	0	1	221	156	2.2	0.9	1.0	2.0	1.0	0	0	1	1	0	0	0	0
76	10K	0	1	202	145	2.3	1.1	1.5	1.0	1.0	0	0	1	1	0	0	0	0
77	2L	0	1	367	296	2.1	1.8	2.5	4.0	5.0	0	0	1	1	0	0	0	0
78	3L	0	1	349	279	2.1	1.8	2.0	5.0	5.0	1	0	0	0	0	0	0	1
79	4L	0	1	325	255	3.4	2.1	1.5	5.0	5.0	1	0	0	0	1	0	0	0
80	5L	3	1	300	225	2.3	1.8	1.0	7.0	5.0	0	0	1	0	1	0	0	0
81	6L	3	1	279	201	2.5	0.9	0.5	11.0	5.0	1	0	0	0	1	0	0	0
82	7L	2	1	259	177	2.1	0.3	0.0	17.0	1.0	1	0	0	0	0	0	0	1
83	8L	3	1	243	168	1.5	0.2	0.5	7.0	1.0	1	0	0	1	0	0	0	0
84	9L	1	1	227	159	2.1	0.4	1.0	3.0	1.0	0	0	1	1	0	0	0	0
85	10L	2	1	210	150	1.8	0.7	1.5	1.0	1.0	0	0	1	0	0	1	0	0
86	11L	1	1	195	142	1.5	0.6	2.0	0.0	1.0	0	0	1	0	1	0	0	0
87	12L	0	1	184	138	1.1	0.5	2.5	0.0	1.0	1	0	0	0	1	0	0	0
88	13L	0	1	173	134	1.4	0.5	3.0	0.0	1.0	0	0	1	0	1	0	0	0
89	4M	0	1	340	267	3.2	1.3	1.5	6.0	5.0	0	1	0	0	1	0	0	0
90	5M	3	1	311	239	3.4	0.7	1.0	7.0	5.0	0	1	0	0	1	0	0	0
91	6M	3	1	283	210	2.9	0.2	0.5	11.0	5.0	1	0	0	0	0	0	0	1
92	7M	4	1	261	181	2.0	0.4	0.0	17.0	5.0	1	0	0	0	0	0	0	1
93	8M	3	1	246	173	1.6	0.6	0.5	7.0	1.0	0	0	1	0	0	0	0	1
94	9M	1	1	234	164	1.1	1.3	1.0	3.0	1.0	0	0	1	0	1	0	0	0
95	10M	2	1	218	153	2.5	1.3	1.5	1.0	1.0	0	0	1	0	1	0	0	0
96	11M	2	1	200	142	1.7	0.6	2.0	0.0	1.0	1	0	0	0	1	0	0	0
97	12M	0	1	187	136	1.3	0.3	2.5	0.0	1.0	0	0	1	0	1	0	0	0
98	13M	0	1	176	131	1.3	0.3	3.0	0.0	1.0	0	0	1	0	1	0	0	0
99	5N	1	1	317	257	3.2	0.6	1.0	7.0	5.0	0	0	1	0	1	0	0	0
100	6N	2	1	290	230	2.9	1.3	0.5	12.0	5.0	1	0	0	0	1	0	0	0
101	7N	3	1	269	206	1.8	1.3	0.0	18.0	5.0	0	1	0	0	0	0	0	1
102	8N	2	1	253	193	1.9	1.1	0.5	7.0	1.0	1	0	0	0	0	0	0	1
103	9N	3	1	240	180	1.1	1.1	1.0	3.0	1.0	0	0	1	0	1	0	0	0
104	10N	4	1	223	164	2.9	-0.3	1.5	1.0	1.0	0	0	1	0	1	0	0	0
105	11N	2	1	203	148	1.7	0.0	2.0	0.0	1.0	0	0	1	0	1	0	0	0
106	12N	1	1	189	139	1.5	0.0	2.5	0.0	1.0	0	0	1	0	1	0	0	0
107	13N	0	1	177	131	1.1	0.0	3.0	0.0	1.0	1	0	0	0	0	0	0	1
108	5O	1	1	328	270	2.8	2.0	1.0	8.0	5.0	0	0	1	0	1	0	0	0
109	6O	4	1	304	245	2.7	1.9	0.5	13.0	5.0	0	0	1	0	1	0	0	0
110	7O	5	1	282	222	2.3	1.6	0.0	20.0	5.0	1	0	0	0	0	0	0	1
111	8O	2	1	262	203	2.3	0.9	0.5	8.0	5.0	1	0	0	0	0	0	0	1
112	9O	3	1	244	186	1.8	0.8	1.0	3.0	1.0	0	0	1	0	1	0	0	0
113	10O	3	1	225	169	2.6	0.8	1.5	1.0	1.0	1	0	0	0	1	0	0	0
114	11O	2	1	205	151	1.8	0.6	2.0	0.0	1.0	1	0	0	0	0	0	0	1
115	12O	2	1	190	141	1.8	0.3	2.5	0.0	1.0	1	0	0	0	0	0	1	0

(Table 1 cont'd)

116	130	1	1	176	132	1.5	-0.4	3.0	0.0	1.0	0	0	1	0	0	0	1	0
117	6P	2	1	318	261	2.6	1.3	0.5	14.0	5.0	0	0	1	0	1	0	0	0
118	7P	2	1	293	233	3.2	0.8	0.0	21.0	5.0	0	0	1	0	1	0	0	0
119	8P	3	1	268	211	2.5	0.3	0.5	8.0	5.0	0	1	0	0	1	0	0	0
120	9P	2	1	250	195	1.5	0.5	1.0	3.0	5.0	1	0	0	0	0	0	0	1
121	10P	2	1	232	181	2.6	0.9	1.5	1.0	1.0	1	0	0	0	1	0	0	0
122	11P	2	1	212	164	2.0	0.9	2.0	0.0	1.0	1	0	0	0	1	0	0	0
123	12P	2	1	194	148	2.1	0.6	2.5	0.0	1.0	0	1	0	0	0	0	1	0

X(1): above sea-level(m), X(2): altitude of impermeable layer(m), X(3): gradient of topography of vertical direction to the route(°), X(4): gradient of topography of parallel direction to the route(°), X(5): horizontal distance from the route(km), X(6): incidence angle to the route(°), X(7): infiltration coefficient(10^{-4} , cm/sec).

coefficient. Columns Z(1) through Z(8) are indicator variables, the first three of which denote the state of land utilization and the last five the state of the soil: Z(1), Z(2) and Z(3) indicate a rice field, a town field and a mulberry field, respectively, and for each observation point only one component of the (0,1)-vector (Z(1),Z(2),Z(3)) takes the value 1. Similarly, Z(j) indicates the j-th state of the soil, $j = 4, 5, 6, 7, 8$. Thus, between the columns of the matrix X there exist two linearly dependent relations, $X(0) = \sum_{i=1}^3 Z(i)$ and $X(0) = \sum_{i=4}^8 Z(i)$, and hence, a methodology using a generalized linear regression model is necessary to estimate the effects of X(i) and Z(j) on the water withdrawal.

In the statistical climatology, an application of generalized linear model will provide us with a new methodology.

2. LINEAR RESTRICTIONS AND CONDITIONAL BLUE

In this section, we state some known results for the generalized linear model. Chipman (1976) gave the following definition to obtain a conditional BLUE of β :

Definition 1. An $m \times k$ ($m \geq k-r$) matrix L is said to be complementary to the $n \times k$ matrix X if the following two conditions are satisfied: (1) $\text{rank}(X) + \text{rank}(L) = k$, and (2) $uX + vL = 0$ implies $uX = vL = 0$, u and v being vectors of respective orders n and m. Furthermore, L is said to be polar to X, if the condition (2) is replaced by a stronger condition: (2)' $XL' = 0$. If L is complementary to X and c is in the column space of L, then the equation $Lc = 0$ is called a set of complementary linear restrictions for the generalized linear model.

The conditional BLUE of β subject to a set of complementary linear restrictions $L\beta = c$, and the variance-covariance matrix of $\hat{\beta}$ are given, respectively, by

$$\hat{\beta} = X^{\dagger}y + L^{\dagger}c, \quad \text{Var}(\hat{\beta}) = \sigma^2 X^{\dagger}(X^{\dagger})', \quad (2.1)$$

where $X^{\dagger} = (X'X + L'L)^{-1}X'$ and $L^{\dagger} = (X'X + L'L)^{-1}L'$, and in general, A^{\dagger} denotes

a generalized inverse of A satisfying $AA^+A = A$, $A^+AA^+ = A^+$ and $(AA^+)' = AA^+$. Here, it should be noted that the coefficient of determination in model fitting, R^2 , and the fitted value \hat{y} are fixed independently of the form of complementary linear restrictions imposed, while the BLUE $\hat{\beta}$ and $\text{Var}(\hat{\beta})$ may vary depending on the forms of the restrictions. To overcome this difficulty, we shall impose one more condition which provides us with a more useful conditional BLUE: Oohashi, Hongo and Yamaki (1979) considered an optimum linear restriction minimizing the trace of $\text{Var}(\hat{\beta})$, which will be restated below.

Definition 2. A set of complementary linear restrictions $L\beta = c$ and the matrix L are said to be optimum if L and $\hat{\beta}$ satisfy the following two conditions: (1) L is $(k-r) \times k$ and complementary to X , and (2) $\text{Var}(\hat{\beta})$ has a minimum trace.

Form this definition, we readily have the following results:

Theorem If a $(k-r) \times k$ matrix L is polar to X , then $L\beta = c$ is optimum.

Corollary Assume that L satisfies the condition (1) of definition 2. Let the nonzero eigenvalues of $X'X$ be $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_r$, while the eigenvalues of $X'(X^-)$ ' be $\mu_1 \geq \mu_2 \geq \dots \geq \mu_k$. Then it holds that

$$\mu_i \geq 1/\lambda_i \quad \text{for } i = 1, \dots, r; \quad \mu_i = 0 \quad \text{for } i = r+1, \dots, k,$$

where all the equalities hold when L is polar to X .

Let M be the class of all conditional BLUEs subject to a set of complementary linear restrictions, and N be that of all conditional BLUEs subject to a set of optimum linear restrictions. Then, $N \subset M$ and, by the above stated results, any BLUE in N minimizes the trace of $\text{Var}(\hat{\beta})$. Further, $\hat{\beta}$ in N is known to be of the form $\hat{\beta} = X^+y + L^+c$, where X^+ and L^+ designate the Moore-Penrose generalized inverse, from which we can see that $\hat{\beta}$ is not unique because there are many choices of c .

3. SOME EXAMPLES OF OPTIMUM L

With the aid of the above theorem, we can derive some practical forms of L corresponding to certain types of constant matrices X . These forms were already shown by Oohashi, Hongo and Yamaki (1979), but it will be beneficial to restate them:

(1) For the application to our present problem, i.e., to the matrix in Table 1, an optimum L can be chosen as

$$L = \begin{pmatrix} d_1 & 0 & 0 & 0 & 0 & 0 & 0 & -d_1 & -d_1 & -d_1 & 0 & 0 & 0 & 0 & 0 \\ d_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -d_2 & -d_2 & -d_2 & -d_2 & -d_2 \end{pmatrix}$$

where d_i ($i = 1, 2$) are arbitrary constants. In practical computation, $d_i = 1$ ($i = 1, 2$) and $c = [0, 0]'$ may be used.

In general, when X is composed of k_1 continuous variables and k_2 classes of

dummy variables with k_1 and k_2 non-negative, an optimum L can be obtained in a similar manner.

(2) Let $X = [x_{ij}]$ ($n \times k$) be of the form :

$$X = \left(\begin{array}{c|ccc} 1 & s_1 & s_2 & \dots & s_p & t_1 & t_2 & \dots & t_q \\ 1 & s_{i_1} & s_{i_2} & \dots & s_{i_p} & t_{j_1} & t_{j_2} & \dots & t_{j_q} \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & s_{v_1} & s_{v_2} & \dots & s_{v_p} & t_{w_1} & t_{w_2} & \dots & t_{w_q} \end{array} \right), \quad p = k_1, q = k_2, k = 1 + k_1 + k_2,$$

where s_1, \dots, s_p are all distinct and each row of the second submatrix is a permutation of the first row, and similarly for the second submatrix. Then, an optimum L is given by

$$L = \left(\begin{array}{cccc|cccc} s & -c_1 & -c_1 & \dots & -c_1 & 0 & 0 & \dots & 0 \\ t & 0 & 0 & \dots & 0 & -c_2 & -c_2 & \dots & -c_2 \end{array} \right)$$

$\underbrace{\hspace{10em}}_{k_1} \qquad \underbrace{\hspace{10em}}_{k_2}$

where $s = c_1 \sum_{i=1}^{k_1} s_i$, $t = c_2 \sum_{j=1}^{k_2} t_j$ and c_1 and c_2 are arbitrary non-zero constants.

(3) For a general X , one may construct an optimum L as

$$L = [\ell_1, \ell_2, \dots, \ell_{k-r}]',$$

by using the eigenvectors ℓ_i ($i = 1, 2, \dots, k-r$) corresponding to zero eigenvalues of $X'X$, but this is sometimes useless in practical application due to the large rounding error in calculating the eigenvectors.

In practical computation, the above forms of optimum L are very useful to obtain the conditional BLUE, and it also gives us a simple computational method of the Moore-Penrose generalized inverse of X .

4. COMPUTATIONAL RESULTS

Our problem is to explain the indices of the ground water withdrawal by the data of the topography, the infiltration coefficient, the land utilization and the soil. The values of R^2 , trace of $\text{Var}(\hat{\beta})$ and $\hat{\sigma}^2$ (unbiased estimate of σ^2) are tabulated in Table 2, in the case of original data y (Case(1)) and of transformed data $\ln(1+y)$ (Case(2)).

TABLE 2.

Estimated values.

	Case(1)	Case(2)
R^2	0.49051	0.54069
$\text{tr Var}(\hat{\beta})$	0.67564	0.13237
$\hat{\sigma}^2$	0.83065	0.16274

TABLE 3.

(a) Correspondence between y and \hat{y} in Case (1).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2										0	0	0				
3					0	0			0	0	0	0				
4				0	0	0	0	0	0	0	0	0	0			
5			0	0	0	0	0	0	0	0	0	3	3	1	1	
6			.28	.01	.34	.10	.45	.51	.68	1.08	1.03	0.99	2.47	2.14	1.43	
7		1	.94	.45	.55	.71	.77	1.14	1.73	1.24	1.64	2.17	2.59	2.42	1.81	2.25
8	0	0	0	2	2	1	1	1	2	1	1	2	4	3	5	2
9	.98	.65	.70	1.35	1.31	1.32	1.26	1.59	2.00	2.20	2.50	3.10	2.59	2.59	2.56	3.04
10	0	0	1	4	1	1	2	1	1	1	1	3	3	2	2	3
11	.39	.92	.93	1.03	1.16	.85	1.23	1.34	1.67	1.32	1.57	2.10	1.85	2.25	2.07	2.81
12	0	0	1	0	0	1	1	0	2	2	0	1	1	3	3	2
13	.26	-.14	1.14	.96	1.16	.81	.95	.87	.85	.59	1.09	1.36	1.06	1.88	1.71	1.64
14			0	0	0	1	1		1	1	0	2	2	4	3	2
15			-.04	-.02	1.52	.60	1.12		1.28	.87	.75	.98	1.04	2.10	1.87	1.92
16			0		0							1	2	2	2	2
17			-.06		-.18							.97	1.33	1.34	1.30	1.45
18												0	0	1	2	2
19												1.11	.83	1.06	1.10	1.26
20												0	0	0	1	
21												.61	.62	.96	.80	

(b) Correspondence between y^* and \hat{y}^* in Case (2).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1																
2										0	0	0				
3					0	0			0	0	0	0				
4					-.37	-.33			-.16	-.05	.16	.39				
5				0	0	0	0	0	0	0	0	0	0			
6			0	0	0	0	0	0	0	0	0	1.39	1.39	.69	.69	
7			.17	.10	.26	.13	.30	.33	.31	.53	.47	.47	1.24	1.11	.74	
8			.69	.69	0	0	.69	.69	1.10	1.10	.69	1.39	1.39	1.10	1.61	1.10
9			.56	.34	.34	.42	.47	.65	.86	.69	.75	1.02	1.24	1.19	.90	1.12
10	0	0	0	1.10	1.10	.69	.69	.69	1.10	.69	.69	1.10	1.61	1.39	1.79	1.10
11	.50	.43	.46	.75	.71	.76	.70	.87	1.07	1.17	1.28	1.46	1.17	1.21	1.18	1.49
12	0	0	.69	1.61	.69	.69	1.10	.69	.69	.69	.69	1.39	1.39	1.10	1.10	1.39
13	.14	.48	.48	.54	.73	.56	.74	.80	.97	.77	.90	1.09	.91	1.12	1.00	1.42
14	0	0	.69	0	0	.69	.69	0	1.10	1.10	0	.69	.69	1.39	1.39	1.10
15	.08	-.10	.63	.65	.62	.56	.60	.55	.53	.39	.65	.78	.54	.99	.97	.80
16			0	0	0	.69	.69		.69	.69	0	1.10	1.10	1.61	1.10	1.10
17			-.04	-.02	.81	.42	.67		.76	.56	.49	.62	.56	1.12	.65	1.02
18			0		0							.69	1.10	1.10	1.10	1.10
19			-.02		-.09							.53	.68	.72	.65	.76
20												0	0	.69	1.10	1.10
21												.57	.45	.57	.68	.78
22												0	0	0	.69	
23												.35	.34	.46	.55	

$$(y^* = \ell n(1 + y))$$

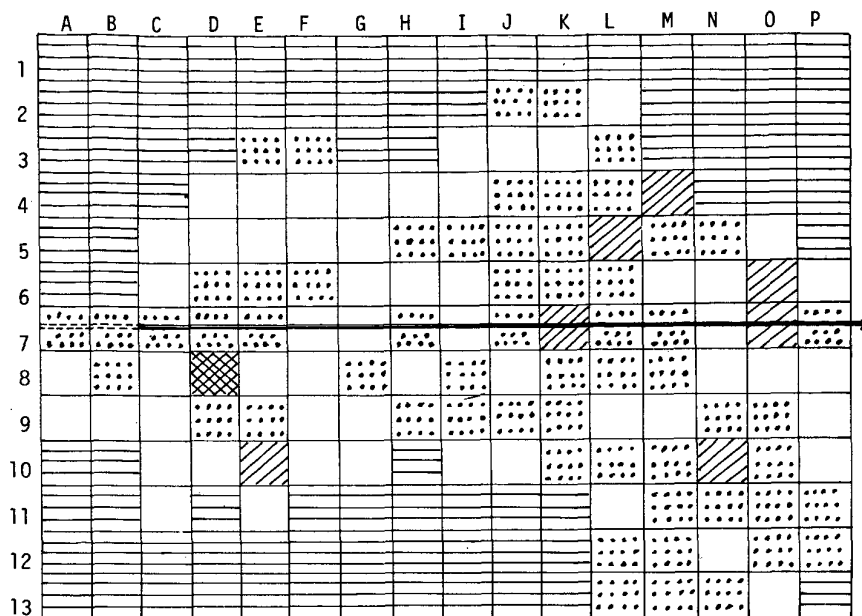


Fig. 2(a). Correspondence map in case (1).

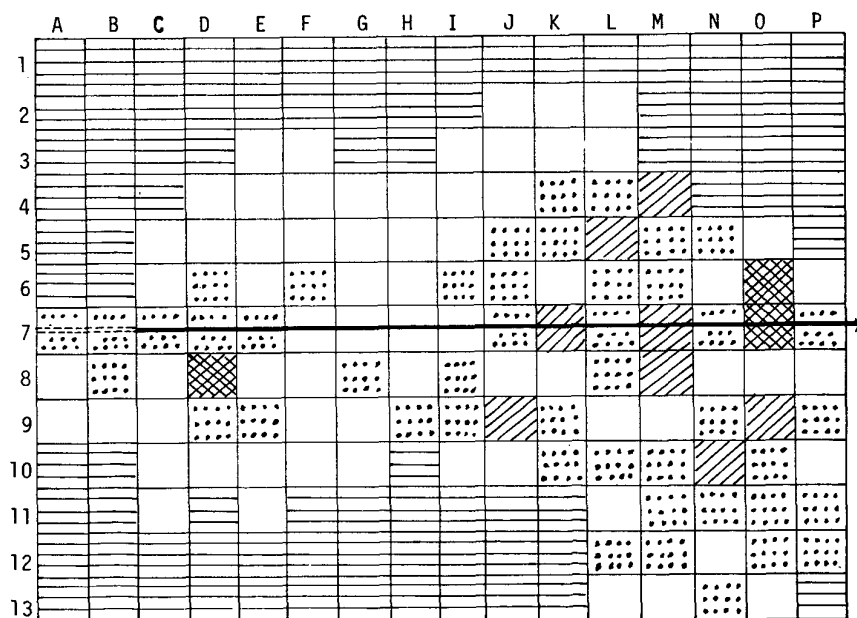










Fig. 2(b). Correspondence map in case (2).

Table 3(a) and (b) are the correspondence between the observed (upper line) and the estimated (lower line) for the case (1) and (2), respectively, and their respective maps are shown in Fig. 2(a) and (b), in which the patterns are defined as follows:

in Fig.2(a)		in Fig.2(b)	
	$\{(y, \hat{y}) \mid y - [y + 0.5] = 0\}$		$\{(y^*, \hat{y}^*) \mid y^* \in I_k \text{ and } \hat{y}^* \in I_k \text{ for some } k\}$
	$\{(y, \hat{y}) \mid y - [y + 0.5] = 1\}$		$\{(y^*, \hat{y}^*) \mid y^* \in I_k \text{ and } y^* \in I_{k-1} \text{ or } I_{k+1} \text{ for some } k\}$
	$\{(y, \hat{y}) \mid y - [y + 0.5] = 2\}$		$\{(y^*, \hat{y}^*) \mid y^* \in I_k \text{ and } y^* \in I_{k-2} \text{ or } I_{k+2} \text{ for some } k\}$
	$\{(y, \hat{y}) \mid y - [y + 0.5] = 3\}$		$\{(y^*, \hat{y}^*) \mid y^* \in I_k \text{ and } y^* \in I_{k-3} \text{ or } I_{k+3} \text{ for some } k\}$

where $[\cdot]$ designates the Gauss symbol, and the intervals I_k are defined by

$$I_k = [\ln(1 + k - 0.5), \ln(1 + k + 0.5)) , \quad k = 0, 1, 2, 3, 4, 5.$$

In Table 2 the value of R^2 in case (2) is greater than that in case (1). However, we cannot assert that the model in case (2) is better than case (1), because the correspondence between the observed and the estimated is shown to be better in case (1) than in case (2), as is seen in Fig.2(a) and (b).

Table 3 (a) and (b) are summarized respectively in Table 4 (a) and (b): In Table 4(a), the (i, j) -th element denotes the number of such cases that $y = i$ and $[\hat{y} + 0.5] = j$, and in Table 4(b) that $y^* = \ln(1+i)$ and $\hat{y}^* \in I_j$.

TABLE 4.

(a) Summarized result of Table 3(a)

$i \backslash j$	-1	0	1	2	3	4	5	Total
0	4	22	24	2	0	0	0	52
1	0	1	23	6	1	0	0	31
2	0	0	14	8	2	0	0	24
3	0	0	1	7	3	0	0	11
4	0	0	1	2	1	0	0	4
5	0	0	0	0	1	0	0	1
Total	4	23	63	25	8	0	0	123

(b) Summarized result of Table 3(b)

$i \backslash j$	0	1	2	3	4	5	Total
0	33	18	1	0	0	0	52
1	1	26	3	1	0	0	31
2	1	15	6	2	0	0	24
3	0	3	7	1	0	0	11
4	0	2	2	0	0	0	4
5	0	0	1	0	0	0	1
Total	35	64	20	4	0	0	123

A comparison between the sums of tridiagonal elements shows that the arrangement (a) is slightly closer to a diagonal matrix than (b).

5. CONCLUDING REMARKS

We have applied a new method to a particular example to get the results stated above. The correspondence between y and \hat{y} is rather satisfactory, though the use of linear model for our sample seems to be slightly rough. To level up the accuracy in the present estimation problem, further study of the hypotheses testing problem in generalized linear model and a better choice of effective variables will be needed, which are left open. The method we have presented here will have a wider applicability in climatological field and in other fields of science.

ACKNOWLEDGEMENTS

The authors are grateful to Dr. J. Gould and Dr. S. Ikeda for their kind advice and comments.

REFERENCES

- Chipman, J.S., 1964. On least squares with insufficient observations. J. Amer. Statist. Assoc. 59:1078-1111.
- Chipman, J.S., 1976. Estimation and aggregation in econometrics. In: Nashed, M.Z. (ed.) Generalized Inverse and Applications. Academic P.:449-769.
- Oohashi, T., Hongo, S. and Yamaki, N., 1979. An optimum linear restriction in the estimation problem for the generalized linear model. Aoyama Computer Science 7:1-8.
- Pringle, R.M. and Rayner, A.A., 1971. Generalized Inverse Matrices with Application to Statistics. Griffin.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

APPLICATION OF THE DISCRIMINANT ANALYSIS IN METEOROLOGY

G. DER-MEGREDITCHIAN

Meteorologie Nationale (EERM), Seine (France)

ABSTRACT

Der-Megreditchian, G. Application of the discriminant analysis in meteorology.
Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

Needs for the methods of statistical analysis of multidimensional data have been increasing for these years in meteorological science. Discriminant analysis plays an important role in meteorology in connection with forecasting and/or prediction of meteorological phenomena.

The present article briefly reviews the methods in discriminant analysis with emphasis on its application to meteorology. Some applications are also shown.

INTRODUCTION

The early applications of statistical methods in meteorology were in the use of descriptive statistics, e.g., calculation of means and variances, data smoothing, fitting theoretical models to empirical distributions, etc. Today, extensive use is made of the statistical methods of analysis of multidimensional data, say, principal components, linear and nonlinear multiple regression, factor analysis, canonical correlation, discriminant analysis, and so on, in which high speed computers play a crucial role.

Discriminant analysis, in particular, plays a prominent role because of a continuing need of the meteorologists to forecast atmospheric phenomena.

Statistical forecasting makes use of a stochastic model of weather, in which predictand $Y = \{y_1, y_2, \dots, y_m\}$ and predictors $X = \{x_1, x_2, \dots, x_n\}$ are random vectors simultaneously extracted during the random experiment.

Regression analysis is used if the predictand is a continuous variable, while discriminant analysis is applied if the predictand is a discrete variable indicating the occurrence of some atmospheric phenomena.

Meteorologists have for some time made use of an empirical discriminant analysis using an elaboration of some indices using two scalar predictors (Galway, Showalter, Telpher) and even three predictors (Molenat) for the forecasting of such meteorological phenomena as thunderstorms, hail, etc.

Some examples are given in later section.

1. THEORETICAL MODEL OF DISCRIMINANT ANALYSIS

The theoretical model of discriminant analysis can be explained in the following manner.

Let $\Omega = \sum_{i=1}^k A_i$ be a partition of a certain event Ω into k mutually exclusive subevents A_i 's.

Suppose a vector $X = \{x_1, \dots, x_n\}$ of n variables x_i called predictors contains information about the outcome of the random experiment. Define the cost $c[A_i:A_j]$ of predicting the outcome A_i when in fact A_j occurs. A decision function $g(X)$ is formulated for all outcomes X of the random experiment so as to predict an outcome A_j .

Given a quality criterion $Q[c]$ defined by the cost matrix $\{c[A_i:A_j]\}$, several strategies are possible, among which two of frequent use are minimizing the mathematical expectation of cost $E(c)$, and the minimax strategy which minimize the maximum cost.

In parametric discriminant analysis, if event A_i occurs, then it is assumed that the vector of predictors X was selected from a population characterized by a probability density function $f_{A_i}(x)$. Nature thus chooses the phenomena A_i with prior (climatic) probability p_i and the predictors vector according to the density $f_{A_i}(x)$.

If $k = 2$, then the "optimal" decision rule may be formulated by means of discriminant function

$$g(X) = \frac{p_2 f_{A_2}(X) c[A_1:A_2]}{p_1 f_{A_1}(X) c[A_2:A_1]},$$

in the following manner: if $g(X) > 1$, we forecast A_2 ; if $g(X) < 1$, we forecast A_1 .

2. PARAMETRIC DISCRIMINANT ANALYSIS - NORMAL DISTRIBUTION -

For multidimensional Gaussian populations we have

$$f_{A_i}(x) = (2\pi)^{-n/2} |V_{(i)}|^{-1} \exp\left[-\frac{1}{2}(x-\mu_{(i)})' V_{(i)}^{-1} (x-\mu_{(i)})\right],$$

where $\mu_{(i)} = \mu_{X(i)} = E_{A_i}[X]$ and $V_{(i)} = V_{XX(i)} = E_{A_i}[(X-\mu_{(i)})(X-\mu_{(i)})']$. It is convenient to use a discriminant function

$$w(X) = \ln g(X),$$

which gives the threshold value 0.

For the case $k = 2$, three cases are possible:

(a) If $V_{(1)} \neq V_{(2)}$, the discriminant function is quadratic:

$$w(x) = \frac{-1}{2} [(x-\mu_{(2)})' V_{(2)}^{-1} (x-\mu_{(2)}) - (x-\mu_{(1)})' V_{(1)}^{-1} (x-\mu_{(1)})] + \ln \frac{p_2 c[A_1:A_2] |V_{(1)}|^{1/2}}{p_1 c[A_2:A_1] |V_{(2)}|^{1/2}}.$$

(b) If $V_{(1)} = V_{(2)}$, the discriminant function is linear:

$$u(x) = (x-\mu_{(+)})' V_{(-)}^{-1} \mu_{(-)} + \ln \frac{p_2 c[A_1:A_2]}{p_1 c[A_2:A_1]}$$

with

$$\mu_{(+)} = \frac{1}{2} [\mu_{(1)} + \mu_{(2)}], \quad \mu_{(-)} = \mu_{(2)} - \mu_{(1)} \quad \text{and} \quad V = V_{(1)} = V_{(2)}.$$

(c) In case (a) above, although the optimal rule is quadratic, it is still preferable to use the best linear rule, defined by the Anderson-Bahadur (1962) method.

It is important to have a measure or index of information for each group of predictors which characterize the probability of decisional errors. The most common indices are the Mahalanobis distance $\Delta_n^2[A_1, A_2]$, the Kullback divergence $J_n[A_1, A_2]$ and the Bhattacharya distance $B_n[A_1, A_2]$, the last two of which are defined respectively by the formulas:

$$J_n[A_1, A_2] = \int_{R^n} f_{A_2}(x) \ln[f_{A_2}(x)/f_{A_1}(x)] dx + \int_{R^n} f_{A_1}(x) \ln[f_{A_1}(x)/f_{A_2}(x)] dx,$$

$$B_n[A_1, A_2] = -\ln \int_{R^n} [f_{A_1}(x) \cdot f_{A_2}(x)]^{1/2} dx.$$

In the present case of Gaussian populations these become:

$$J_n[A_1, A_2] = \frac{1}{2} \{ \text{tr}[V_{(2)} - V_{(1)}] [V_{(1)}^{-1} - V_{(2)}^{-1}] + \mu_{(-)}' [V_{(1)}^{-1} + V_{(2)}^{-1}] \mu_{(-)} \},$$

and

$$B_n[A_1, A_2] = \frac{1}{4} [-(V_{(1)}^{-1} \mu_{(1)} + V_{(2)}^{-1} \mu_{(2)})' (V_{(1)}^{-1} + V_{(2)}^{-1})^{-1} (V_{(1)}^{-1} \mu_{(1)} + V_{(2)}^{-1} \mu_{(2)}) + \mu_{(1)}' V_{(1)}^{-1} \mu_{(1)} + \mu_{(2)}' V_{(2)}^{-1} \mu_{(2)}] - \ln \left[|V_{(1)}^{-1} V_{(2)}^{-1}|^{1/4} / \left\{ \frac{1}{2} |V_{(1)}^{-1} + V_{(2)}^{-1}| \right\}^{1/2} \right].$$

If the discrimination function is linear, then the divergence becomes the Mahalanobis distance:

$$\Delta_n^2[A_1, A_2] = \mu_{(-)}' V_{(-)}^{-1} \mu_{(-)}.$$

3. SELECTION OF PREDICTORS

The need to sort out the useful information and reject the uninteresting or even spurious variables leads us to choose the group of k "best" predictors. We distinguish the following cases, 1 - 6:

1. Exhaustive selection method.

For each group of predictors, $\{x_{i_1}, \dots, x_{i_k}\}$, we compute the value of a specific index of information. We examine all the groups of k predictors. Unfortunately, this is possible only for very small number of predictors, say $k \leq 15$.

2. Progressive selection.

Here we distinguish the forward procedure in which we choose successively the best predictors, and the backward procedure in which we eliminate successively the worst predictors.

The actual calculations are facilitated the index of information is the Mahalanobis distance $\Delta_n^2[A_1, A_2]$, by using the following formula which gives the increase of Δ_n^2 when the number of predictors increases from k for the vector X to $(k+1)$ for the vector $\{X, x\}$:

$$\Delta_{k+1}^2[A_1, A_2] - \Delta_k^2[A_1, A_2] = \frac{[m_{x(-)} - v_{xX} v_{XX}^{-1} v_{X(-)}]^2}{\sigma_x^2 - v_{xX} v_{XX}^{-1} v_{Xx}}.$$

It allows us at step number $(k+1)$ of the selection process to replace $(N-k)$ inversions of matrix of order " $k+1$ ", by only one inversion of a matrix of order " k ", where N is the number of potential predictors.

3. Improved progressive selection.

At the step " k " of progressive selection we obtain the indices i_1, \dots, i_k of the best group of predictors. After that at the step " $k+1$ " we obtain the index i_{k+1} of the additional best predictor. At this time we examine successively each of the already chosen predictors x_{i_1}, x_{i_2}, \dots , when the others are fixed. An existing computer program permits us to perform this operation completely 5 times.

4. Random selection.

Progressive selection is not of course an optimal procedure. One may show that sometimes we can not find in this manner the best set of predictors, although the procedure is suboptimal in some sense. This is the reason we must use an entirely different algorithm, which uses the principle of exhaustive selection in some form.

By a random process we find " k " different integers i_1, \dots, i_k and we compute for the vector $\{x_{i_1}, \dots, x_{i_k}\}$ the value of the information index $Q_k[x_{i_1}, \dots, x_{i_k}]$. We repeat this operation and save the best scores by comparison with the former result. Our program is not expansive in time for several thousands of such random extractions.

At least it has happened in this manner that we have found some "excellent" predictors which escaped the progressive selection.

5. Adaptative random selection.

In this algorithm we want to combine both the advantages of the exhaustive and

the progressive selections. In this way we utilize the so-called punishment-reward principle. In the initial step of selection, each predictor is randomly chosen with a uniform probability $1/N$. For two groups of integers $\{i_1, \dots, i_k\}$ and $\{j_1, \dots, j_k\}$ we define the good ($x(G)$) and the bad ($x(B)$) predictors by comparing the index of information for each group. Each x which is not a good nor a bad predictor is called a remaining predictor ($x(R)$). The at the next step random extraction is performed in accordance with the following probabilities:

$$P_{k+1}[x(G)] = P_k[x(G)] + \delta P, \quad P_{k+1}[x(B)] = P_k[x(B)] - \delta P, \quad P_{k+1}[x(R)] = P_k[x(R)].$$

The correction δP is made in such a way that probabilities $P_k(x)$ become neither negative, nor greater than one. The application of this method to real data seems to show that it is a very good modification of the selection of scalar predictors.

6. Selection of random fields (or vectors).

In meteorological practice it may occur that among our predictors we have some meteorological fields or vectors. Then to obtain the forecasting scheme we will be interested in not destroying the physical sense of our predictors and performing the selection among those fields (or among those vectors) and not among the components of those fields (or vectors) in order to find the most informative of them.

In that way we associate with each field X a scalar variable z which, in the case of linear discrimination, contains the same information about process as the field X . In other words we have

$$\Delta_z^2[A_1, A_2] = \Delta_X^2[A_1, A_2].$$

Here we use the linear transformation

$$z = F_1' X,$$

where $F = [F_1, \dots, F_i, \dots, F_n]$ is a matrix having the following properties:

$$F' V_{XX} F = I_n, \quad F' (\mu_{(-)} \mu_{(-)}') F = L.$$

Here $L = [\ell_{ij}]$ is a diagonal matrix with diagonal elements ℓ_{ij} , δ_{ij} being the Kronecker delta. It is easy to see that only $\ell_{ij} \neq 0$.

Thus we have substituted for the field X the corresponding scalar variable z and we can now perform the selection of the scalars z with the help of the usual selection algorithms 1 - 5.

4. THE CHOICE OF THE "OPTIMAL" NUMBER OF PREDICTORS

The actual realization of the forecasting scheme is obtained with the empirical

discriminant function $g(X)$, since the theoretical discriminant function $g(X)$ is always unknown. All the parameters of $g(x)$ are obtained from the available sample of observations; the parameters p_i , $\mu_{X(i)}$, $V_{XX(i)}$ are estimated by \hat{p}_i , $\hat{\mu}_{X(i)}$, $\hat{V}_{XX(i)}$. That is why, when we add a predictor we on one hand obtain additional information, but on the other hand obtain some additional error. The choice of the optimal number of predictors is a compromise between these two contradictory properties.

We shall discuss only the case of linear discrimination whose measure is fully determined by means of the Mahalanobis distance.

Notice first that $\hat{\Delta}_n^2$ is a biased estimator of Δ_n^2 since we have

$$E[\hat{\Delta}_n^2] = \frac{T-2}{T-n-3} \left[\Delta_n^2 + \frac{n}{T_1} + \frac{n}{T_2} \right],$$

T_1 and T_2 being the sizes of samples from the two populations and $T = T_1 + T_2$. The success in correctly allocating the data into the two populations A_1 and A_2 must be estimated by the unbiased estimator

$$\tilde{\Delta}_n^2 = \frac{T-n-3}{T-2} \hat{\Delta}_n^2 - \frac{n}{T_1} - \frac{n}{T_2}.$$

Let us now examine four different ways of solving the problem of "optimal" choice of the number of predictors.

1. Traditional approach.

In the Gaussian case the hypothesis

$$H_0: \Delta_{k+l}^2 - \Delta_k^2 = 0$$

of no significant contribution of the last l predictors to the first k predictors can be tested with the statistic

$$F = \frac{\frac{T-k-l-1}{l} \frac{T_1 T_2 [\hat{\Delta}_{k+l}^2 - \hat{\Delta}_k^2]}{T(T-2) + T_1 T_2 \hat{\Delta}_k^2}},$$

which is distributed under H_0 as a Fisher's F with l and $(T-k-l-1)$ degrees of freedom. One may criticize a stopping rule based upon the above testing, because the Mahalanobis distance is not a monotonic function of the number of steps, so that a decision of no significant increase in information may be changed some steps later.

2. Dunn-Varady-Chourigune approach.

Let $\Pi[A_i:A_j]$ be the probability of a wrong decision when the decision rule is given by $\hat{u}(X)$ and $\Pi(e) = \{\Pi[A_1:A_2] + \Pi[A_2:A_1]\} / 2$ (here we put $p_1 = p_2 = 1$).

It is well known that the estimates of $\Pi(e)$:

$$\hat{\Pi}(e) = \Phi(-\hat{\Delta}/2) \quad \text{and} \quad \tilde{\Pi}(e) = \Phi(-\tilde{\Delta}/2), \quad \text{with} \quad \Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^x e^{-x^2/2} dx,$$

are generally much too optimistic and in fact are only valid for the learning sample.

Let us now introduce the variables $D[A_i:A_j]$ and $D(e)$ by means of the relations:

$$\Pi[A_i:A_j] = \Phi(-\frac{1}{2} D[A_i:A_j]) , \quad \Pi(e) = \Phi(-\frac{1}{2} D(e)) .$$

Dunn and Varady (1966) have obtained the quantiles of the probability distribution of $D[A_i:A_j]$ by means of a simulation procedure using the randomization of Δ .

The corresponding confidence intervals can be obtained from a selection algorithm due to Chourigine's formula which gives an very good fitting of those 95% intervals for $n \leq 10$ and $T_1 = T_2 \leq 500$:

$$\begin{aligned} (\frac{1}{2} \underline{D}[A_i:A_j], \frac{1}{2} \overline{D}[A_i:A_j]) &= \frac{\hat{\Delta}_n^2}{2} - \{ \frac{2n}{T} + (\frac{24}{T} + 0.15) \} \{ 1 + 0.0125 \hat{\Delta}_n^2 \} , \\ (\frac{1}{2} \underline{D}(e), \frac{1}{2} \overline{D}(e)) &= \frac{\hat{\Delta}_n^2}{2} - \{ \frac{2n}{T} + (\frac{16}{T} + 0.1) \} \{ 1 + 0.125 \hat{\Delta}_n^2 \} . \end{aligned}$$

For a stopping rule strategy based on the lower limit \underline{D} of the confidence interval, the compromise which define the "optimal" value of n results from the two contradictory tendencies: for a fixed value of T $\hat{\Delta}_n$ increases with n , but the length of the confidence interval increases also, so we observe first an increase in \underline{D} , then a maximum value and finally a decrease.

3. The Okamoto-Deev approach.

Sedransk and Okamoto (1971) and Deev (1970) have studied the asymptotic behavior of the conditional law of the discriminant function $\hat{u}_{A_i}(X)$. In fact, as we have asymptotically

$$\hat{u}_{A_i} \sim N((-1)^i \Delta^2 / 2, \Delta^2) ,$$

Okamoto has obtained the conditional distribution function of the variable $U_i = [\hat{u}_{A_i} - (-1)^i \Delta^2] / \Delta$ as follows:

$$F_{U_i}(x|A_i) = \{ 1 + L(d, \Delta) + Q(d, \Delta) \} \Phi(x) + O(1/N^3) ,$$

where $\Phi(x)$ is the first term of the asymptotic expansion, $L(d, \Delta)$ the term of order $1/N$, $Q(d, \Delta)$ the term of order $1/N^2$, N being defined as one of the numbers T_1 , T_2 and n .

Deev has obtained an even more interesting result, since the first term of his expansion gives an excellent estimate, which is better than Okamoto's first three terms. If the discriminant function $\hat{u} = [X - \hat{u}_{(+)}] \hat{V}^{-1} \hat{u}_{(-)}$ defines a decision value by comparison with the threshold

$$s = \ln\{p_1^c[A_2:A_1]/p_2^c[A_1:A_2]\} ,$$

then the first term of the asymptotic expansion of the probability of a wrong decision is given by

$$\begin{aligned} \Pi[A_1, A_2] &= P[\hat{u} < s | A_2] \\ &= \Phi \left(\frac{2s(f-n-1) - f[\Delta^2 T_1 T_2 + (n-1)(T_2 - T_1)]}{2(f-n+1)DT_1 T_2} \right), \end{aligned}$$

where

$$D^2 = \frac{f^2(f+1)f_2[\Delta^2 + (n-1)f_1]}{(f-n+1)^2(f-n+2)}, \quad f_1 = T/(T_1 T_2), \quad f_2 = (T+1)/T$$

The utility of such results is obvious. It gives an objective estimate of the quality of discrimination (performed with the empirical $\hat{u}(x)$) on the test sample taking into account the unbalanced sizes of each sample ($T_1 \neq T_2$).

4. Minimization of the empirical average risk.

Vapnik and Tchervokenis (1974) have studied the theoretical average risk associated with a forecast $\tilde{y}(t)$ of $y(t)$:

$$J(a) = \int_{-\infty}^{\infty} [y - \tilde{y}(a)]^2 dP(x, y),$$

and its empirical analog :

$$\hat{J}_{T,m}(a) = \frac{1}{T} \sum_{t=1}^T (y_t - \tilde{y}_t)^2,$$

where m is a parameter representing the complexity of the decisional algorithm. The main result of their work says that with a probability greater than $1-\eta$, η being any given positive number, we can have

$$J(a) \leq \hat{J}_{T,m}(a) \Psi \left[\frac{T}{m}, \frac{\ln(1/\eta)}{T} \right],$$

where the function $\Psi \rightarrow 1$ when $T \rightarrow \infty$.

On the other hand, when m increases, the empirical risk \hat{J} decreases, but the function Ψ decreases.

The results of Vapnik and Tchervokenis for example assures with a probability greater than $1-\eta$ that

$$P_{\text{test}}(e) \leq P_{\text{learn}}(e) + 2 \sqrt{\frac{n[\ln(4T/n) + 1] - \ln(5n/\eta)}{T}},$$

which gives us, for instance, the approximate size of the learning sample necessary to obtain the test sample of given quality.

5. REALIZATION OF DISCRIMINANT SCHEMES

In the forecasting procedure the operation which gives us some difficulties is the inversion of predictors' variance-covariance matrix $V_{XX(i)}$ or V_{XX} .

In the case of linear discrimination, if the matrix V_{XX} is ill-conditioned, we can apply the following procedures.

(a) Ridge regression, in which the inverse of V_{XX} is replaced by the inversion of $V_{XX} + qI_n$. (b) One of the pseudo-inverse procedures of a squared matrix, for example the Moore-Penrose pseudo-inverse. (c) Preliminary linear transformation of the predictors x_i by means of principal component $z_i = C_i'X$. We replace the inverse with a diagonalization procedure, and we obtain a simple form for the discriminant function:

$$u(Z) \approx \sum_{i=1}^k m_{z_i(-)} [z_i - m_{z_i(+)}] / \lambda_i,$$

and for the Mahalanobis distance:

$$\Delta_n^2(Z) \approx \sum_{i=1}^k [m_{z_i(-)}]^2 / \lambda_i,$$

where $\lambda_i = \lambda_i(V)$ denotes the i th eigenvalue of the matrix V_{XX} , and the number k is chosen in such a way that we have $\lambda_{k+1} < q$.

In general one tries to avoid the quadratic discrimination and the adjustment of too many parameters on the learning sample, since in this case the quality of discrimination on this sample is quite illusory and soon disappears on the test sample. Anderson and Bahadur (1962) have studied the "best" linear decision rule $a'X$ in the case when $V_{XX(1)} \neq V_{XX(2)}$ and the optimal discriminant function is quadratic in Gaussian distribution case. The conditional probabilities of a wrong decision are then:

$$P[A_1:A_j] = 1 - \Phi(y_j), \text{ with } y_j = (-1)^j [a' \mu_{X(j)} - s] / [a' V_{XX(j)} a]^{1/2}.$$

Minimizing the probability of error is equivalent to maximizing y_i . Then a one-parametric procedure enables one to find a set of the so-called "admissible" points (y_1, y_2) for which there do not exist other "uniformly better" strategies. So for each $t_i \in (0,1)$ and $t_1 + t_2 = 1$ we have:

$$a_{\text{opt}} = [t_1 V_{XX(1)} + t_2 V_{XX(2)}]^{-1} \mu_{(-)},$$

$$s_{\text{opt}} = a'_{\text{opt}} \mu_{X(1)} + t_1 a'_{\text{opt}} V_{XX(1)} a_{\text{opt}} = a'_{\text{opt}} \mu_{X(2)} - t_2 a'_{\text{opt}} V_{XX(2)} a_{\text{opt}}.$$

Note that this procedure may be generalized to a family of multidimensional laws for which constant value surfaces are concentration ellipsoids.

Different procedures have been proposed for the case of discrete variables. In

particular Saporta has proposed an algorithm based on the utilization of the Tschuprow coefficient C_{ij} as a measure of statistical independence, which has the properties of a correlation coefficient. He introduces formally a new concept, the "partial" Tschuprow coefficient by means of a recursive relationship similar to the classical case for the partial correlation coefficient:

$$C_{13.2} = \frac{C_{13} - C_{12} C_{23}}{\sqrt{(1-C_{12}^2)(1-C_{23}^2)}}.$$

Then it is possible to obtain a selection procedure based on those partial coefficients in the same manner as in the regression case for the maximization of the multiple correlation coefficient. Then it is not difficult to obtain the discriminant although some caution must be exercised.

Other results have been obtained for the case of mixed predictors (quantitative, and qualitative, ordinal or nominal variables).

6. ESTIMATION OF THE FORECAST'S QUALITY

Conditional probabilities of a wrong decision $P[A_1:A_j]$, for which the explicit form is particularly simple in the Gaussian case, put into evidence the dependence of the forecast quality on the choice of the threshold value, $s = \ln\{p_1 c[A_2:A_1]/p_2 c[A_1:A_2]\}$, and consequently on the cost matrix $\{c[A_i:A_j]\}$. In particular, for the case of linear discrimination ($V_{XX(1)} = V_{XX(2)}$) we have

$$P[A_1:A_j] = \Phi\left[(-1)^{i+j} \frac{\Delta}{2} + (-1)^{i+1} \frac{s}{\Delta}\right].$$

The forecast quality is described by the two curves $P[A_1:A_2]$ and $P[A_2:A_1]$, and each individual user may use the point corresponding to his own cost matrix.

We can also obtain similar curves corresponding to the relative conditional frequency of a good (or false) forecast $F[A_1:A_j]$, the absolute conditional frequency $N[A_1:A_j]$, the mean cost of a decision rule \bar{c} , and many of the quality indices discussed by Dice, Sokal and Sneath, Kulzinsky, Rogers, Tanimoto, Yule, Jaccard, etc.

These curves may be expressed as functions of the threshold s for a given value of Δ^2 , and each user may choose his own index and then determine the corresponding optimal threshold using the appropriate curve.

The parameters of the discriminant function are obtained on the learning sample, for which we get an estimate $Q_{\text{learn}}(n, T)$ of the forecast quality which overestimates the quality resulting from the selection of n .

On the test sample we get an estimate of the quality $Q_{\text{test}}(n, S)$ which is more realistic, but the size of the sample S is generally much lower than the size T of the learning sample.

Sometimes we don't have enough observations to divide the sample into two parts

(learning and test). In this case we apply a method which allows us to obtain for the whole sample an estimate $Q_{S.R.}(n, T)$ of the forecast quality comparable to $Q_{test}(n, T-2\tau-1)$. We proceed as follows: Compute the parameters of a sample in which we have eliminated the $2\tau + 1$ observations $X(\theta)$, $\theta \in (t-\tau, t+\tau)$. Then we compute the value of the discriminant function only for the single value $X(t)$ of the predictors. Note that the coefficients of the discriminant function are not re-computed at each time, but may be obtained by means of successive corrections using the matrix formula:

$$(A - XX')^{-1} = A^{-1} + \frac{A^{-1}XX'A^{-1}}{1 - X'A^{-1}X}.$$

7. NON-PARAMETRIC DISCRIMINANT ANALYSIS

In this case we don't consider the random nature of the predictor vector, but use only geometric considerations in the space R^n of observational data. For this purpose a metric (distance) is defined in this space, e.g., Euclidean, Minkowski, Hamming, etc. Then various algorithms may be introduced to obtain discrimination procedures, many of them heuristically.

1. Nearest neighbor method.

If $X(\text{near})$ is the nearest point in R^n to $X(t)$, then the decision is that $X(t)$ belongs to the same class as $X(\text{near})$.

2. Fix-Hodges method (k-nearest neighbor).

We apply the majority rule among the nearest k points to the point $X(t)$. It is a natural extension of the Nearest neighbor method and gives better results in those parts of R^n where points of both classes are available.

The crucial factor here is the choice of the "optimal" value of k , which corresponds to the definition of the "optimal" diameter in the so-called "Ball Regression". For the order of this optimal value Mechalkine (1969) obtained the formula:

$$k_{\text{opt}} \sim [T_1 \cdot T_2]^{2/(4+n)},$$

for slightly different values of T_1 .

3. Average distance method.

Each observation $X(t)$ is allocated to the class A_i if the average distance

$$\overline{d(X(t), A_i)} = \frac{1}{T_i} \sum_{X(\tau) \in A_i} d(X(t), X(\tau))$$

is less than the average distance $\overline{d(X(t), A_j)}$.

4. The kernel method.

Here each class A_i is represented by some point $a_{(i)} \in R^n$, which is called the kernel of the class A_i . It may be the vector of mean values, of modes, or of medians. The class allocation is obtained by using proximity relations between the observation $X(t)$ and the kernel $a_{(i)}$. In this case the discrimination function is

$$g(X) = d(X(t), a_{(2)}) - d(X(t), a_{(1)}) .$$

The advantage of this method is that it could be used for very large samples, since we need to save in central memory only the coordinates of the kernel $a_{(i)}$, and then we need only the value of each $X(t)$ for the class allocation.

Furthermore, the kernel method is directly applicable to an arbitrary number of classes.

Geometric considerations give us many other variations of the heuristic algorithms, which, although non-optimal in the parametric sense, often have the very useful property of robustness, when we pass from the learning sample to the test sample. Non-parametric methods have an advantage over parametric methods, which although "most powerful", "overfit" the learning samples and lose ability to forecast on the test samples.

Of course, the assumptions necessary for the applicability of parametric methods are very often not easy to verify. Current practice suggests that a good compromise is to obtain simultaneously both parametric and non-parametric solutions thus exhibiting several possible discrimination decisions.

8. 'METEOROLOGICAL EXAMPLES OF THE APPLICATION OF DISCRIMINANT ANALYSIS

The application of a statistical forecasting (or decision) model is based on the relationships synchronous or asynchronous between predictors and predictands.

The synchronous connections are of course stronger than the asynchronous, and may be applied by using the outputs of hydrodynamical models (the deterministic forecasts) as predictors. Two variations are then possible: the first is the so-called "Perfect- Prog" method for which the correlations are computed on the observational sample between the true values of $X(t)$ and the values of $y(t)$. The second is the so-called "MOS" (Model Output Statistics) method for which the correlations are computed on the sample of hydrodynamical model's output between the deterministic forecasts $\tilde{X}(t)$ and the predictand $y(t)$. Each method has its own advantages and defects.

We applied the two methods for the forecasting of several meteorological phenomena: precipitation occurrence for 7 stations in France, avalanche's occurrence for an hundred stations in Savoie, SO_2 pollution at Rouen, frost' occurrence etc.

For the avalanche's problem we have elaborated a model called "Edelweiss" for the statistical forecasting of the avalanche's risk probability for 120 stations

in the Savoie region of France. We have completed the learning and test samples, and we are in the process of implementating the model.

For the precipitation forecasting we use the "Amethyste" hydrodynamical model described by the French Meteorological Office by Rousseau and we used the statistical adaptation of the model's outputs for 7 French stations in both his "Perfect Prog" and "MOS" variants. Shortrange forecasts are given in operational form systematically for 24, 48, 72 and 36 hours and we have statistics on some aspects of the model's performance for the first 6 months.

For the SO₂ pollution at Rouen we elaborate a very short-range forecasting model (3-hours) based on local and synoptical predictors.

For the frost occurrence forecast we applied a lot of various discriminant algorithms in a small sample of 225 observations which shows an excellent separation of the two populations for a set of 6 predictors.

9. FURTHER DEVELOPMENTS AND CONCLUSIONS

Discriminant analysis seems to be a fastgrowing branch of statistical science. Many studies have been made in the search for better algorithms: most powerful, most robust, with less restrictive constraints, applicable to non-normal case, with predictors of various nature (qualitative and quantitative). New methods of selection of useful predictors have been studied, so that it is possible to eliminate the spurious results, i.e., the bad predictors. It is in fact a crucial problem, because we need objective methods to determine what kind of information we need to use for the forecasts.

Many interesting studies have been made in the field of quality estimation, that is the extrapolation on the test sample of the various estimation of forecast accuracy obtained on the learning sample.

In Meteorology and Climatology the applications of discriminant analysis are numerous and varied: for the forecasting of occurrence of atmospheric phenomena, for different diagnostic process, for qualitative forecasting of continuous variables, for objective measuring the accuracy of deterministic forecasts, etc.

It is very important to emphasize the multidimensional aspects of the discriminant analysis, because only the collective properties of a set of meteorological variables determine the meteorological situations determining the occurrence or the non-occurrence of meteorological phenomena.

I believe that this statistical method will be soon a current tool in the hands of the meteorologists.

REFERENCES

- Anderson, T.W., 1957. An Introduction to Multivariate Statistical Analysis. John Wiley, New York.

- Anderson, T.A. and Bahadur, R.R., 1962. Classification into two multivariate normal distributions with different covariance matrices. *Ann. Math. Statist.* 33.
- Bois, P. and Ohled, C. Analyse des données nivoclimatiques en vue de la prévision des avalanches.
- Chouriquine, 1969. Le choix des paramètres pour la classification de deux populations normales. *Méthodes statistiques de classification* N1, 6. Moscou.
- Deev, A.D., 1970. Representation of statistics of discriminant analysis, and asymptotic expansion when space dimensions are comparable with sample size. *Dokl. Akad. Nauk, SSSR*, 195:759-762 (in Russian).
- Der Megreditchian, G., 1969. Un nouveau procédé de réalisation des schémas de discrimination et de régression (en russe.). *Météor. et hydro.* 7.
- Der Megreditchian, G. et Lukijanov, L., 1969. Quelques particularités de l'application de l'analyse discriminante linéaire à la prévision (en russe.). *Ann. du Centre Hydrometeor. de l'URSS*, 44.
- Der Megreditchian, G., 1972. Une méthodologie décisionnelle statistique pour la prévision des phénomènes atmosphériques dangereux. Document interne EERM.
- Der Megreditchian, G., 1973. Méthodes statistiques de prévision par classes en Météorologie. *La Météorologie* V-26.
- Der Megreditchian, G., 1975. Approche statistique du problème d'évaluation des risques d'avalanche. *La Météorologie* V-3.
- Dunn, O.J. and Varady, P.V., 1966. Probabilities of correct classification in discriminant analysis. *Biometrics* 22.
- Facy, L. and Der Megreditchian, G., 1973. La pollution atmosphérique. Organisation mondiale de la santé PNUD/ROM/71/512.
- Fisher, R.A., 1938. The statistical utilization of multiple measurements. *Ann. Eug.* 4.
- Fix, E. and Hodges J.L., 1952. Discriminatory analysis: Non-parametric discrimination. USAF Sch. of Avia. Medecine, Randolph Field, Texas. Rep. 4 & 11.
- Foley, 1971. Considerations of sample and feature size. *IEEE Trans. on Comp.* C 20-12.
- Fukanaga, Kessel. 1971. Estimation of classification error. *IEEE Trans. on Comp.* C 20 - 12/
- Gnedenko, B.V., 1970. Cours de la théorie des probabilités. Moscou.
- Hills, M., 1966. Allocation rules and their error rates. *J. Roy. Stat. Soc.*, B, 28.
- Kullback, S., 1958. *Information Theory and Statistics*. John Wiley, New York.
- Lachenbruch, P.A. and Mickey, M.R., 1968. Estimation of error rates in discriminant analysis. *Technometrics* 10:1-11.
- Méchaline, L.D., 1969. Méthodes locales de classification. Recueil "Méthodes statistiques de classification" (en russe.) Editions M.O.U., Moscou.
- Nagy, G., 1968. State of the art in pattern recognition. *Proc. of IEE*, 56.
- Nilson, N.J., 1965. *Learning Machines*. McGraw Hill, New York.
- Romed, 1973. Méthodes et programmes d'analyse discriminante. Dunod.
- Sebestyan, G.S., 1962. *Decision Making Process in Pattern Recognition*. New York.
- Sedransk, N. and Okamoto, M., 1971. Estimation of the probabilities of misclassification for a linear discriminant function. *Ann. Inst. Stat. Math.*, Tokyo, 23:419-435.
- Sonetchkin, D., 1972. Déchiffrement météorologique des images satellites (en russe.) *Ann. du Centre Hydrometeor. de l'URSS*, Moscou.
- Sorum, M.J., 1971. Estimating the conditional probabilities of misclassification. *Technometrics* 13.
- Vapnik, Tchervokénis, 1974. *Théorie de Reconnaissance des formes*. Moscou.

REGIONAL CLASSIFICATION OF EAST AFRICAN RAINFALL STATIONS INTO HOMOGENEOUS GROUPS
USING THE METHOD OF PRINCIPAL COMPONENT ANALYSIS

L. OGALLO

Dept. of Meteorology, Univ. of Nairobi, Nairobi (Kenya)

ABSTRACT

Ogallo, L., Regional classification of East African rainfall stations into homogeneous groups using the method of principal component analysis. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

An attempt has been made in the present study to use the empirical orthogonal functional method of principal component analysis to classify East African annual rainfall stations into homogeneous regional groups. The annual rainfall records used were from 86 stations distributed all over East Africa during the common period 1931-75.

The results from the analysis indicated that communality was greater than 70% at all stations except four. These four stations were Kigoma, Kasulu, Kagondo and Kilindoni. The lowest communality of 57% was observed at Kigoma. The Kaiser's criterion indicated that the cut-off value for the eigenvalues was at the eigenvalue number 16. The sixteen eigenvectors accounted for 80.6% of the total rainfall variance, of which 50.3% were explained by the first three orthogonal vectors.

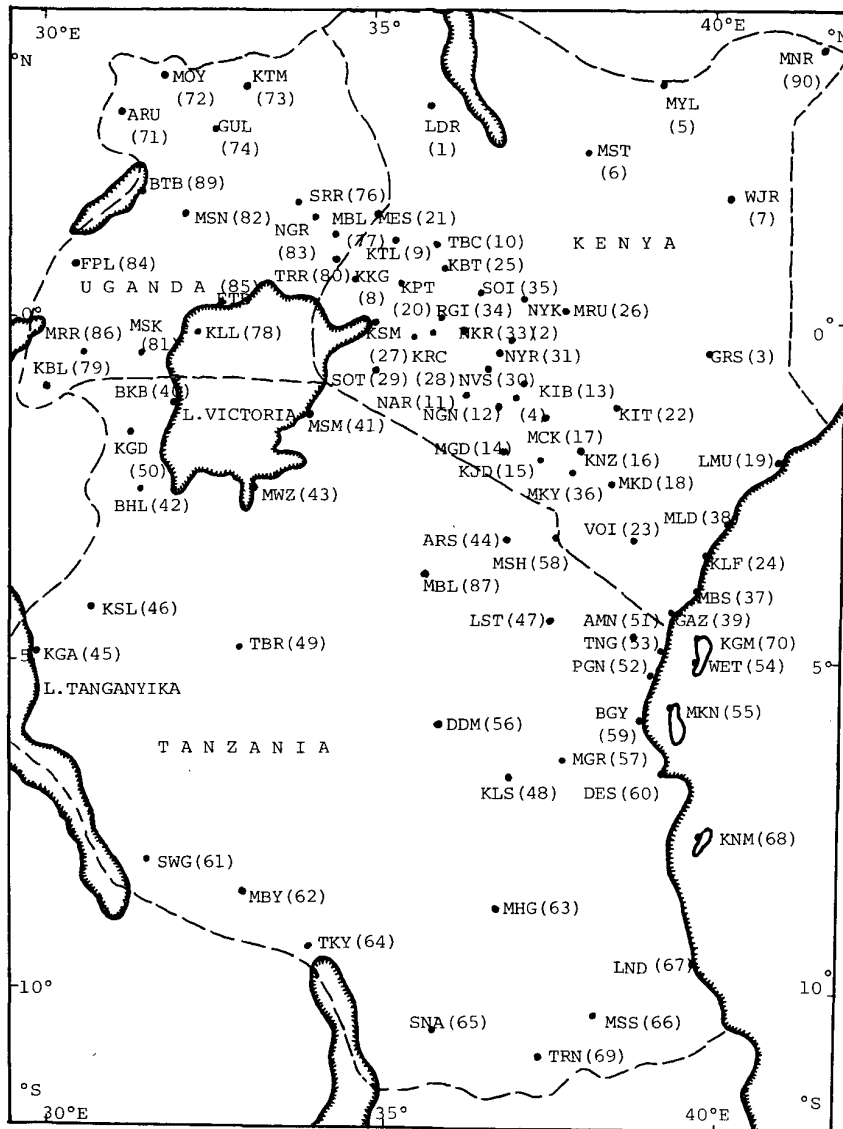
Fourteen regional patterns were discernible from the spatial patterns formed by the major eigenvectors.

1. INTRODUCTION

Empirical orthogonal functional methods have been applied to climatological records by many authors in attempts to reduce the dimensionality of the basic data being processed, and to describe some climatological patterns. Such work include those by Veitch (1965), Craddock (1965), Craddock et al. (1969,1970), Gregory (1975), and Dyer (1977). The theories of the empirical orthogonal functions have been discussed by Burt (1952), WMO (1966), Harman (1967), Rummel (1970), Kim et al. (1970), Craddock (1973), Preisendorfer and Barnett (1977), Child (1978), and many others.

In the present study the method of principal component analysis has been used in the attempt to group the East African rainfall stations into homogeneous regional groups.

The term East Africa refers to three countries namely Kenya, Uganda and Tanzania. The region is enclosed by latitudes 4°N - 11°S and longitudes 30°E - 42°E. The spatial pattern of the stations used is presented in Figure 1, while Table 1 gives their locations. The annual rainfall records used were for the common period 1931-75. The rainfall records were obtained from Kenya Meteorological Department.



Country & Stations	Station Code Name	Station Code Number	Latitude O 'E	Longitude O 'E	Elevation (Meters)
KENYA					
Garissa	GRS	3	0 26S	39 38	128
Gazi	GAZ	39	4 25S	39 31	46
Kabarnet	KBT	25	0 30N	35 45	2043
Kabete	NRB	4	1 15S	36 44	1891
Kajiado	KJD	14	1 52S	36 48	1738
Kakamege	KKG	8	0 17N	34 45	1555
Kapsabet	KPT	20	0 12N	35 07	1999
Kericho	KRC	28	0 23S	35 17	1982
Kiambu	KIB	13	1 11S	36 50	1767
Kilifi	KLF	24	3 40S	39 51	3
Kisumu	KSM	27	0 06S	34 45	1146
Kitale	CTL	9	0 54N	34 55	1829
Kitui	KTI	22	1 22S	38 01	1177
Konza	KNZ	16	1 44S	37 08	1655
Lamu	LMU	19	2 16S	40 54	9
Lodwar	LDR	1	3 07N	35 37	566
Londiani	LNI	75	0 10S	35 35	2317
Machakos	MCK	17	1 31S	37 16	1646
Magadi	MGD	15	1 53S	36 17	613
Makindu	MKD	18	2 17S	35 50	1000
Makuyu	MKY	36	0 55S	37 10	1540
Malindi	MLD	38	3 13S	40 07	3
Marsabit	MST	6	2 19N	37 59	1345
Meru	MRU	26	0 03N	37 39	1570
Mombasa	MBS	37	4 04S	39 42	16
Mt.Elgon	MES	21	1 08N	34 45	2226
Moyale	MYL	5	3 32N	39 03	1113
Naivasha	NVS	30	0 43S	36 26	1901
Nakuru	NKR	33	0 17S	36 04	1851
Nanyuki	NYK	2	0 05N	37 10	2104
Narok	NAR	11	1 08S	35 50	1890
Ngong	NGN	12	1 20S	36 40	2043
N.Kinangop	NKP	32	0 34S	36 38	2631
Nyeri	NYR	31	0 26S	36 57	1829
Rongai	RGI	34	0 11S	35 51	1890
Solai	SOI	35	0 07S	36 06	1829
Sotik	SOT	29	0 40S	35 05	1824
Tambach	TBC	10	0 36N	35 32	1829
Voi	VOI	23	3 24S	38 34	560
Wajir	WJR	7	1 45N	40 04	244
TANZANIA					
Amani	AMN	51	5 06S	38 38	911
Arusha	ARS	44	3 23S	36 41	1372
Bagamoyo	BGY	59	6 25S	38 55	9
Biharamulo	BHL	42	2 38S	31 19	1479
Bukoba	BKB	40	1 20S	31 49	1144
Dar es Salaam	DES	60	6 49S	39 18	9
Dodoma	DDM	56	6 10S	35 46	1120
Kigoma	KGA	45	4 52S	29 38	777
Kilosa	KLS	48	6 50S	37 00	491
Kagondo	KGD	50			1372
Kilindoni	KNM	68	10 00	39 43	9
Kigomasha	KGM	70	4 52	39 41	15
Kasulu	KSL	46	4 34	30 06	1320
Lindi	LND	67	10 00	39 43	9

(TABLE 1 continued)

Lushoto	LST	47	4 47S	38 17	1396
Mahenge	MHG	63	8 41S	36 43	1107
Masasi	MSS	66	10 42S	38 49	457
Mbeya	MBY	62	8 56S	33 28	1759
Morogoro	MGR	57	6 51S	37 40	579
Moshi	MSH	58	3 21S	37 20	813
Musoma	MSM	41	1 30S	33 48	1148
Mwanza	MWZ	43	2 31S	32 54	1131
Mkokotoni	MKN	55	5 52S	39 15	9
Pangani	PGN	52	5 26S	38 59	9
Songea	SNA	65	10 42S	35 40	1166
Sumbawanga	SWG	61	7 57S	31 36	19
Tabora	TBR	49	5 02S	32 49	1266
Tanga	TNG	53	5 04S	39 06	9
Tukuyu	TKY	64	9 15S	33 38	1616
Tundura	TRN	69	11 06S	37 22	701
Wete	WET	54	5 04S	39 43	18
UGANDA					
Arua	ARU	71	3 03N	30 35	1280
Entebbe	EBT	85	0 03N	32 27	1146
Fort Portal	FPL	84	0 40N	30 17	1539
Gulu	GUL	74	2 45N	32 20	1106
Kabale	KBL	79	1 15S	29 59	1871
Ka;angala	KLL	78	0 20S	32 19	1158
Kitgum	KTM	73	3 17N	32 53	937
Masaka	MSK	81	0 20S	31 44	1313
Masindi	MSN	82	1 41N	31 43	1146
Mbale	MBL	77	1 06N	34 11	1220
Mbarara	MRR	86	0 37S	30 39	1443
Moyo	MOY	72	3 41N	31 44	1036
Ngora	NGR	83	1 27N	33 46	1128
Serere	SRR	76	1 31N	33 27	1139
Tororo	TRR	80	0 43N	34 10	1226

2. METHODS OF ANALYSIS

The observed values of the annual rainfall at several locations (stations) for the set of years were subjected to principal component analysis by generating a correlation data matrix between the set of the stations for the set of period (S-mode). This method has been applied by Gregory (1975) and Ryer (1977) to delimitate the regional patterns of the annual rainfall over United Kingdom and South Africa, respectively.

In the attempts to delineate the homogeneous regional patterns, the major eigenvectors which were significantly correlated with each station were noted, and the regional classification was finally based on the spatial patterns formed by principal components (eigenvectors). The Kaiser's criterion (Kaiser (1959)) was used to determine the cut-off value for the eigenvalues. These eigenvectors were further subjected to the orthogonal varimax and oblique rotations. Rotation of these

hypothetical vectors often remove certain ambiguities that are sometimes evident in the direct solutions. The orthogonally rotated components were further presented graphically in a two dimensional space of the reference axes to display visually more informations about the variables.

3. RESULTS AND DISCUSSIONS

The results of the analysis indicated that the final communality was as high as 95% at a number of stations. It was greater than 70% at all stations except four. These four stations were Kigoma, Kasulu, and Kagondo in north-western Tanzania, and Kilindoni in Mafia Island. The lowest communality of 57% was observed at Kigoma. Kigoma is situated in the eastern shore of Lake Tanganyika and is blocked to the east by high ground. The characteristics of rainfall at Kigoma has been noted to have no similarity to those of most situations with the same type of annual rainfall regime (Tomstt (1975)). Communality indicates the proportion of the total variance of the annual rainfall at each station that is explained by the common empirical orthogonal vectors. The results indicate that at least 70% of the total variance of the annual rainfall at each station is accounted for by the common factors which apply over the rest of the stations. This may be an indication of the influence of some common rain generating functions, and it seems to suggest that the influence of the unique properties of the individual stations were generally of little significance.

TABLE 2.

Results of the principal component analysis. (Only the first 20 components are presented.)

Eigenvalue Number	Eigenvalue Before Rotation	% of Total Variance Extracted	Cumulative Variance %
1	29.8	33.6	33.6
2	9.6	11.1	44.7
3	4.8	5.6	50.3
4	3.3	3.9	54.2
5	3.2	3.7	57.9
6	2.9	3.3	61.2
7	2.6	3.1	64.3
8	2.4	2.8	67.1
9	2.1	2.4	69.5
10	1.8	2.1	71.6
11	1.5	1.8	73.4
12	1.4	1.6	75.0
13	1.3	1.5	76.5
14	1.3	1.5	78.0
15	1.2	1.4	79.4
16	1.1	1.2	80.6
17	0.9	1.1	81.8
18	0.9	1.1	82.9
19	0.8	0.9	83.7
20	0.7	0.8	84.5

From Table 2 , the Kaiser's criterion indicates that the cut-off value for the eigenvalues is at the eigenvalue number 16. These sixteen eigenvectors account for 80.6% of the total variance of the rainfall of which only 33.6% was extracted by the first eigenvector. The first three eigenvectors together accounted for 50.3% of the total variance, indicating that no few common factors could be found that can account for most of the variance of the annual rainfall in East Africa.

Craddock (1965) noted that the first eigenvector only could extract as high as 92.28% of the total variance of the monthly temperatures over Central England, while Veitch (1965) observed that 75% of the total variance of the Australian pressure field could be accounted for by the first three principal components. Craddock and Flood (1969) found that these three first components could extract 65% of the total variance of the 500 mb geopotential fields over the Northern Hemisphere. On subjecting annual rainfall to empirical orthogonal analysis, Gregory (1975) observed that the first three components could account for 68% of the total variance of the annual rainfall over United Kingdom. Dyer (1976) found that only 47.23% of the total variance of the South African annual rainfall could be extracted by the first three eigenvectors with the first vector accounting for 27.95% .

In order to classify the rainfall stations into some homogeneous regional groups, the eigenvectors which were significantly correlated with each station were noted. The regional grouping was finally based on the spatial patterns formed by three major eigenvectors. The fourteen homogeneous groups which were discernible from the orthogonal varimax solutions are presented in Figure 2, while the major principal components for the various regions are given in Table 3. Closely identical regional patterns were also observed from the results of the direct solutions, oblique rotations, and the graphical solutions. No cluster analysis was performed but the high degree of association between the stations in the various groups were confirmed by scanning the correlation matrix.

The results of the analysis indicated that central highlands of Kenya and most of Uganda were highly correlated with the first eigenvector, while coastal regions especially northern Tanzania were significantly correlated with the second component. The fifth component was prominent in the southern highlands of Tanzania. In other regions more than one components were prominent. No distinct regional patterns could be delineated around Lake Victoria due to the limited data available around the lake, but the characteristics of the lake stations (Region N) were noted to tend to those of the stations in the neighbouring regions.

Figure 3 gives the relief map of East Africa. Most of Uganda rise to over 1000 m. In the west of Uganda are the highlands which lead to Ruwenzori mountains, and to the eastern border with Kenya lie Mt. Elgon. To the south are the highlands which are continuous from north western Tanzania. Apart from north eastern region, most of Uganda receive substantial amount of rainfall throughout the year. A trough of

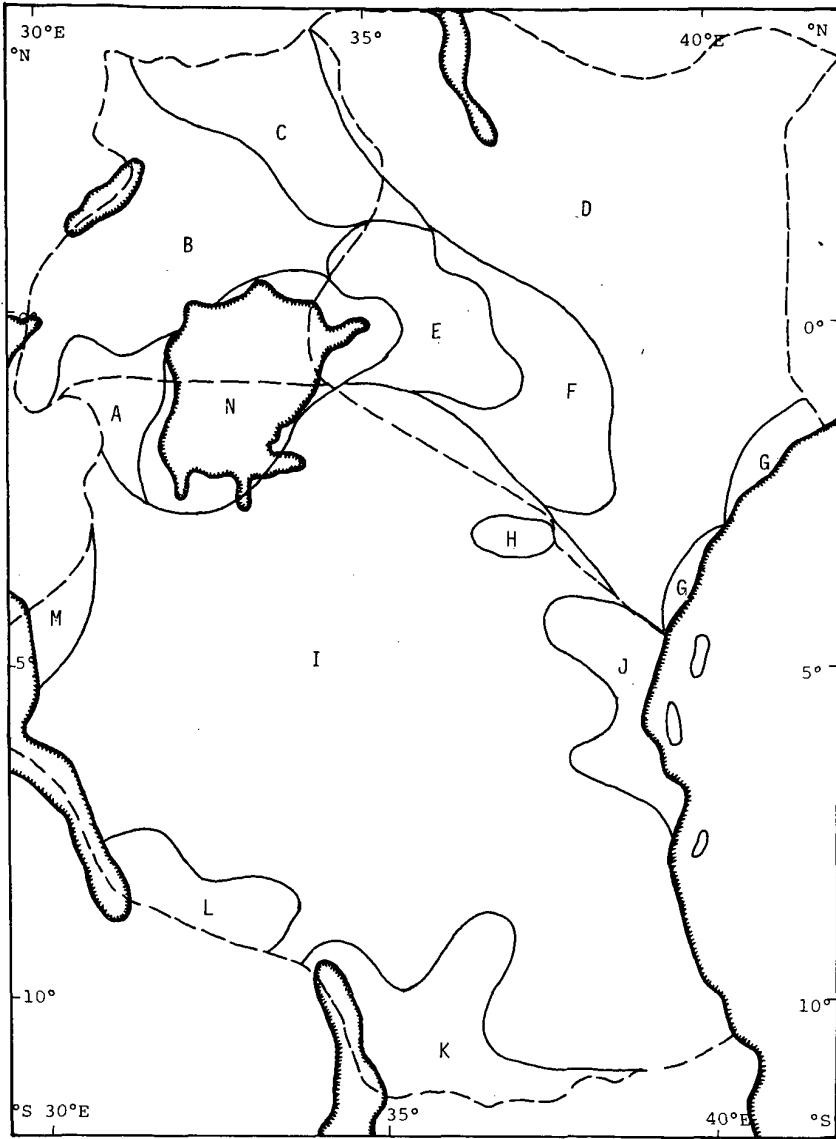


Fig. 2. Regional classification from the orthogonal varimax rotation.

low pressure extends over Uganda and Lake Victoria throughout the year. The trough has been observed to move towards the centre of the lake at night giving an inflow into the regions of the trough (Asnani (1979)).

Figure 4 gives the spatial distribution of the mean annual rainfall in East Africa.

TABLE 3.

The major eigenvectors observed in the various regional stations.

Region	The Major Eigenvevtors	Some General Characteristics of the Eigenvectors
A	Ix & I	
B	I	Loading on Component I > 0.7
C	I & II	
D	IV	Component IV is not dominant but is the only major component.
E	I	Component II is positive giving the only difference to Region B.
F	I	Component I is not dominant, but is the only major component. The loading on Component I lies between 0.5 & 0.7.
G	I & II	
H	I	Component I is not dominant but is the only major component.
I	II & III	
J	II	Loading on Component II > 0.7.
K	V	Component V is not dominant, but is the only major component.
L	III	Component III is not dominant, but is the only major component.
M	III & V	
N	Lake Victoria region. No distinct regional patterns could be delineated but the characteristics of the lake stations were close to those of the bordering regions.	

Only three regions of Kenya have rainfall over 1000 m. These are the central highlands, the western parts of Kenya, and a narrow coastal strip. These select the influence of topography, moisture sources from Lake Victoria and the moist Congo airmass, and the effect of Indian Ocean. Due to the variations of latitude, exposure, and geographical positions, there are also variations in rainfall in the central highlands. The central highlands is divided into two by the rift valley. The western highlands especially their western slopes receive amount of rainfall which appear to increase with highest, notably in the central latitudes (Brown et al. (1973)). The stations in the rift valley are generally drier than in the eastern highlands. In the eastern ranges the rainfall regimes become more seasonal.

Most of Tanzania has one dry and one wet season. The double highland formations observed in the central highlands of Kenya is discernible in the southern border of Tanzania. In the eastern border with Kenya are the highlands which lead to Mt.

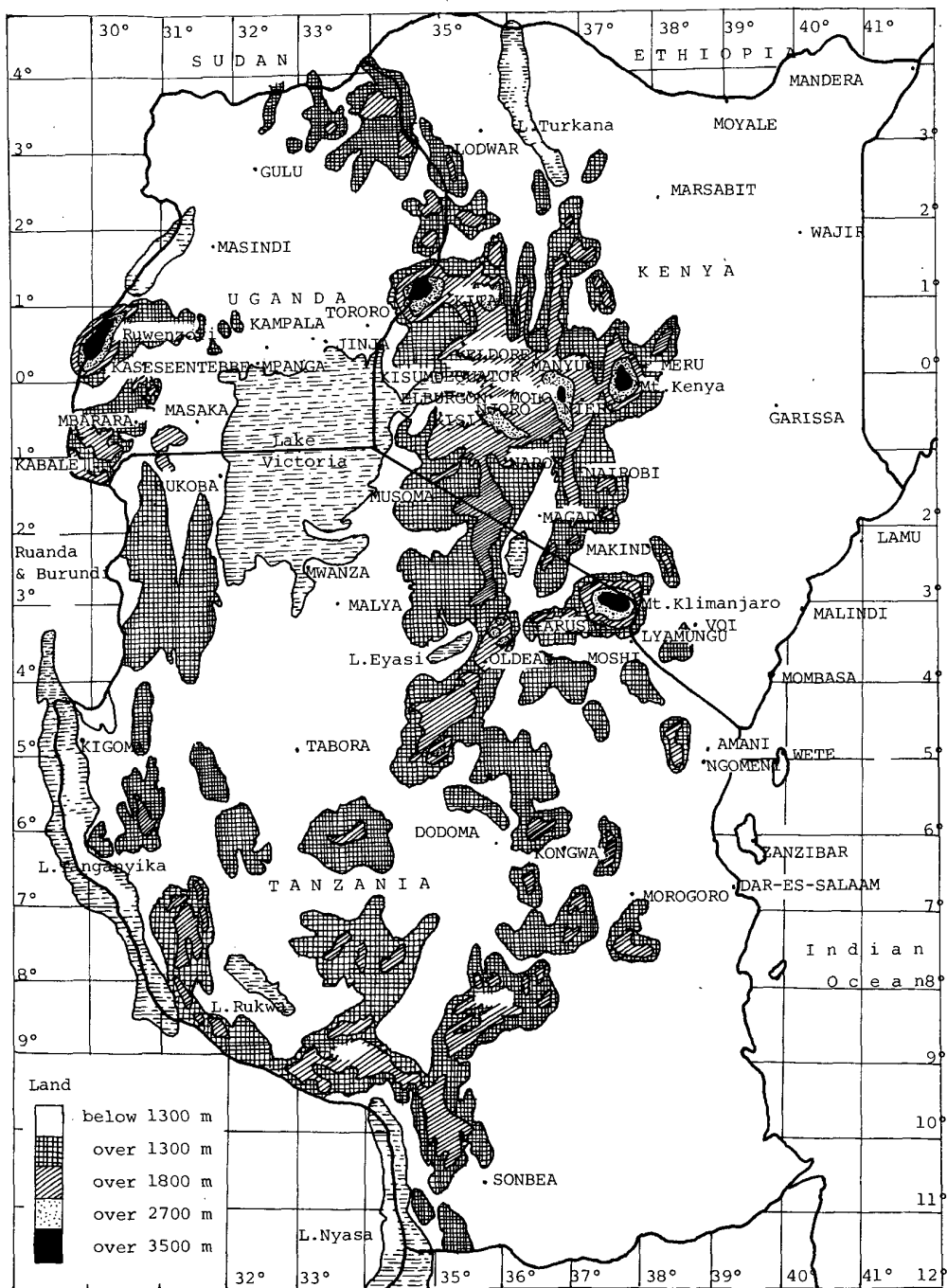


Fig. 3. Relief map of East Africa

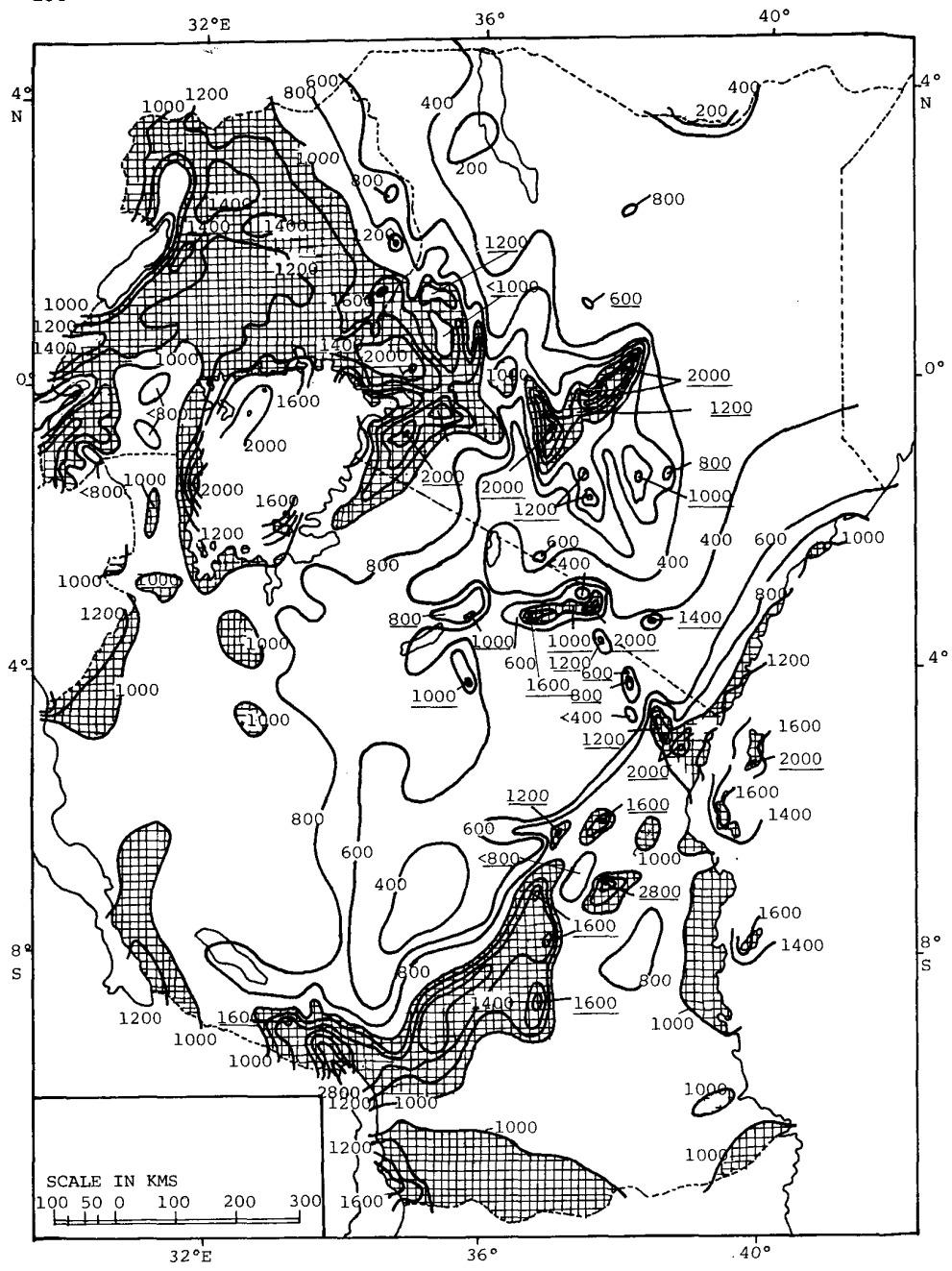


Fig. 4. Mean annual rainfall

Kilimanjaro. South east of these lie patches of high ground which include the Lushoto Mountains. To the west of Lake Victoria are the highlands which extend to southern Uganda. The seasonal distribution of rainfall is generally bimodal in the highlands with the exception of the southern highlands of Tanzania.

The synoptic flow over East Africa is mainly easterly throughout the year. Additional northerly current is observed during the north hemispheric winter, while the influence of the southerly component is experienced during the south hemispheric winter.

The annual rainfall of East Africa exhibit strong seasonality, but the regional grouping of East Africa based on seasonal distributions has been observed to be complicated (Potts (1971), Griffiths(1972), Brown et al.(1973)). The seasonal variations of rainfall depend largely on the seasonal migrations of the ITCZ which are related to the sun's movements. The ITCZ is very diffuse in East Africa due to the diversity of topography. Topography and exposure largely control the amount of rainfall.

4. CONCLUSIONS

The results of the analysis indicate that the method of principal component analysis was capable of grouping the annual rainfall of East Africa into some recognizable regional patterns. Fourteen regional patterns were discernible from the spatial patterns formed by the major orthogonal vectors.

ACKNOWLEDGEMENTS

The author indebted to Professor S. Gregory (Sheffield Univ., U.K.) for giving valuable suggestions and for making the facilities available for the work. The author is also very grateful to the Inter-University Council for providing the financial support. To Professor G.C. Asnani and Dr. P.M.R. Kiangi for their useful discussions. The manuscript was typed by Maria Nyawade while the diagrams were drafted by Mr. J. Munyi to whom the author finally extend his sincere gratitude.

REFERENCES

- Asnani, G.C. and J.H. Kinuthia, 1962. Diurnal variation of precipitation in East Africa. E.A. Met. Dept., Memo 8 : 58pp.
- Brown, L.H. and Coheme, J., 1973. A study of the agroclimatology of the highlands of Eastern Africa. WMO Tech. Note 125 : 197pp.
- Burt, C., 1952. Tests of significance in Factor Analysis. Br. J. Psych. 5:109-113.
- Child, D., 1978. Essentials of Factor Analysis. Holt, Rinehart & Winston : 107pp.
- Craddock, J.M., 1965. A meteorological application of principal component analysis. Statist. 15 : 143-156.
- Craddock, J.M., 1973. Problems and prospects for eigenvector analysis in meteorology. Statist. 22 : 133-145.
- Craddock, J.M. and Flood, C.R., 1969. Eigenvectors for representing the 500 mb. geopotential surface over the Northern Hemisphere. Quart. J. R. Met. Soc. 95:576-593.

- Craddock, J.M. and Flintoff, S., 1970. Eigenvector representation of Northern Hemispheric fields. *Quart. J. R. Met. Soc.* 96:124-129.
- Dyer, T.G.J., 1977. The assignment of rainfall stations into homogeneous groups, an application of principal component analysis. *Quart. J. R. Met. Soc.* 103:1005-1013.
- EAMD, 1962. Climatic seasons of East Africa. *E.A. Met. Dept.* 8:1-4.
- Gregory, S., 1975. On the delimitation of regional patterns of recent climatic fluctuations. *Weather* 30:276-287.
- Griffiths, J.F., 1972. *Climates of Africa*. World Survey of Climatology 10, Elsevier, 604 pp.
- Harman, H.H., 1967. *Modern Factor Analysis*. Chicago Univ.P., 469 pp.
- Kaiser, H.F., 1959. Computer program for varimax rotation in factor analysis. *Psyc. Meas.* 19:413-420.
- Kim, J. and Nie, N.H., 1970. *Factor Analysis*. SPSS, McGraw Hill: 208-246.
- Potts, A., 1971. Application of harmonic analysis to the study of East African rainfall data. *J. Trop. Geo.* 34:31-42.
- Preisnerdorfer, R.W. and Barnett, T.P., 1977. Significance test for empirical orthogonal functions. In: *Proc. 5th Conf. on Prob. and Stat. on Phys. Sc. , Amer. Met. Soc.* : 169-194.
- Rummel, R.J., 1970. *Applied Factor Analysis*. Northwestern Univ.P..
- Veitch, L.G., 1969. The description of the Australian pressure fields by principal components. *Quart. J. R. Met. Soc.* 91:184-195.
- WMO, 1966. *Climatic Change*. World Met. Org. Tech. Note 79: 79pp.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

ON A MATHEMATICAL MODEL OF CARBON DIOXIDE CONCENTRATIONS IN THE MID TROPOSPHERE

J. GOULD, F.A. AHRENS and C.S. HONG

Dept. Math., Claremont Grad. Sch., Claremont, California

ABSTRACT

Gould, J., Ahrens, F.A. and Hong C.S., On a mathematical model of carbon dioxide concentrations in the mid troposphere. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1., 1979

A model describing the spatiotemporal behavior of carbon dioxide, $C(\theta, \psi, r, t)$, in the mid troposphere is proposed. The analytic model considers the diffusive and advective transport of CO_2 due to the stationary and nonstationary sources and sinks (anthropogenic, land biotic, and oceanic) at the surface of the earth. For the axially symmetric case mathematical methods to solve for $C(\theta, \psi, r, t)$ based on empirical evidence are given.

1. INTRODUCTION

Carbon dioxide is one of the tracer gases used in the study of global atmospheric mixing. The CO_2 concentration is observed to contain systematic variations due to atmospheric transport and to sources and sinks which exist mainly at the surface of the earth. Herein we combine some classical mathematical techniques with some empirical evidence in proposing a model of CO_2 concentration in a region of the troposphere. In proposing the model we intend to bring together some of the body of results and, since the model embodies some simplifications, to contribute to the direction of future research.

The underlying hypothesis is that the observed spatiotemporal variations of CO_2 concentration in the mid troposphere can be explained by a model of atmospheric transport from sources and sinks at the surface level via advection and turbulent diffusion. The model enables a decomposition into components. We shall address a mathematical basis for these components and solution methods. The advantages of this mathematical modeling approach lie in the natural decomposition, the numerical simplicity which may provide physical insight without a massive computer effort, and the provision to examine the effects of potential changes in the future source/sink behavior.

2. THE DIFFUSION MODEL

Let $C(\theta, \psi, r, t)$ denote the concentration of CO_2 in units g-cm^{-3} at time t and position (θ, ψ, r) , where θ is longitude, ψ is colatitude, and r is distance from the center of the earth. The concentration must, at each instant, satisfy a material balance taken over any volume element. Assuming that the principal axes of the turbulent diffusivity tensor K coincide with the axes of the spherical coordinate system, atmospheric incompressibility, and that molecular diffusion is negligible when compared to turbulent diffusion, then C must satisfy

$$\begin{aligned} \frac{\partial C}{\partial t} + u_r \frac{\partial C}{\partial r} + \frac{u_\psi}{r} \frac{\partial C}{\partial \psi} + \frac{u_\theta}{r \sin \psi} \frac{\partial C}{\partial \theta} &= \frac{1}{r^2} \frac{\partial}{\partial r} [K_r r^2 \left(\frac{\partial C}{\partial r} + \frac{\gamma C}{r^2} \right)] \\ &+ \frac{1}{r^2 \sin \psi} \frac{\partial}{\partial \psi} [K_\psi \sin \psi \frac{\partial C}{\partial \psi}] + \frac{1}{r^2 \sin^2 \psi} \frac{\partial}{\partial \theta} [K_\theta \frac{\partial C}{\partial \theta}] + S, \end{aligned} \quad (1)$$

where (u_r, u_ψ, u_θ) represent advection and S is the rate of generation of CO_2 through a unit area. The term $\gamma C/r^2$ incorporates a virtual pressure gradient due to gravity; $\gamma = 7.74 \times 10^{11} \text{ cm}$. In the course of reducing (1), assumptions will be made based on empirical evidence, intuitive appeal, and/or mathematical expedience.

The major sources and sinks of CO_2 are the oceans, the land biota, and anthropogenic oxidation; atmospheric CO_2 is relatively inert. Restricting our attention to the region above $2 \times 10^5 \text{ cm}$, we neglect terrain variations and consider the sources and sinks to be exclusively on the surface of a spherical earth with radius $a = 6.366 \times 10^8 \text{ cm}$. While this restricts validity of the model to regions which do not have significant surface height compared to the height of the tropopause, on a large scale the earth is very nearly spherical. This value of $2 \times 10^5 \text{ cm}$ is suggested in Tverskoi (1965) and is supported by Bischof (1965), Bischof and Bolin (1966) and Garratt and Pearman (1973). Thus we may delete S from (1) to incorporate it as part of the boundary condition at $r = a$.

While the tropopause forms an ellipsoidal envelope of the earth with height varying from $10 \times 10^5 \text{ cm}$ to $19 \times 10^5 \text{ cm}$, for mathematical convenience we shall assume that $H = 10 \times 10^5 \text{ cm}$, the height of the tropopause, is independent of temporal and spatial coordinates. Accordingly, we define the region of interest of our model, the mid troposphere, to be between $2 \times 10^5 \text{ cm}$ and $10 \times 10^5 \text{ cm}$. The variations of CO_2 concentration across the tropopause have been studied in Bischof (1965, 1971, 1973) and Bolin and Bischof (1966, 1970). Because of the temperature inversion, the vertical mixing just above the tropopause is much less intense than below. This causes the tropopause to behave as a damping layer for CO_2 flux (Bolin and Bischof (1970)); this may be introduced into the model as a boundary condition at $r = a + H$

by defining a leakage coefficient α which relates CO_2 flux to the gradient across the tropopause ; the value α lies in a neighborhood of $2.79 \times 10^3 \text{ cm-s}^{-1}$. Let \bar{C} denote the CO_2 concentration just above the tropopause. The amplitude of seasonal variation of \bar{C} is small compared to that of C in the mid troposphere ; also \bar{C} contains little spatial variation. Therefore, we assume that a spatially independent and aperiodic value of \bar{C} shall suffice for calculating the gradient across the tropopause. The nonstationary behavior of \bar{C} shall, however, remain as a source of variation.

One dimensional diffusion models with constant $K_r = 2.5 \times 10^5 \text{ cm-s}^{-1}$ (Bolin and Bischof (1970), Bolin and Keeling (1963)) or constant $K_\psi = 3 \times 10^{10} \text{ cm}^2\text{-s}^{-1}$ (Bolin and Keeling (1963), Junge and Czeplak (1968)) have been constructed with reasonable success. While no similar work exists for K_θ , from geometrical consideration we anticipate that $K_\theta = K_\psi$ at the equator and generalize this result for mathematical convenience.

The annual averaged drift velocity in the north-south direction is estimated to be $u_\psi = 10^{-1} \text{ cm-s}^{-1}$ (Hoffert (1974)). Thus, the transport of CO_2 in the north-south direction due to advection is negligible compared to that due to diffusion. There is no consistent averaged radial drift pattern ; we expect u_r to be near 0. So the radial diffusion of CO_2 is dominated by turbulent diffusion transport. The annual averaged drift velocity in the east-west direction has magnitudes of typically 3000 cm-s^{-1} , 250 cm-s^{-1} , and 1000 cm-s^{-1} above the northern hemisphere, the equator, and the southern hemisphere, respectively (Mintz (1954)). Thus, we regard diffusive transport in the east-west direction as negligible in comparison to advective transport. Hence, we adopt a model of axial symmetry (i.e. $\partial C / \partial \theta = 0$) because of the predominance of east-west advection and the absence of source/sink data with longitudinal dependence. As shall become evident, the parameter θ may be suppressed in our axially symmetric model.

We recognize two types of sources and sinks, one whose intensity is independent of the concentration of CO_2 present and the other whose intensity is dependent on the concentration of CO_2 present. While photosynthetic rates on a diurnal time scale depend on the concentration of CO_2 present, this becomes a negligible effect on a seasonal or secular time scale. Hence, we identify the land biota, as well as anthropogenic oxidation, source/sink behavior as independent of the concentration of CO_2 present. We now decompose this CO_2 independent source/sink distribution S into two components, $\sigma(\theta, \psi, t)$ which is stationary and $\eta(\theta, \psi, t)$ which is non-stationary with respect to time ; so $S(\theta, \psi, t) = \sigma(\theta, \psi, t) + \eta(\theta, \psi, t)$, where

$$\eta(\theta, \psi, t) = \frac{1}{T} \int_{t-T}^t S(\theta, \psi, z) dz \quad (2)$$

and T is the fundamental period of the periodic constituents of the stationary

components. Since diurnal variations are negligible, $T = 1$ year. We identify the stationary component σ as corresponding to land biotic source/sink behavior. We now assume that the spatial pattern of the nonstationary source/sink distribution, which we identify as primarily anthropogenic oxidation, remains invariant over the interval of study. The fact that the rates of CO_2 increase in ppm-yr^{-1} at various observatories (Keeling and Pales (1965), Keeling and Brown (1965), Bolin and Bischof (1966,1970), Bolin (1973), Keeling et al. (1976) agree) provides indirect evidence toward the validity of this assumption. So $\eta(\psi, t) = \eta_1(t) L(\psi)$.

The exchange of CO_2 across the ocean-atmosphere boundary, however, does depend on the presence of atmospheric CO_2 via the gradient in partial pressures of CO_2 across the ocean surface (Keeling (1965)). The partial pressures at the ocean surface (in ppm), P_w , depends also on the ocean surface pH and temperature at the location, however, we shall select a representative value constant over time which averages the empirical relation. The flux across the ocean-atmosphere boundary is $-W(\rho C - P_w)$, where W is the coefficient of gas exchange that relates CO_2 flow at the surface to the gradient in CO_2 partial pressures and ρ is the conversion factor from g-cm^{-3} to ppm. Although ρ varies with temperature and pressure, we say $\rho = 5.1 \times 10^8 \text{ ppm-cm}^{-3}\text{-g}^{-1}$. The coefficient W depends on temperature and surface roughness, which results from air and water circulation. In a laboratory W has been determined (Kanwisher (1963)) in experiments which did not involve breaking waves and violent winds. From estimates of the average CO_2 flux entering the atmosphere from the ocean between 30°N and 30°S and the average gradient of partial pressures across this boundary (Keeling (1965)), we fix $W = 2.6 \times 10^{-11} \text{ g-yr}^{-1}\text{-cm}^{-2}\text{-ppm}^{-1}$. We generalize this result to the entire globe and neglect the variations in the fraction of the surface of the earth covered by ocean in any latitude band; this mathematical convenience reflects remaining uncertainties in the determination of the W function.

according to these various simplifications and assumptions, we obtain the following boundary value problem (3-6) :

$$\frac{\partial C}{\partial t} = \frac{K_r}{r^2} \frac{\partial}{\partial r} \left[r^2 \frac{\partial C}{\partial r} + \gamma C \right] + \frac{K_\psi}{r^2 \sin \psi} \frac{\partial \sin \psi}{\partial \psi} \frac{\partial C}{\partial \psi} \quad (3)$$

$$-K_r \frac{\partial C}{\partial r} \Big|_{r=a} = S - W(\rho C - P_w) \Big|_{r=a} \quad (4)$$

$$K_r \frac{\partial C}{\partial r} \Big|_{r=a+H} = \alpha(\bar{C} - C) \Big|_{r=a+H} \quad (5)$$

$$\frac{\partial C}{\partial \psi} \Big|_{\psi=0} = \frac{\partial C}{\partial \psi} \Big|_{\psi=\pi} = 0 \quad (6)$$

The geometry of the spherical coordinate system yields boundary condition (6).

The principal sources and sinks of atmospheric CO_2 are the land biota, the oceans, and anthropogenic oxidation. The above natural decomposition leads to corresponding stationary and nonstationary components of atmospheric CO_2 concentration. Accordingly, we may use source/sink data to estimate σ and η , and then find the induced stationary component C_s and nonstationary component C_a of mid tropospheric CO_2 concentration using (3-6) by the following methods. Finally, we obtain $C(\theta, \psi, r, t) = C_s(\psi, r, t) + C_a(\psi, r, t)$ because of the linearity of the governing equations.

Diurnal Variation. As a matter of preanalysis, let us first consider a vertical column of air with no net horizontal exchange and $\alpha = \gamma = 0$. the governing partial differential equation and its boundary conditions are

$$\frac{\partial C}{\partial t} = K_r \frac{\partial^2 C}{\partial r^2} \quad (7)$$

$$K_r \frac{\partial C}{\partial r} \Big|_{r=a+H} = 0 \quad (8)$$

$$-K_r \frac{\partial C}{\partial r} \Big|_{r=a} = S_j e^{i\omega_j t}, \quad j = 1, 2 \quad (9.j)$$

with $\omega_1 = 1.72 \times 10^{-2} \text{ day}^{-1}$ for a source with period 1 year (seasonal variation) and $\omega_2 = 2 \text{ day}^{-1}$ for a source with period 1 day (diurnal variation). By solution of problems (7,8,9.j) we find that the relative decay of the amplitude of the diurnal variation to the decay of the amplitude of the seasonal variation decreases from 7×10^{-2} at $2 \times 10^5 \text{ cm}$ to 3×10^{-6} at $10 \times 10^5 \text{ cm}$. Accordingly, in the mid troposphere the amplitude of diurnal variations is negligible compared with that of diurnal variations; hence, in the mid troposphere we shall ignore diurnal variations.

3. AXIALLY SYMMETRIC NONSTATIONARY SOLUTION

The nonstationary component C_a of atmospheric CO_2 concentration is governed by the boundary value problem

$$\frac{\partial C_a}{\partial t} = \frac{K_r}{r^2} \frac{\partial}{\partial r} \left[r^2 \frac{\partial C_a}{\partial r} + \gamma C_a \right] + \frac{K_\psi}{r^2 \sin \psi} \frac{\partial}{\partial \psi} \sin \psi \frac{\partial C_a}{\partial \psi} \quad (10)$$

$$-K_r \frac{\partial C_a}{\partial r} \Big|_{r=a} = (\eta - w\rho C_a) \Big|_{r=a} \quad (11)$$

$$K_r \frac{\partial C_a}{\partial r} \Big|_{r=a+H} = \alpha (\bar{C} - C_a) \Big|_{r=a+H} \quad (12)$$

$$\frac{\partial C_a}{\partial \psi} \Big|_{\psi=0} = \frac{\partial C_a}{\partial \psi} \Big|_{\psi=\pi} = 0. \quad (13)$$

Using a separation of variables scheme, we seek solutions of the form $C_a(\psi, r, t) = \Psi(\psi)R(r)T(t)$. It follows that the only solutions for $\Psi(\psi)$ are $\{P_n(\cos\psi), n = 0, 1, 2, \dots\}$, where P_n is the Legendre polynomial of degree n (MacRobert (1967)). From the separation it also follows that $T(t)$ must satisfy the ordinary differential equation

$$\frac{dT}{dt} - \nu T = 0 \quad (14)$$

where ν is some complex constant. Solutions to (14) are of the form $T(t) = T_0 e^{\nu t}$. Since the nonstationary component is characterized by pure exponential behavior with no sinusoidal effects, ν must be real.

Following Hoffert (1974), we assume that the zonally (latitude band) averaged nonstationary source/sink distribution can be written in the product form

$$\eta(\psi, t) = \bar{\eta}(t_0) e^{\lambda(t-t_0)} L(\psi). \quad (15)$$

That is, $\eta_1(t) = \bar{\eta}(t_0) e^{\lambda(t-t_0)}$, where $\bar{\eta}(t_0)$ is the averaged global anthropogenic flux of CO_2 at reference time $t_0 = 1950$. Based on Baes et al. (1976), $\bar{\eta}(t_0) = 3.15 \times 10^{-9} \text{ g CO}_2\text{-cm}^{-2}\text{-day}^{-1}$ and the growth constant $\lambda = 1.18 \times 10^{-4} \text{ day}^{-1}$. From Hoffert (1974), $L(\psi)$ is a normalized latitudinal distribution such that $\int_0^\pi L(\psi) \sin\psi d\psi = 2$. Thus, adopting Hoffert's results,

$$L(\psi) = \frac{2 f(\psi)}{\sin \psi} \quad (16)$$

where $f(\psi)$ is the fraction of energy consumption per unit latitude. This empirical description of $L(\psi)$ can be approximated by a mixture of Legendre polynomials

$$L(\psi) = \sum_{n=0}^m L_n P_n(\cos\psi), \quad (17)$$

where $\{L_n, n = 0, 1, \dots, m\}$ are suitably chosen constants. It is also desirable to choose L_0, \dots, L_m so that $L(\psi)$ is never negative on the interval $[0, \pi]$; $L_0 = 1$ guarantees the normality condition.

It is tempting to conclude that since $\eta(\psi, t)$ increases exponentially with growth parameter λ , the globally averaged CO_2 concentration just above the tropopause, \bar{C} , is also exponentially increasing with the same growth parameter λ . This turns out not to be the case. Because of many active sinks (mainly oceanic) at the surface, $\bar{C}(t)$ increases less rapidly.

An improved empirical description is obtained using a mixed exponential form

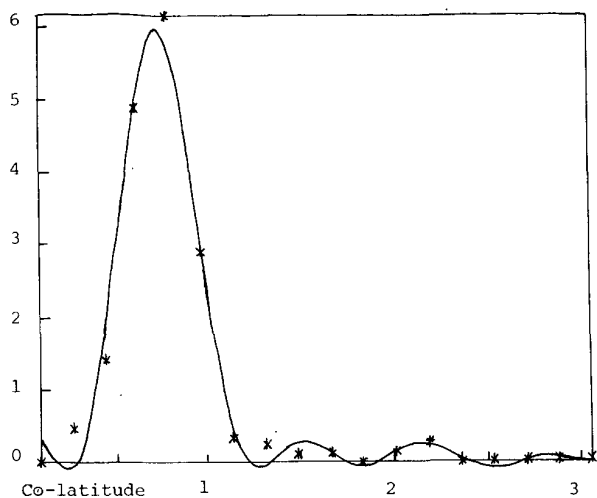


Fig. 1. Normalized latitudinal source/sink distribution profile

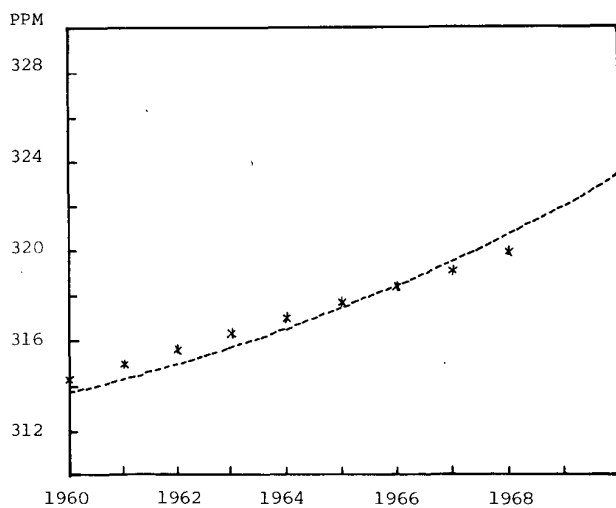


Fig. 2. An empirical model of $\bar{C}(t)$ using $l = 1$.

$$\bar{C}(t) = \sum_{p=0}^l \bar{C}_p e^{\lambda_p(t-t_0)} \quad (18)$$

where $\lambda_0 = \lambda$.

It is now evident from the above discussion that the general solution of the

nonstationary component is given by

$$C_a(\psi, r, t) = \sum_{p=0}^l \sum_{n=0}^m P_n(\cos\psi) \bar{R}_{np}(r) e^{\lambda_p(t-t_0)} \quad (19)$$

Each function $\bar{R}_{np}(r)$ is the solution to a single variable boundary value problem

$$r^2 \bar{R}_{np}'' + (2r + \gamma) \bar{R}_{np}' - \frac{1}{K_r} [\lambda_p r^2 + n(n+1)K_\psi] \bar{R}_{np} = 0 \quad (20)$$

$$-K_r \bar{R}_{np}'(a) = I(p) \bar{\eta}(t_0) L_n - W \rho \bar{R}_{np}(a) \quad (21)$$

$$K_r \bar{R}_{np}'(a+H) = \alpha [I(n) \bar{C}_p - \bar{R}_{np}(a+H)], \quad (22)$$

where $I(p)$ and $I(n)$ are unity if the argument is 0 and 0 if the argument is nonzero.

4. AXIALLY SYMMETRIC STATIONARY SOLUTION

The Fourier representation of the concentration independent stationally source/sink component is

$$\sigma(\psi, t) = \operatorname{Re} \sum_{j=0}^{\infty} \sigma_j(\psi) e^{ij\omega t}, \quad (23)$$

where $\sigma_j(\psi)$ is a complex valued function embodying the phase behavior of the j th harmonic $\operatorname{Re} \sigma_j(\psi) e^{ij\omega t}$ and $\omega = 1.72 \times 10^{-2} \text{ day}^{-1}$ is the fundamental frequency of seasonal variation. In order to empirically describe these harmonics we may use standard Fourier technique on the atmospheric release/uptake data reported by latitude band and month in 10^{14} gC in Machta (1974); each of these values, however, needs to be first normalized by dividing by the respective area of the latitude band and multiplied by 1.205×10^{-1} to convert to the basic time unit of 1 day and since 1 gC yields 3.67 gCO_2 . We observe that from this data we obtain $\sigma_0(\psi) = 0$. We discard those functions $\sigma_j(\psi)$ which do not differ significantly from zero and approximate each of the remainder by a mixture of Legendre polynomials

$$\sigma_j(\psi) = \sum_{n=0}^m B_{nj} P_n(\cos\psi), \quad (24)$$

finding the parameter B_{nj} by standard regression techniques. Note that $B_{n0} = 0$ for all n .

Denote by C_j the solution due to the j th harmonic; by virtue of the principle of superposition from the linearity of the model (3-6), we may write the stationary solution

$$C_s(\psi, r, t) = \operatorname{Re} \sum_{j=0}^{\infty} C_j(\psi, r, t). \quad (25)$$

Each harmonic C_j of stationary mid tropospheric CO_2 concentration is governed by the boundary value problem

$$\frac{\partial C_j}{\partial r} = \frac{K_r}{r^2} \frac{\partial}{\partial r} [r^2 \frac{\partial C_j}{\partial r} + \gamma C_j] + \frac{K_\psi}{r^2 \sin \psi} \frac{\partial}{\partial \psi} \sin \psi \frac{\partial C_j}{\partial \psi} \quad (26)$$

$$-K_r \frac{\partial C_j}{\partial r} \Big|_{r=a} = \operatorname{Re} \sigma_j(\psi) e^{ij\omega t} + I(j) W P_w - W P C_j \Big|_{r=a} \quad (27)$$

$$K_r \frac{\partial C_j}{\partial r} \Big|_{r=a+H} = -\alpha C_j \Big|_{r=a+H} \quad (28)$$

$$\frac{\partial C_j}{\partial \psi} \Big|_{\psi=0} = \frac{\partial C_j}{\partial \psi} \Big|_{\psi=\pi} = 0. \quad (29)$$

Using the same separability scheme, as above, for each C_j , we obtain the stationary solution of the form

$$C_s(\psi, r, t) = \operatorname{Re} \sum_{j=0}^{\infty} e^{ij\omega t} \sum_{n=0}^{\infty} P_n(\cos \psi) \tilde{R}_{nj}(r). \quad (30)$$

The functions $P_n(\cos \psi)$ are, as above, the solutions of Legendre's equation which is derived in the separation. Each function $\tilde{R}_{nj}(r)$ is the solution to a single variable boundary value problem

$$r^2 \tilde{R}_{nj}'' + (2r + \gamma) \tilde{R}_{nj}' - \frac{1}{K_r} [ij\omega r^2 + n(n+1)K_\psi] \tilde{R}_{nj} = 0 \quad (31)$$

$$-K_r \tilde{R}_{nj}'(a) = B_{nj} + I(j) W P_w - W P \tilde{R}_{nj}(a) \quad (32)$$

$$K_r \tilde{R}_{nj}'(a+H) = -\alpha \tilde{R}_{nj}(a+H). \quad (33)$$

Here also $I(j)$ is the indicator of whether $j = 0$. Upon finding the significant terms, the summation (30) should be truncated to contain only those.

5. OBTAINING $C(\theta, \psi, r, t)$

Taking advantage of the linearity of the governing partial differential equation (3), our two components C_a and C_s , which satisfy their respective problems,

(10-13) and (26-29), sum to a solution of the master problem (3-6) for mid tropospheric CO_2 concentration

$$C(\theta, \psi, r, t) = C_a(\psi, r, t) + C_s(\psi, r, t)$$

$$= \sum_{p=0}^L \sum_{n=0}^m P_n(\cos\psi) \bar{R}_{np}(r) e^{\lambda_p(t-t_0)} + \text{Re} \left\{ \sum_{j=0}^J e^{i j \omega t} \sum_{n=0}^N P_n(\cos\psi) \tilde{R}_{nj}(r) \right\}, \quad (34)$$

where \bar{R}_{np} and \tilde{R}_{nj} satisfy their respective boundary value problems.

This method of solution is conceptually more vivid than the usual pure numerical solution methods. It also turns out to be computationally more feasible, requiring less machine memory and processor time. The functions \bar{R}_{np} and \tilde{R}_{nj} are numerically calculated with almost as much ease as the exponential, trigonometric, and Legendre functions. The form of their boundary value problems (20-22 or 31-33) is linear. Any solution to a linear single variable boundary value problem is of the form $a_1 Y_1 + a_2 Y_2$, where a_1 and a_2 are constants and Y_1 and Y_2 are independent solutions of the second order linear differential equation (20 or 31). Without any loss of generality these independent solutions may be specified according to the initial conditions

$$\begin{pmatrix} Y_1(a) & Y_1'(a) \\ Y_2(a) & Y_2'(a) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \quad (35)$$

Tables of values of the functions Y_1 and Y_2 are obtained on the interval $[a, a+H]$ via some numerical integration scheme (e.g. Runge-Kutta) applied to the two independent initial value problems (Hildebrand(1956)). Upon substitution of the form $a_1 Y_1 + a_2 Y_2$ into the boundary conditions (21-22 or 32-33), a system of linear equations results from which a_1 and a_2 can be found. It is not known with mathematical certainty whether this system is always nonsingular; however, for most physically well-posed problems, a unique solution can be expected. The above procedure must be repeated for each pair (n, p) or (n, j) which produces a significant component of (34).

6. DISCUSSION

The above model is an interpretation of global scale processes and empirical evidence. While it extends prior models by simultaneously considering variations according to (ψ, r, t) within an axially symmetric framework, several assumptions and simplifications were made for its construction; thus, this model should be viewed within a sequence of progressively more appropriate semi-analytic descriptions of the distribution of CO_2 concentration in the mid troposphere.

In approaching the various subproblems and in making assumptions and simplifications based on empirical evidence, intuitive appeal, or mathematical expedience, we have identified several areas where the analysis presented here could be validated or extended; it would yield additional interesting results if more data were available. By restricting the domain to the equatorial region, we may obtain solutions which include longitudinal variations in order to shed light on the prior assumption that the east-west advection suppresses longitudinal variations sufficiently to justify the axially symmetric model. If there are significant longitudinal variations, within any latitude band the CO_2 concentration will form standing wave patterns subject to random fluctuations due to transient source/sink behavior. Additional data is required to investigate the spatial dependence of the flux across the ocean-atmosphere boundary due to latitude variations of the proportion of the surface covered by ocean and due to other concomitant meteorological factors.

Computational results suitable for comparison with Bolin and Keeling (1963) and others shall be forthcoming in a later communication. Mathematical attention must be focused on finding the optimal nonnegative Legendre polynomial approximation to the normalized nonstationary source/sink distribution (17). Other empirical parameters (18,24) require evaluation. Some tuning of the model by a sensitivity analysis of vaguely specified parameters will be an important refinement of the results.

ACKNOWLEDGEMENT

We gratefully acknowledge the financial support of General Dynamics Corporation to the Mathematics Clinic of Claremont Graduate School and Harvey Mudd College during which much of the analysis was performed. Many persons associated with the Mathematics Clinic have contributed to the task of atmospheric CO_2 modeling; thanks are due to Dr. C. Coleman, Dr. S. Busenberg, Dr. A. Jones, Dr. J. Spanier, T. Jenkins, D. Cline, B. Halford, N. Morgan, and T. Allen. We also thank Dr. C. D. Keeling of the Scripps Institute of Oceanography for his advice and suggestions. We also wish to acknowledge the U.S. Office of Naval Research for the opportunity to present these results.

REFERENCES

- Baes, C. F., Goeller, H. E., Olson, J. S. and Rotty, R. M., 1976. The global carbon dioxide problem. Nat. Tech. Inf. Service, U.S. Dept. of Comm., ORNL-5194.
- Bischof, W., 1973. Carbon dioxide concentration in the upper troposphere and lower stratosphere, III. Tellus 25:305-308.
- Bischof, W., 1971. Carbon dioxide concentration in the upper troposphere and lower stratosphere, II. Tellus 23:558-561.
- Bischof, W., 1965. Carbon dioxide concentration in the upper troposphere and lower stratosphere, I. Tellus 17: 398-402.
- Bischof, W. and Bolin, B., 1970. Variations of the carbon dioxide content of the atmosphere in the northern hemisphere, Tellus 22: 431-442.

- Bischof, W. and Bolin, B., 1966. Space and time variations of the carbon dioxide content of the troposphere and lower stratosphere. *Tellus* 18: 155-159.
- Bolin, B. and Keelin, C.D., 1963. Large-scale atmospheric mixing as deduced from the seasonal and meridional variations of carbon dioxide. *J. Geophys. Res.* 68(13): 3899-3920.
- Brown, C.W. and Keeling, C.D., 1965. The concentration of atmospheric carbon dioxide in Antarctica. *J. Geophys. Res.* 70(24):6077-6085.
- Garratt, J.R. and Pearman, G.I., 1973. Carbon dioxide concentration in the atmospheric boundary layer over southeast Australia. *Atmos. Environ.* 7: 1257-1266.
- Garratt, J.R. and Pearman, G.I., 1972. Global aspects of carbon dioxide. *Search* 3 : 3.
- Hildebrand, F.B., 1956. *Introduction to Numerical Analysis*. McGraw-Hill, New York.
- Hoffert, M.A., 1974. Global distributions of atmospheric carbon dioxide in the fossil fuel era: a projection. *Atmos. Environ.* 8: 1225-1249.
- Junge, C.E. and Czeplak, G., 1968. Some aspects of the seasonal variation of carbon dioxide and ozone. *Tellus* 20 : 422-434.
- Kanwisher, J., 1963. Effect of wind on CO_2 exchange across the sea surface. *J. Geophys. Res.* 68: 3921-3927.
- Keeling, C.D., 1977. *Conversations*.
- Keeling, C.D., 1965. Carbon dioxide in surface waters of the Pacific Ocean 2. Calculation of the exchange with the atmosphere. *J. Geophys. Res.* 70(24):6099-6102.
- Keeling, C.D., Adams, C.A., Ekdahl, C.A. and Guenther, P.R., 1976. Atmospheric carbon dioxide variations at the south pole. *Tellus* 28: 552-564.
- Machta, L., 1974. Global-scale atmospheric mixing. *Adv. Geophys.* 18:33-56.
- MacRobert, T.M., 1967. *Spherical Harmonics*. Pergamon P., Oxford.
- Mintz, Y., 1954. The observed zonal circulation of the atmosphere. *Bull. Amer. Meteor. Soc.* 35:208-214.
- Pales, R.C. and Keeling, C.D., 1965. The concentration of atmospheric carbon dioxide in Hawaii. *J. Geophys. Res.* 70(24): 6053-6076.
- Rotty, R.M., 1976. Global carbon dioxide production from fossil fuels and cement, A.D. 1950 - A.D. 2000. *ONR Conf. on the Fate of Fossil Fuel Carbonates*, Honolulu.
- Tverskoi, P.N., 1965. *Physics of the Atmosphere*. U.S. Dept. of Comm., Springfield, Va.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

SOME NEW WORLDWIDE CLOUD COVER MODELS

S.T.BEAN and P.N.SOMERVILLE

Dept. Stat., Univ. of Central Florida, Orlando, Florida

ABSTRACT

Bean, S.T. and Somerville, P.N. Some new worldwide cloud cover models. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

Using daily measurements of day and night infrared, and incoming and absorbed solar radiation obtained from a TIROS satellite over a period of approximately 45 months, and integrated over 2.5 degree latitude longitude grids, the proportion of cloud cover over each grid each day was derived for the entire period. For each of four three month periods, for each grid location, estimates a and b of the two parameters of the best fit beta distribution were obtained. The (a,b) plane was divided into a number of regions. All the geographical locations whose (a,b) estimates were in the same region in the (a,b) plane were said to have the same cloud cover type for that season. For each season, the world is thus divided into separate cloud cover types.

Using estimates of mean cloud cover, for each season the world was again divided into separate cloud cover types. The process was repeated for standard deviations.

For each season three separate cloud cover models were thus obtained using the criteria of shape of frequency distribution, mean cloud cover and variability of cloud cover.

INTRODUCTION

The purpose of this study was to develop a model for worldwide cloud cover using a satellite data set containing infrared radiation measurements. Other cloud cover models exist (Barnes et al., 1968; Falls, 1974; Greaves et al., 1971). These early cloud models used primarily ground-based cloud observations. The satellite data set containing Day IR, Night IR, Incoming, and Absorbed solar radiation measurements on a 2.5-degree latitude-longitude grid covering a 45 month period of record has recently become available. There was originally a 2-year period of similar data on an NMC grid. The first step was to convert these infrared data to estimates of cloud cover. The statistical analysis of classification of cloud region characteristics was then performed.

There are several reasons for desiring a cloud model based on satellite data. The ground-based data are much more limited in scope. Some fairly large areas of the world have either no data or very sparse data, and models using ground-based observations necessitate a number of assumptions, including on occasion that a region is essentially like its antipodal location. A good worldwide cloud cover model is

needed for the purpose of studying the relationship between cloudiness, precipitation, and the Earth radiation budget.

CONVERSION OF SATELLITE IR MEASUREMENTS TO CLOUD COVER

A major initial task was to derive cloud cover estimates from the satellite infrared data. The method used in this investigation follows the suggestions obtained through personal communications with Thomas I. Gray, Jr. (1978).

Albedo is defined as the reflective power, or the fraction of incident light that is reflected by a surface or body. Included in the satellite data are the amount of incoming solar radiation, I_{in} , and the amount of absorbed solar radiation, I_{ab} . The satellite observed albedo A , is estimated by

$$A = (I_{in} - I_{ab}) / I_{in} \quad (1)$$

If the Earth's surface absorbed all solar radiation, then the cloud cover might be taken simply as 1 minus albedo (assuming also that clouds reflect all solar radiation). Different parts of the Earth's surface, however, have differing radiances. For example, the albedo of the ocean is approximately 5 percent (95 percent of the solar radiation being absorbed), while the Sahara desert reflects approximately 40 percent of the solar radiation reaching it.

To determine cloud cover, we needed to obtain the background radiation of the region of the Earth of interest. To do this, for a given season and a specific location, we calculated A from eqn(1) for every day of a season and observed the minimum value, A_{min} . This minimum value should occur on the day of least (hopefully near zero) cloud cover. If r is the reflectance of the clouds and x is the fraction of cloud cover, then the basic formula may be written as

$$A = x \cdot r + (1-x) \cdot A_{min}$$

from which we have the fraction of cloud cover, x , as

$$x = (A - A_{min}) / (r - A_{min}) \quad (2)$$

This formula requires a knowledge of r which varies.

A way to estimate the cloud reflectance r , is by observing the difference between the Earth's surface temperature and the temperature equivalent of the satellite-observed daytime infrared reading (denoted by IR_D). The radiance of the IR_D by Stefan's law is equal to $5.75 \cdot 10^{-8} T^4$ (watts/m²), where T is the temperature equivalent in degrees Kelvin. Putting $z = (\text{surface temperature} - T)$, (units degree Kelvin), the following relationship has been observed:

$$r = -0.000265z^2 + 0.0295z + 0.10 \quad (3)$$

A surface temperature of 30°C for latitudes within 25 degrees of the equator and -5°C for latitudes within 25 degrees of the pole was used. Interpolations were used for intermediate latitudes.

Using this method, the proportion of cloud cover was calculated for each day of the year over the 4-year period covering the entire globe. We considered the data for the four seasons separately, and we developed a separate cloud cover model for Winter, Spring, Summer and Fall, where Winter consists of the months December, January, and February, etc.

Because this investigation is based on derived cloud cover estimates and may be subject to criticism, it is noted that ground-based cloud observations are also estimates as well as cloud cover obtained by satellite photography. We make this conjecture: Those variables which are not well defined in the IR to cloud cover conversion procedure will have small contributions to climatic modelling of the clouds over the entire season. For a specific day and area the preceding procedure may not be entirely satisfactory for synoptic cloud cover analysis. It should be noted that because of the orbit of the TIROS satellite the estimates in the near polar regions are degraded somewhat.

A PROBABILISTIC MODEL FOR CLOUD COVER

The proportion of cloud cover over any grid square is a random variable which has some probability distribution associated with it. Falls (1974) found that the Beta distribution could be used to represent the probability distribution of the proportion of cloud cover. Henderson and Sellers (1978) has also found the Beta distribution useful as a model for the probability distribution of the proportion of cloud cover. The Beta probability density function with parameters a and b is given by:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1} (1-x)^{b-1}$$

for $0 \leq x \leq 1$, $a > 0$, $b > 0$.

The Beta probability density function can assume a variety of shapes. It can be mound shaped, U-shaped, or J-shaped with varying amounts of skewness. Table 1 shows the relationship between the a and b parameters and the shape of the frequency curve.

TABLE 1.

Shapes of the Beta probability density function for different a and b parameters

Shape	Parameters	
Mound	$a > 1$	$b > 1$
J	$a > 1$	$b < 1$
Reverse J	$a < 1$	$b > 1$
U	$a < 1$	$b < 1$
Uniform	$a = 1$	$b = 1$

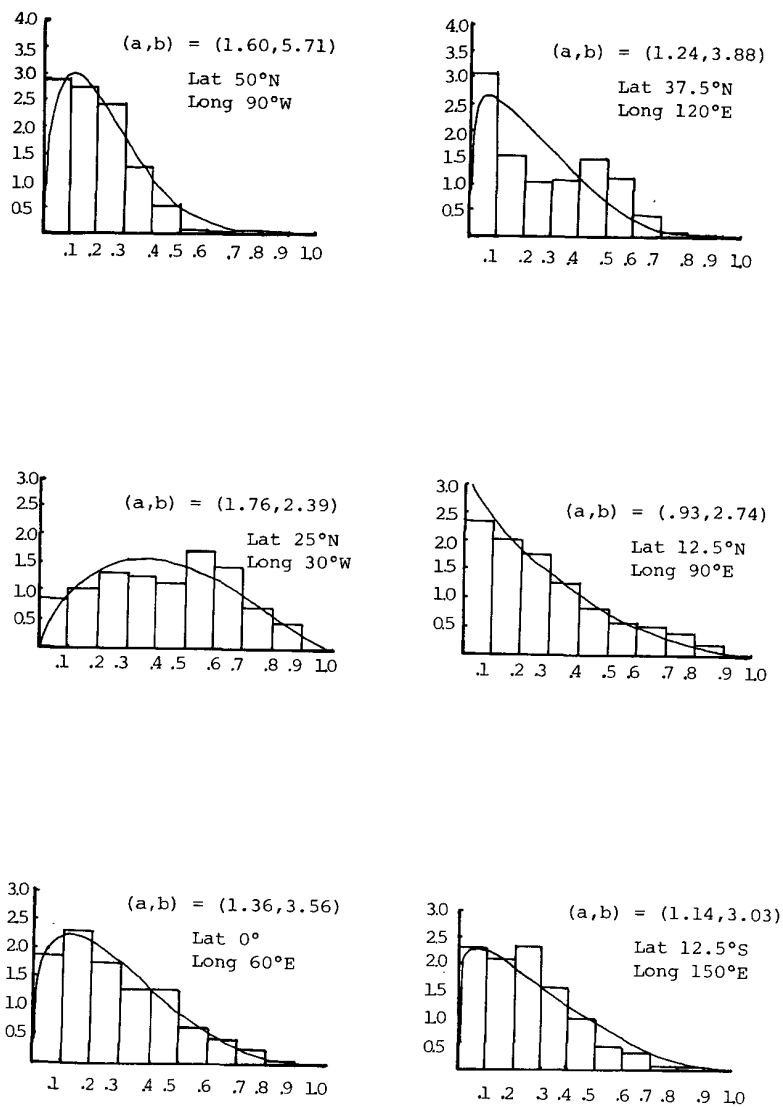


Fig. 1. Histograms of the proportion of cloud cover with Beta curves superimposed.

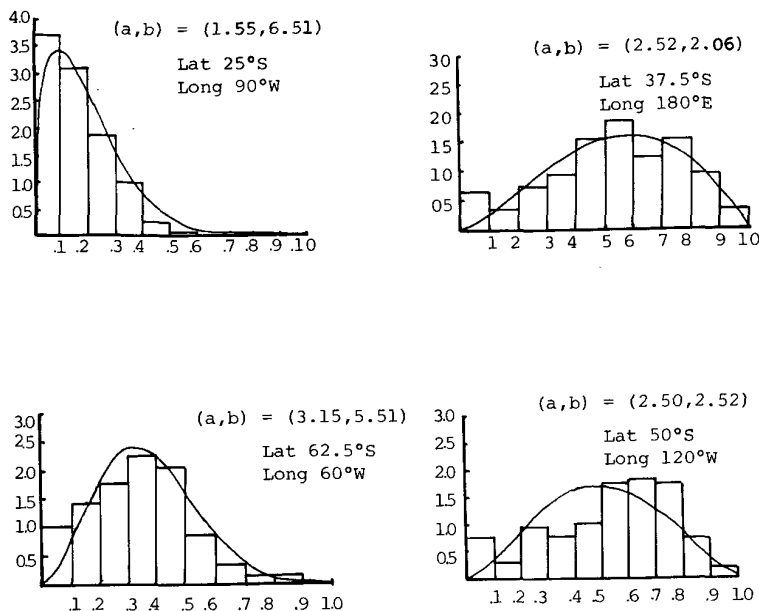


Fig. 2. Histograms of the proportion of cloud cover with Beta curves superimposed.

The question arises as to how well the Beta distribution actually fits the cloud cover data at hand. To answer this question we first estimated the a and b parameters using the methods of moments. These estimates are given by

$$\hat{b} = (1 - \bar{x})[\bar{x}(1 - \bar{x}) - s^2]/s^2, \quad \hat{a} = \bar{x} \hat{b} / (1 - \bar{x}),$$

where \bar{x} is the sample mean proportion of cloud cover, and s^2 is the sample variance of the proportion of cloud cover. Next, we constructed histograms for the cloud cover at several grid locations. A grid location was selected at random from latitude circles 12.5° apart beginning with 50°N latitude and extending to 62.5°S latitude. A histogram was constructed for each selected location on the basis of the cloud cover data for the Winter quarter (December, January, February) for the four years of data. Each of the histograms was constructed on the basis of approximately 350 cloud cover values. The corresponding Beta curves are shown superimposed on the histograms in Figures 1-2.

The (a,b) parameters of the Beta distribution give a good deal of information about the cloud cover characteristics of a given location as illustrated in the previous table and figures. Thus, we used the estimated parameters (a,b) in determining regions of homogeneous cloud cover.

As a first step, it is instructive to consider the frequency histograms of the calculated (a,b) values for each of the four seasons. There are 10,224 grid points

over the globe, and each grid point has an (\hat{a}, \hat{b}) pair associated with it for each season. The following frequency histograms (Figures 3-6) of the 10,224 (\hat{a}, \hat{b}) pairs give some indication of where the parameter pairs are falling in the (a, b) plane.

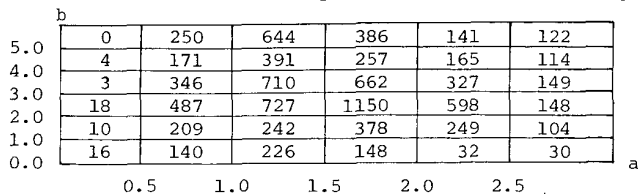


Fig.3. Winter frequency histogram for global (\hat{a}, \hat{b}) values (December, January, February)

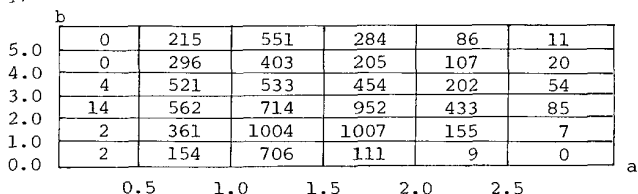


Fig.4. Spring frequency histogram for global (\hat{a}, \hat{b}) values (March, April, May)

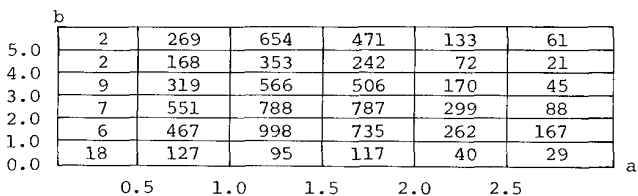


Fig.5. Summer frequency histogram for global (\hat{a}, \hat{b}) values (June, July, August)

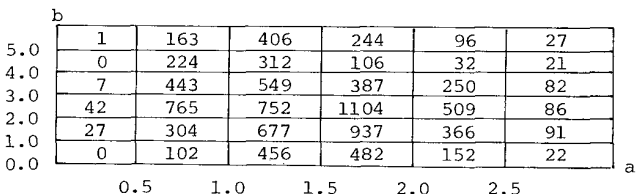


Fig.6. Fall frequency histogram for global (\hat{a}, \hat{b}) values (September, October, November)

The numbers in the above frequency histograms indicate how many (\hat{a}, \hat{b}) parameters over the entire globe fall in the specified block. For example, there are 644 (\hat{a}, \hat{b}) pairs globally such that $1.0 < \hat{a} \leq 1.5$ and $\hat{b} > 5.0$ in the Winter quarter.

The 36 regions in the (a, b) plane from the frequency histogram for global (\hat{a}, \hat{b}) values form a basis for determining homogeneous cloud cover regions. Grid points on the globe which have (\hat{a}, \hat{b}) parameters falling in the same block have very similar cloud cover characteristics. This concept leads to a preliminary cloud cover model.

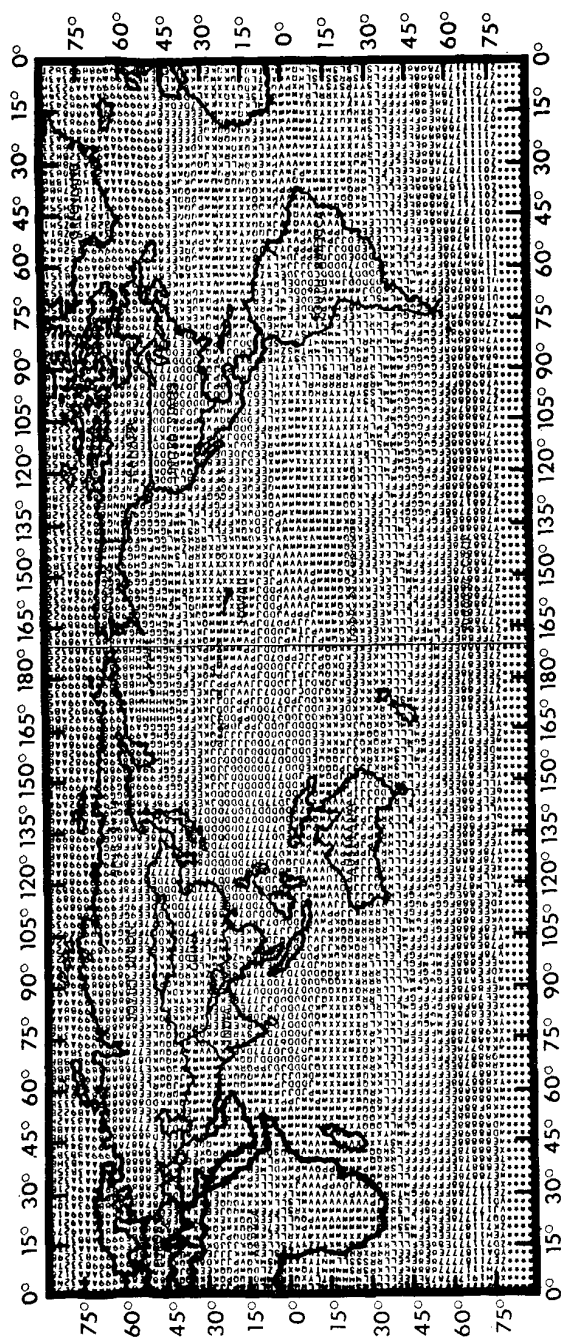


Fig. 7. Global cloud cover classification on (a,b) parameters of the Beta distribution
Summer (June, July, August)

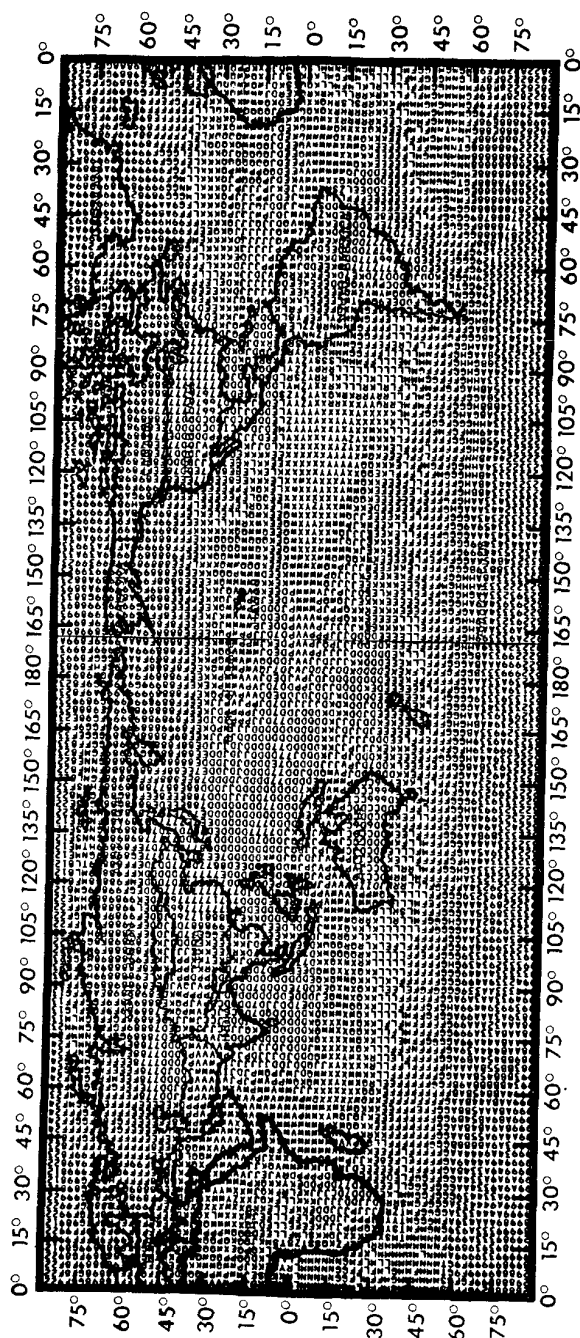


Fig. 8. Global cloud cover classification based on (a,b) parameters of the Beta distribution
Fall (September, October, November)

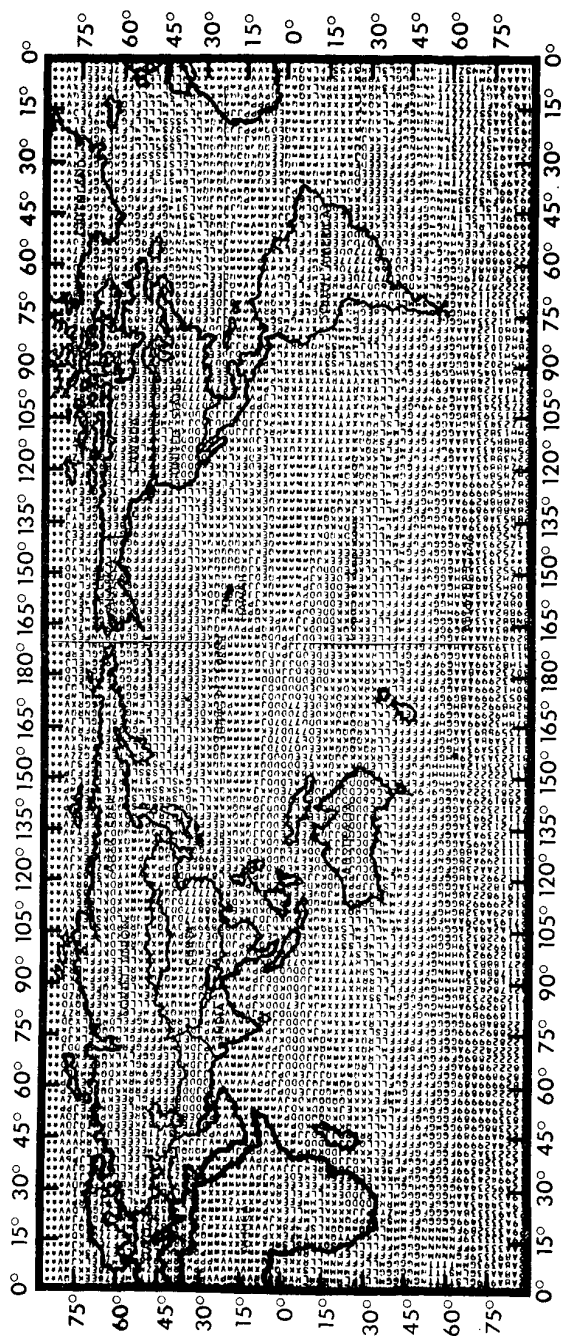


Fig. 9. Global cloud cover classification based on (a,b) parameters of the Beta distribution Winter (December, January, February)

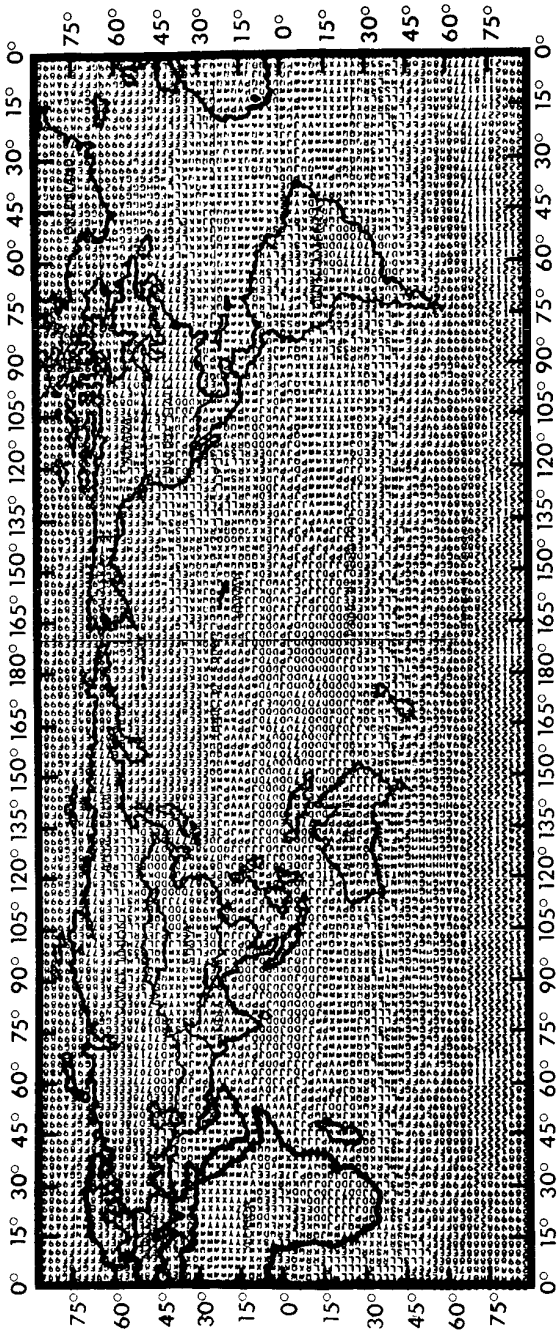


Fig. 10. Global cloud cover classification based on (a,b) parameters of the Beta distribution Spring (March, April, May)

Each grid point on the globe is assigned to one of 36 groups depending on the block in which (a,b) falls. A FORTRAN program was used to label each grid point with 0 - 9 or A - Z depending on the group the grid point fell into. These labels were printed in a rectangular array maintaining the latitude and longitude position of each point. This procedure results in a map of the globe containing a large number of contiguous regions which have the same basic cloud cover characteristics. The resulting maps are shown in Figures 7-10 with global maps superimposed. These maps have been simplified by combining several of the 36 blocks and recording the maps. The record maps are shown in Figures 15-18. The key for the original and maps is given in Figure 11. The key for the record maps is given in Figure 12.

	b	U	V	W	X	Y	Z	
5		O	P	Q	R	S	T	
4		I	J	K	L	M	N	
3		C	D	E	F	G	H	
2		6	7	8	9	A	B	
1		0	1	2	3	4	5	a
0								
		0.5	1.0	1.5	2.0	2.5		

Fig. 11. Code for original homogeneous cloud cover regions map.

The maps may be interpreted according to the distributional characteristics of the various regions. Typical Beta frequency curves for the 12 recorded regions are given in Figures 13-14.

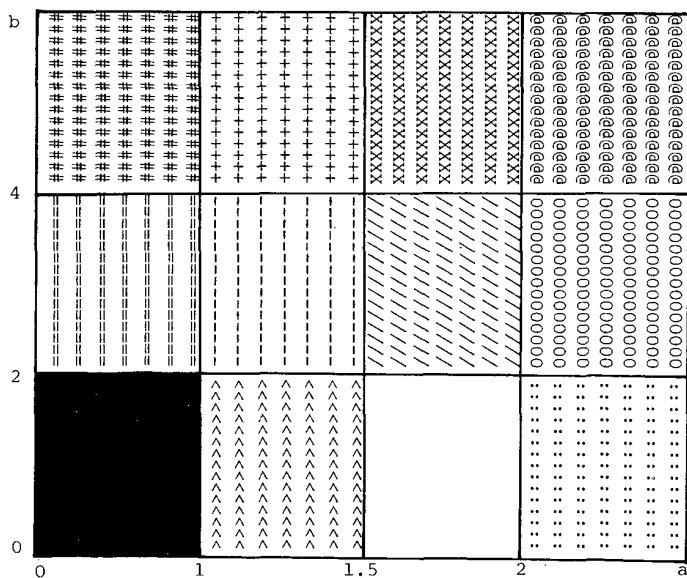


Fig. 12. Key for recorded homogeneous cloud cover regions map.

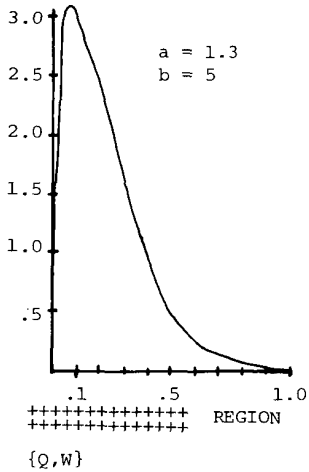
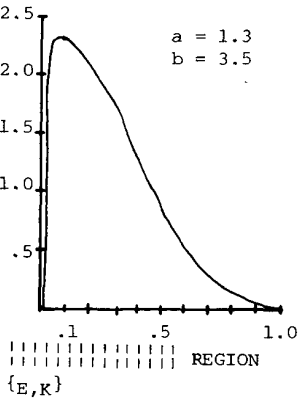
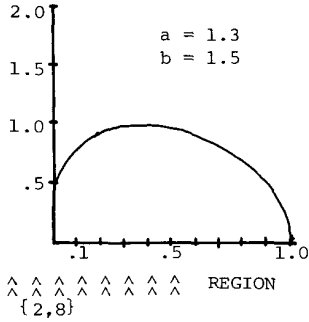
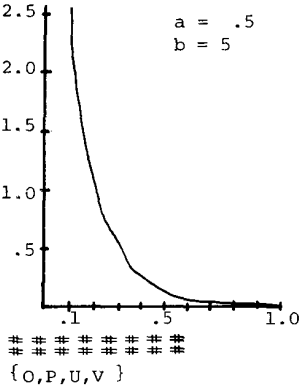
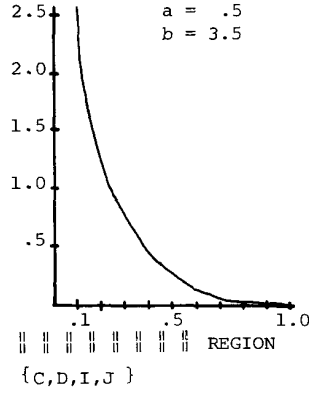
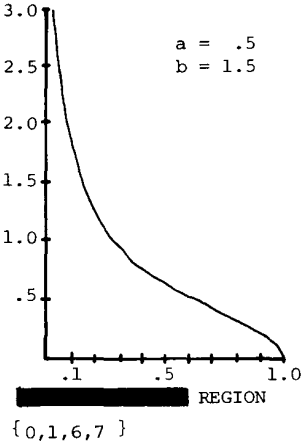


Fig. 13. Typical Beta frequency curves corresponding to 12 regions in Fig.12.

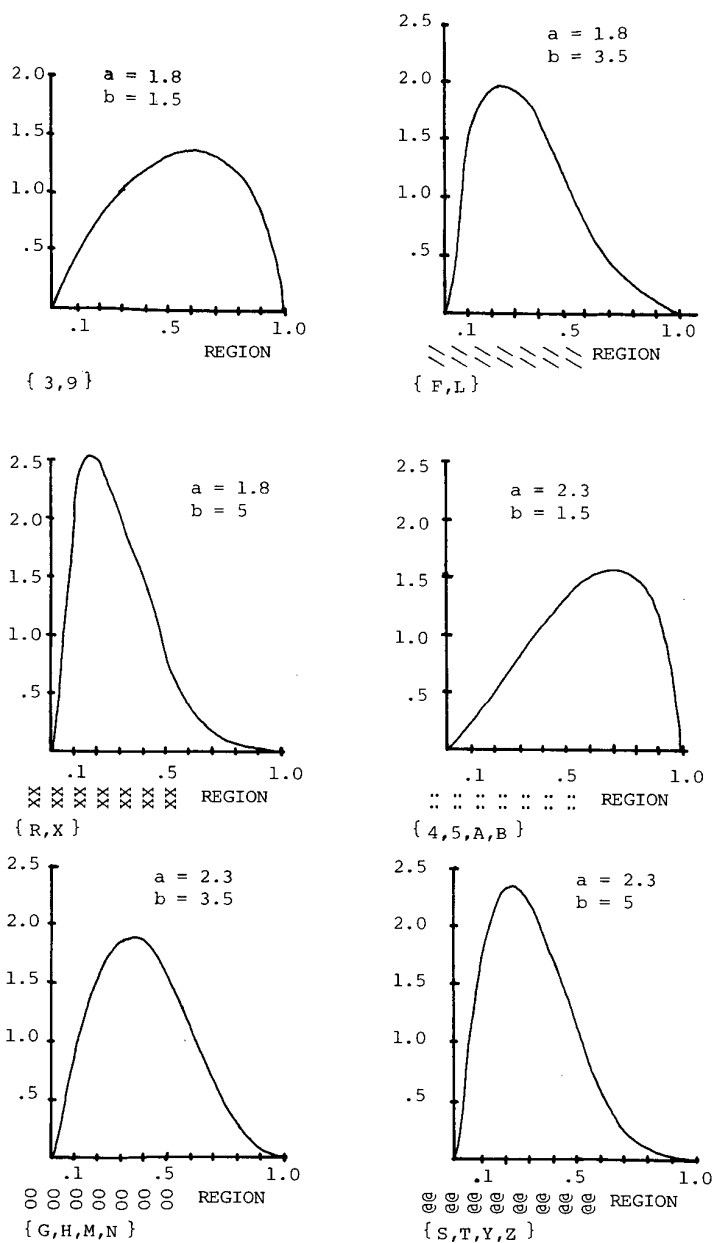


Fig. 14. Typical Beta frequency curves corresponding to 12 regions in Fig.12.

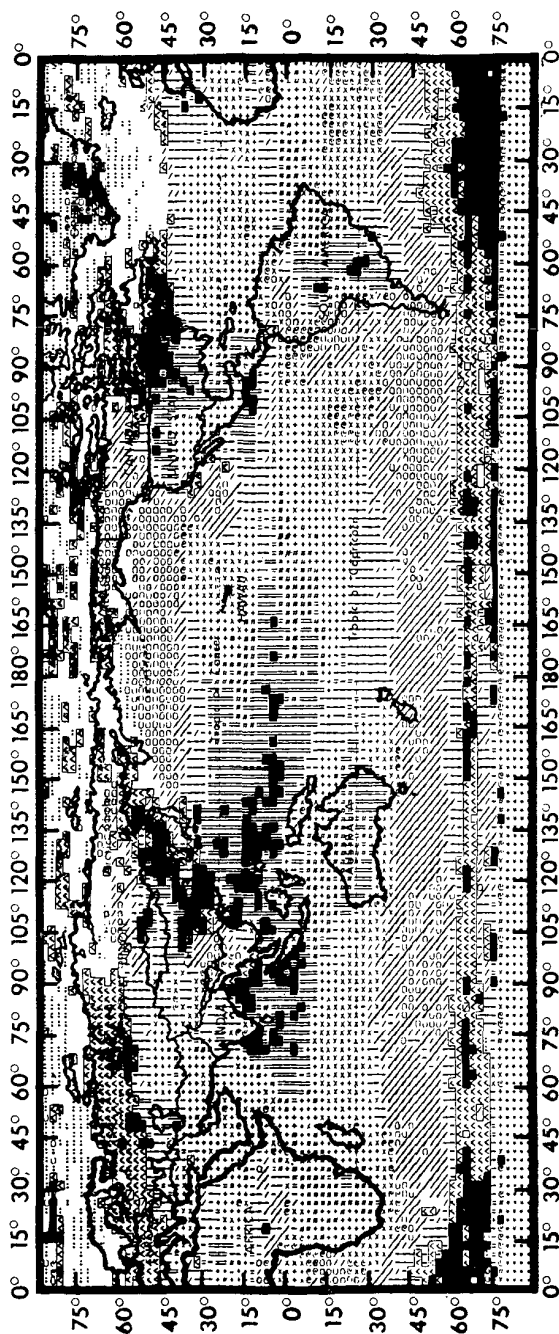


Fig. 15. Global cloud cover classification based on (a,b) parameters of the Beta distribution Summer (June, July, August)

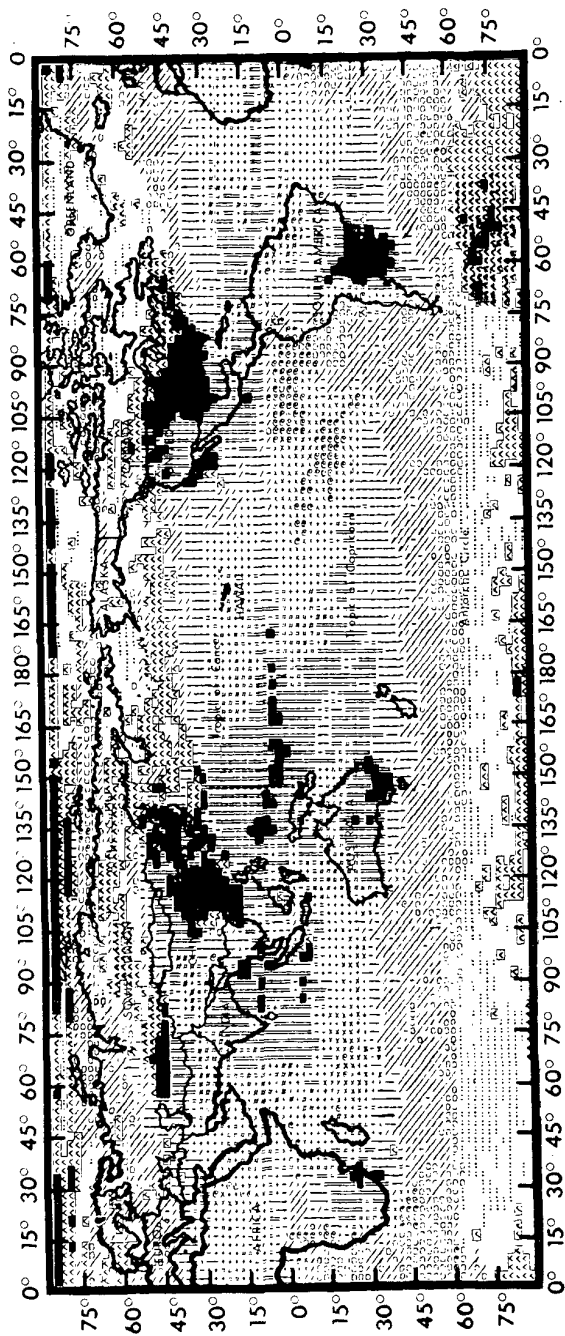


Fig. 16. Global cloud cover classification based on (a,b) parameters of the Beta distribution
Fall (September, October, November)

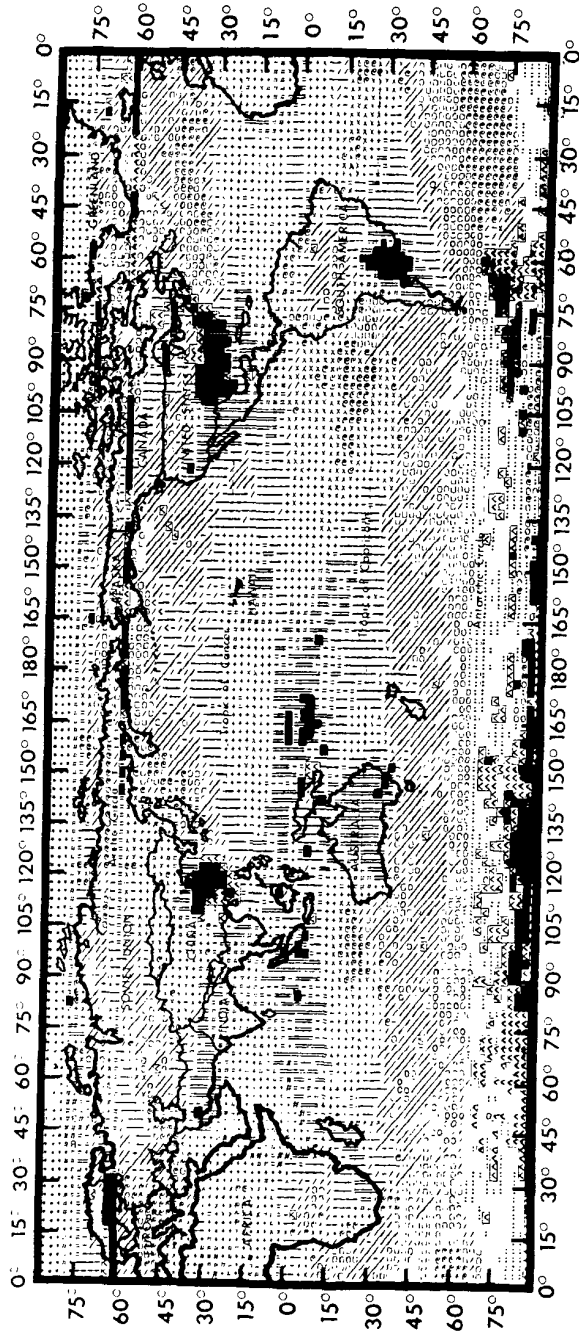


Fig. 17. Global cloud cover classification based on (a,b) parameters of the Beta distribution
Winter (December, January, February)

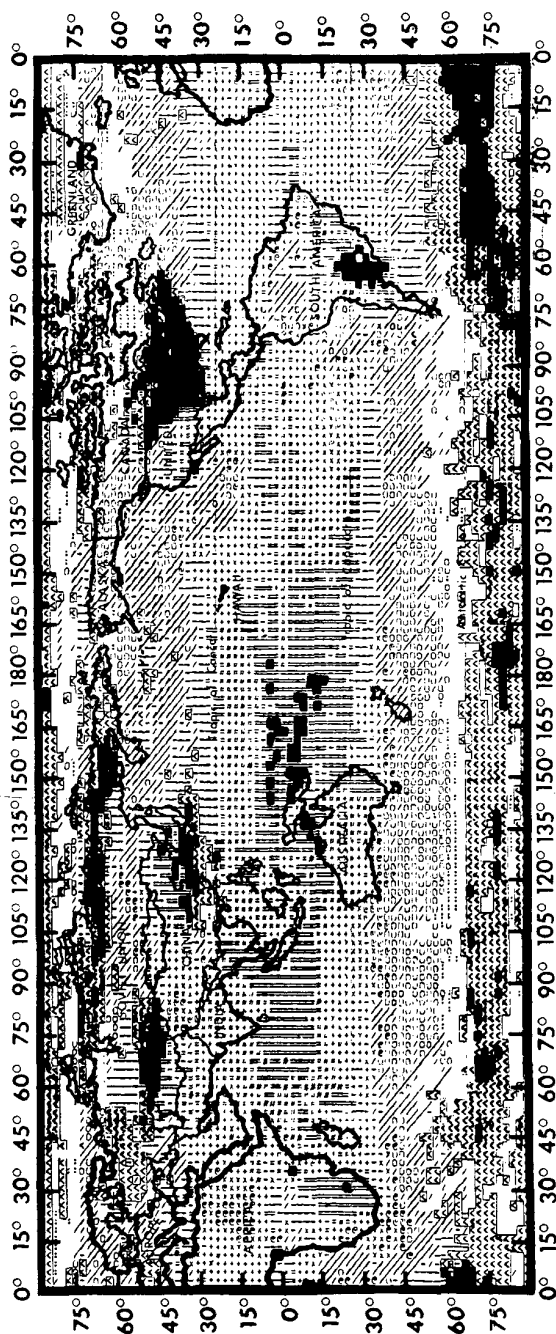


Fig. 18. Global cloud cover classification based on (a,b) parameters of the Beta distribution
Spring (March, April, May)

WORLD WIDE CLOUD COVER MODEL USING MEAN AND STANDARD DEVIATION CONTOUR LINES

One may wish to determine mean and standard deviation of the proportion of cloud cover at any given location. To do this we have developed contour maps which give the mean or standard deviation. The methodology is similar to that used to determine homogeneous cloud cover regions. We first calculated the mean and standard deviation of cloud cover for each grid point for a particular season. To obtain a contour map for the mean cloud cover for a particular season we used a FORTRAN program to print the first digit of the mean cloud cover for each grid point on the globe maintaining the latitude and longitude position. Contour lines were then traced on a transparent map overlaying the printout. Separate maps were constructed in the same way for the standard deviation of cloud cover. These maps are shown in Figures 19-26.

The maps in Figures 19-26 may be used as a general guide to the cloud cover characteristics for any place on the globe. Also, the mean and standard deviation of the proportion of cloud cover obtained from these maps may be used (in the absence of the previously given maps in Figures 7-10) to obtain estimates for the parameters (a,b) in the Beta distribution.

CONCLUSIONS AND RECOMMENDATIONS

The current cloud cover model illustrates a useful objective methodology for cloud cover classification. The developed maps can be useful for those who need some information regarding the cloud cover characteristics for any particular location on the globe. This is particularly useful in that the practicing climatologist can obtain a great deal of cloud cover information without going through large volumes of data.

There are problems with the current cloud cover model. The most obvious of these is the lack of data. To make a good climatological model a reasonably long record length is required. The satellite data available for this study comprise approximately 44 months. This is sufficient for some model development, but a longer period of record would be desirable. Also, temporal persistence cuts down on the actual number of independent cloud cover observations. Another problem is that the cloud cover used in the model is derived. This is not to say that the cloud cover values used are not accurate, but the derived cloud cover should be compared with some independently observed cloud cover. It should be noted that any type of cloud cover measurement will have some degree of error associated with it.

The topic of error analysis in the estimation of the parameters (a,b) was not covered for several reasons. There are several types of errors which made the problem quite complex. First, as previously mentioned, there is an unknown error component in the derived cloud cover itself which may vary from place to place. Also, there is error in recorded satellite measurements. The cloud cover values have a temporal correlation which is not fully known. The short length of the record certainly

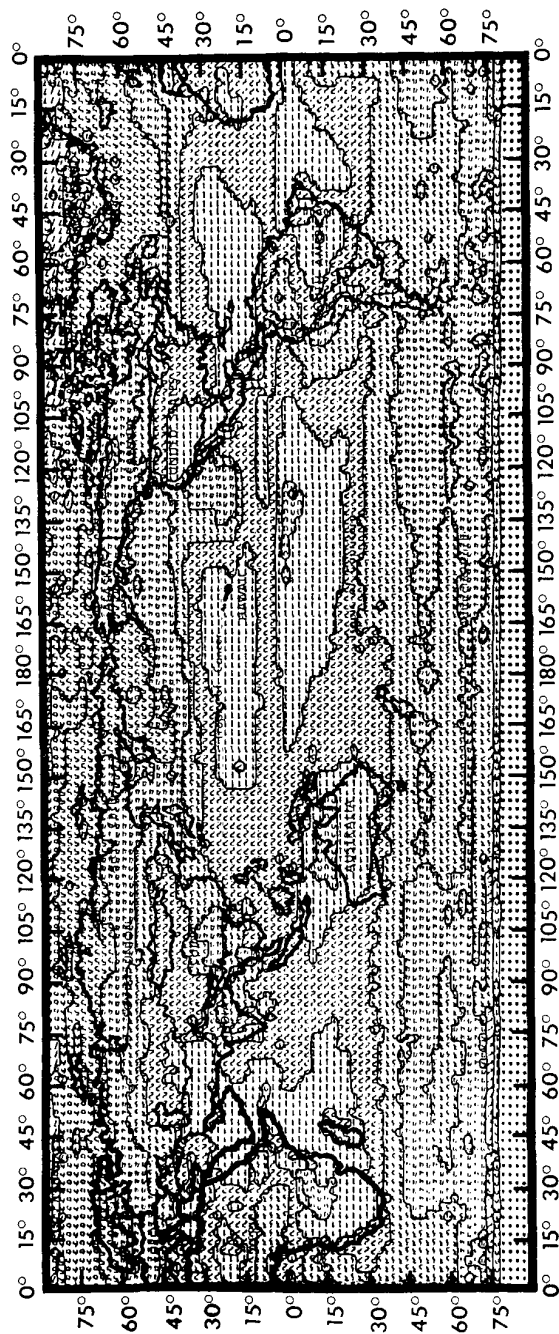


Fig. 19. Mean cloud cover — Digit printed $\times 10$ = Mean percent cloud cover
Summer (June, July, August)

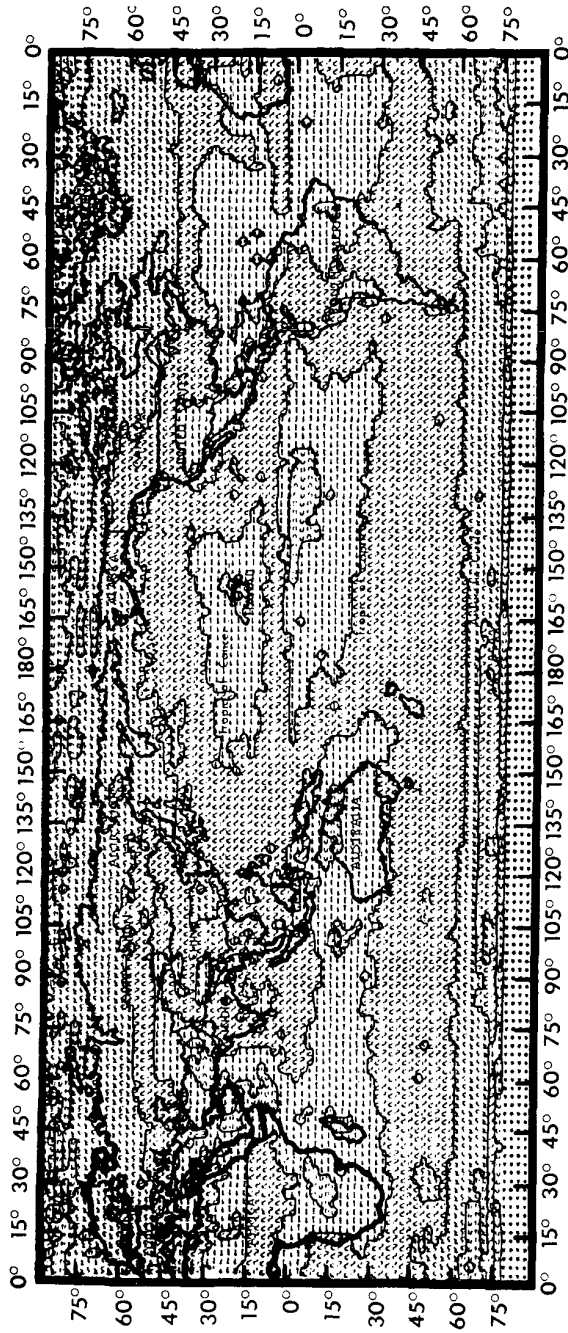


Fig. 20. Standard deviation cloud cover - Digit printed $\times 10 =$ Standard deviation
Summer (June, July, August)

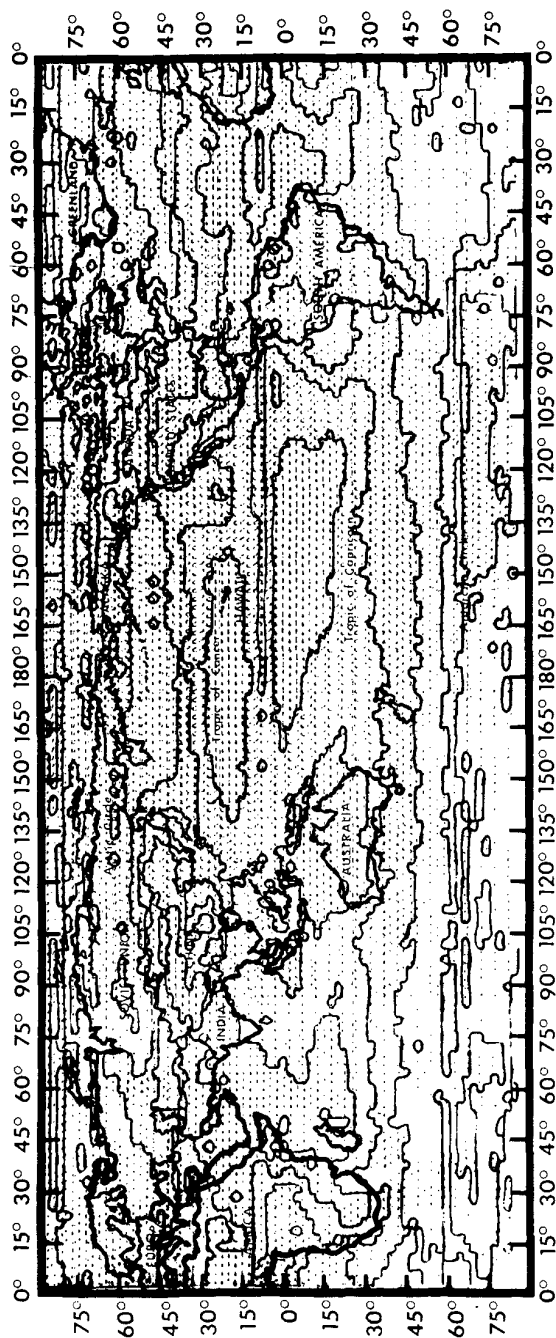


Fig. 21. Mean cloud cover
Fall (September, October, November)

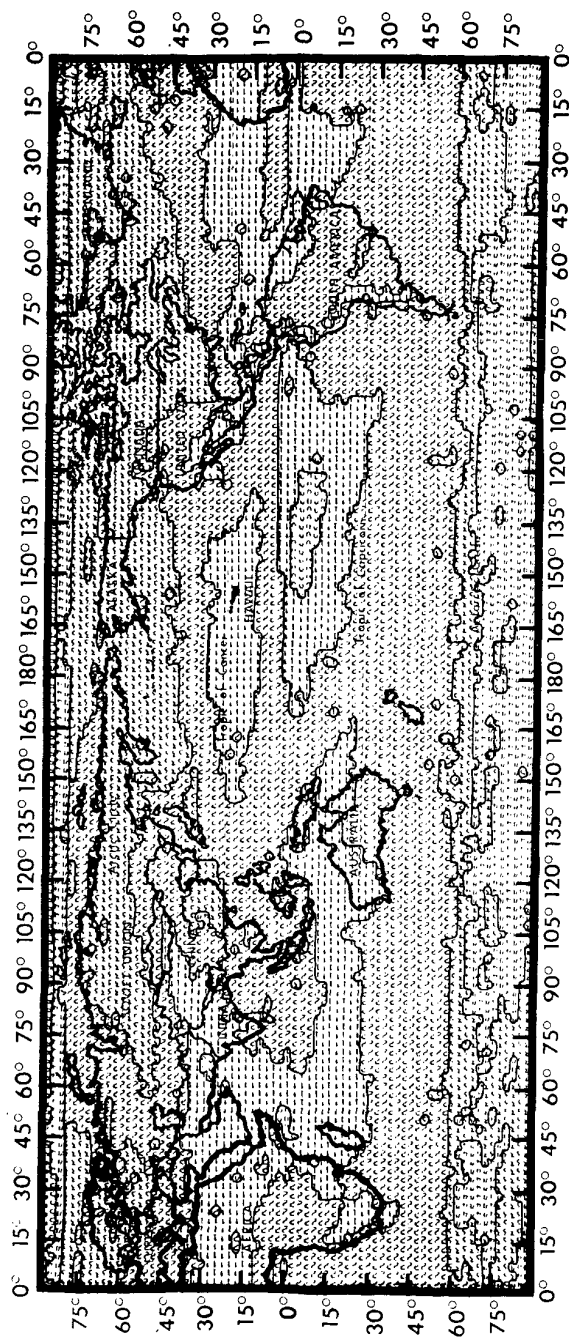


Fig. 22. Standard deviation of cloud cover
Fall (September, October, November)

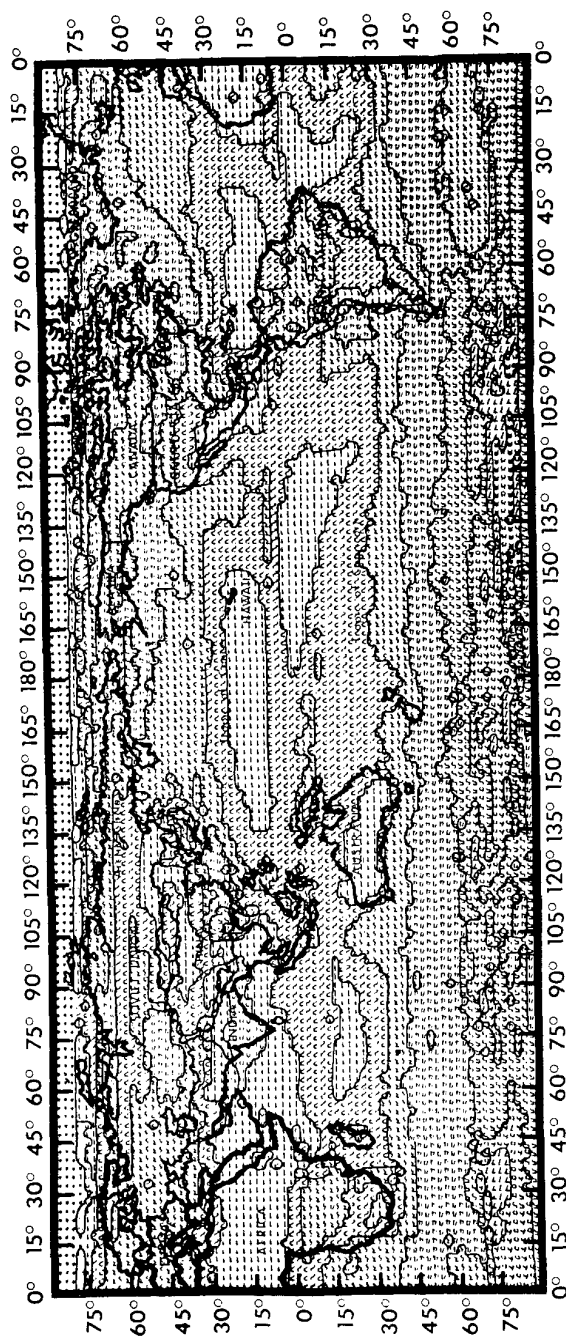


Fig. 23. Mean cloud cover
Winter (December, January, February)

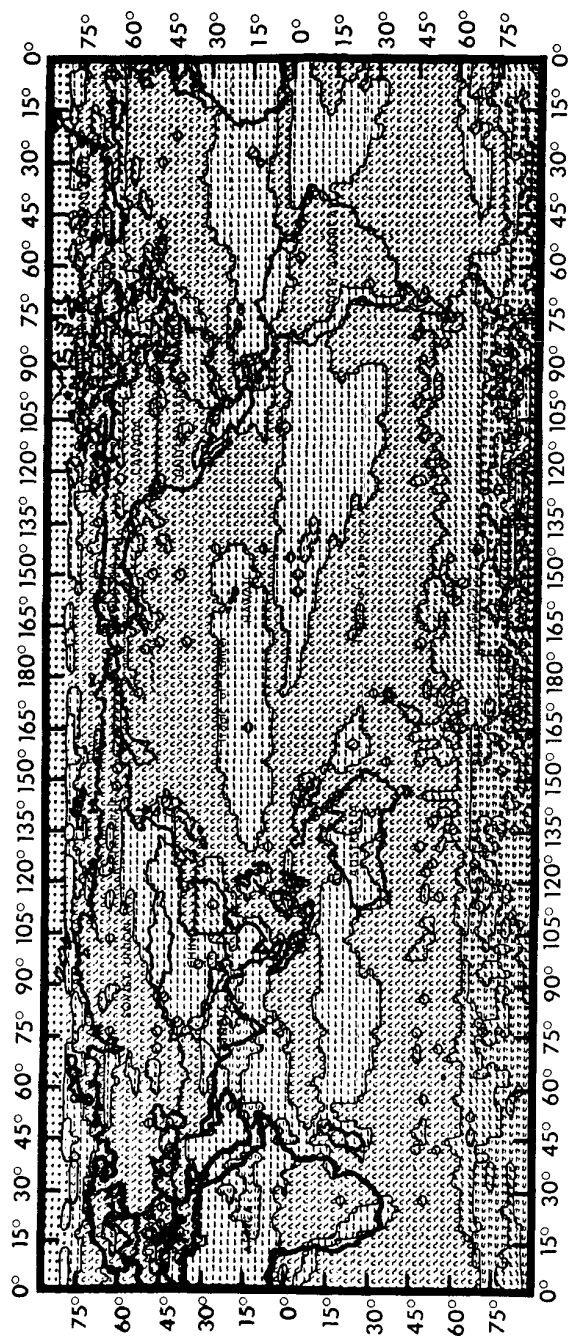


Fig. 24 . Standard deviation of cloud cover
Winter (December, January, February)

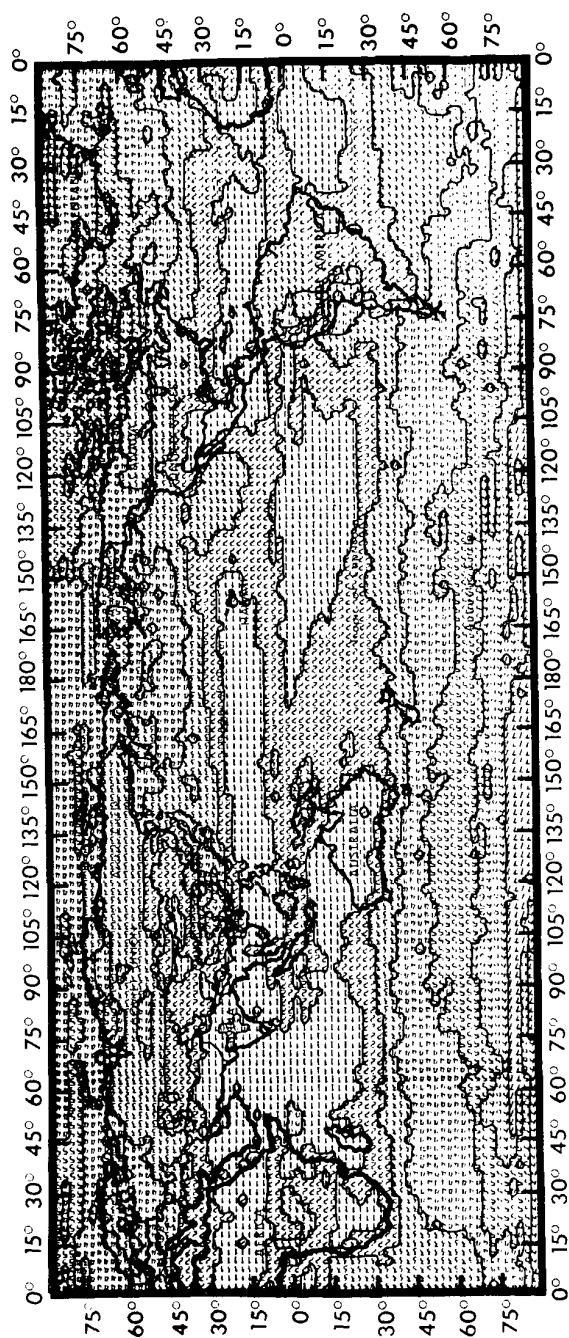


Fig. 25. Mean cloud cover
Spring (March, April, May)

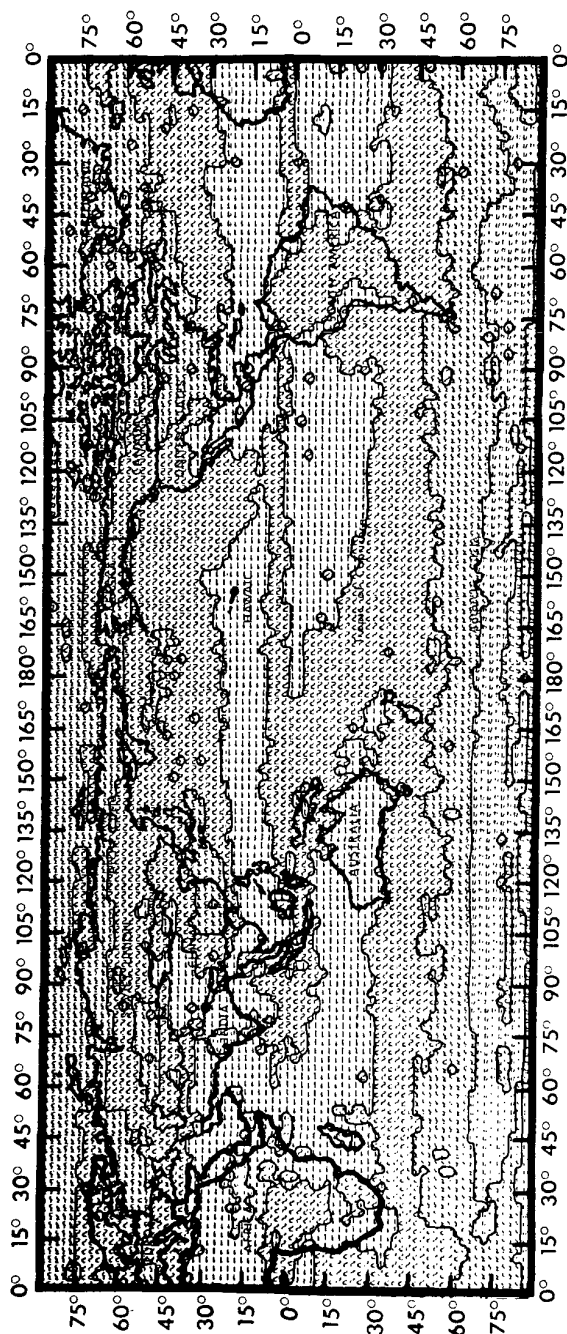


Fig. 26. Standard deviation of cloud cover
Spring (March, April, May)

contributes to the error (the degree of error depends on the temporal persistence of cloud cover and the long term cyclic behavior of cloud cover). Another factor is that the satellite measurements were taken for a given grid location at the same time each day.

It is worth remarking that even with the above-mentioned errors, it is still possible to use the data at hand to classify areas of the globe into homogeneous cloud cover regions. The methodology is sound, and the presently developed models should give a good overall picture of global cloud cover characteristics.

The above comments lead to three major recommendations. First, the derived cloud cover should be verified using some independent measurements. Second, spatial and temporal persistence of cloud cover need to be investigated. Third, the model should be updated with a longer length record.

ACKNOWLEDGEMENTS

This research was sponsored in part under Contract NAS-8-33071 National Aeronautics and Space Administration, Marshall Space Flight Center.

The authors are indebted to O.E. Smith, Marshall Space Flight Center, for introducing the problem, and for many helpful discussions and comments.

The efforts of Roy Jenne and his coworkers at NCAR in furnishing the data sets for this study are gratefully acknowledged.

The assistance of Chris Maukenon in developing many of the computer programs used in the investigation was invaluable.

The conscientious and painstaking efforts of Sarah Autrey and Debbie Waitt in many elements of the project and especially their efforts in developing the graphics have resulted in significant contributions and are greatly appreciated.

The authors are grateful to Cindy Sloan for her careful and conscientious efforts in typing the report.

REFERENCES

- Barnes, J.C., Glasen, A.H., Sherr, P.E. and Willand, J.H., 1968. Worldwide cloud cover distribution for use in computer simulations. NASA Contractor Report CR-61226.
- Falls, L.W., 1974. The beta distribution: a statistical model for world cloud cover. J. Geophys. Res. 79:1261-1264.
- Gray, Thomas I. Jr., 1978. National Environmental Satellite Service, NOAA, Washington D.C. Personal Communication.
- Greaves, J.R., Spiegler, D.B. and Willand, J.H., 1971. Development of a global cloud cover model for simulating Earth-viewing space missions. NASA Contractor Report CR-61345.
- Henderson-Sellers, A., 1978. Surface type and its effect upon cloud cover: A climatological investigation. J. Geophys. Res. 83:5057-5062.

VARIABILITY OF NORTHERN HEMISPHERE MEAN SURFACE AIR TEMPERATURE DURING RECENT TWO HUNDRED YEARS

R. YAMAMOTO

Geophy. Inst., Kyoto Univ., Kyoto (Japan)

ABSTRACT

Yamamoto, R. Variability of northern hemisphere mean surface air temperature during recent two hundred years. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1 , 1979

Variability of the hemispherical or global mean surface air temperature during recent 100 years has been estimated by several authors. Most of the works indicate no reliability of the spatial averaging, which is indispensable to confirm any time changes. In the present paper, the optimum interpolation method is applied to the northern hemisphere network of the seasonal mean surface air temperature anomaly, and the values at regularly distributed grid points are obtained, together with the errors, for each season from 1876 through 1975. The grid point values and their errors thus obtained can give the zonal and hemispherical mean as well as longitudinal profile along a latitude circle, together with their errors.

The changes of the northern hemisphere mean temperature obtained in the present paper is clearly smaller than those in the previous works, though the main warming from the 1880s and the cooling from the 1940s are found similarly to those in the previous works. Possibility of underestimate in our present analysis and overestimate in the previous works is discussed.

The network of temperature observations was very sparse before the decade of 1880s. Possibility of the application of the optimum interpolation to such a sparse network is examined. It is noticed that the field of correlation which is requisite for the optimum interpolation has an appreciable temporal change in long distance, which means that the estimation of large-scale field of the temperature changes in the period before 1880s decade should be attempted by some other method than the optimum interpolation.

1. INTRODUCTION

Interest has been increasing in the change and variability of the climate, in particular connection with increasing vulnerability of human life to the climatic conditions (WMO, 1978). We are not free from anxiety about a possibility that a drastic change of climate may occur in the near future, associating with the changes of climate-controlling factors such as increasing concentration of the carbon-dioxide in the atmosphere (Kellogg, 1977).

It is important for the scientists to study the mechanism of, and to acquire some insight into the climatic changes. Analysis of the past climatic data may be probably one of the promising approaches for the problem, in addition to numerical model experiments (e.g., Manabe and Wetherald (1975)).

The climatic element such as surface air temperature, in general, suffers from local variability, which appears to be a noise in detecting statistically the effects of global scale changes of the climate-controlling factors. It is reasonable to expect that the effects may appear appreciably in the changes of global or hemispherical mean climate, in which the local variability is smoothed out by spatial averaging.

Our main concern in the present paper is directed to variability of the northern hemisphere mean surface air temperature. Some mentions are given on the available data of the temperature in Section 2. In Section 3, some problems involved in the results of the previous workers are pointed out. The optimum interpolation procedure of data analysis, which can easily give an evaluation of the error of the estimation, is briefly described in Section 4. The change of the northern hemisphere mean temperature for recent 100 years is presented in Section 5. Section 6 is devoted to discuss the possibility of estimating the large-scale temperature change in the older period of 100 years, and some concluding remarks are given in the final section.

2. DATA OF THE SURFACE AIR TEMPERATURE

The surface air temperature is most closely related to the human life, and the data network is the most abundant among the various climatic elements. However, the air temperature has remarkable diurnal variation and is vulnerable to local topographic effects, and its spatial representativeness is less than the other climatic elements such as the barometric pressure. Therefore, special attention should be paid to the analysis of the network data, even if the time mean data is treated (Mitchell (1963)).

Fig. 1 shows the time series of annual mean surface air temperature for the 100 years (1876-1975) at Hakodate (Japan, 41°49'N, 140°45'E), and those of the temperature anomaly of zonal mean along 40°N and of the northern hemisphere mean. Estimation procedure of the last two values will be described in Section 4. Thin curves in this figure represent the long-term trends, which is determined by

$$T = \bar{T} + \sum_{k=1}^4 (A_k \cos[2\pi kt/200] + B_k \sin[2\pi kt/200])$$

where T is the annual mean temperature or its anomaly, \bar{T} the 100 year mean, and t is the year. Fourier coefficients A_k and B_k are determined by the least square method. The RMSEs of the annual mean from the long-term trend are estimated as about 0.6°C, 0.1°C and 0.06°C, respectively. This implies that the values averaged over large area is appropriate for detecting statistically the effects of global scale change of climate-controlling factors such as the carbon-dioxide concentration and increase of the stratospheric aerosols due to volcanic eruptions.

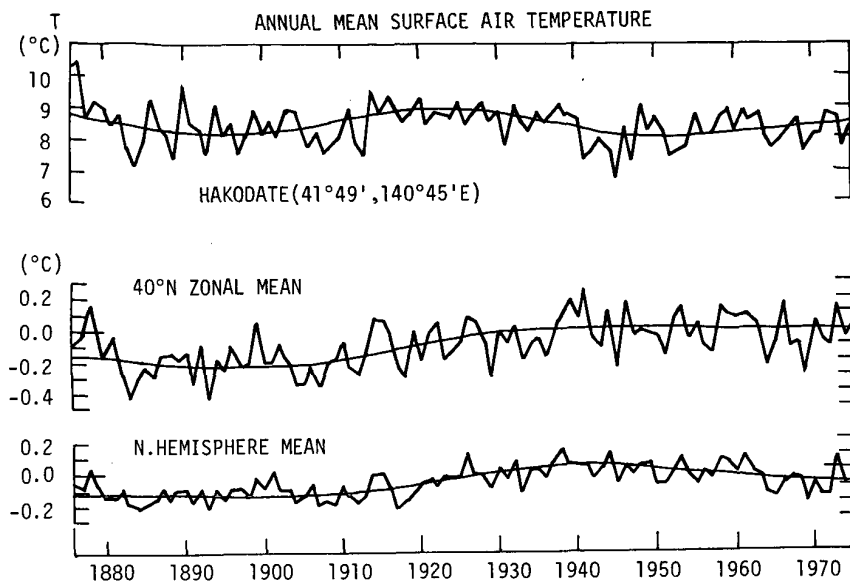


Fig. 1. Time series of annual mean surface air temperature at Hakodate ($41^{\circ}49'N$, $140^{\circ}45'E$), and the zonal mean of temperature anomaly at $40^{\circ}N$ and the northern hemispheric mean (thick curves) and their long-term trends (thin curves).

TABLE 1.

Total number of stations north of $25^{\circ}S$ with available data of monthly mean surface in the air temperature, in the NCAR archives.

	$85^{\circ}N$ $-65^{\circ}N$	$65^{\circ}N$ $-45^{\circ}N$	$45^{\circ}N$ $-25^{\circ}N$	$25^{\circ}N$ $-5^{\circ}N$	$5^{\circ}N$ $-5^{\circ}S$	$5^{\circ}S$ $-25^{\circ}S$	TOTAL
1656-1675		1					1
1676-1695		1					1
1696-1715		1					1
1716-1735		1					1
1736-1755		2	1				3
1756-1775		6	1				7
1776-1795		9	3				12
1796-1815		9	3				12
1816-1835		10	4	1			15
1836-1855		10	6	1			17
1856-1875	5	16	14	7			42
1876-1895	7	68	76	35	2	9	197
1896-1915	10	81	95	51	7	21	265
1916-1935	17	105	128	68	10	39	367

The longest series of the data of the surface air temperature measured instrumentally is found in Central England since the year of 1659 (Manley (1974)). However, the number of stations with available temperature data increased quite slowly, as shown in Table 1, which is referred to the archives of NCAR, U.S.A. Before the last

quarter of the 19th Century, almost all the stations with data are located within the north mid-latitudes, and there were few available data in the tropics until the decade of 1870. Therefore, all the previous works to estimate the northern hemisphere mean temperature were undertaken for the period after the last quarter of the 19th Century. In the present paper, two periods of 100 years or 1776-1875 and 1876-1975 are separately taken.

3. ANALYSIS OF THE NORTHERN HEMISPHERE MEAN TEMPERATURE FOR RECENT 100 YEARS

Pioneer work on the global or hemispherical mean temperature was made by Willett (1950). His analysis was based upon a presumption that the 5 year mean of the seasonal mean temperature at a station may have significant representativeness over broad area. Dividing the whole globe into 18 latitudinal bands of 10 degree width, Willett determined the mean temperature of each band by averaging the data at stations located within the band, taking precaution against the overweight in region of dense network such as Europe and North American Continents. Succeeding to the Willett method of analysis, Mitchell (1961,1963) has obtained the change of the northern hemisphere mean temperature with special attention to the homogeneity of the data. Similar analyses were made by Reitan (1974) and Brinkmann (1976) for more recent decades. Budyko (1969,1977) has calculated the change of the annual mean temperature averaged over the northern hemisphere from maps of the temperature anomaly for each month.

Inspecting these results, Lamb(1977) made the following mention : " margins of error of the estimate must be presumed greatest in the extensive ocean areas, but repetitions of the calculations in different laboratories indicate that the main warming from the 1880s and the cooling from the 1940s are beyond doubt! The results by Mitchell (1963) and Budyko(1969) are shown in Fig.2, which is reproduced from Robock (1978).

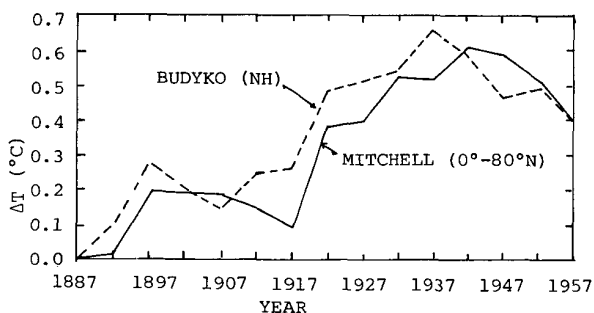


Fig. 2. The 5-year mean of surface air temperature anomaly reported by Mitchell(1963) and Budyko(1969), reproduced from Robock(1978).

Although the general tendencies of the warming from the 1880s and of the cooling from the 1940s are commonly seen in both of the results, there are some disagreements, e.g., some delay of several years in Mitchell's result. On the other hand, it is essentially necessary to assess how much errors are included in the estimated hemispherical average, for confirmation of the actual changes. These situations suggest us to re-compute the hemispherical mean surface air temperature, paying special attention to the assessment of the errors of estimation. The present author and his collaborator (Yamamoto and Hoshiai (1979a)) have adopted the optimum interpolation method, which makes possible to evaluate the error involved in the analysis rather easily.

4. OPTIMUM INTERPOLATION METHOD

Spatial mean of the temperature can be easily calculated when the data at regularly distributed grid points are given. The values and the errors at these grid points can be estimated by applying the optimum interpolation technique to the network of stations located irregularly, see Yamamoto and Hoshiai (1979,a,b). This technique, which is usually used in objective analysis of numerical weather prediction, has been developed by Gandin (1963).

Optimum interpolation technique gives a value of deviation T'_g from the time mean \bar{T}_g (hereafter, $\bar{\cdot}$ designates the time mean, i.e., the expectation w.r.t.time) at an arbitrary grid point by a linear combination of the observed data \hat{T}'_i at n stations, $i = 1, 2, \dots, n$, within the range of appreciably positive correlation with the grid point g :

$$T'_g = \sum_{i=1}^n \hat{T}'_i P_i + I_g \quad (1)$$

where P_i is the weighting factor and I_g the interpolation error. Employment of the anomaly data diminishes the influence of the difference of the station altitude. The observed deviation \hat{T}'_i consists of the true deviation T'_i and the observational error ϵ_i , the latter of which includes the effects of local irregularities. ϵ_i and T'^2_i are assumed to be homogeneous within the correlated range, and will be denoted by ϵ^2 and σ^2 , respectively. The value of ϵ^2 can be estimated with the aids of the structure functions, as shown later.

Under the condition that the value of \bar{I}_g^2 should be minimum, the following equations which determine the P_i are derived:

$$\sum_{j=1}^n \mu_{ij}^1 P_j + \lambda^2 P_i = \mu_g^1, \quad i = 1, 2, \dots, n, \quad (2)$$

where $\lambda^2 \equiv \epsilon^2/\sigma^2$, and μ_j^1 and μ_g^1 are the correlation coefficients between the i -th and j -th stations, and the i -th station and the grid point g , respectively.

Then, the value of I_g^2 is reduced to E_g^2 :

$$E_g^2 = \sigma^2 (1.0 - \sum_{i=1}^n P_i \mu_g^i) . \quad (3)$$

The details of this method can be found in the textbook by Gandin (1963).

In practice of the optimum interpolation, it is prerequisite to know the dependencies of the structure function and of the correlation function upon the distance between a pair of stations. The former is used to estimate the observational error.

Under the assumptions that the observational error ϵ_i is random and independent of the true deviation T_i' and ϵ_j ($j \neq i$), the structure function B_j^i between the i -th and the j -th stations is expressed as follows:

$$\begin{aligned} B_j^i &= \overline{(T_j' - T_i')^2} = \overline{[(T_j' - T_i') + (\epsilon_j - \epsilon_i)]^2} \\ &= \overline{(T_j' - T_i')^2} + \overline{\epsilon_j^2 + \epsilon_i^2} = \overline{(T_j' - T_i')^2} + 2\epsilon^2 . \end{aligned} \quad (4)$$

Then, we have

$$\lim_{i \rightarrow j} B_j^i = 2\epsilon^2 . \quad (5)$$

An example of distance dependency of the structure function for the annual mean temperature is given in Fig.3, which is referred to the data sample of 60 years at stations in western Europe network. Extrapolation to zero distance gives the observational error of about 0.13°C .

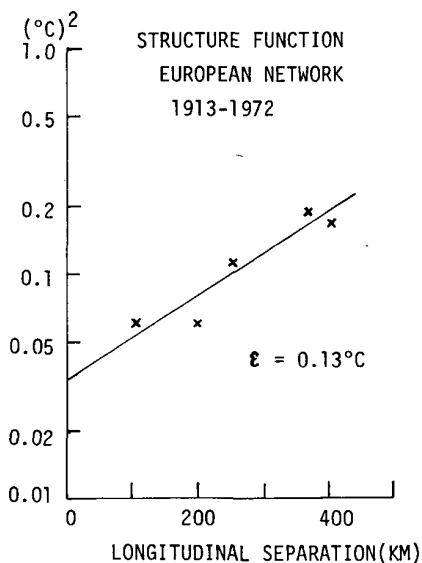


Fig.3. Dependency of the structure function of the annual mean temperature upon longitudinal separation. Data are at stations in European network for 60 years (1913-1972). Extrapolation to zero separation gives the value of $2\epsilon^2$.

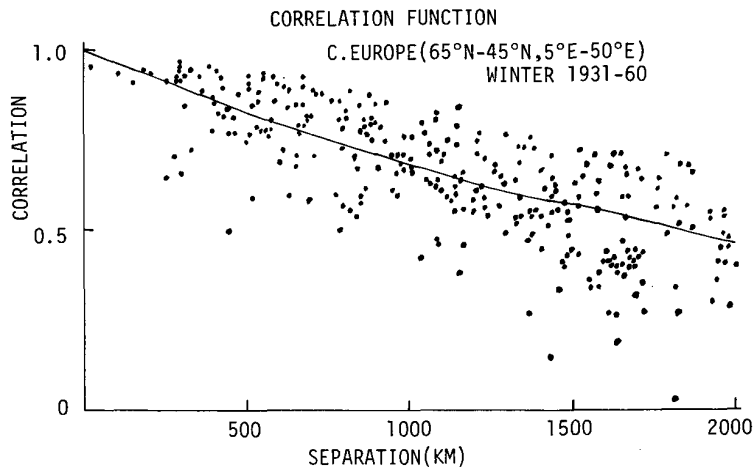


Fig. 4. Correlation coefficients of 3-month mean temperature in central Europe, computed from 30 year data (1931-1960).

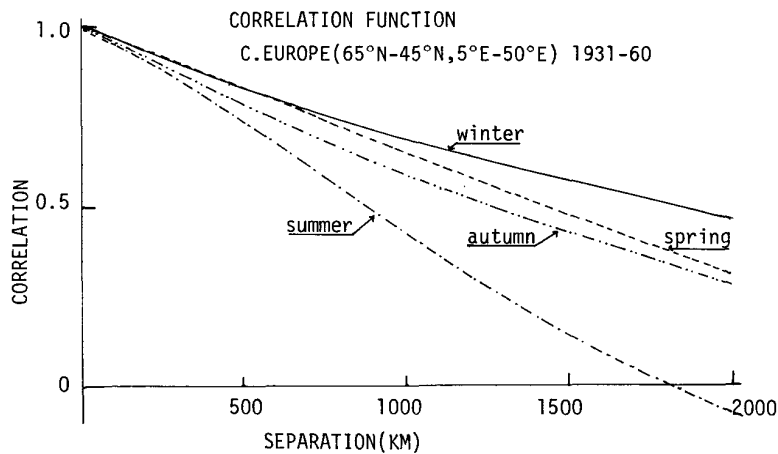


Fig. 5. Regression curves of correlation functions of 3-month mean temperature for 30 year period (1931-1960) in central Europe.

The correlation function is, in general, dependent on the direction and the distance between the pair of stations. Here, isotropy of the correlation is assumed, although the directional dependency should be taken into account in more advanced analysis. Concerning the dependency of the function upon the distance, we obtain a regression curve from a set of μ_j^i . An example of the correlation of 3-month mean

temperature is given in Fig.4, and the regression curves for each season in Fig.5. At distance of 1000 km, the correlation has a value of 0.67 in winter and of 0.43 in summer. The values μ_g^i in eqns. (2) and (3) is determined from the regression curves.

The values of σ^2 and μ_g^i are estimated from the data sample of 30 years (1931-1960). The 95% confidence intervals for these quantities are determined, and the upper limit for σ^2 and the lower limit for μ_g^i are taken in estimation of the interpolation error E_g by the eqn. (3). Because the distribution of T' may be assumed to be of Gaussian type, the value of E_g thus obtained is roughly equal to the 68% confidence limit for the error. If no available data exists within the range of appreciably positive correlation with the grid point concerned, we have $E_g^2 = \sigma^2$, and the value T'_g falls in the range of $\pm\sigma$ with a probability of about 68%.

The observed data at stations used for interpolation of one grid point value are not utilized for other grid point values. This limitation on data employment keeps the interpolation error of one grid point independent of the other, and simplifies the computation of spatial mean error estimate. The zonal mean value of the temperature deviation and its estimation error can be easily calculated from these interpolated values at the grid points of equal spacing along the latitude circle. Thus, the 68% confidence limit for the error of zonal mean is estimated as $\{\sum_{g=1}^G E_g^2 / G\}^{1/2}$, where G is the number of grid points used for the zonal mean calculation. Latitudinal averaging of the zonal mean gives the mean over the hemisphere or latitudinal belts, in which case the error becomes generally smaller than that of the zonal mean. In a similar way, time averaging also diminishes the error.

5. RE-COMPUTATION OF THE CHANGE OF THE NORTHERN HEMISPHERE MEAN TEMPERATURE DURING RECENT 100 YEARS

We have attempted to re-compute the change of the northern hemisphere mean temperature during recent 100 years from 1876 through 1975, using the observed data at 367 stations north of 25°S, the locations of which are shown in Fig.6 (Yamamoto and Hoshiai (1979b)). The optimum interpolation technique is applied to the network data of the seasonal mean (3-month mean) temperature deviation from the 30 year mean (1931-1960) of each seasonal mean temperature. The temperature deviation and the interpolation error are estimated at grid points on 10° latitude and 30° longitude (45° longitude only at 80°N) intersections from the equator to 80°N.

These grid point values give the hemispherical distribution of the temperature anomaly for each season from 1876 to 1975. The detailed descriptions of the results will be found in Hoshiai's paper (1980). Some mentions on the errors are given here. Magnitude of the interpolation error depends upon the latitude, season and the number of stations with available data. The interpolation error at the grid point of 70°N, 0° Long. is, for example, 2.05°C and 1.64°C in 1880 and 1970 winter,

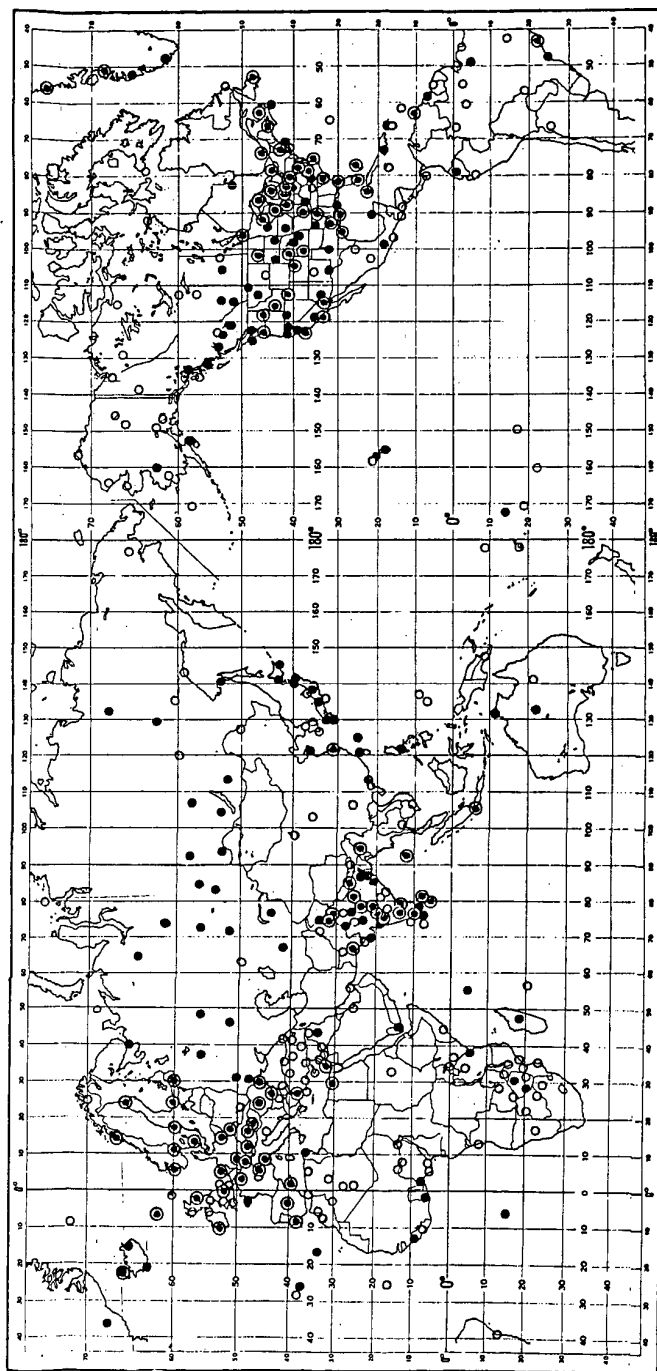


Fig. 6. Locations of the 367 stations adopted for the re-computation of the change of the northern hemisphere mean temperature during recent 100 years. Encircled dot signifies the stations with data before 1876, dot the stations with data before 1900, and open circle the stations without data in 1901.

and 1.25°C and 1.04°C in 1880 and 1970 summer, respectively. The errors at $0^{\circ}\text{Lat. } 180^{\circ}\text{Long.}$ is 0.56°C in winter and 0.48°C in summer, irrespective of the year.

The zonal mean of the seasonal mean temperature anomaly and its error are calculated from the values of 12 grid points (8 grid points along 80°N only). Latitudinal averaging of the zonal mean values with areal weighting gives the latitudinal band and the hemispherical ones. The 5-year running means averaged over the latitudinal bands of 30° width are shown in Fig.7. For the bands of $90^{\circ}\text{N}-60^{\circ}\text{N}$, $60^{\circ}\text{N}-30^{\circ}\text{N}$ and $30^{\circ}\text{N}-\text{the equator}$, the errors are $0.11-0.12^{\circ}\text{C}$, $0.04-0.05^{\circ}\text{C}$ and 0.03°C , respectively. In polar band ($90^{\circ}\text{N}-60^{\circ}\text{N}$), the minimum temperature in the 1910s decade and the maximum in the 1930s are significant, and the warming from the 1910s to the 1940s has the rate of about $0.6^{\circ}\text{C}/20$ years. In mid-latitude ($60^{\circ}\text{N}-30^{\circ}\text{N}$), the minimum temperature in the 1880s and the maximum in the 1940s are found. Although the changes in tropics ($30^{\circ}\text{N}-\text{the equator}$) are small, the maximum appears appreciably in the 1960s decade.

Table 2a and 2b show the northern hemisphere mean temperature anomaly from the 30 year mean (1931-1960) and the 68% confidence limit for the error, for each season of the 100 years (1876-1975).

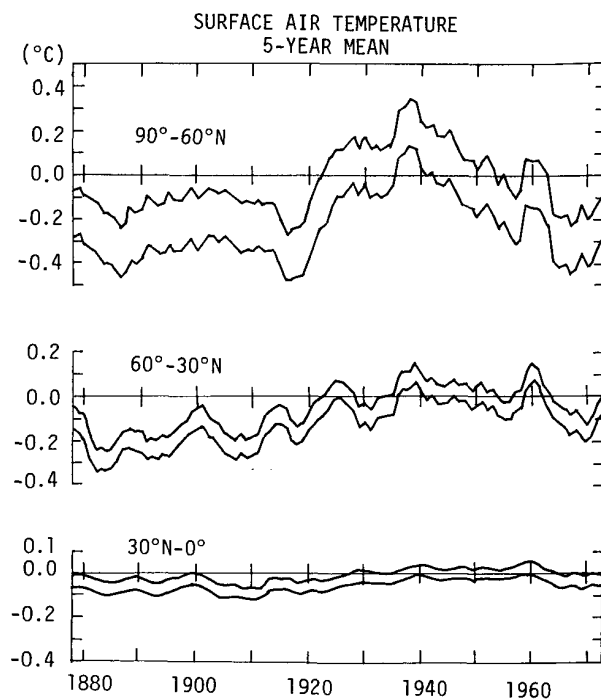


Fig.7. The 5-year running mean surface air temperature anomaly averaged over 30° width latitude band, expressed by the 68% confidence interval.

TABLE 2a.

Seasonal mean of the surface air temperature anomaly from the 30 year mean (1931-1960) averaged over the northern hemisphere, in the unit of °C.

YEAR	WN	SP	SM	AT	YEAR	WN	SP	SM	AT
1876	0.00	-0.11	-0.02	-0.16	1926	0.24	0.10	0.00	0.08
1877	-0.17	-0.13	-0.02	-0.10	1927	-0.07	-0.13	-0.01	0.08
1878	0.04	0.06	-0.01	0.00	1928	0.06	-0.06	-0.07	0.01
1879	-0.08	-0.03	-0.08	-0.14	1929	-0.26	-0.07	-0.02	-0.04
1880	-0.20	-0.12	-0.08	-0.22	1930	-0.02	0.04	0.03	-0.00
1881	-0.20	-0.08	-0.11	-0.22	1931	-0.18	-0.03	0.07	-0.02
1882	-0.03	-0.09	-0.13	-0.26	1932	0.16	0.06	0.02	-0.05
1883	-0.46	-0.16	-0.08	-0.12	1933	-0.22	-0.20	-0.03	-0.12
1884	-0.17	-0.29	-0.18	-0.24	1934	-0.08	-0.01	-0.01	0.07
1885	-0.30	-0.14	-0.15	-0.20	1935	0.19	-0.06	-0.00	-0.10
1886	-0.23	-0.14	-0.10	-0.21	1936	-0.23	-0.01	0.06	0.00
1887	-0.16	-0.08	-0.13	-0.12	1937	0.17	-0.06	0.07	0.16
1888	-0.30	-0.14	-0.09	-0.12	1938	0.03	0.23	0.06	0.20
1889	-0.21	-0.02	-0.06	-0.18	1939	0.09	0.03	0.03	0.04
1890	-0.11	-0.09	-0.14	-0.09	1940	0.12	0.12	-0.02	0.01
1891	-0.29	-0.15	-0.11	-0.15	1941	0.17	0.05	0.04	-0.04
1892	-0.07	-0.18	-0.11	-0.13	1942	-0.13	-0.03	-0.04	0.02
1893	-0.64	-0.06	-0.09	-0.14	1943	-0.03	-0.02	-0.05	0.10
1894	-0.07	-0.09	-0.11	-0.18	1944	0.33	0.06	-0.03	0.06
1895	-0.37	-0.11	-0.11	-0.09	1945	-0.22	0.01	-0.05	-0.05
1896	-0.13	-0.20	-0.02	-0.12	1946	0.02	0.08	-0.03	0.01
1897	-0.19	-0.06	-0.04	-0.09	1947	-0.23	0.12	-0.01	0.09
1898	-0.08	-0.33	-0.06	-0.14	1948	0.07	0.05	0.00	-0.01
1899	0.03	-0.09	-0.10	0.03	1949	0.15	0.02	-0.05	0.07
1900	-0.30	0.00	-0.01	-0.04	1950	-0.21	0.02	-0.08	-0.07
1901	-0.03	0.04	0.01	-0.08	1951	-0.25	-0.01	-0.03	0.01
1902	0.01	-0.14	-0.16	-0.20	1952	0.18	-0.09	0.01	-0.12
1903	-0.03	-0.10	-0.20	-0.10	1953	0.05	0.17	0.10	0.04
1904	-0.22	-0.21	-0.17	-0.10	1954	-0.09	-0.10	0.00	0.09
1905	-0.22	-0.19	-0.11	-0.06	1955	-0.03	-0.17	-0.02	-0.08
1906	-0.18	-0.00	-0.04	-0.12	1956	-0.33	-0.20	-0.14	-0.24
1907	-0.19	-0.22	-0.24	-0.16	1957	-0.08	-0.05	0.04	-0.01
1908	-0.14	-0.15	-0.15	-0.25	1958	0.23	0.05	-0.02	0.05
1909	-0.31	-0.26	-0.10	-0.08	1959	0.18	0.13	0.03	-0.09
1910	-0.11	-0.04	-0.12	-0.19	1960	0.13	-0.14	0.05	-0.05
1911	-0.27	-0.17	-0.09	-0.09	1961	0.32	0.02	0.04	-0.03
1912	-0.08	-0.12	-0.20	-0.35	1962	0.13	0.08	-0.04	-0.09
1913	-0.19	-0.14	-0.15	-0.07	1963	-0.07	-0.07	-0.04	0.11
1914	0.20	-0.06	-0.11	-0.11	1964	-0.13	-0.15	-0.10	-0.17
1915	0.02	0.04	-0.04	0.00	1965	-0.21	-0.11	-0.18	-0.13
1916	0.08	-0.15	-0.12	-0.11	1966	-0.14	-0.10	-0.01	-0.10
1917	-0.42	-0.31	-0.09	-0.14	1967	-0.24	0.09	-0.07	0.01
1918	-0.41	-0.17	-0.15	-0.03	1968	-0.15	0.26	-0.14	-0.18
1919	-0.21	-0.13	-0.10	-0.13	1969	-0.52	-0.16	-0.07	-0.08
1920	-0.11	0.07	-0.04	-0.20	1970	-0.01	-0.08	-0.06	-0.19
1921	-0.08	0.08	-0.02	-0.18	1971	-0.20	-0.23	-0.11	-0.04
1922	-0.15	0.04	-0.08	-0.15	1972	-0.21	-0.05	-0.10	-0.23
1923	-0.07	-0.12	-0.09	0.06	1973	0.17	0.12	0.01	-0.01
1924	-0.07	-0.08	-0.02	-0.01	1974	-0.12	0.01	-0.07	-0.12
1925	-0.04	-0.03	-0.04	0.00	1975	0.07	0.13	-0.02	-0.05

WN - Winter (December, January, February), SP - Spring (March, April, May)
 SM - Summer (June, July, August), AT - Autumn (September, October, November)

TABLE 2b.

The 68% confidence limit for the error in the estimation of the northern hemisphere mean temperature anomaly given in Table 2a, in the unit of °C.

YEAR	WN	SP	SM	AT	YEAR				
1876	0.19	0.13	0.09	0.12	1926	0.15	0.12	0.09	0.11
1877	0.17	0.13	0.09	0.12	1927	0.15	0.12	0.09	0.11
1878	0.17	0.13	0.09	0.12	1928	0.15	0.12	0.09	0.11
1879	0.17	0.13	0.09	0.12	1929	0.15	0.12	0.09	0.11
1880	0.17	0.13	0.09	0.12	1930	0.15	0.12	0.09	0.11
1881	0.16	0.12	0.09	0.12	1931	0.14	0.12	0.09	0.11
1882	0.16	0.12	0.09	0.12	1932	0.14	0.12	0.09	0.11
1883	0.16	0.12	0.09	0.12	1933	0.14	0.12	0.09	0.11
1884	0.16	0.12	0.09	0.12	1934	0.14	0.12	0.09	0.11
1885	0.16	0.12	0.09	0.12	1935	0.14	0.12	0.09	0.11
1886	0.16	0.12	0.09	0.12	1936	0.14	0.12	0.09	0.11
1887	0.16	0.12	0.09	0.12	1937	0.14	0.12	0.09	0.11
1888	0.16	0.12	0.09	0.12	1938	0.14	0.12	0.09	0.11
1889	0.16	0.12	0.09	0.12	1939	0.14	0.12	0.09	0.11
1890	0.16	0.12	0.09	0.12	1940	0.14	0.12	0.09	0.11
1891	0.16	0.12	0.09	0.12	1941	0.14	0.12	0.09	0.11
1892	0.16	0.12	0.09	0.12	1942	0.14	0.12	0.09	0.11
1893	0.16	0.12	0.09	0.12	1943	0.14	0.12	0.09	0.11
1894	0.15	0.12	0.09	0.12	1944	0.14	0.12	0.09	0.11
1895	0.15	0.12	0.09	0.12	1945	0.14	0.12	0.09	0.11
1896	0.16	0.12	0.09	0.12	1946	0.14	0.12	0.09	0.11
1897	0.15	0.12	0.09	0.12	1947	0.14	0.12	0.09	0.11
1898	0.15	0.12	0.09	0.12	1948	0.14	0.12	0.09	0.11
1899	0.15	0.12	0.09	0.12	1949	0.14	0.12	0.09	0.11
1900	0.15	0.12	0.09	0.12	1950	0.14	0.12	0.09	0.11
1901	0.15	0.12	0.09	0.12	1951	0.14	0.12	0.09	0.11
1902	0.15	0.12	0.09	0.12	1952	0.14	0.12	0.09	0.11
1903	0.15	0.12	0.09	0.12	1953	0.14	0.12	0.09	0.11
1904	0.15	0.12	0.09	0.12	1954	0.14	0.12	0.09	0.11
1905	0.15	0.12	0.09	0.12	1955	0.14	0.12	0.09	0.11
1906	0.15	0.12	0.09	0.12	1956	0.14	0.12	0.09	0.11
1907	0.15	0.12	0.09	0.12	1957	0.14	0.12	0.09	0.11
1908	0.15	0.12	0.09	0.12	1958	0.14	0.12	0.09	0.11
1909	0.15	0.12	0.09	0.12	1959	0.14	0.12	0.09	0.11
1910	0.15	0.12	0.09	0.12	1960	0.14	0.12	0.09	0.11
1911	0.15	0.12	0.09	0.12	1961	0.14	0.12	0.09	0.11
1912	0.15	0.12	0.09	0.12	1962	0.15	0.12	0.09	0.11
1913	0.15	0.12	0.09	0.12	1963	0.15	0.12	0.09	0.11
1914	0.15	0.12	0.09	0.12	1964	0.15	0.12	0.09	0.11
1915	0.15	0.12	0.09	0.12	1965	0.15	0.12	0.09	0.11
1916	0.15	0.12	0.09	0.12	1966	0.15	0.12	0.09	0.12
1917	0.15	0.12	0.09	0.12	1967	0.15	0.12	0.09	0.12
1918	0.15	0.12	0.09	0.12	1968	0.15	0.12	0.09	0.12
1919	0.15	0.12	0.09	0.12	1969	0.15	0.12	0.09	0.12
1920	0.15	0.12	0.09	0.12	1970	0.15	0.12	0.09	0.11
1921	0.15	0.12	0.09	0.12	1971	0.15	0.12	0.09	0.11
1922	0.15	0.12	0.09	0.12	1972	0.15	0.12	0.09	0.11
1923	0.15	0.12	0.09	0.12	1973	0.15	0.12	0.09	0.11
1924	0.15	0.12	0.09	0.12	1974	0.15	0.12	0.09	0.11
1925	0.15	0.12	0.09	0.12	1975	0.15	0.12	0.09	0.11

The computed change of 5-year running mean of the northern hemisphere mean temperature deviation is given in Fig.8, which is shown with the 68% confidence intervals of the error (0.03°C). General tendencies of the warming from the 1880s and of the cooling from the 1940s are found, similarly to those of Mitchell's and Budyko's results. However, the range of the temperature change from the 1880s minimum to the 1940s maximum in the present work is at most 0.3°C and is clearly less than the ones by Mitchell and Budyko shown in Fig.2. (Fig.8 is on next page.)

Fig. 9 shows the 30 year running means of the northern hemisphere mean temperature in the upper panel, and the standard deviation computed from the data sample of 30 years, in the lower panel. The general tendencies of the warming from the 1890s and the cooling from the 1940s are more clear in the both seasons. The range from the

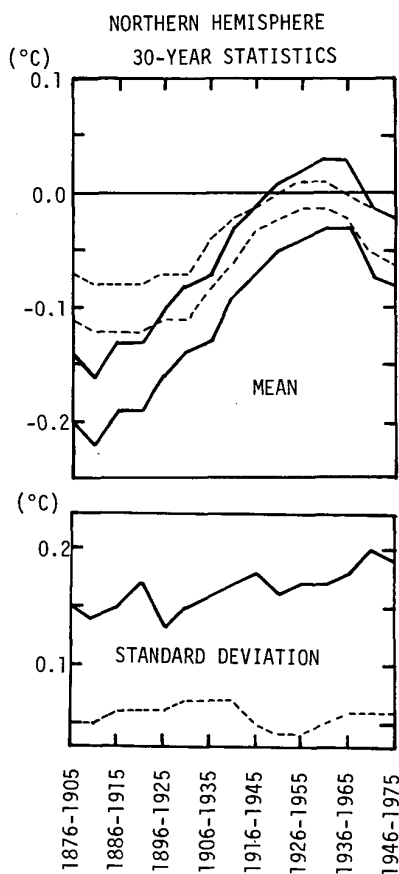


Fig. 9. Mean and standard deviation of the northern hemisphere mean surface air temperature in winter (solid curves) and in summer (broken curves), computed from data sample of 30 year.

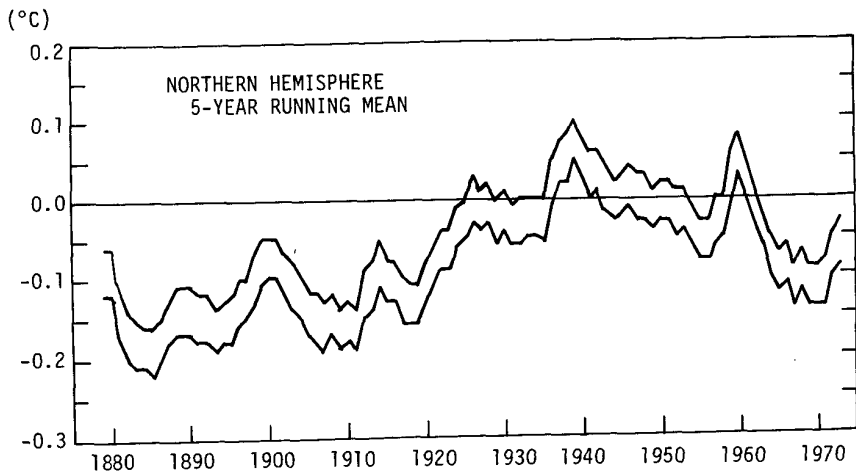


Fig. 8. The 5-year running mean of the northern hemisphere mean surface air temperature anomaly from 30-year mean (1931-1960), expressed by the 68% confidence intervals.

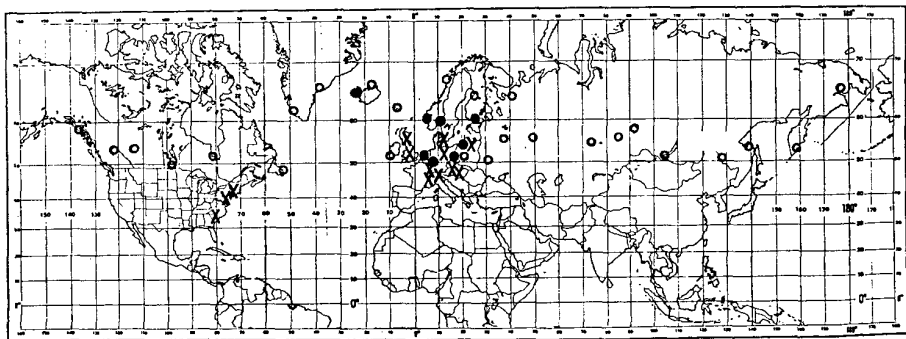


Fig. 10. Locations of stations taken for the temperature estimation in 200 year period. The 13 stations with data before 1800 (group A) are indicated by crosses. The 8 stations with data before 1851 Group B) are shown by dots, and the 26 stations with data before 1876 (group C) are shown by open circles, all of which are located between 47.5°N and 67.5°N.

maximum to the minimum for winter is at least 0.3°C and evidently smaller than that by Mitchell's and Budyko's ones. It is noticeable for the standard deviation of the winter temperature to have a tendency of gradual increase.

The optimum interpolation technique involves a possibility of underestimation of spatial mean. This interpolation technique gives a value of zero deviation to a grid point which has no observed data within the correlated range. Such situation occurs in region of sparse network, particularly over the Pacific and Atlantic Oceans and Arctic region. An underestimation of variability of the spatial average should be brought about in such cases. Such inevitable underestimation in sparse network region may perhaps be avoided by use of normalized weighting factors instead of P_i determined by eqn (2). Gandin (1963) shows that the errors in optimum interpolation with normalized weighting factors are larger than that with unnormalized factors adopted here.

On the other hand, we can not overlook a possibility that the change of the northern hemisphere mean temperature be probably overestimated in Mitchell's results. According to a recent analysis by Barnett (1978), the temperature variability over the land is 2-6 times larger than that over the oceans. In Willett-Mitchells' procedure, no particular attention is paid to the oceanic regions, and overestimation should be more or less made. The detailed discussions on the reliability in estimation of hemispherical mean temperature will be found in a paper by one of the authors (Hoshiai(1980)).

6. TEMPERATURE CHANGES IN THE PERIOD FROM THE LAST QUARTER OF THE 18TH CENTURY TO THAT OF THE 19TH CENTURY

The period from the last quarter of the 18th century to that of the 19th century is much more interesting than the period of recent 100 years treated in the previous sections, because of the following reasons: In this period, the last phase of the Little Ice Age showed the significant features. Some remarkable changes of temperature should appear due to several eruptions of volcanoes with severity equal to or greater than that of Krakatoa eruption in 1883 (Lamb(1970)). And the year of 1816 was abnormally cool, memorized as the year without summer (Hughes(1979)). These situations make us to examine the possibility of the optimum interpolation to very sparse network.

The number of stations with available data of instrumentally observed temperature before 1800 is only 13 in the NCAR archives of data, and the locations are confined in western Europe and north-eastern United States, as shown in Fig.10. First, the behaviours of the correlation functions are examined. The 8 stations with data before 1851 (B group), all of which are located between 47.5°N and 67.5°N , are supplemented, in addition to the 13 stations with data before 1800 (A group).

The correlation coefficients of the annual mean temperature between pairs of

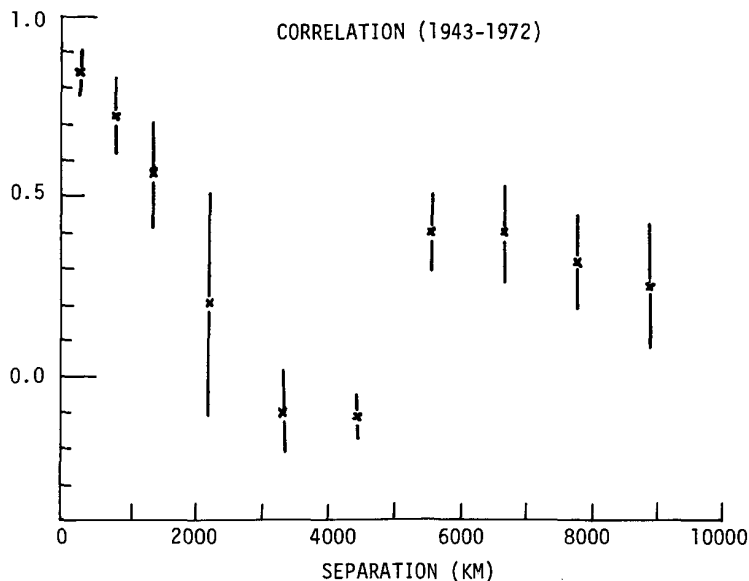


Fig. 11. Correlation of the annual mean temperature between pair of stations shown in Fig.10. These are computed from the 30 year (1943-1972) data and averaged over the separation range of each 1100 km (550 km in small range only). The standard deviations are shown by vertical bars.

computed from the data sample of 30 year (1943-1972) are given in Fig.11. Dependency of the correlation on the separation between stations is seen clearly, and the correlation has, for short separation, a value greater than 0.5. Such comparatively large values of correlation are taken in the analysis for recent 100 years described in the previous sections. Although the correlations do almost vanish at distance from 3000 to 4500 km, the correlations are appreciably positive at more distant separation. These positive correlations at distance of about 5500-8000 km make perhaps possible to apply fruitfully the optimum interpolation to very sparse network. If the fields of correlation would have no temporal change from decade to decade, use of the weighting factors P_1 which are determined from the recent data sample, and of the temperature anomaly data at stations of group A in the 18th Century gives the value at stations of group B and C in the 18th Century.

Fig. 12 shows how much the correlation coefficients change from one 30-year period to other, for two ranges of separation. There is no remarkable change for small distance, and this justifies the treatment in the previous sections, where the correlation functions determined from the 30-year data sample of 1931-1960 are utilized for the 100 year of 1876-1975.

A remarkable dip of the correlation in the period of 1853-1882 is found for large distance. This implies shortening of characteristic scale of annual mean temperature

distribution, which may probably be associated with change of the atmospheric circulations. This shows that the use of correlation functions determined from the data in different period of years is appropriate for long distance, in some cases. Also, the estimation of the field of the temperature changes in the period from the last quarter of the 18th Century to that of the 19th Century should be attempted by other method rather than the optimum interpolation method.

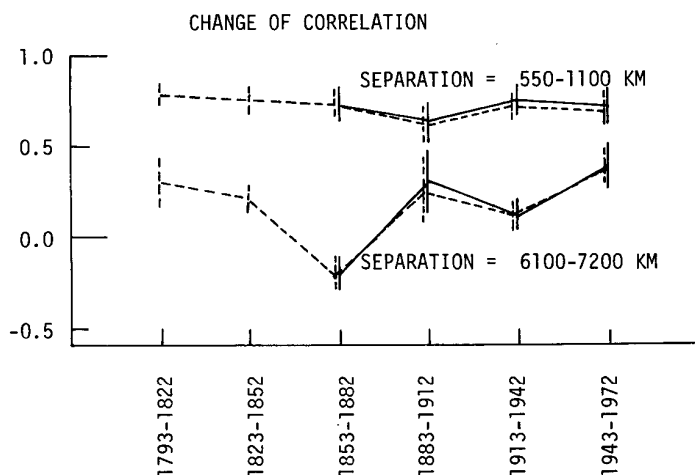


Fig. 12. Temporal change of the correlation of the annual mean temperature, computed from the 30 year data of the indicated period. Broken curves are obtained from the data at stations of group A, and the solid curves from those of groups A and B.

7. CONCLUDING REMARKS

Change of the northern hemisphere mean temperature is estimated by applying the optimum interpolation technique, assessing the errors of estimation. The technique of optimum interpolation has certainly a possibility of underestimating the spatial average. However, the magnitude of the underestimation may be possibly equal to or less than the error estimated here. Some discussions are given concerning the quantitative disagreement among the results by other authors and ours.

It is possible to apply the optimum interpolation technique to other climatic elements, with no essential modification. In contrast to the temperature which is vulnerable to diurnal variation and local effects, the barometric pressure is, in general, insensitive to such effects and has good representativeness over broad area.

The application of the technique to the hemispherical pressure field will probably produce fruitful results in the past 200 years.

ACKNOWLEDGEMENTS

The work presented here was supported financially by the scientific fund from the Ministry of Education of Japan. Most of the results were obtained in collaboration with Mr.M.Hoshiai of Aichigakuin University, and the computations were performed by using the computer at the Data Processing Center of Aichigakuin University. The author's thanks are also due to Miss.S.Tsubata for her nice arrangements of the manuscript.

REFERENCES

- Barnett,T.P.,1978. Estimating variability of surface air temperature in the northern hemisphere. *Mon.Wea.Rev.* 106:1353-1367.
- Brinkmann,W.A.R., 1976. Surface temperature trend for the northern hemisphere-updated. *Quat.Res.* 6:355-358.
- Budyko,M.I., 1969. The effect of solar radiation variations on the climate of the earth. *Tellus* 21:611-619.
- Budyko,M.I., 1977. On present-day climatic changes. *Tellus* 29:193-204.
- Gandin,L.S., 1963. Objective Analysis of Meteorological fields. Israel Program for Scientific Translations, Jerusalem.
- Hoshiai,M., 1980. To be published.
- Hughes,P., 1979. The year without a summer. *Weatherwise* 32:108-111.
- Kellogg,W.W., 1977. Effects of human activities on global climate. WMO Tech.Note 156, 47pp.
- Lamb,H.H.,1970. Volcanic dust in the atmosphere: with a chronology and assessment of its meteorological significance. *Phil.Trans.Roy.Soc.,London* 266:425-433.
- Lamb,H.H.,1972. Climate - Present, Past and Future. Vol. 1 ; 1977. Vol. 2., Methuen, London.
- Manabe,S. and Wetherald,R.T., 1975. The effects of doubling the CO₂ concentration on the climate of the general circulation model. *J.Atmos.Sci.* 32:3-15.
- Manley,G., 1974. Central England temperatures:monthly means 1659 to 1973. *Quart. J.Roy.Meteor.Soc.* 100:389-405.
- Mitchell,J.M.,Jr., 1961. Recent secular changes of global temperature. *Ann.New York Acad.Sci.* 95:235-250.
- Mitchell,J.M.,Jr., 1963. On the world wide pattern of secular temperature change. *Changes of Climate. Arid Zone Research* 20:161-181. UNESCO,Paris.
- Mitchell,J.M.,Jr., 1975. A reassessment of atmospheric pollution as a cause of long-term changes of global temperature. In: Singer,S.F.et al. (eds.) *The Changing Global Environment*. Dordrecht.
- Reitan,C.H., 1974. A climatic model of solar radiation and temperature change. *Quart. Res.* 4:25-38.
- Robock,A., 1978. Internally and externally caused climatic change. *J.Atmos.Sci.* 35: 1111-1122.
- Willett,H.C., 1950. Temperature trends of the past century. *Centenary Proc.Roy.Meteor. Soc.* 195-206.
- WMO, 1979. Declaration of the World Climate Conference. WMO Bull. 28:124-126.
- Yamamoto,R. and Hoshiai,M., 1979a. Recent change of the northern hemisphere mean surface air temperature estimated by optimum interpolation. *Mon.Wea.Rev.* 107.
- Yamamoto,R. and Hoshiai,M.,1979b. Fluctuations of northern hemisphere mean surface air temperature during recent 100 years, estimated by optimum interpolation. Presented at IAMAP Symp.on Climatic Fluctua. of the Past Century from Atmos. and Oceanic Observ., IUGG General Assembly, Canberra, Dec.1979.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

THE FOUR-YEAR CYCLE IN ATMOSPHERIC AND SOLAR PHENOMENA

K. TAKAHASHI

Resou. Counc. of Scie. and Techn. Agency, Tokyo (Japan)

ABSTRACT

Takahashi, K., The four-year cycle in atmospheric and solar phenomena. Proc. 1-st Intren. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

The 4-year cycle of atmospheric oscillation is analyzed by means of harmonic analysis. The result indicates little variation in 4-year-cycle phase angles taken at different periods of time. This means that the 4-year cycle is a stable period like diurnal and annual changes. It also implies that the 4-year cycle is caused by astronomical influences. In fact, variability of geomagnetic disturbance — one index of solar activities — shows a 4-year cycle. When geomagnetic disturbance variability increases, the temperature rises in high latitudes, as does atmospheric pressure. This phenomenon can be explained by supposing that the solar corpuscular flow increase enhances mixing of air masses between north and south.

INTRODUCTION

Every year, Japan suffers damage from typhoons and heavy rains. The author has noticed that storm damages for leap years are generally smaller than those for other years. The yearly loss of lives due to storms in Japan in the period 1915-1972 is listed in Table 1; the figures in the first column indicate loss of lives in leap years, while those in subsequent three columns are the yearly loss of lives in each of subsequent three years. The numbers of victims are obviously smaller in leap years except 1948. Understandably, such periodic recurrence may be accidental, but such a probability is very small. It is far more likely that victims are actually fewer in leap years.

Such an investigation suggests that a 4-year cycle might exist in the atmospheric phenomena, because leap years occur every four years. Many researches have been done on the periodicity of meteorological phenomena, but little on the 4-year cycle.

This paper analyzes the 4-year cycle in the annual mean meteorological elements at stations over the northern hemisphere, and investigates the mode of 4-year cycle in the same area.

Although the amplitudes are very small, the 4-year cycle is notably stable, and is also found in yearly solar activity changes. The 4-year cycle in the atmosphere, therefore, seems stimulated by solar radiation influence.

TABLE 1.

Yearly loss of lives due to storms.

Lag of year	0	1	2	3
Leap year				
1912	-	-	-	93
1916	153	1372	174	159
1920	224	1379	448	193
1924	319	379	801	643
1928	345	335	176	298
1932	411	306	3245	884
1936	341	306	1741	140
1940	75	593	1229	246
1944	364	3528	6936	1950
1948	1162	997	781	1387
1952	387	4984	8850	381
1956	299	1142	1626	5548
1960	280	736	237	575
1964	305	318	578	603
1968	366	183	175	343
1972	637	-	-	-
Mean	378	1183	1928	896

DETECTION METHODS

First, to briefly review detection methods, the 4-year cycle analysis in meteorological elements was done mainly by means of the so-called Schuster's method.

Let $\theta(t)$ be annual mean values of a meteorological element and 4-year harmonics in this time series be expressed by

$$\theta(t) = A \cos(2\pi t/4 + \phi) \quad (1)$$

where A is the amplitude and ϕ is the phase angle. The A and ϕ are calculated by the relations:

$$A^2 = \left\{ \frac{2}{N} \sum_{n=1}^N \theta(n\Delta t) \cos(2\pi n\Delta t/4) \right\}^2 + \left\{ \frac{2}{N} \sum_{n=1}^N \theta(n\Delta t) \sin(2\pi n\Delta t/4) \right\}^2, \quad (2)$$

$$\tan \phi = \sum_{n=1}^N \theta(n\Delta t) \sin(2\pi n\Delta t/4) / \sum_{n=1}^N \theta(n\Delta t) \cos(2\pi n\Delta t/4),$$

where Δt is a fixed time interval, say, a year.

If $\theta(t)$ is a random time series, the expectation of A is:

$$\epsilon = 2\sigma_{\theta} / \sqrt{N} \quad (3)$$

where σ_{θ} is the standard deviation of $\theta(t)$.

Accordingly, if the calculated A is larger enough than ϵ , the 4-year cycle is statistically significant, but A is comparable to or less than ϵ , the 4-year cycle is insignificant. The probability that A/ϵ is above 1.5 and 2 is 10.5 %

and 1.8 %, respectively, provided the original time series be a random number. These values will give us the critical values in significance test.

These criteria, however, appear too severe to the present case, because a time series of annual means of meteorological element is never of random numbers but of long periodic changes, i.e., predominated by climatic changes. Fig. 1(a) shows the mean time spectrum of the temperature obtained by an 80-year harmonic analysis at 18 stations over the northern hemisphere. It is found that the amplitude of low harmonics, that is, an amplitude of several ten year harmonics, predominates. Accordingly, these long period components must be reduced from σ_θ to detect the 4-year cycle. Then, the existence of the 4-year cycle will be quite probable if A/ε 1.5. Examples of significant 4-year cycles are shown in Table 2: the value of A/ε is larger than 2 for the temperature at two stations, Surgut and Upernivik; the existence of the 4-year cycle is obvious at these two stations. It may also be concluded that a 4-year cycle exists in the atmosphere, though it is hidden by other disturbances at most of the stations.

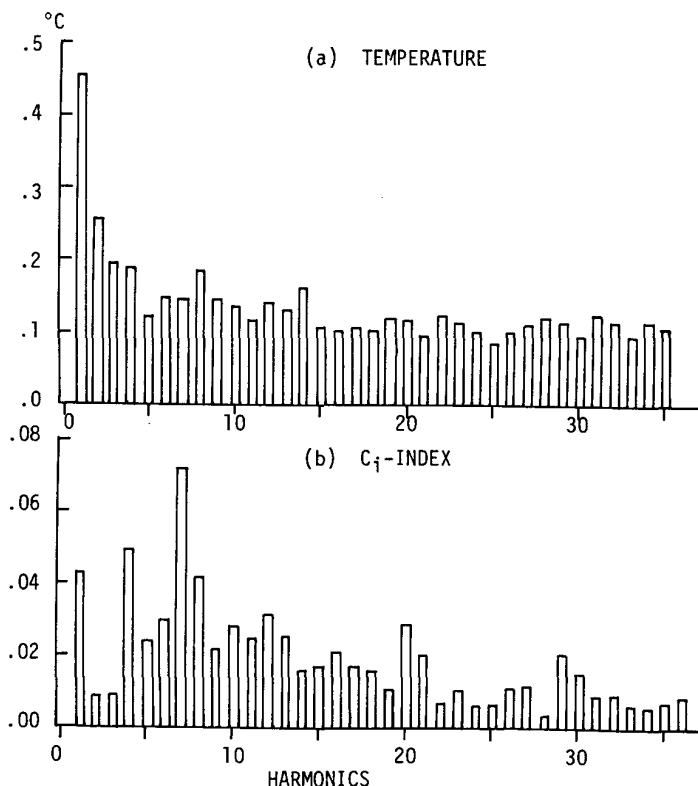


Fig. 1. Spectrum of 80-year harmonic analysis.

TABLE 2.

Significant 4-year cycle in the meteorological elements

Element	Amplitude A	Expectation ϵ	A/ ϵ	Period of analysis
Temp. at Surgut	0.52°F	0.25°F	2.1	1881-1960
Temp. at Upernivik	1.12°F	0.25°F	4.5	1885-1950
Temp. at Sydney	0.26°F	0.15°F	1.7	1859-1950
Prec. at Calcutta	110 mm	69 mm	1.6	1881-1960
Perc. at Alexandoria	51 mm	14 mm	3.6	1890-1940
Pres. at Honolulu	0.16 mmHg	0.10 mmHg	1.6	1881-1950
Pres. at Copenhagen	0.34 mmHg	0.20 mmHg	1.7	1881-1960

DETECTION OF THE FOUR-YEAR CYCLE BY MEANS OF THE PHASE ANGLE

Schuster's method is sometimes inconvenient for detecting the 4-year cycle when the amplitude is very small compared to other disturbances. Phase angles of the 4-year harmonics are examined to correct the shortcomings of the Schuster method: If a stable 4-year cycle really exists, the phase angle ϕ in equation (1), calculated for different time-intervals with the same time-origin, will remain the same, whereas if the 4-year cycle does not exist, the phase angle would not.

Examples of the results of the 4-year harmonic analysis for successive 32-year time-intervals are shown in Table 3, in which it is seen that the phase angle are almost the same, with a few exceptions, for different time-intervals. In this way, a stable 4-year cycle in meteorological elements has been detected with the method mentioned above, too. It may then be concluded the 4-year cycle is stable over quite a long period, 100 years or more.

TABLE 3.

Four-year harmonic analysis at different time-intervals.

Period of analysis \ Element	Copenhagen	New Heaven	Sydney	Rome	Calcutta	Sunspot
	Temp. A ϕ	Temp. A ϕ	Temp. A ϕ	Prec. A ϕ	Prec. A ϕ	Number A ϕ
1759-1790	0.40 210°	— —	— —	— —	— —	6.3 340°
1791-1822	0.44 260	0.26 260°	— —	661 200°	— —	3.1 290
1823-1854	0.34 340	0.18 200	— —	56 260	128 30°	5.4 220
1855-1886	0.35 230	0.10 220	0.12 210°	47 30	70 100	2.5 190
1887-1918	0.25 280	0.31 320	0.13 290	54 40	210 140	4.8 150
1919-1950	0.20 220	0.25 220	0.09 270	46 320	100 170	1.2 190

DETECTION OF THE FOUR-YEAR CYCLE IN SOLAR ACTIVITY

As was seen before, the 4-year cycle in meteorological elements is stable, which give us a suggestion that the cycle would be stimulated by astronomical influences, - solar activity, for instance.

To check the 4-year cycles in solar activity, the spectrum of 80-year harmonic analysis of C_1 -index — a geomagnetic disturbance index or a solar activity index — are shown in Fig. 1(b). It is well known that 11-year cycles predominate in solar activity changes, which is clearly seen in the spectrum in Fig. 1(b). We can also see in the figure that a 4-year cycle in the C_1 -index exists, which is proved to be statistically significant by Schuster's method. The list of phase angles of 4-year cycles for the sun spot number shown in the last column of Table 3 also indicates the existence of a stable 4-year cycle in solar activity.

The author has shown elsewhere that a 9.592-month cycle of tidal forces on the sun appears through conjunction and opposition of the earth and Venus. This period is detectable in the change of the solar constant. Amplitude A of the 9.592-month cycle in the solar constant change analysed for the period, September 1923- December 1947 is $0.001 \text{ cal/cm}^2.\text{min}$, while the expectation is $0.00056 \text{ cal/cm}^2.\text{min}$. Hence, $A/\epsilon = 1.8$ and the existence of the cycle is significant.

Fig. 2 shows a 9.592-month periodogram for the K_p -index between 1937 and 1973. The curves are roughly parallel and the amplitude is greater than the standard error. Solar activity seems to decrease due to strong tidal forces on the sun.

These results indicate the existence of the 9.592-month cycle in solar activity changes. If this cycle is approved, a 3.983-year cycle can be easily derived by coupling it with a 1-year cycle. This length is almost equal to 4 years. The 4-year cycle, then, may originate from the influence of the conjunction of the earth and Venus.

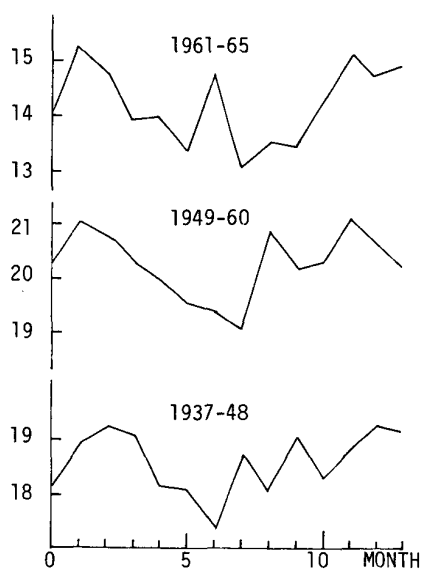


Fig. 2. 9.592-month periodogram for the K_p -index.

FOUR-YEAR PERIDOGRAM ANALYSIS OF ZONAL MEAN PRESSURE AND TEMPERATURE

The 4-year cycle in the atmosphere is by no means a local phenomenon. It is a global one. We shall examine the mode of the 4-year cycle on the earth by periodogram analysis of zonal mean values.

Table 4 shows the results of a 4-year periodogram analysis for zonal mean pressure, temperature and indices of solar activity. The figures in the table indicate the deviation from the normal.

TABLE 4.

4-year periodogram analysis of zonal mean pressure and temperature (Period of analysis is about 80 years).

Corresponding year Element	1950	1951	1952	1953
Pressure				
North of 60 N°	-2	12	-5	-4 0.01 mmHg
50-60	-13	10	3	3
40-50	6	-1	-3	-1
30-40	2	-2	-1	4
20-30	3	-2	-1	4
0-20	-4	1	5	0
Temperature				
North of 60 N°	-15	27	-13	-1 0.01°C
50-60	20	-23	1	1
40-50	27	-10	-11	-6
30-40	5	3	2	-10
20-30	0	3	-5	2
0-20	-2	-3	-2	8
Solar activity				
C _i -index	0.012	0.021	-0.002	-0.030
Sunspot	-4.0	-0.5	0.40	3.0 cal/cm ² .min
Solar constant	0.008	-0.0005	-0.0001	0.0001

Pressure changes in the table at various latitudes show that pressure change features are divided into three regions — polar, middle and low latitudes, whose boundaries are 50°N and 20°N. Pressure changes in the middle latitudes are in an opposite phase to those of polar and low latitudes. The amplitude of the change is large in high latitudes and minimum at 30°N. These features correspond to the 3 cells of general circulation.

Pressure change at high latitudes is found to be roughly in the same phase as that of the C_i-index. In other words, the C_i-index increase corresponds to a zonal mean pressure increase in polar regions.

The mode of the 4-year periodogram analysis for zonal mean temperature shows a similar latitudinal distribution to that of the pressure, though the boundaries of three regions are 60°N and 20°N. The 4-year cycle is most distinct at the 54°N belt

and is in the opposite phase with those in the polar and low latitude regions. The temperature in the polar region is high when the C_1 -index is high.

These results can be explained by the hypothesis that air mass exchange between the polar and middle latitudes increases when a strong invasion of the corpuscular flow from the sun occurs. The present analysis, therefore, suggests that the 4-year cycle in the atmosphere originates in the 4-year cycle of solar activity.

FEATURES OF FOUR-YEAR CYCLE OSCILLATION OVER THE NORTHERN HEMISPHERE

As a result of applying the 4-year harmonic analysis to the meteorological elements at stations over the northern hemisphere and investigating the distribution of amplitudes and phase angles, Table 5 shows the frequency distribution of phase angles of air pressure, temperature and precipitation at northern hemisphere stations.

TABLE 5.

Frequency distribution of phase angles of 4-year cycle (1908-39).

Lower limit of phase angle (°)	0	30	60	90	120	150	180	210	240	270	300	330
Pressure	3	4	5	2	1	1	5	16	4	2	1	3
Temperature	0	3	1	4	3	7	8	5	5	8	1	0
Precipitation	5	6	2	3	4	3	4	4	1	2	4	1

The frequency distribution for pressure has two maxima, one distinct peak at about 220° and another at about 50°. Temperature frequency distribution, on the other hand, has one vague maximum at about 220°, while that for precipitation has no distinct maximum.

Next, phase angles and amplitudes for the temperature are plotted on the northern hemisphere map and contours drawn as shown in Fig. 3. Phase angles of about 180° are distributed on the high latitude area with three legs and phase angles of about 240° surround this area. Notable is that such a three-wave structure in the west-lies predominated in the unusual weather year of 1963.

Fig. 4 shows the phase angle distribution for pressure. Clearly, the northern hemisphere is divided roughly into two kinds of domains: one around a phase angle about 50° and the other around 220°, corresponding to the two maxima in the phase angle frequency distribution. This pattern shows that a 4-year standing oscillation is stimulated in the pressure field. The oscillation pattern, however, is not simple compared with the distribution pattern of annual precipitation over the northern hemisphere. It is notable that the 220° area corresponds to a climatically rainy area, while the 50° area does to a climatically dry one.

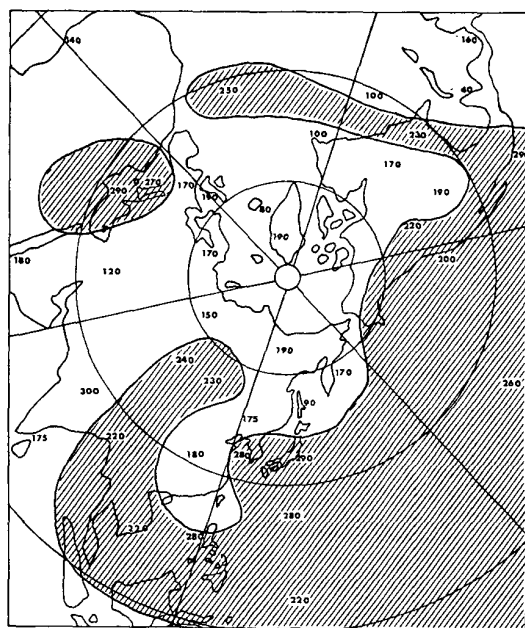
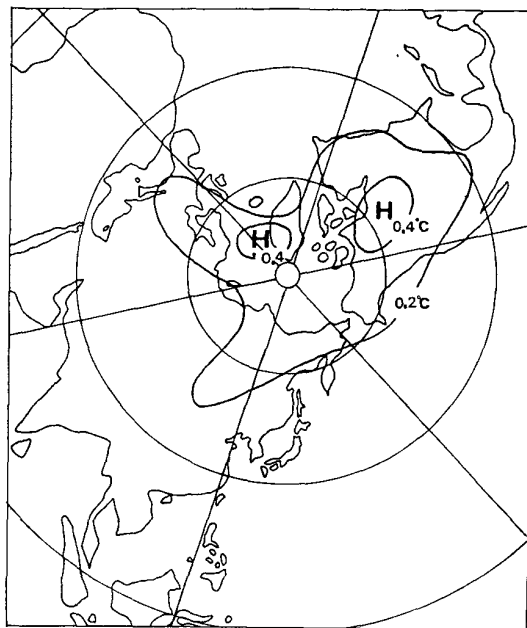


Fig. 3. Distribution of amplitude and phase angles of 4-year temperature oscillation.

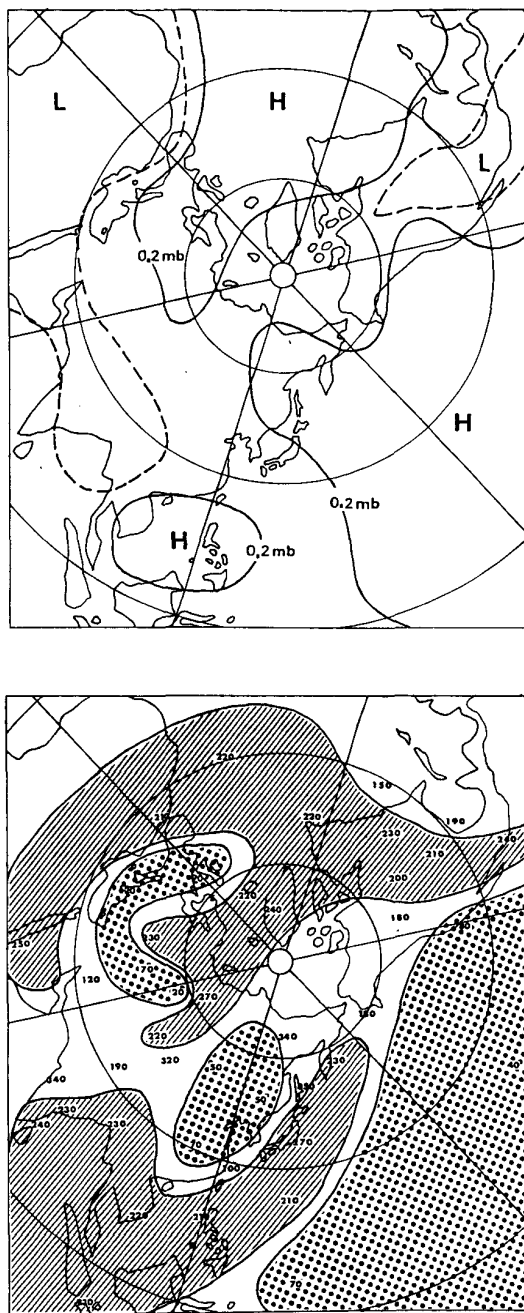


Fig. 4. Distribution of amplitude and phase angles of 4-year pressure oscillation.

CONCLUSION

The above analyses illustrate the existence of 4-year cycle stationary oscillation with a little phase angle change in the atmospheric oscillation over the northern hemisphere. The cause for the oscillation is structually unclear yet, but it seems to originate in the 4-year solar activity cycle and 4-year corpuscular flow change cycle. Furthermore, a stable 5-year cycle exists in northern hemisphere oscillation and this seems to arise from periodic change of the solar constant, which will be discussed in future.

CLASSIFICATION OF MONSOON CLIMATES AND STABILITY OF THEIR MOISTURE REGIME

V.P.SUBRAHMANYAM¹ and H.S.RAM MOHAN²

1 Dept. Meteor. & Oceanog., Andhra Univ., Waltair (South India)

2 Dept. Marine Sci., Univ. of Cochin, Cochin (South India)

ABSTRACT

Subrahmanyam, V.P. and Ram Mohan, H.S. Classification of monsoon climates and stability of their moisture regime. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

India is a typical example of a monsoonal country in the sense that the prevailing wind circulation over the country reverses almost exactly by 180° in the conventional summer and winter seasons. But what matters most for its categorization is not so much the wind directional change, as the precipitation regime on which the national economy of the country vitally depends.

Köppen appears to be the first (and perhaps the only) climatologist that attempted a quantitative delineation of the monsoon climates as an intermediate category in the tropical climates between the tropical rainforest (Af) on the one hand, and the savanna (Aw) with pronounced seasonal dryness on the other. In the present paper, a new approach towards delineation of monsoonal climates based on water balance concept has been proposed (mainly from the point of view of agriculture) employing an Index of Moisture Availability, defined as the ratio of actual evapotranspiration for individual months to the climatic annual actual evapotranspiration divided by twelve.

But one of the most intriguing features of all climates in general, and of the monsoon climates in particular, is their highly unstable moisture regime, which arises mainly on account of the (yet) uncertainties in the monsoonal circulation whose exact onset, withdrawal and progress over the country are usually unpredictable and seem to have no periodicity whatsoever. This particular characteristics of the monsoon climates is illustrated in the paper through water balance diagrams shifts of a few representative stations, observed through year-to-year fluctuations in moisture indices following imbalances in their water budgets.

It is suggested that this type of analysis on a shorter-term (weekly, for instance) basis for the monsoon season in individual climatic zones would be very useful to the planner in agricultural and hydrological project designs and maintenance.

The term 'Monsoon' which was originally used to describe the seasonal alternation of winds over the arabian Sea is now generally applied to quasi-stationary disturbances in the average zonal circulation, particularly in the tropics, arising from temperature and humidity differences between the airmasses originating over the continental and oceanic regions of the northern and southern hemispheres. Though monsoons are thus essentially seasonal winds blowing in opposite directions in summer and winter, it is mainly the abundant rainfall associated with them that is of great

economic importance on account of its impact on the food production from agriculture in India and other countries of Southeast Asia. In fact, to many, the Indian Southwest or summer monsoon simply implies rainfall alone. It is no wonder then that extensive research work has been carried out on the several aspects of the monsoons in general and of the Southwest monsoon of the Indian region; in particular, the space-time variations and distribution of rainfall, the disturbances and storms associated with them, the breaks in and failures of the monsoonal circulation and the forecasting of monsoon rains. All such studies have revealed that monsoon climates are a unique category by themselves, the distribution and variability of their different constituent features comprising a complex matrix and controlling the characteristics of the circulation as a whole. Even though the large variations in rainfall in space and time as determined by the different combinations of atmospheric processes produce a variety of climatic types, from perhumid to arid, the effect and rhythm of the monsoonal rainfall regime are clearly reflected in most of them.

There have been numerous schemes of climatic classification based on different parameters to suit different purposes but outstanding among them are those by botanists, plant geographers and ecologists who attempted correlations of the climatic provinces with the nature and distribution of vegetation types. Wladimir Köppen (1900) was about the first among them to succeed in evolving a quantitative scheme of classification of world climates based mainly on critical temperatures for the growth and maintenance of different kinds of vegetation. He himself revised it later (Köppen, 1918) with greater attention to temperature, rainfall and their seasonal characteristics. Further modifications of the scheme have continued (Köppen, 1931, 1936, Trewartha, 1968) with revision of the boundary limits as new data became available. The Köppen system in its present form has five major categories of climate the first of which is designated as 'A' (Tropical) having an average temperature of 64.4°F (18°C) or higher for the coolest month of the year. Within this category are subdivisions defined by specific values of precipitation. Of relevant interest here is the 'Am' (Tropical Rainforest) having no dry season (mean precipitation of the driest month being 2.4", i.e., 6 cms. or more) and the 'Aw' (Tropical Savanna) with pronounced winter dryness (at least one month in the winter season receiving less than 2.4", i.e., 6 cms. of precipitation). According to Köppen, in the tropical monsoon climate, the precipitation of the driest month is less than 2.4" or 6 cms. but equal to or greater than $(3.94 - r/25)$ inches where 'r' is the average annual precipitation in inches. Thus, Köppen appears to be the only climatologist so far to have classified the monsoon climates into a separate category. Mizukoshi (1971) has described the regional divisions of Monsoon Asia using this classification.

However, as is well known, the value and validity of a climatic classification scheme are determined largely by the purpose for which it is intended. When vegetation is the primary concern, the moisture effectivity of a climate is governed by

the relative efficiencies of precipitation and temperature which must, therefore, be studied together but not individually, as Köppen did. Thus, both the thermal and hygric factors have to be simultaneously considered to determine whether water supply by precipitation is greater or less than the water needed for the fullest development of vegetation. It is with this end in view that Thornthwaite (1943) made a rational contribution to climatological literature by introducing the concept of Potential Evapotranspiration (P.E.), otherwise as the water need. When the magnitudes of precipitation and potential evapotranspiration are equal there is neither water surplus nor water deficit while when precipitation exceeds the water need month after month the soil moisture is raised eventually to its field capacity value and later water surplus results. On the other hand, when precipitation falls below the water need soil moisture is gradually depleted for evapotranspirational purposes, the Actual Evapotranspiration (A.E.) decreases below the potential limit and water deficiency occurs. For computing the water balance parameters on a monthly basis Thornthwaite (1948) evolved an elegant book-keeping procedure which was modified later by Thornthwaite and Mather (1955) after accumulation of more data.

The moisture regime of climate according to the Thornthwaite scheme of classification (Thornthwaite 1948, Thornthwaite and Hare 1955, Carter and Mather 1966) is based on the water balance procedure outlined above. Classification of the climates of the Indian region according to this scheme (both the thermal and moisture regimes) was made for the first time by Subrahmanyam (1956). Yet, the peculiarities of the monsoonal regime of climate are not brought out in this scheme.

In the present paper, therefore, an attempt has been made to extend the water balance concepts to delineate the monsoon climates as a separate category. Water balances of more than 250 stations in India, Sri Lanka, Burma, Pakistan and Bangla Desh have been worked out on a climatic basis. Actual Evapotranspiration (A.E.) values obtained from the book-keeping procedure have been made use of to define an Index of Moisture (I_{MA}) as the ratio of the A.E. for any individual month of the year to the average monthly value, i.e., the total annual value of A.E. divided by 12. The maximum and minimum values of the climatic indices of moisture availability so obtained on a monthly basis are noted and the difference between them, expressed as a percentage of the climatic annual value of A.E. divided by 12 has been designated as the I_{MA} -Range. The calculations of the indices of moisture availability for two typical stations - Calicut (Köppen's Am in South India and Bikaner (Bw) in the north are shown in Table 1 and values for a few selected stations are given in Table 2.

Analysis of results for all the stations in the Indian region has led to the conclusion that stations within the domain of intense monsoonal circulation and associated rainfall (like Calicut) have I_{MA} -Range values below 15% while for other stations (non-monsoonal) the values are well above 15% and sometimes even exceed 100% (as for Bikaner). Stations with lower ranges of the index seem to have a fairly

TABLE 1.

Monthly values of actual evapotranspiration and indices of moisture availability.

Station	J	F	M	A	M	J	J	A	S	O	N	D	Annual	Annual/12
CALICUT (Am)	Actual Evapotranspiration, cms.												138.2	11.52
	7.8	5.4	5.2	10.4	16.8	14.5	12.7	12.9	13.3	14.3	13.6	11.3		
	Indices of Moisture Availability													
	0.68	0.47	0.45	0.90	1.46	1.26	1.10	1.12	1.15	1.24	1.18	0.98		
BIKANER (BW)	Actual Evapotranspiration, cms.												29.1	2.43
	0.7	0.7	0.6	0.5	1.5	3.1	8.5	9.1	3.3	0.5	0.1	0.5		
	Indices of Moisture Availability													
	0.29	0.29	0.25	0.21	0.62	1.28	3.50	3.74	1.36	0.21	0.04	0.21		
I_{MA} -Range for Calicut = $\frac{1.46 - 0.45}{11.52} \times 100 = 8.77 \%$														
I_{MA} -Range for Bikaner = $\frac{3.74 - 0.04}{2.43} \times 100 = 152.26 \%$														

uniform distribution of the monthly values of A.E. through the year. This is not so much because precipitation is uniform, since in monsoon climates it is highly seasonal, but because of the gradual soil moisture accretion that takes place during the wet season, i.e., when precipitation exceeds P.E.. The soil moisture thus accumulated becomes available for evapotranspirational use during the lean months and, therefore, the A.E. values of these months are raised well above the precipitation figures. On the other hand, at stations where the ranges are higher (even greater than 500% in desert regions), the soil seldom attains field capacity even during the wettest month; even if the field capacity is reached, the soil moisture is as rapidly depleted as it is raised so that the A.E. values decline to very low figures in the subsequent prolonged dry season. Interestingly, Köppen's scheme for classifying monsoon climates appears to pay due attention to these facts though the role of soil has not been explicitly mentioned therein.

The precipitation during the wet season could be due to the monsoonal (summer or winter) circulation, or on account of depressions or cyclones or might as well be due to orographic effects. A monsoon type of climates from the point of view of agricultural vegetation may, therefore, be defined as one in which precipitation has such a high seasonal concentration that is adequate to sustain fairly uniform levels of A.E. throughout the year. Accordingly, all stations having I_{MA} -Ranges lower than 15% may be categorized as belonging to the monsoon climates and the monsoon months are those in which the indices of moisture availability are above 1; at Calicut thus the monsoon months are from May to November.

The monsoon climates of the Indian region under this scheme are shown in Fig.1. The south-west coast of India, a part of the Coromandel coastal tract and its hinterland, north Orissa, Bengal, Bangla Desh, parts of Assam and the western coastal strip of Burma all come under this category. Also, Sri Lanka and the Andaman and Nicobar islands in the Bay of Bengal are monsoonal in their climate. Significantly, all

TABLE 2.

Ranges on indices of moisture availability for a few selected stations in the Indian region

MONSOON CLIMATES		NON-MONSOON CLIMATES	
Station	I_{MA} -Range (%)	Station	I_{MA} -Range (%)
Cochin	6.49	Tiruchirapalle	25.66
Madras	11.24	Visakhapatnam	20.45
Trincomalee	7.44	Bombay	20.76
Chittagong	9.72	Veeraval	92.86
Victoria Point	5.93	Hissar	84.27
Port Blair	4.88	Nokkundi	792.46
Rangoon	10.70	Karachi	231.13
Sibsagar	14.42	Jodhpur	132.56

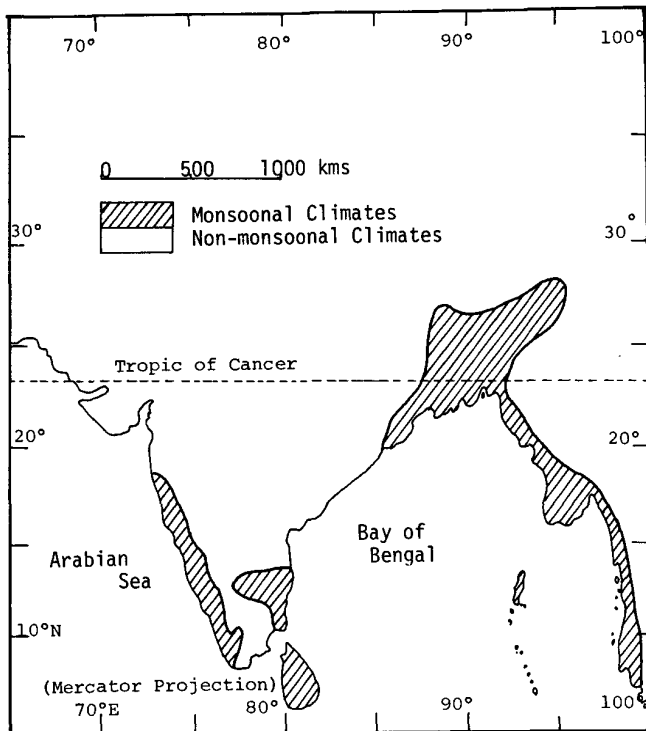


Fig. 1. Monsoon climates of Southeast Asia

stations belonging to Am and Af climates and some stations in the Aw, As, Cw, Cs and Cf categories of the Köppen system come under the presently proposed monsoonal type of climate. No station in the 'B' (dry) group of Köppen climates finds a place

here. Another interesting observation is that the monsoon stations, as delineated now, belong to one or the other among the climatic categories from perhumid to dry subhumid according to the Thornthwaite scheme of classification; very few come under the semi-arid group while none is arid.

Though on a climatic basis monsoon climates have fairly uniform values of A.E., one of the most intriguing aspects of the monsoonal moisture regime is its high stability. This arises mainly on account of the random fluctuations in the monsoon circulation whose onset, establishment, progress and withdrawal over the country are usually unpredictable. The wide fluctuations in the rainfall amounts following such changes in the monsoonal circulation invariably produce variations in water balance, occasionally of such magnitudes that the very climatic types of the station are shifted from their normal category into the drier or wetter direction. In the present context under such circumstances a monsoon station may become non-monsoonal in character in a dry year and during a wet year a normal non-monsoonal station may exhibit monsoonal characteristics. Such shifts in climate, though temporary, are of great interest to the applied climatologist, for their magnitude and frequency not only reflect the extent of seasonality of climate of a station but also determine the climatological potentialities of a region for agricultural and hydrological development.

For making an analytical study of the above aspect, two stations coming under the category of monsoon climates - Mangalore on the south-west coast of peninsular India and Balasore on the north-east coast - have been chosen. Mangalore is affected by the Arabian Sea branch of the Indian Southwest or summer monsoon while Balasore comes under the influence of the Bay of Bengal branch. Thornthwaite Moisture Index (I_m) values of these stations for individual years have been plotted (Figs. 2 and 3) and years of extreme climatic shifts have been worked out and compared graphically with their respective normal pictures (Figs. 4 and 5). The climatic shifts at the stations during the study period (1901-1964) are summarized in Table 3 (Subrahmanyam and Sarma 1973).

The climate of Mangalore (Perhumid, $I_m = 101.9\%$, I_{MA} -Range = 11.51%) may be seen to be rather conservative being most of the time close to its normal climatic regime. On 34 occasions out of 64 years it struck to its perhumid category though experiencing a number of migrations onto the more humid side (Fig.2 (a)) and on 29 occasions it exhibited shifts onto the drier side. Perhumid climatic zones are generally characterized by low variabilities and a high stability of their moisture regime so that even a small variation in water deficit or water surplus from the normal would lead to pronounced imbalances in their water budgets. Fig.3(a) shows the water balance charts of Mangalore for the normal year as well as for the extreme years on the drier and wetter sides.

During the normal year Mangalore receives 342.8 cms. of rainfall and has a water

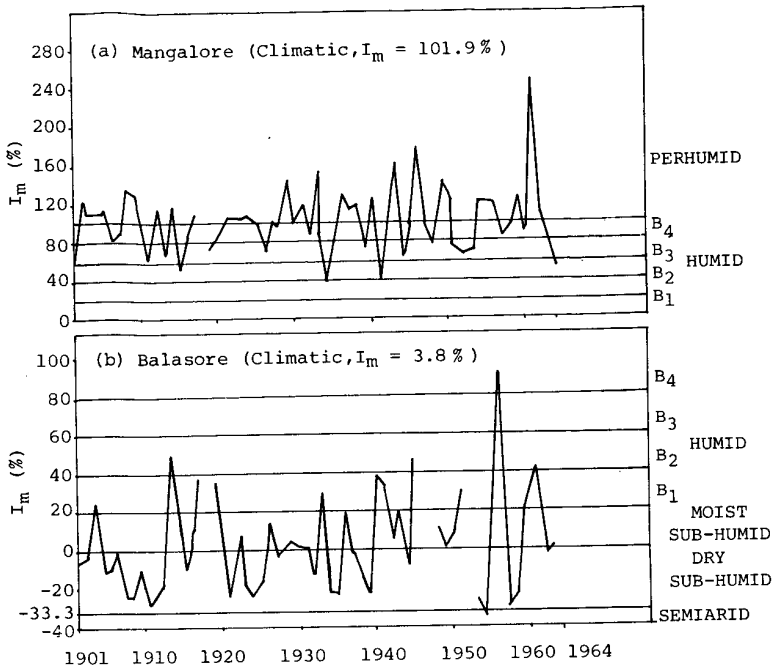


Fig. 2. Climatic shifts : (a) Mangalore , (b) Balasore

TABLE 3.

Climatic shifts in monsoon climates

Station	Number of climatic shifts								
	Perhumid	Humid				Subhumid		Semi-arid	Arid
	A	B ₄	B ₃	B ₂	B ₁	Moist C ₂	Dry C ₁	D	E
MANGALORE (A)	34	14	11	2	2	-	-	-	-
BALASORE (C ₂)	-	1	-	4	9	15	29	1	-

surplus of 216 cms. and a deficit of 43.3 cms. (Table 4). But during the wet year (1961) when its climate became more humid due to an increase of precipitation to 583.8 cms. its water surplus rose to 464.2 cms. which was twice the normal. In the dry year (1934) the station experienced less than 65% of its normal rainfall and so its water surplus decreased by almost half while the water deficit was higher than the normal by about 12%.

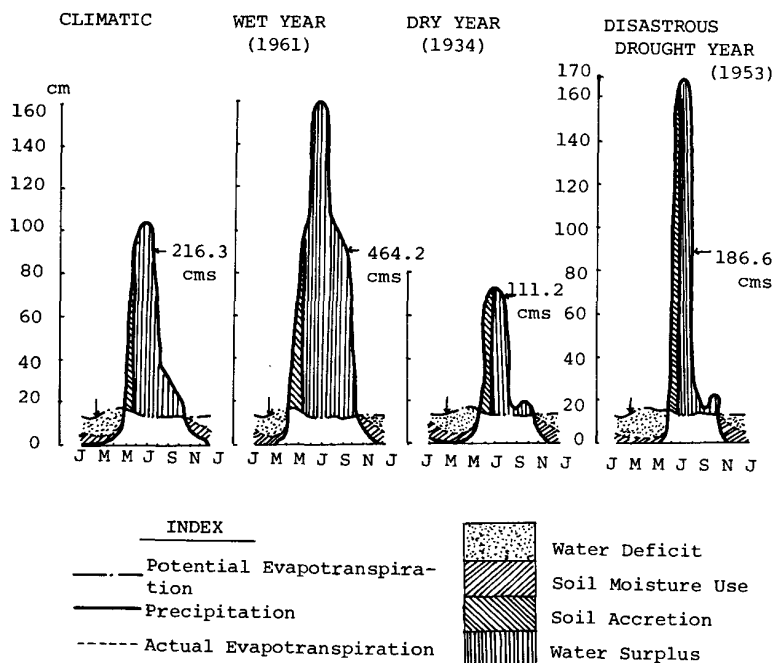


Fig. 3(a). Water balances of Mangalore.

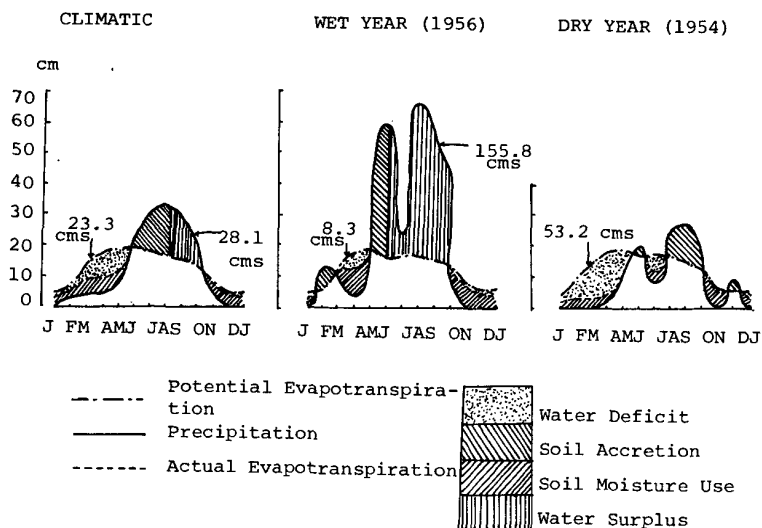


Fig. 3(b). Water balances of Balasore.

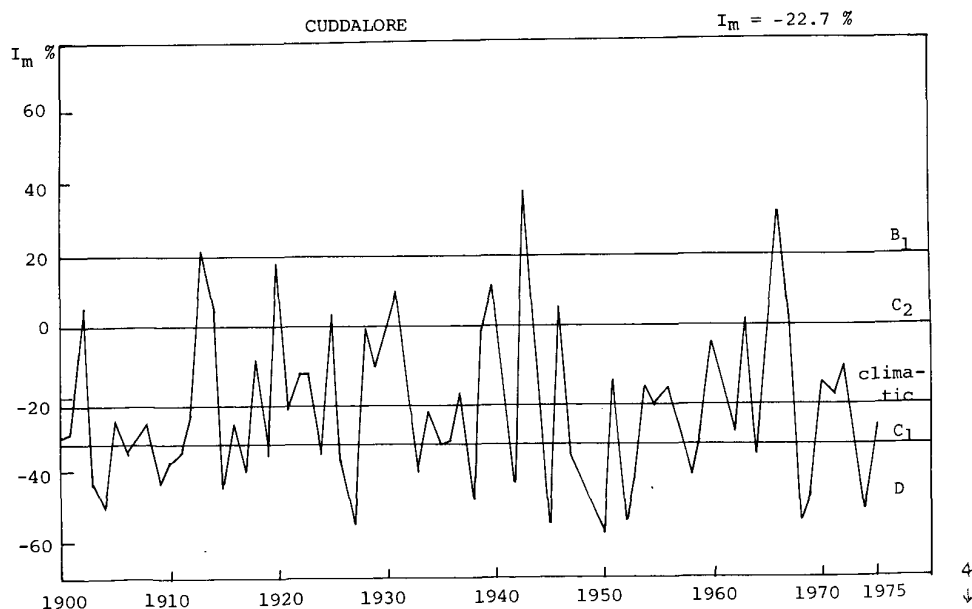
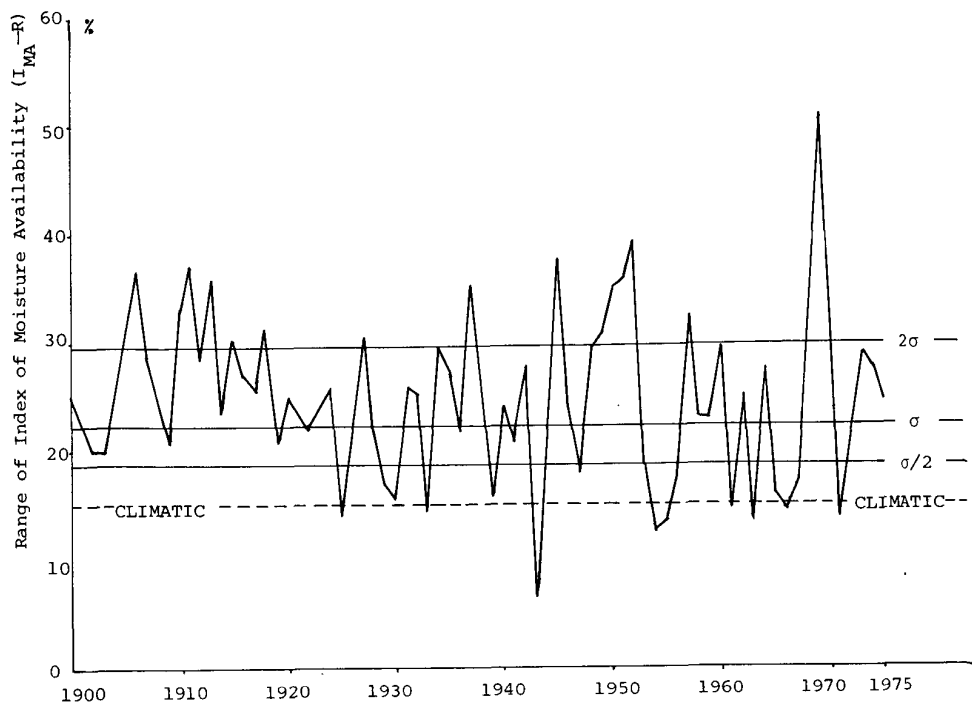


Fig.4. Yearly march of range of index of moisture availability at Cuddalore (upper); Climatic shifts in dry subhumid zone (lower).

TABLE 4.

Comparative water balance data in monsoon climates (after Subrahmanyam and Sarma 1973)

Year	Water need cms.	Precipitation cms.	Water surplus cms.	Water deficit	Moisture regime
(a) MANGALORE					
Normal year	169.8	342.8	216.3	43.3	A (Perhumid)
Wet year (1961)	166.7	583.8	464.2	44.7	A (Perhumid)
Dry year (1934)	168.6	219.5	111.2	48.5	B ₁ (Humid)
Disastrous drought year (1953)	171.9	289.4	186.6	67.6	B ₃ (Humid)
(b) BALASORE					
Normal year	156.6	162.4	28.1	22.3	C ₂ (Moist subhumid)
Wet year (1956)	154.1	295.3	155.8	8.3	B ₄ (Humid)
Dry year (1954)	158.0	115.7	0.0	53.2	D (Semi-arid)

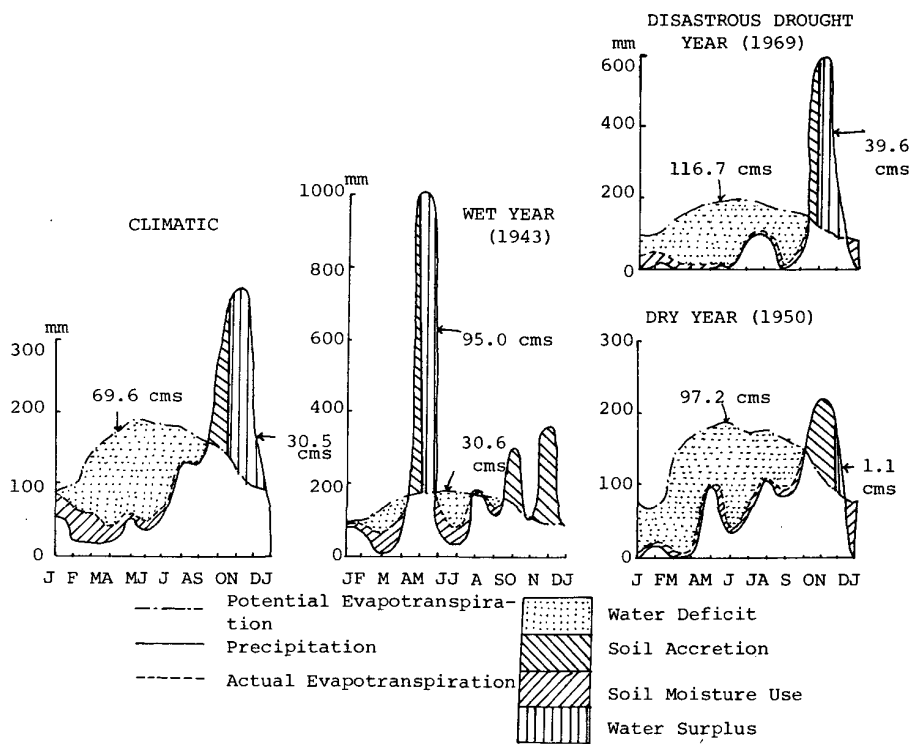


Fig.5. Water balances of Cuddalore

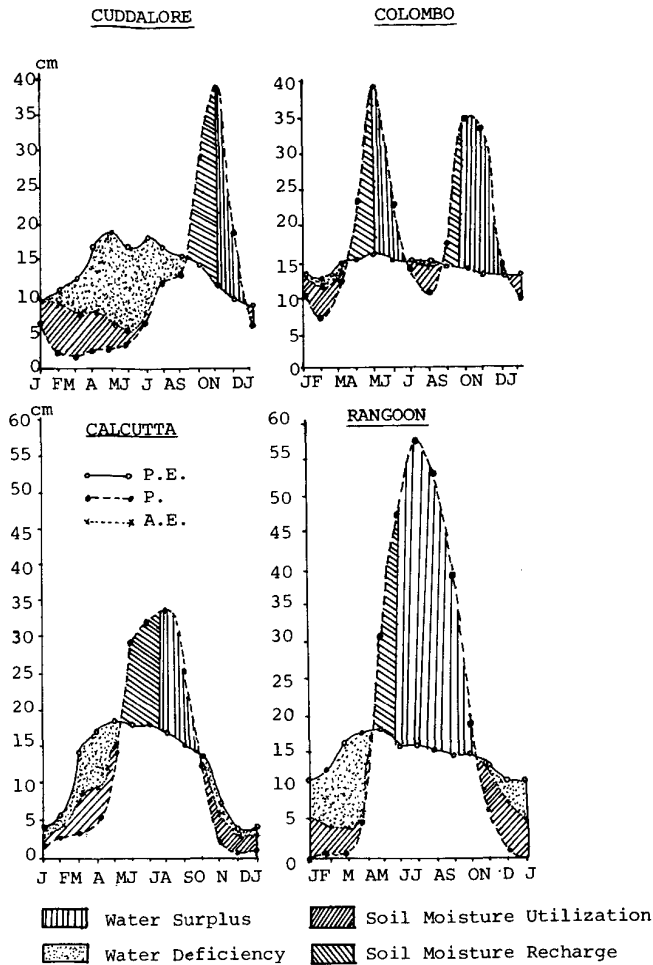


Fig. 5. (cont'd)

On the other hand, the climate of Balasore (Moist Subhumid, $I_m = 3.8\%$, I_{MA} -Range = 11.52%) is characterized by wide variations in moisture regime from year to year (Fig.2(b)). It remained in its normal climatic category only on 15 occasions but shifted towards drier regime on 30 occasions and 14 times alone onto the humid side. Such frequent and violent fluctuations are characteristic of the subhumid climates, the so-called 'border-line category' of the climatic spectrum. During the wet year (1956) rainfall here was 295.3 cms., an increase by about 80% over the normal of 162.4 cms., following a very active summer monsoon (Fig.3(b)); consequently the water

surplus too rose enormously to about 6 times the normal value of 28.1 cms. In the dry year which was also a disastrous drought year (1954) water deficit was about 2.5 times the normal of 22.3 cms. because rainfall was deficient by about 28% (Table 4).

TABLE 5.

Comparative water balance data for Cuddalore.

Year	Water need cms.	Precipitation cms.	Water surplus cms.	Water deficit cms.	I_{MA} -Range %
Normal year	172.5	133.4	30.5	69.6	14.92
Wet year (1943)	170.4	235.2	95.0	30.6	7.21
Dry year (1950)	167.8	86.5	1.1	97.2	35.20
Disastrous drought (1969)	172.5	102.9	39.6	116.7	50.97

In this connection it is interesting to consider the imbalance in the water budgets for the years of extreme climatic shifts observed through year-to-year fluctuations in the I_{MA} -Range values. Cuddalore (Dry subhumid, $I_m = 22.7\%$) another typical monsoon station on the Coromandel coastal tract of South India affected by the Northeast monsoon has been chosen for this study. Yearly water balance data of Cuddalore (Ram Mohan 1978) for the period 1901-1975 have been used to calculate the ranges of the indices of moisture availability plotted graphically in Fig.4. The normal I_{MA} -Range value here is 14.92%, the monsoon months being August to December. In 1943 the I_{MA} -Range had its lowest value (7.21%) showing that the station had experienced a vigorous Northeast monsoon during the year when the rainfall exceeded the normal by about 75%. As a result, the water surplus increased by more than three times and the deficit decreased sharply by about 50% (Fig.5) compared to the normal values. On the otherhand, the I_{MA} -Range was highest (50.97%) in 1969, the disastrous drought year, the deviation being more than twice the standard deviation from the normal. During this extremely dry year water deficiency shot up by about 70% which pushed the climate of the station into the arid category and making its I_{MA} -Range value non-monsoonal. Interestingly, however, during this year the water surplus was also higher because of late receipt of rains in December (Table 5). This underlines the fact that the stability of the moisture regime of climate is governed not only by the amount of precipitation but by its distribution in time as well, an aspect of great ecological importance in drought climatology. This feature is further supported by the situation in 1950 (dry year) when, though the precipitation was more deficient, the I_{MA} -Range value rose only to 35.20%; this is because rainfall was less erratic and better-distributed during the year and so the A.E. values were more uniform. Another significant fact that came to light from the

present study is that the wet, dry and disastrous drought years delineated earlier using the Thornthwaite I_m values seem to coincide exactly with the years of extreme shifts for the I_{MA} -Range values too; this again emphasizes the utility and versatility of the index of moisture availability as an ecoclimatic tool for analytical studies in monsoon climatology.

During the study period, Cuddalore deviated onto the drier (non-monsoonal) side by more than twice the standard deviation on 16 occasions, between one standard deviation and two standard deviations on 32 occasions and between half-a-standard deviation and one standard deviation on 12 occasions; for 16 years, however, it remained in its normal monsoon climatic category. As already remarked earlier in connection with Balasore, such large and frequent fluctuations are typical of the sub-humid (buffer) climates which have high instability and poor conservatism. Cuddalore, specifically, tended to go into drier situations more often than into the humid direction and this conclusion is quite in tune with the general finding of instability of the dry subhumid climates of Tamilnadu (in South India) using the I_m values for studying climatic shifts (Ram Mohan 1978).

As the Index of Moisture Availability thus appears to be such a useful parameter for delineating periods and zones of monsoon climates, it is suggested that this type of analysis on a weekly or bi-weekly basis in different monsoon regions of the world would be very helpful to the planners in the design and maintenance of water project systems in agriculture or hydrology.

REFERENCES

- Carter, D.B. and Mather, J.R., 1966. Climatic classification for Environmental Biology, Publ. in Clim., Drexel. Inst. Tech. XIX, 4:341-352.
- Koppen, W., 1900. Versuche einer Klassikation der Klimate. Geogr. Zeitschr. 6 : 593-611, 657-679.
- Koppen, W., 1931. Grundriss der Klimakunde. Gruyter and Co., Berlin.
- Koppen, W., 1936. Das geographische System der Klimate. In: Handbuch der Klimatologie. Gebruder Borntraeger, Berlin, Vol. I, Part C.
- Mizukoshi, M., 1971. Regional divisions of Monsoon Asia by Koppen's classification of climate - A climatological approach. In: Yoshino, M.M. (ed.) Water Balance of Monsoon Asia. Univ. Hawaii P., Honolulu :259-273.
- Ram Mohan, H.S., 1978. A study on the water balance and drought climatology of Tamilnadu. Unpubl. Ph.D. thesis submitted to Andhra Univ., Waltair, S.India.
- Subrahmanyam, V.P., 1956. Climatic types of India according to the rational classification of Thornthwaite. Ind. J. Met. and Geophy., 7: 1-12.
- Subrahmanyam, V.P. and Sarma, A.A.L.N., 1973. Studies in drought climatology. Part I : Moist climates of South India. Trop. Ecol. 14 : 129-137.
- Thornthwaite, C.W., 1943. Problems in the classification of climates. Geogr. Rev. 33:233-235.
- Thornthwaite, C.W., 1948. An approach toward a rational classification of climate. Geogr. Rev. 38:55-94.
- Thornthwaite, C.W. and Hare, F.K., 1955. Climatic classification in Forestry. Unasylva 9:50-59
- Thornthwaite, C.W. and Mather, J.R., 1955. The Water Balance. Publ. in Clim., Drexel

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

SUITABLE PROBABILITY MODEL FOR SEVERE CYCLONIC STORMS STRIKING THE COAST AROUND THE BAY OF BENGAL

D.A.MOOLEY

Indian Inst. Trop. Met., Pune - 411005 (India)

ABSTRACT

Mooley, D.A., Suitable probability model for severe cyclonic storms striking the coast around the Bay of Bengal. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

Severe cyclonic storm is a natural calamity. When it is close to the coast or when it strikes the coast there is disruption in communications, damage to property and crops, and loss of life. All the 141 severe cyclonic storms which struck the Arakan Coast of Burma and the coasts of Bangla Desh, east India and Sri Lanka during the period 1877-1977 have been considered in this study. These are low pressure systems in which the associated wind is 48 knots or more. Only severe cyclonic storms have been considered since they cause much more damage than that caused by cyclonic storms and also since it was felt advisable to confine the analysis to systems of practically the same intensity. Swed and Eisenhart's runs test for runs above and below the median to detect trend or oscillation and Mann-Kendall Rank Statistic test for randomness were applied to the time interval between successive severe cyclonic storms which struck the coast. The results of these tests suggest that this interval can be generally taken to be random. Thus, the event of the coast being struck with a severe cyclonic storm is seen to be random in time continuum.

Applying the test for adequacy of the Poisson distribution, it is found that Poisson distribution is adequate. Hence the same was fitted to the data on the number of severe cyclonic storms striking the coast in a year for the whole period, and to the data for the component periods, 1891-1964 and 1877-1964. The goodness-of-fit as tested by Chi-square test is found to be very good in all the three cases. The number of severe cyclonic storms crossing the coast around the Bay in a year is a Poisson-distributed variable.

The Poisson distribution was also fit to the number of severe storms striking the coast during October-November and to the number of severe storms striking Bangla Desh - north Arakan Coast during the year. In both of these cases, the fit as tested by Chi-square test is found to be very good. Thus even when we consider a part of the whole coast around the Bay or a part of the year, the Poisson distribution gives a very good fit.

1. INTRODUCTION

Low pressure areas form over the Bay of Bengal. Occasionally the lows move into the Bay from the east as remnants of typhoons or depressions from the China Sea. Some of the lows over the Bay intensify into depressions and some of the depressions further intensify into cyclonic storms. A few of the cyclonic storms develop further into severe cyclonic storms. The wind speed associated with severe cyclonic storms

is 48 knots or more. These develop over the Bay mostly in the months from April to December and strike any section of the coast around the Bay. The enormous destruction of life and property, leading to disruption of the economy of the affected districts is caused by the extremely strong winds, torrential rains and tidal wave associated with a severe cyclonic storm. Severe cyclonic storms only have been considered since these cause much more havoc than that caused by cyclonic storms. Another reason for considering severe cyclonic storms only is to confine the analysis, to the extent possible, to systems of practically the same intensity. In this study, the suitability of the Poisson probability model for the severe cyclonic storms striking the coast around the Bay has been examined.

2. DATA SOURCES AND DISCUSSION OF DATA

All the severe cyclonic storms (to be referred hereafter as severe storms) which hit the coast around the Bay of Bengal (as shown in thick in Fig. 1) during the period 1877-1977, have been considered. Hereafter, this coast will be referred to as the coast. The relevant data were obtained from the tracks of storms and depressions over the Bay of Bengal and the Arabian Sea published by the India Meteorological Department (1964) for the period 1961-70, as well as from India Weather Review, Annual Summary, for these years. Information for the period 1971-76 was extracted from the articles giving accounts of storms and depressions by Das et al (1972, 1973), Alexander et al (1974, 1976, 1977) and Pant et al (1978) in the Indian Journal of Meteorology/Hydrology and Geophysics. Information for 1977 was obtained from the account of storms and depressions prepared by the Deputy Director General of Observatories (Forecasting), Pune-5.

The number of severe storms which struck the coast in each of the years during the period 1877-1977 is shown in Figure 2. The total number during the period was 141. It has been mentioned in the publication by India Meteorological Department (loc. cit) that for the period 1891-1960, the information on storms and depressions was obtained from India Weather Review, Annual Summary and information for the prior period was obtained from the papers by Eliot (1885, 1888) and from "Reports on Meteorology of India" for the years 1886 to 1890, and that the series of storms prepared from the accounts given in India Weather Review, Annual Summary, can be considered as one obtained from a homogeneous series of data.

A careful examination of Figure 2 shows that through the number of severe storms in a year during the period 1877-1890 is not generally different from that during the period 1891-1964, there is a relatively higher frequency of no severe storm striking the coast during the former period, but during the period 1965-77, the number in each year is generally much higher than that for the period 1891-1964. While no specific trend is revealed, it is seen that the mean and variance for the

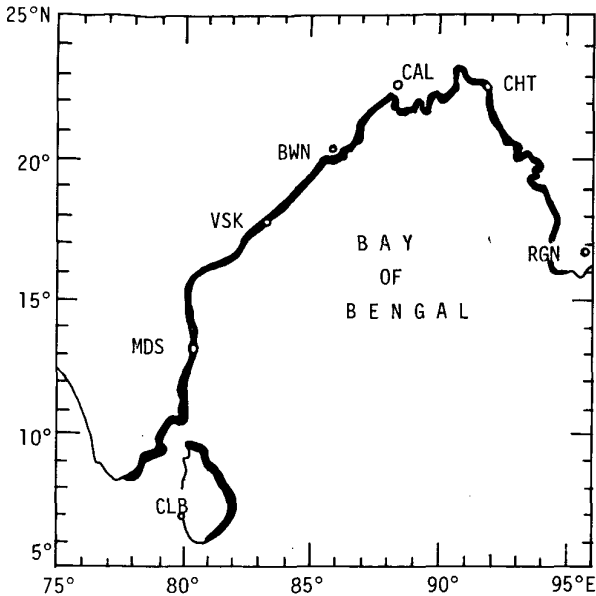


Fig. 1. Map showing the Bay of Bengal. The coastline considered is shown in thick line.

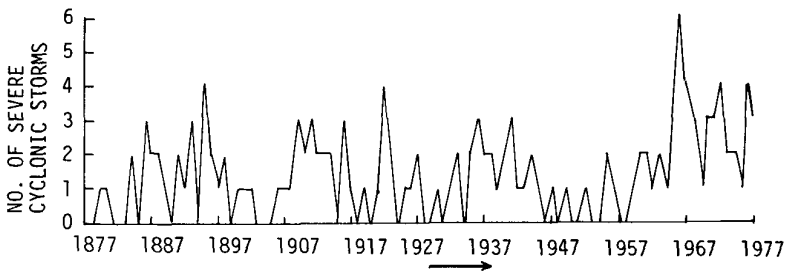


Fig. 2. Number of severe cyclonic storms striking coastal belt around Bay of Bengal.

period 1965-1977 are much higher. Table I gives the mean and variance for the periods 1891-1964, 1877-1964, 1965-1977 and 1877-1977. It is clearly seen that the mean and the variance for the period, 1965-77 are generally more than two times of those for the long periods. In fact, even if we take any 13-year period prior to 1965, and obtain mean and variance for the same it is seen that for the period 1886-1898, the mean is highest, being 1.78, and for the period 1910-1922, the variance is highest, being 1.81. It can thus be inferred that the period 1965-77 is characterised by a much higher mean and variance of the number of severe storms crossing the coast

as compared to those for any 13-year period prior to 1965. The period 1965-1977 is thus very unusual in the record of severe cyclonic storms striking the coast.

TABLE I

Mean and variance of the number of severe cyclonic storms striking the coast around Bay in a year.

Period	Mean	Variance
1891-1964	1.22	1.087
1877-1964	1.16	1.091
1965-1977	3.00	2.583
1877-1977	1.40	1.530

TABLE II

Number of cyclonic storms which formed over the Bay, the number which intensified into severe storms over the Bay and the number which struck the coast as severe cyclonic storms during 13-year periods.

Period	Number of storms which formed over Bay	Number of storms which intensified into severe storms over Bay	Number of severe storms which struck coast	Efficiency of intensification of storms into severe storms over Bay	Ratio of severe storms which struck coast to storms which formed over Bay
1877-1899	49	13	12	0.26	0.24
1890-1902	56	19	18	0.34	0.32
1903-1915	64	19	17	0.30	0.27
1916-1928	61	19	16	0.31	0.26
1929-1941	70	24	19	0.34	0.27
1942-1954	46	13	8	0.28	0.17
1952-1964	45	21	13	0.47	0.29
1965-1977	70	44	39	0.63	0.56
1882-1894	71	16	16	0.23	0.23
1884-1896	72	22	22	0.31	0.31
1886-1898	74	24	23	0.32	0.31
1924-1936	74	16	13	0.22	0.18
1932-1944	71	27	22	0.38	0.31

The number of severe storms crossing the coast in different 13-year periods has been examined with reference to the number of storms which formed over the Bay and the number of cyclonic storms which intensified into severe cyclonic storms over the Bay. Table II gives these figures for the successive 13-year periods and for the periods 1952-1964 and 1965-1977. The table also gives information for other 13-year periods for which the number of storms which formed over the Bay was higher than that which formed during the period 1965-1977. It also gives the efficiency of intensification of storms into severe storms over Bay and the ratio of severe

storms which struck coast to the storms which formed over the Bay. A careful examination of Table II brings out that prior to 1965 there were a number of 13-year periods when the number of storms which formed over the Bay was nearly equal to or higher than that which formed over the Bay during the period 1965-1977, and as such there is nothing unusual about the number of storms which formed over the Bay during the 13-year period 1965-1977. However, the efficiency of intensification of storms into severe storms over the Bay, which generally varied from 0.25 to 0.35 for the different 13-year periods prior to 1965 increased sharply to 0.63 during the period 1965-1977. This would suggest that the meteorological conditions over and near the Bay during the period 1965-1977 were more often favourable for intensification of storms into severe storms. The values in the last column clearly show that the number of severe storms which struck the coast during the period 1965-1977 was unusually large. The period 1965-1977 is thus very unusual from the viewpoint of intensification of storms into severe storms of the Bay of Bengal.

3. RANDOMNESS OF SEVERE STORM STRIKING THE COAST

To examine whether the striking of the coast by a severe storm is a random or a non-random event, Mann-Kendall rank statistic test for randomness and Swed and Eisenhart's runs test for runs above and below the median to detect trend or oscillation, as recommended by WMO (1966), were applied to the time interval between successive epochs of severe storms striking the coast. The tests were applied to (i) the whole series of 140 intervals, (ii) two equal components of the whole series, (iii) four equal components of the whole series, and (iv) two samples each of size 24 and commencing from two randomly chosen epochs of severe storms striking the coast. The results of these tests are given in Table III. It can be seen from the results of Mann-Kendall test that the latter half of the interval series only shows non-randomness significant at 5 percent level, the value of the test statistic being slightly smaller than the value significant at 1 percent level. The Swed and Eisenhart's test for the same sample shows a value not significant but close to 95 percent confidence value, suggesting oscillation. However, no non-randomness is indicated by the two tests for the sample consisting of the last 35 intervals. This is perhaps due to the fact that this sample is for the period 3 January 1966 to 19 November 1977 which is contained in the period of unusually higher mean and variance, 1965-1977. The non-randomness significant at 5 percent level has been introduced by the period 1965 to 1977 which has mean and variance much higher than those for the period prior to 1965.

The series of time intervals between successive severe storms striking the coast during the period 1877 to 1964 does not reveal any significant non-randomness. The mean and the variance have increased rather sharply after 1964.

TABLE III

Results of Mann-Kendall rank statistic test and Swed and Eisenhart's runs test applied to time interval between successive severe cyclonic storms crossing the coast.

Sample Size	Period	Mann-Kendall test		Swed & Eisenhart's test	
		τ	$\tau_{0.95}$	R	$R_{0.05} / R_{0.95}$
140	1877-1977	-0.125*	± 0.112	77	60/81
70	20 May 1879	0.006	± 0.160	36	28/43
	-27 May 1936				
70	27 May 1936	-0.214*	± 0.160	42	28/43
	-19 Nov. 1977				
35	20 May 1879	-0.044	± 0.232	19	12/23
	- 6 Dec. 1909				
35	6 Dec. 1909	0.136	± 0.232	20	12/23
	-27 May 1936				
35	27 May 1936	-0.025	± 0.232	18	12/23
	- 3 Jan. 1966				
35	3 Jan. 1966	0.069	± 0.232	19	12/23
	-19 Nov. 1977				
24	23 Sept. 1911	0.065	± 0.286	12	8/17
(Random)	-29 Nov. 1930				
24	4 Oct. 1936	0.220	± 0.286	14	8/17
(Random)	-31 Oct. 1960				

τ is Mann-Kendall test statistic, R is number of runs above and below the median. Suffix to these statistics denotes the level of confidence. Asterisk denotes significance at 5 percent level.

4. PROBABILITY MODEL

The probability of a severe storm striking the coast on any day is very low. In this situation, when we consider one-year period, we may expect that Poisson distribution may fit data on severe storms. Thom (1966) has given a criterion for adequacy of the Poisson distribution. According to this criterion, if $P(X_{n-1}^2 > X_{n-1}^2) > 0.05$, where n is the number of years of data, $X_{n-1}^2 = n \frac{\sum Y^2}{\sum Y} - \sum Y$, and Y is the number of severe storms in a year, then Poisson distribution is adequate. The Poisson distribution is characterised by the property that its mean and variance are equal. The values of X_{n-1}^2 for the data for the periods 1891-1964, 1877-1964, and 1877-1977 are 63.8, 80.7 and 109.3 respectively and the corresponding values of $P(X_{n-1}^2 > X_{n-1}^2)$ are 0.75, 0.67 and 0.25 respectively. The criterion is thus satisfied and Poisson distribution is adequate for all the three periods. Table I also shows that mean and variance are fairly close to each other. The Poisson probability model was, therefore, fitted to the data for the three periods 1891-1964, 1877-1964 and 1877-1977, and the goodness-of-fit was tested by Chi-square test. Table IV shows the goodness-of-fit. The fit is very good for all the three periods; however, it appears to be slightly better for the periods 1891-1964 and

1877-1977. The number of severe cyclonic storms striking the coast in a year is thus a Poisson-distributed variable. On the basis of the Poisson model, the probabilities of one, two, three, four severe storms striking the coast around the Bay in a year are, 0.345, 0.242, 0.113, and 0.039 respectively.

TABLE IV

Goodness-of-fit of Poisson probability model to the number of severe cyclonic storms striking the coast around the Bay of Bengal in a year for different periods.

Period	No. of severe storms striking coast in a year	Observed frequency	Frequency on Poisson hypothesis	Contribution to Chi-square
1891-1964	0	21	21.78	0.028
	1	26	26.67	0.017
	2	19	16.30	0.447
	3	6	6.73	0.079
	4	2	2.0	0.107
	≥ 5	0	0.52	
<hr/>				
$P(\chi^2 > 0.678) = 0.88; \chi^2 = 0.678 \quad (\text{d.f.3})$				
1877-1964	0	28	27.58	0.006
	1	29	32.00	0.281
	2	22	18.57	0.634
	3	7	7.18	0.004
	4	2	2.08	0.168
	≥ 5	0	0.59	
<hr/>				
$P(\chi^2 > 1.093) = 0.77; \chi^2 = 1.093 \quad (\text{d.f.3})$				
1877-1977	0	28	25.01	0.358
	1	31	34.90	0.436
	2	24	24.36	0.005
	3	12	11.35	0.037
	4	5	3.96	0.071
	≥ 5	1	1.42	
<hr/>				
$P(\chi^2 > 0.907) = 0.83; \chi^2 = 0.907 \quad (\text{d.f.3})$				

The fit of the Poisson distribution to the number of severe storms striking the coast in a season, and to the number of severe storms striking a section of the coast in a year has also been examined. The season considered is October-November and the section considered is Bangla Desh - North Arakan Coast (from 22°N 89°E to 20°N 93°E). The results are given in Table V. The fit is seen to be very good. Poisson distribution can thus be applied to severe storms striking the whole coast or a section of the coast in a year or a season.

TABLE V

Goodness-of-fit of Poisson distribution to the number of severe cyclonic storms striking (A) the coast during October-November, (B) Bangla Desh - North Arakan Section of the coast in a year.

Period	Coast	No. of severe cyclonic storms striking	Observed frequency	Frequency on Poisson hypothesis
October - November	Whole Coast	0	57	54.24
		1	30	33.73
		2	10	10.50
		3	3 }	2.53
		≥ 4	1 }	
<hr/>				
$P(\chi^2 > 1.423) = 0.56; \chi^2 = 1.423 \quad (\text{d.f.2})$				
Year	Bangla Desh - North Arakan (22°N 89°E -20°N 93°E)	0	69	67.97
		1	25	26.97
		2	6	5.35
		≥ 3	1	0.71
		<hr/>		
$P(\chi^2 > 0.306) = 0.67; \chi^2 = 0.306 \quad (\text{d.f.1})$				

5. CONCLUDING REMARKS

The number of severe storms striking the coast around the Bay or a section of the same, in a year or a season, is a Poisson-distributed variable. The probabilities obtained on the basis of the Poisson probability model could be used for planning funds to mitigate the hardships resulting from these natural calamities.

ACKNOWLEDGEMENT

The author is grateful to the Director for the Facilities to pursue this work. He would like to convey his thanks to Mrs. S.P. Lakade for typing the manuscript.

REFERENCES

- Das, P.K., George C.A. and Jambunathan R., 1972. Cyclones and depressions of 1971- Bay of Bengal and Arabian Sea. Ind. J. Met. and Geophys., 23: 453-466.
- Das, P.K., George C.A. and Jambunathan R., 1973. Cyclones and depressions of 1972- Bay of Bengal and Arabian Sea. Ind. J. Met. and Geophys., 24: 327-344.
- Eliot J., 1885. Account of the southwest monsoon storms generated in the Bay of Bengal during the period 1877-1881. Indian Met. Memoirs, Vol. II.
- Eliot J., 1888. List and brief account of the southwest monsoon storms generated in the Bay of Bengal during the period 1882-1886. Indian Met. Memoirs, Vol. IV.

- George Alexander, George C.A. and Jambunathan R.,1974. Cyclones and depressions of 1973- Bay of Bengal and Arabian Sea. Ind. J. Met. and Geophys., 25: 347-362.
- George Alexander, Bhaskar Rao,N.S. and Jambunathan R.,1976. Cyclones and depressions of 1974- Bay of Bengal and Arabian Sea. Ind. J. Met., Hydrol. and Geophys., 27: 113-126.
- George Alexander, Srinivasan V. and Jambunathan R.,1977. Cyclones and depressions of 1975- Bay of Bengal and Arabian Sea. Ind. J. Met., Hydrol. and Geophys., 28: 3-20.
- India Met.Dept.,1964. Tracks of storms and depressions over the Bay of Bengal and the Arabian Sea (1877-1960), and subsequent supplement to this (1961-70).
- Pant P.S., Srinivasan V., Ramkrishnan A.R. and Jambunathan R.,1978. Cyclones and depressions in the Indian Seas, in 1976. Ind. J. Met., Hydrol. and Geophys., 29: 613-628.
- Thom H.C.S.,1966. Some methods of climatological analysis. Technical Note No. 81, WMO - No. 199 T.P. 103: 31.
- WMO,1966. Climatic change. Technical Note No. 79, WMO - No. 195 T.P. 100, 80 pp.

RAINFALL INTENSITY-DURATION-RETURN PERIOD EQUATIONS AND NOMOGRAPHS OF INDIA

RAM BABU, K.G.TEJWANI, M.C.AGARWAL and L.S.BHUSHAN

Central Soil & Water Conserv. Research & Training Inst., Dehra Dun (India).

ABSTRACT

Babu,R.,Tejwani,K.G.,Agarwal,M.C. and Bhushan,L.S., Rainfall intensity-duration-return period equations and nomographs of India. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

Rainfall intensity-duration-return period equations and nomographs are required for design of soil conservation and runoff disposal structures and for planning flood control projects. Rainfall intensity-duration-return period equations and nomographs have been developed for 42 stations situated in the northern, central, eastern, western and southern zones of India. With the help of either the equations and nomographs, the intensity for any desired duration and return period (or frequency) may be determined. The deviation in the rainfall intensity obtained by the two methods (equations and nomographs) has been observed to be less than 8%. Looking into simplicity in use, quickness and precision in results obtained, the nomographs appear to be the most handy tool for field workers.

The zonal equations and nomographs for northern, central, eastern, western and southern zones of India were also developed. The zonal equations and nomographs developed for various zones compared fairly well with the equations and nomographs for individual stations falling in that zone. Only in sporadic cases, the variations in estimated rainfall intensities by use of the station equation and the zonal equation was noticed up to $\pm 50\%$ in northern, eastern and southern zones. While in 37 stations out of 42 the deviation was below 30% suggesting the usefulness of zonal equations and nomographs.

INTRODUCTION

Rainfall is one of the most important factors responsible for soil erosion. The characteristics of rain storms amount, intensity and duration play an important role in determining the rate of soil erosion. Greater is the intensity of rainfall, greater kinetic energy it possesses. The kinetic energy of rainfall dislodges soil particles and splashes them in suspension in runoff. Among other factors the amount of runoff is determined by rainfall intensity, duration and amount. A rainfall of longer duration reduces the infiltration capacity of soil. As a result a long duration rain storm produces considerable runoff regardless of its intensity. The capacity of a runoff conveyance system is usually based on a certain depth of rainfall to be expected during a selected period of time. Farm terraces, culverts, bridges and flood control structures are thus designed on the basis of safely conveying runoff expected from rain storms of specified frequency, intensity and duration.

The significance of rainfall intensity, duration and frequency analysis is also important from economic considerations. An over designed structure involves excessive cost and under designed structure will be unsafe and also involves high recurring expenditure on repair, maintenance and replacement. An intermediate design would provide a structure with reasonable initial and maintenance cost values.

Rainfall intensity-duration-return period equations and nomographs on regional basis are required in the country for design of soil conservation and runoff disposal structures and for planning flood control projects. Such relationships and nomographs have been developed at a few stations scattered over one or other part of the country (Gupta et al. (1975) , Raghunath et al. (1969), Khullar et al. (1975) and Senapati et al. (1976)) but no serious efforts have been made to develop such a tool for a region or the country as a whole.

For understanding the rainfall characteristics of a station long period records of automatic rain gauge are needed. At present such records are available for only a limited number of stations. If general relationships and nomographs could be developed for various zones in India (northern, central, eastern, western and southern), they may prove to be reliable in determining intensity, duration and frequency of rainfall of a particular station in these zones for design purposes.

Based on observed data for 42 stations, the intensity-duration-return period equations and nomographs for individual station as well as for different zones of India (northern, central, eastern, western and southern) have been developed and discussed.

DATA COLLECTION

To derive prediction equations for intensity-duration-frequency and also for development of nomographs, the continuous recorded rainfall data were collected from the Indian Meteorological Department, Poona for 39 stations situated in northern, central, eastern, western and southern zones of India. Due to the non-availability of data for long periods for all the stations under study, 15 years records for 35 stations and 9-13 years for 4 stations, have been used for various durations. The rainfall intensity values for Dehra Dun (Gupta et al. (1969)) and Agra (Tejwani et al. (1975)) were taken from the published record. The Indore data was obtained through Indo-UK Dry Farming Project (ICAR) , Indore. The locations of these stations along with zonal boundaries are shown in Fig. 1. The data for all the stations was tested for reliability using the procedure of Ogrosky and Mockus (1957) and it was observed that the length of record for all the stations are adequate and hence could be used for frequency analysis.

ANALYTICAL PROCEDURE

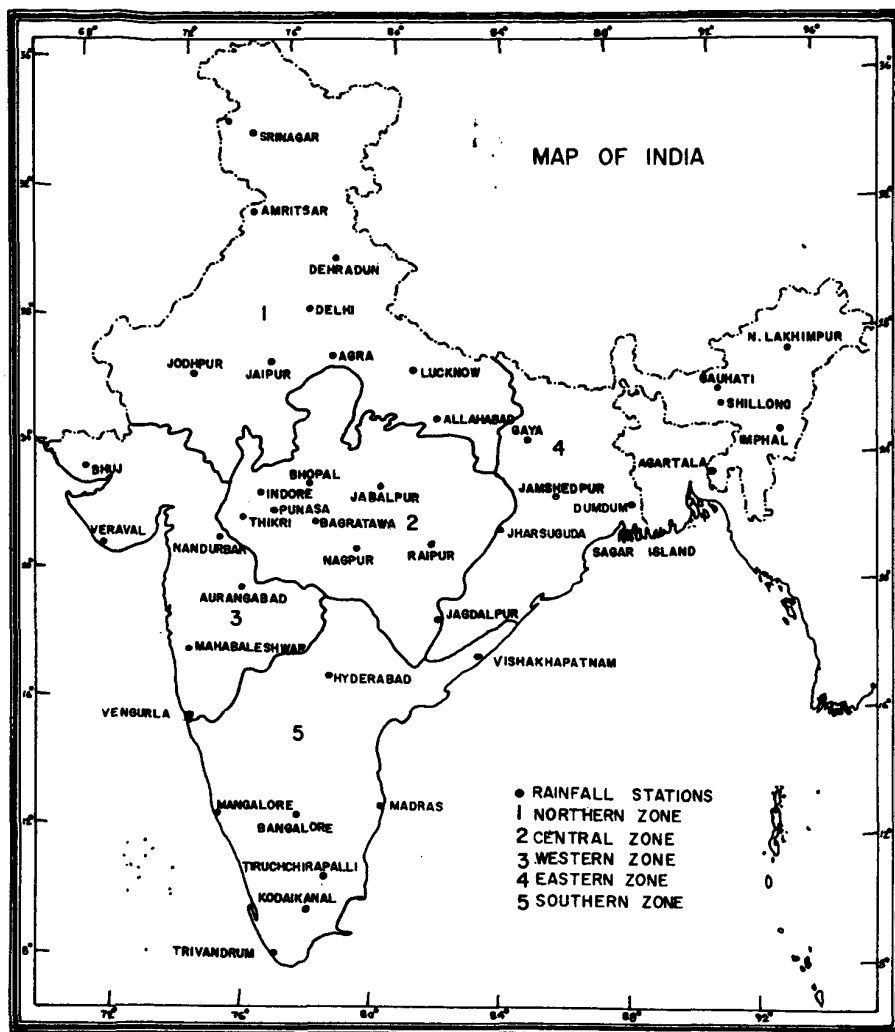


Fig. 1. Map of India showing the locations of the recording rain gauge stations and zonal boundaries.

Various formulae have been advanced connecting the three parameters – rainfall intensity, duration and return period (Frevert et al. (1955), Linsley et al. (1949), Skurlow (1960), Nemec (1973), Gupta et al. (1969), Raghunath et al. (1969), Khullar et al. (1975) and Senapati et al. (1976)). The formula is of the general form:

$$I = KT^a / (t+b)^d \quad (1)$$

where I = Intensity of rainfall (cm/hr), T = Return period (years), t = Duration

(hours); K , a , b and d are constants.

Eqn. 1 was used for developing intensity-duration-frequency relationships.

Method of Frequency Analysis and Development of Frequency Lines

Various methods have been proposed for frequency analysis and there are several theoretical interpretations or reasonings for the preference of one method or the other (Chow(1964)). Mathematical or graphical methods are generally used for frequency analysis. When the records are of short duration, the sampling error would be large. A rigid mathematical treatment is not justified when the data are available for less than 30 years (Dalrymple(1960)). As our data are of short period of about 15 years, graphical methods have been employed. Gumbel extreme value technique was applied for computation of return period values and the frequency lines were plotted after computing the plotted points by 'Computed method' suggested by Ogrosky and

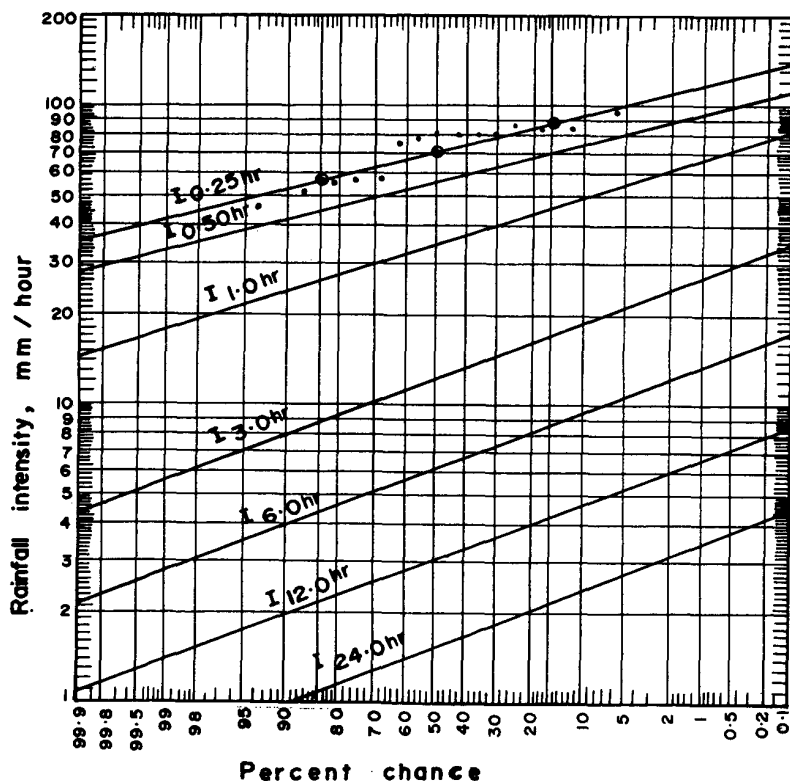


Fig. 2. Frequency distribution of rainfall intensities for various durations - Hyderabad.

Mockus (1957). Frequency lines for 15 minutes, 30 minutes, 1 hour, 3 hours, 6 hours, 12 hours and 24 hours intensity data were developed and plotted on log-normal probability paper (see Fig. 2).

Deriving Equation for Intensity-Duration-Return Period

The intensity-duration-return period equation, eqn.1 can be expressed by taking logarithms on both sides as:

$$\log I = \log K + a \log T - d \log(t+b) \quad (2)$$

or

$$\log I = \log K_1 - d \log(t+b) \quad (3)$$

where

$$\log K_1 = \log K + a \log T. \quad (4)$$

In order to evaluate the coefficients a , b , d and K from general expression for frequency curves, the following steps are involved:

Step I . On log-log paper the values of rainfall intensity for each individual duration were plotted on the y-axis and the return period (or recurrence interval) in years on x-axis (fig. 3). Points were connected by dotted lines for each duration.

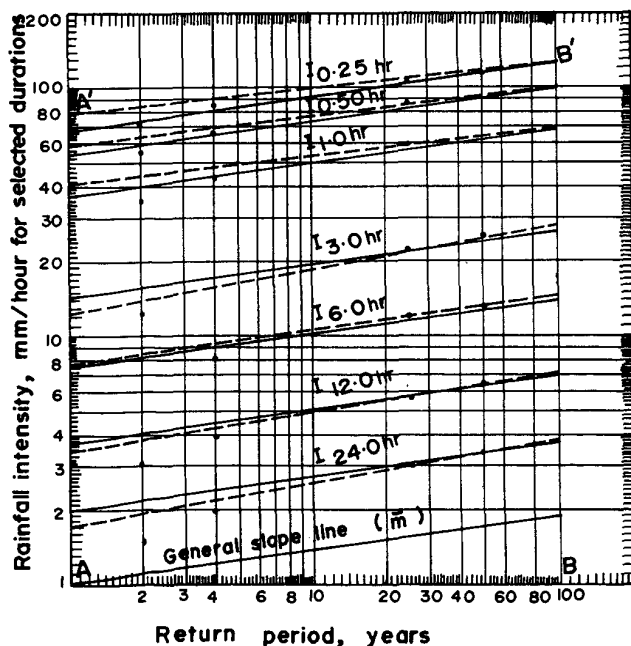


Fig. 3. Rainfall intensities for selected durations and return periods — Hyderabad.

The dotted lines were extended to cut the y-axis against one year recurrence interval.

Step II. The intensity values, I_t at different values of t equal to 2, 5, 10, 25, 50 and 100 years recurrence intervals for each duration were read from Fig. 3.

The mathematical relationship between T and various values of I_t is given by

$$\log I_t = m \log T + C \quad (5)$$

where I_t = maximum intensity for duration t , T = recurrence interval, m = frequency factor for each line (i.e., slope of the frequency line) and C = intercept on y-axis at $T = 1$. These interval equations define the intensity-frequency relationship for any selected duration.

Step III. Then slope (m) of eqn.5 for the durations were determined and then their geometric mean (\bar{m}) was computed. The slope of line (m) represents the exponent a in eqn.1. The geometric mean slope \bar{m} thus determined represents actually T^a in eqn.1.

Step IV. A line representing geometric mean slope, \bar{m} , was drawn (Fig.3) at the base through the origin; solid lines parallel to this mean slope were drawn to have the lines as close as possible to points between 10 to 100 years return periods extending them to cut the y-axis. Rainfall intensities against one year return period for all the selected durations were then read on the y-axis.

Step V. Values of intensity of one year recurrence interval were plotted on the y-axis with selected durations (t) on the x-axis on log-log paper (Fig.4). Since the points so plotted did not fall on a straight line, a suitable constant (b) to time t was added. Thus the equation become:

$$I = K / (t+b)^d \quad (6)$$

This was done by trial and error method so that the deviations were minimum (Fig.4).

Step VI. The eqn.3 written in its logarithmic form is:

$$\log I = \log K - d \log(t+b) \quad (7)$$

or

$$\log I - \log K + d \log(t+b) = 0 \quad (8)$$

The constants K and d in eqn.8 were then solved by the method of least squares; They may be obtained by solving the eqns. 9 and 10:

$$\log K = \frac{\sum \log I \cdot \sum [\log(t+b)]^2 - \sum [\log I \cdot \log(t+b)] \cdot \sum \log(t+b)}{n \sum [\log(t+b)]^2 - [\sum \log(t+b)]^2} \quad (9)$$

and

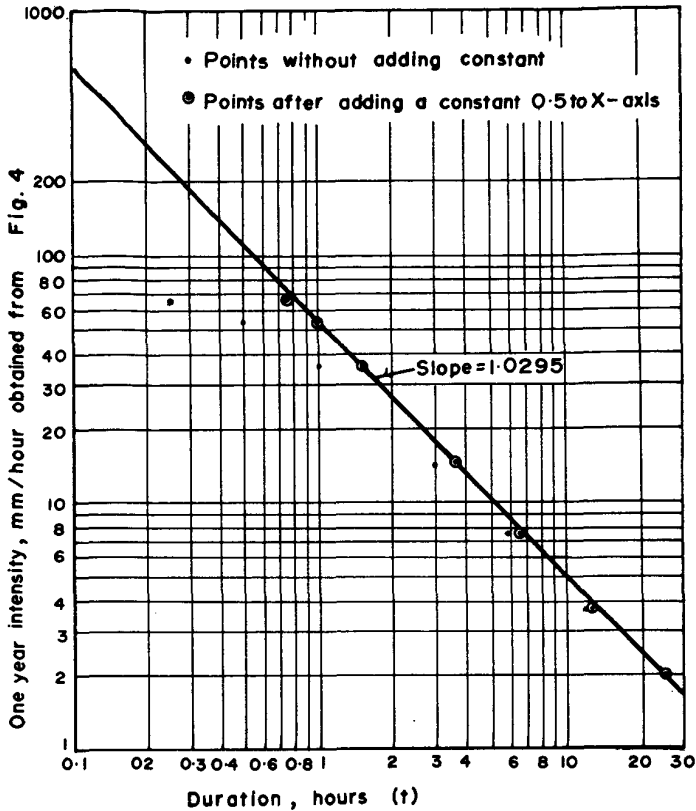


Fig. 4. Fitting of constants b and d in the equation $I = 52.5/(t+b)^d$ - Hyderabad.

$$d = \frac{\sum \log I \cdot \sum \log(t+b) - n \sum [\log I \cdot \log(t+b)]}{n \sum [\log(t+b)]^2 - [\sum \log(t+b)]^2} \quad (10)$$

Thus all the parameters a (Step III), b (Step V) and K and d (Step VI) become known for eqn.1.

Step VII. At this stage, frequency factor T^a obtained in Step III above was included to give finally the intensity-duration-frequency or return period formula

$$I = \frac{K}{(t+b)^d} T^a \quad (11)$$

Development of Nomograph

A nomograph is an alignment chart consists of a set of parallel scales which are

suitably graduated. In the present study, there were only three variables and thus the alignment chart had three parallel scales so graduated that a line which joins values on two scales will intersect the third scale at a value which satisfies the given equation.

In order to design alignment charts for equation of the form

$$f_1(u) + f_2(v) = f_3(w) , \quad (12)$$

the following are required:

- (a) the graduation of scales, which are marked with the values of the variable and on which the distances to the graduations are laid off the proportion to the corresponding values of the function of the variables, and
- (b) the determination of spacing of the parallel scale. The scale equation for determining functional modulus m which is commonly defined as a proportionality multiplier used to bring a range of values of particular function with a selected length for a scale which is given as:

$$m = L / (f(u_2) - f(u_1)) \quad (13)$$

where m = calculated functional modulus, L = length of the scale chosen, $f(u_2)$ = upper limit of the function and $f(u_1)$ = lower limit of the function.

The unknown functional modulus m_w was calculated by

$$m_w = m_u m_v / m_u - m_v \quad (14)$$

where m_u and m_v are the calculated functional moduli.

$$\text{Scale spacing ratio} = m_u / m_v \quad (15)$$

was determined with the help of the equation

$$I = K T^a / (t+b)^d . \quad (16)$$

The limiting values of intensity were determined on the basis of conditions laid down on t and T .

RESULTS AND DISCUSSION

Mathematical Equations

Following the procedure as discussed above, the intensity-duration-frequency relationships for 9 stations of northern zone, 9 stations of central zone, 10 stations of eastern zone, 6 stations of western zone and 8 stations of southern zone of India were developed and are reported in Table 1. The precision of these equations could be recognized after verifying the reliability of any one of the station equation. For example, for Hyderabad, the maximum percent deviation between the rainfall intensity values obtained from developed equation $I = 5.25 T^{0.1354} / (t+0.50)^{1.0295}$ and the observed values obtained from frequency lines from primary data for various

durations and 10,15 and 50 years frequencies ranged from -6.9% to 5.3% , which is quite low (Table 2). Notwithstanding the inherent weakness of an average equation, the developed equations seems to be quite reliable and may be used with confidence.

TABLE 1.

Intensity-duration-return period relationships - India.

<u>Northern Zone</u>			
Station	Equation	Station	Equation
Agra	$I = \frac{4.911 T^{0.1667}}{(t+0.25)^{0.6293}}$	Jodhpur	$I = \frac{4.098 T^{0.1677}}{(t+0.5)^{1.0369}}$
Allahabad	$I = \frac{8.570 T^{0.1692}}{(t+0.5)^{1.0190}}$	Lucknow	$I = \frac{6.074 T^{0.1813}}{(t+0.5)^{1.0331}}$
Amritsar	$I = \frac{14.41 T^{0.1304}}{(t+1.4)^{1.2963}}$	New Delhi	$I = \frac{5.208 T^{0.1574}}{(t+0.5)^{1.1072}}$
Dehra Dun	$I = \frac{6.00 T^{0.2200}}{(t+0.5)^{0.8000}}$	Srinagar	$I = \frac{1.503 T^{0.2730}}{(t+0.25)^{1.0636}}$
Jaipur	$I = \frac{6.219 T^{0.1026}}{(t+0.5)^{1.1172}}$	Northern zone	$I = \frac{5.9143 T^{0.1623}}{(t+0.5)^{1.0127}}$
<u>Central Zone</u>			
Bagra-tawa	$I = \frac{8.5704 T^{0.2214}}{(t+1.25)^{0.9331}}$	Nagpur	$I = \frac{11.4500 T^{0.1560}}{(t+1.25)^{1.0324}}$
Bhopal	$I = \frac{6.9296 T^{0.1892}}{(t+0.50)^{0.8767}}$	Punasa	$I = \frac{4.7011 T^{0.2608}}{(t+0.5)^{0.8653}}$
Indore	$I = \frac{6.9280 T^{0.1394}}{(t+0.50)^{1.0651}}$	Raipur	$I = \frac{4.6830 T^{0.1398}}{(t+0.15)^{0.9284}}$
Jabalpur	$I = \frac{11.3790 T^{0.1746}}{(t+1.25)^{1.1206}}$	Thikri	$I = \frac{6.0880 T^{0.1747}}{(t+1.00)^{0.8587}}$
Jagadalspur	$I = \frac{4.7065 T^{0.1746}}{(t+0.25)^{0.9902}}$	Central zone	$I = \frac{7.4645 T^{0.1712}}{(t+0.75)^{0.9599}}$
<u>Western Zone</u>			
Aurangabad	$I = \frac{6.018 T^{0.1459}}{(t+0.50)^{1.0923}}$	Nandurbar	$I = \frac{4.254 T^{0.2070}}{(t+0.25)^{0.7704}}$
Bhuj	$I = \frac{3.823 T^{0.1267}}{(t+0.50)^{1.0923}}$	Vengurla	$I = \frac{6.863 T^{0.1670}}{(t+0.75)^{0.8683}}$
Mahabaleshwar	$I = \frac{3.483 T^{0.1267}}{(t+0.0)^{0.4853}}$	Veraval	$I = \frac{7.787 T^{0.2084}}{(t+0.50)^{0.8908}}$
		Western zone	$I = \frac{3.974 T^{0.1647}}{(t+0.15)^{0.7327}}$

(table 1 cont'd.)

<u>Eastern Zone</u>			
Agartala	$I = \frac{8.097 T^{0.1177}}{(t+0.50)^{0.8191}}$	Jamshedpur	$I = \frac{6.930 T^{0.1337}}{(t+0.50)^{0.8737}}$
Dum Dum	$I = \frac{5.940 T^{0.1150}}{(t+0.15)^{0.9241}}$	Jharsuguda	$I = \frac{8.596 T^{0.1392}}{(t+0.75)^{0.8740}}$
Gauhati	$I = \frac{7.206 T^{0.1557}}{(t+0.75)^{0.9401}}$	North Lakhimpur	$I = \frac{14.070 T^{0.1256}}{(t+1.25)^{1.0730}}$
Gaya	$I = \frac{7.176 T^{0.1483}}{(t+0.50)^{0.9459}}$	Sagar Island	$I = \frac{16.524 T^{0.1402}}{(t+1.50)^{0.9635}}$
Imphal	$I = \frac{4.939 T^{0.1340}}{(t+0.50)^{0.9719}}$	Shillong	$I = \frac{6.728 T^{0.1502}}{(t+0.75)^{0.9575}}$
		Eastern zone	$I = \frac{6.933 T^{0.1353}}{(t+0.50)^{0.8801}}$

<u>Southern Zone</u>			
Bangalore	$I = \frac{6.275 T^{0.1262}}{(t+0.50)^{1.1280}}$	Mangalore	$I = \frac{6.744 T^{0.1395}}{(t+0.50)^{0.9347}}$
Hydrabad	$I = \frac{5.250 T^{0.1354}}{(t+0.50)^{1.0295}}$	Tiruchirapalli	$I = \frac{7.136 T^{0.1638}}{(t+0.50)^{0.9624}}$
Kodaikanal	$I = \frac{5.914 T^{0.1711}}{(t+0.50)^{1.0086}}$	Trivandrum	$I = \frac{6.762 T^{0.1536}}{(t+0.50)^{0.8159}}$
Madras	$I = \frac{6.126 T^{0.1664}}{(t+0.50)^{0.8027}}$	Vishakhapatnum	$I = \frac{6.646 T^{0.1692}}{(t+0.50)^{0.9963}}$
		Southern zone	$I = \frac{6.311 T^{0.1523}}{(t+0.50)^{0.9465}}$

(I = intensity (cm/hr); T = return period (year); t = duration (hour)).

Further on the basis of equations for individual stations, zonal equations were also developed. From the equations of individual stations, the intensity for any desired duration and frequency (or return period, or recurrence interval) can be determined for that location and the zonal equation may be used for any location falling in that zone.

Nomographs

On the basis of intensity-duration-frequency relationships developed for 42 stations (Table 1) situated in northern, central, eastern, western and southern zones of India nomographs were prepared for all these stations. A nomograph of one such station (Hyderabad) has been shown in Fig.5. From the nomographs, the rainfall intensity for any desired duration between 10 to 100 years frequency (or return

TABLE 2.

Comparison among calculated and observed intensities of rainfall (cm/hr) and the present deviation - Hyderabad.

Duration mins; hrs	i_{cal}			i_{obs}			σ_i		
	frequency, years			frequency, years			frequency years		
	10	25	50	10	25	50	10	25	50
15 mins.	9.6	10.9	12.0	9.6	11.0	11.5	0.0	-0.9	4.3
30 mins.	7.2	8.1	8.9	7.6	8.6	9.3	-5.3	-5.8	-4.3
1 hr.	4.7	5.4	5.9	5.0	5.8	6.3	-6.0	-6.9	-6.3
3 hrs.	2.0	2.2	2.5	1.9	2.2	2.5	5.3	0.0	0.0
6 hrs.	1.0	1.2	1.3	1.0	1.2	1.3	0.0	0.0	0.0

i_{cal} = calculated intensity of rainfall (cm/hr) from developed equation

i_{obs} = observed intensity of rainfall (cm/hr) from the frequency lines from primary data

σ_i = percent deviation of observed values from the frequency lines from those calculated by the developed equation

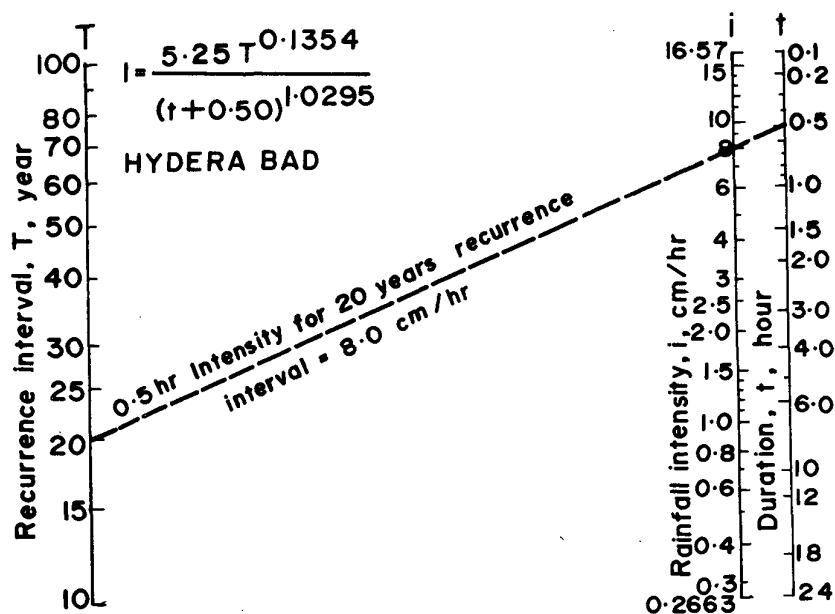
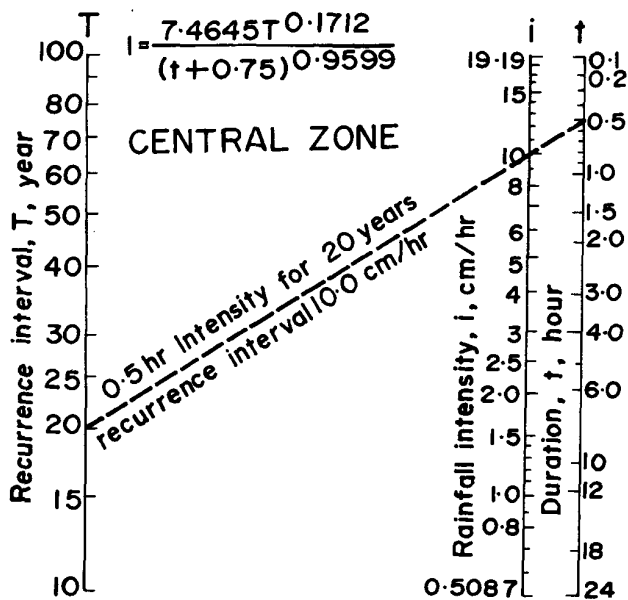
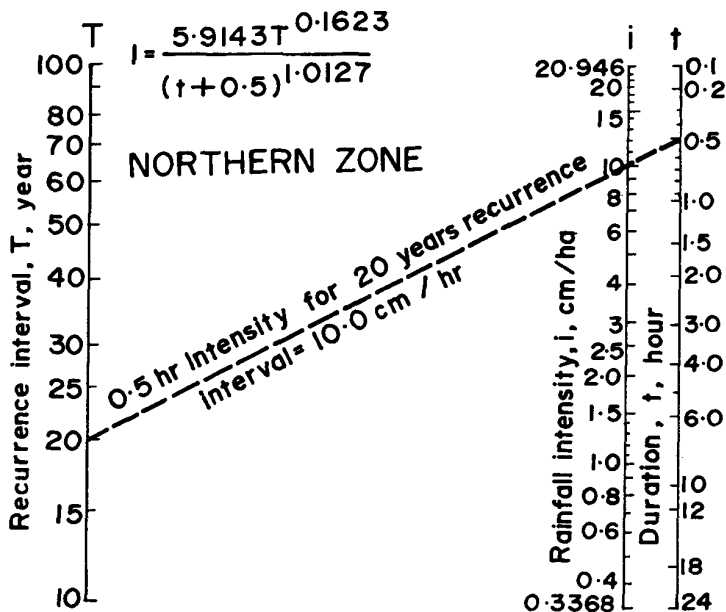
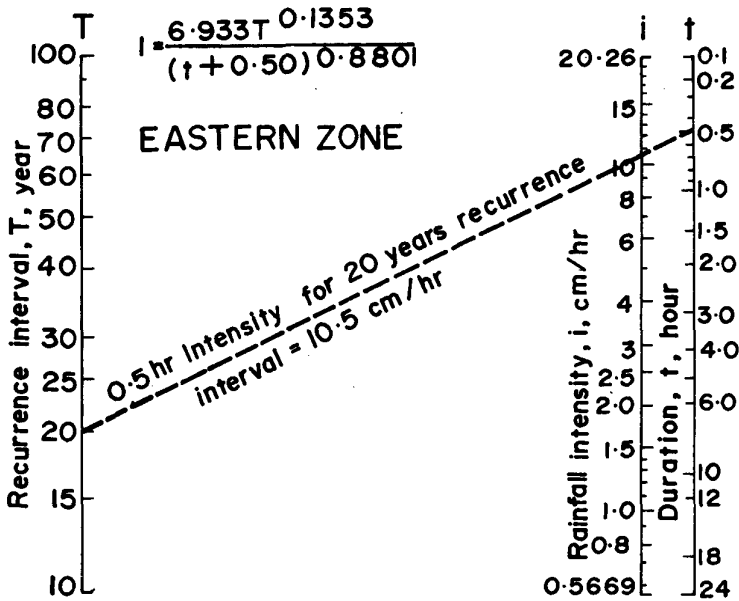
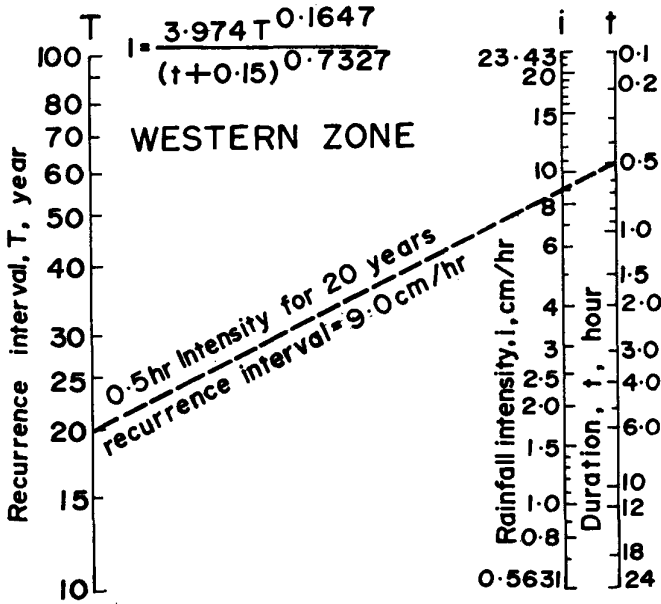


Fig. 5. Nomograph for solving intensity-duration-return period for recurrence interval equation, $I = 5.25 T^{0.1354} / (t+0.50)^{1.0295}$ - Hyderabad.

period) could be directly read for that location. Zonal nomographs for all the five zones (Fig.6) were also developed which may be used for determining intensity for any duration and recurrence interval for any location in the zone.





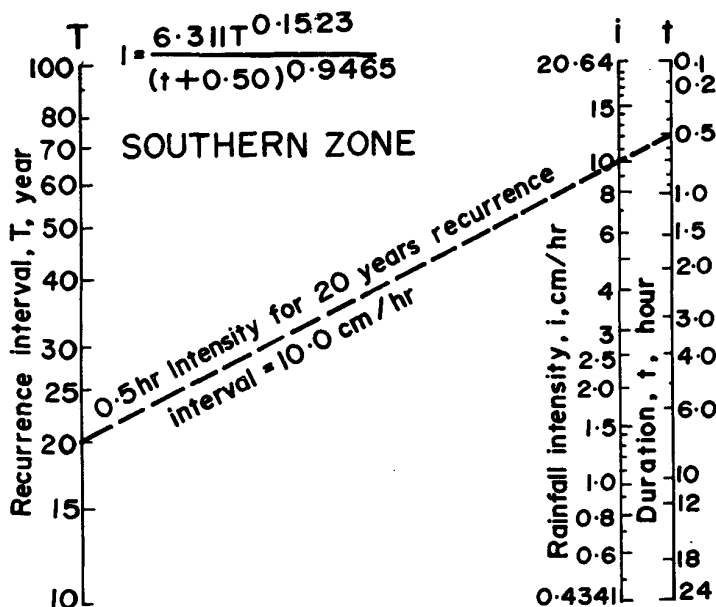


Fig. 6. Nomographs of intensity-duration-return period (or recurrence interval) equations for Northern, Central, Western, Eastern and Southern zones of India.

Comparison of Mathematical and Nomographic Solutions

Percent deviation of rainfall intensity values observed from nomographs and those calculated from corresponding mathematical equations for various durations and 10, 25, and 50 years frequencies showed that maximum deviation between the nomographic solutions and mathematical equations ranges from -5.3% to 3.5% in case of Allahabad (northern zone); -5.3% to 4.3% in case of Jabalpur (central zone); -7.9% to 5.9% in case of Dum Dum (eastern zone); -5.0% to 7.7% in case of Mahabaleshwar (western zone) and -6.9% to 8.0% in case of Kodaikanal (southern zone). The deviations for other 37 stations were still less.

While comparing the rainfall intensity values at various frequencies and durations obtained from the developed equations and from the observed values obtained from the probability charts, it is observed that the maximum deviations range from -18.2% to 14.3% for Amritsar (northern zone); -15.5% to 19.6% for Raipur (central zone); -17.6% to 13.6% for Imphal (eastern zone); -14.6% to 16.7% for Nandurbar (western zone) and -15.0% to 19.0% for Kodaikanal (southern zone). Thus the variations lie within the accepted limit ($< 20\%$). On further scrutiny it is observed that the nomographic solutions are more precise for predicting intensity of rainfall of various durations and frequencies. Looking into simplicity in use, quickness and precision

in results obtained, nomographs appear to be the most handy tool for field workers.

When the rainfall intensity of 15, 30 and 60 minutes durations for 10, 25 and 50 years frequency obtained for different locations within a zone was compared with the values obtained by the zonal equations, it appears that, in general, the deviation occurs between $\pm 20\%$ to 30% . However, the deviation was noticed upto 50% at some stations in the northern, eastern and southern zones. Such high variation occurs only at those places where rainfall occurs either with a too low or too high intensity. This indicates the limitations of zonal equations. It is, therefore, suggested that the zonal equations are best suited for locations where intermediate intensity rainfall is received which is always true for any equation or nomographs developed for a region or the country as a whole.

ACKNOWLEDGEMENT

The authors are grateful to Dr. O.P.Gautam, Director-General, Indian Council of Agricultural Research, New Delhi, Dr. V.V. Dhruva Narayana, Head, Division of Engineering, Central Soil Salinity Research Institute, Karnal and to Mr. D.C.Das, Deputy Commissioner (Soil Cons.), Govt. of India, New Delhi for their keen interest in the project and very valuable suggestions. Authors wish to express their thanks to Mr. N.S.Rawat, Nirmal Kumar and R.Tandon for data transfer and calculations and also Mr. M.S.Nayal and K.S.Virmani for preparing charts and drawings.

The authors are grateful to Director-General, Indian Meteorological Department, New Delhi, for allowing the use of recording data.

REFERENCES

- Chow, V.T., 1964. Handbook of Applied Hydrology. McGraw Hill, New York.
- Dalrymple, T., 1960. Flood-frequency analyses. U.S.Geol.Survey Water Supply-Paper 1543-A, U.S.Dept.Interior. 80pp.
- Frevert, R.K., Schwab, G.P., Edminster, T.W. and Barnes, K.K., 1955. Soil and Water Conservation Engineering. John Wiley and Sons, New York.
- Gupta, S.K., Dalal, S.S. and Ram Babu, 1968. Analysis of point rainfall data of Dehra Dun. Irrig.Power J. 25(3):291-330.
- Khullar, A.K., Das, D.C. and Ram Babu, 1975. Station nomograph and one hour rainfall for intensity duration return period computation in India. Soil Cons.Digest 3(2): 1-9.
- Linsley, R.K., Kohler, M.A. and Paulhus, J.L.H., 1949. Applied Hydrology. McGraw Hill, New York.
- Nemec, J., 1973. Engineering Hydrology. Tata McGraw Hill Pub.Com.Ltd., New Delhi.
- Ogrosky, H.O. and Mockus, V., 1957. National Engineering Handbook. Sec.4. Hydrology Supp.A.18-11 to 14. Soil Cons.Serv., U.S.Dept.Agric..
- Raghunath, B., Das, D.C., Srinivas and Lakshmanan, V., 1969. Rainfall intensity-duration-return period analysis and simplified methodology. The Harvester 11(2): 86-92.
- Senapati, P.C., Shakya, S.K. and Nema, J.P., 1976. Nomograph of intensity, duration and recurrence interval of rainfall at Bombay (Colaba). Irrig. and Power J. 33(4):525-528.
- Skurlow, J., 1960. A nomograph for estimation of rainfall intensity and runoff. J.

Soil Cons.Serv.New South Wales 16(2):126-136.

Tejwani,K.G., Gupta,S.K. and Mathur,H.N.,1975. Soil and Water Conservation Research 1956-71. Indian Council of Agricultural Research, New Delhi.

Reprinted from:
Statistical Climatology. Developments in Atmospheric Science, 13
edited by S. Ikeda et al.

© Elsevier Scientific Publishing Company, 1980

PROBABILITY MODEL FOR THE CALAMITOUS BEHAVIOUR OF THE SUMMER MONSOON OVER INDIA

D.A.MOOLEY and B.PARTHASARATHY

Indian Institute of Tropical Meteorology, Pune-411 005, (India)

ABSTRACT

Mooley, D.A. and Parthasarathy, B., Probability model for the calamitous behaviour of the summer monsoon over India. Proc. 1-st Intern. Conf. on Stat. Climat., held in Tokyo, Nov.29-Dec.1, 1979

The behaviour of the summer monsoon over India is often friendly and helpful to the economy of the country. However, occasionally it is calamitous and it breaks the back of the Indian economy, and leads to large-scale sufferings of the people. The calamitous behaviour of the summer monsoon (May-October) has been examined in this study. The percentage of the country's area with monsoon rainfall deficiency of 20 percent or more is defined as the Monsoon Deficiency Index (MDI), and that with monsoon rainfall excess of 20 percent or more, as Monsoon Excess Index (MEI). The sum of these two indices is defined as the Monsoon Vagaries Index (MVI). MDI, MEI and MVI for the country have been worked out for each of the years during the period 1871-1978. MVI series was subjected to statistical analysis. The results of this analysis show that MVI is a Gamma-distributed Variable. The 85th and 90th percentiles of this Gamma distribution are 52 percent and 58 percent respectively. The criterion of $MVI > 55$ percent has been adopted for defining the calamitous calamitous behaviour of the monsoon. The criterion of 55 percent is equivalent to the 87th percentile of the Gamma model fitted to the MVI series. Thus the probability of MVI exceeding 55 percent is about 0.13. The years of calamitous behaviour of the monsoon have been identified. The application of Mann-Kendall rank statistic test and Swed and Eisenhart's test for runs above and below the median to the time interval between successive calamitous behaviour does not bring out any significant non-randomness and the occurrence of these calamities can be taken to be random in time continuum. In view of the low probability of the calamitous behaviour per year, Poisson model could be expected to fit the occasions of calamitous behaviour in a five-year period. Tham's criterion for adequacy of Poisson distribution is found to be satisfied. Poisson model was fitted to the data on calamitous behaviour and the goodness-of-fit was tested by Chi-squared test. The fit has been found to be very good. The Poisson distribution is a limiting case of the Binomial distribution and in the situation of transition, both the distributions may show good fit. In view of this, the Binomial distribution was also fitted. The Binomial fit is seen to be much better.

INTRODUCTION

The Indian economy is largely dependent on the summer monsoon. Indian budget has often been referred to as a gamble in monsoon. The behaviour of the monsoon is often helpful to the economy of the country. However, occasionally, its high erratic behaviour results in a calamity and the economy gets disrupted (Mooley (1975, 1976), Ramdas (1976)). The normal human activities get affected very adversely,

leading to large-scale sufferings of the people. In this study, the calamitous behaviour of the monsoon as defined by a specific criterion, has been examined during the period 1871-1978 to find out whether the occasions of such behaviour exhibit randomness and whether they follow any probability law.

DATA AND METHODOLOGY

The onset of the summer monsoon is earlier than normal and its withdrawal is later than normal in some years. In addition, in some parts of the country, the rain which occurs in May and October is useful for preliminary agricultural operations. In view of this situation, rainfall during the period May to October has been considered as summer monsoon rainfall. Hereafter, summer monsoon will be referred to as monsoon. A large portion of the meteorological subdivision, Jammu and Kashmir is hilly and the number of rain gauge stations is very inadequate. The meteorological subdivision of Himachal Pradesh is a hilly area. The subdivision of west Uttar Pradesh has got hilly portion in northwest. In hilly areas, the representativeness of a rain gauge station is small. The subdivisions of Arabian Sea Islands and Bay Islands consist of a few island stations and as such are extremely small. In view of these reasons, the four sub-divisions, Jammu and Kashmir, Himachal Pradesh, Arabian Sea Islands and Bay Islands, and the hilly portion of West Uttar Pradesh have not been considered. Hereafter, the area of the country will refer to the area of India as indicated in Figure 1.

Monthly rainfall data of all the available rain gauges for the period 1871-1978 have been utilised. Prior to 1901, the number of rain gauges was about 100. Prior to 1901, the number of rain gauges was about 350 and during the period 1901-70, it was 2000 to 3000. After 1970, the rainfall data of about 350 observatory stations have been used. Areal average monsoon rainfall has been obtained for each of the subdivisions for each of the years. The departure of monsoon rainfall from normal, i.e. long-period mean, has been worked for each of the subdivisions and for each of the years and this has been expressed as percentage of the normal.

Monsoon Deficiency Index (MDI), defined as percentage area of the country with percentage rainfall departure of ≤ -20 , and Monsoon Excess Index (MEI), defined as percentage area of the country with percentage rainfall departure of ≥ 20 , have been computed for each year. The Monsoon Vagaries Index (MVI), defined as the sum of MDI and MEI, has also been obtained for each of the years. MVI is shown in Figure 2.

CRITERION FOR CALAMITOUS BEHAVIOUR OF THE MONSOON

The distribution of MVI is skewed. Gamma distribution was fitted to MVI and

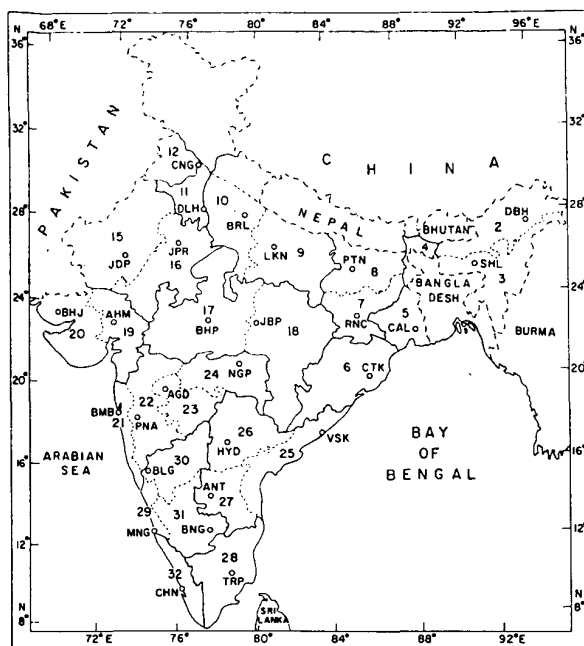


Fig. 1. Sub-divisions considered are given below and shown in the map by their numbers.

2 North Assam	12 Punjab	23 Marathwada
3 South Assam	15 West Rajasthan	24 Vidarbha
4 Sub-Himalayan West Bengal	16 East Rajasthan	25 Coastal Andhra Pradesh
5 Gangetic West Bengal	17 West Madhya Pradesh	26 Telangana
6 Orissa	18 East Madhya Pradesh	27 Rayalaseema
7 Bihar Plateau	19 Gujarat Region	28 Tamil Nadu
8 Bihar Plains	20 Saurashtra & Kutch	29 Coastal Karnataka
9 East Uttar Pradesh	21 Konkan	30 North Karnataka
10 West Uttar Pradesh (Plains)	22 Madhya Maharashtra	31 South Karnataka
11 Haryana		32 Kerala

the goodness-of-fit was tested by variance ratio test and Chi-square test. Both of these tests show that the fit of the Gamma model to MVI is very good. The Maximum Likelihood estimates of the parameters of the Gamma distribution are, shape para-

meter $\hat{g} = 3.50$ and scale parameter, $\hat{b} = 9.71$. The probability density function is given by $p(x) = x^{2.5} e^{-x/9.71} / [(9.71)^{3.5} \Gamma(3.5)]$ for $x > 0$. The 85th and 90th percentiles of this distribution are 53.0 and 58.3 percent respectively. The criterion of 55 percent has been adopted for identifying the calamitous behaviour of the monsoon. The criterion of 55 percent corresponds to 87th percentile of the Gamma distribution fitted to MVI. Thus the probability of MVI exceeding 55 percent is 0.13. The years with MVI exceeding 55 percent have been identified as years with calamitous behaviour of the monsoon.

TESTS FOR RANDOMNESS OF THE CALAMITOUS BEHAVIOUR

The years with calamitous monsoon behaviour and the corresponding MVI are given in Table 1. To test whether the occasions of the calamitous behaviour show any significant non-randomness, Mann-Kendall rank statistic test for randomness and Swed and Eisenhart's test for runs above and below the median, as recommended by WMO(1966 a,b) were applied to the time interval between successive occasions of the calamitous behaviour of the monsoon.

The value of the Mann-Kendall rank statistic is 0.013, whereas the value significant at 5 percent level is outside the interval ± 0.392 . Thus the test does not bring out any significant non-randomness.

The number of runs above and below the median is 8. According to tables by Owen (1962), a value of or less would suggest a trend significant at 5 percent level, while a value of 11 or more would suggest oscillation significant at 5 percent level. The number of runs lies between these two limits. Thus neither significant trend nor significant oscillation is suggested by Swed and Eisenhart's runs test.

Both the tests show that there is no significant nonrandomness in the time interval series, and the interval can be taken to be random. The occurrence of the calamitous behaviour of the monsoon appears to be a random event in time continuum.

PROBABILITY MODEL FOR CALAMITOUS BEHAVIOUR OF THE MONSOON

To plan funds for mitigating the hardships due to the calamitous behaviour of the monsoon, we would like to know the probability of 1,2,3 such occasions in a five-year period. For this purpose, we have to obtain a probability model which would show a good fit to the number of occasions of calamitous monsoon behaviour in a five-year period. Since the mean probability of such occasions per year is low, it is expected that Poisson distribution may show a good fit. The probability mass function of the Poisson distribution is given by

$$P(x) = e^{-m} \frac{m^x}{x!}, \quad \text{for } x = 0, 1, 2, 3, \text{ etc.}$$

Table 1

Years in which the behaviour of the monsoon was calamitous and the corresponding MVI.

Year	MVI	Year	MVI	Year	MVI
1871	58.1	1911	55.6	1941	63.5
1877	62.7	1916	58.6	1956	59.1
1878	56.6	1917	65.7	1961	62.9
1894	56.8	1918	75.7	1965	64.4
1899	77.2	1933	62.3	1972	66.1

Table 2

Goodness-of-fit of the Poisson distribution to the number of occasions of calamitous behaviour of monsoon in a five-year period.

No. of years of calamitous behaviour in a five-year period.	Observed frequency	Frequency on Poisson hypothesis	Contribution to Chi-square
0	10	10.49	.024
1	8	7.28	0.071
2	2	2.53)	0.016
)	
≥ 3	1	0.70)	
$\chi^2 = 0.111$ (d.f.1)			

Table 3

Goodness-of-fit of the Binomial distribution to the number of occasions of calamitous behaviour of the monsoon in a five-year period.

No. of occasions of calamitous behaviour of the monsoon in a five-year period.	Observed frequency	Frequency on Binomial hypothesis	Contribution to Chi-square
0	10	9.93	0.0005
1	8	7.96	0.0002
2	2	2.62)	0.0040
)	
≥ 3	1	0.49)	
$\chi^2 = 0.005$ (d.f.1)			

= 0, otherwiss

x is the number of events, and m is the mean number of events.

Thom (WMO 1966b) has given a criterion for adequacy of the Poisson distribution. This criterion is, $P(\chi_{n-1}^2 > \chi_{n-1}^2) > 0.05$,

$$\text{where } \chi_{n-1}^2 = \frac{n \sum Y^2}{\sum Y} - \sum Y,$$

n, the number of five-year periods and Y is the number of occasions of calamitous monsoon behaviour in a five-year period.

$$\chi_{n-1}^2 = 20.0 \text{ (d.f.20) and } P(\chi_{20}^2 > 20.0) = 0.46. \quad \text{Thus Poisson distribution}$$

is adequate in this cass. This distribution was fitted to the data, and the goodness-of-fit was tested by Chi-square test. The results are given in Table 2. It can be seen that the fit is very good.

The Poisson distribution is a limiting case of the Binomial distribution when the chance of success or failure is low and the number of trials is large, the mean number of successes or failures in n trials remaining finite. In the situation of transition from the Binomial to the Poisson, both the distributions might show good fit. In view of this, the Binomial distribution was also fitted to the data. The probability mass function of the Binomial distribution is given by

$$P_n(x) = \binom{n}{p} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, 3, \text{ etc.}$$

$$= 0, \text{ otherwise}$$

Where p is the mean probability of success, and n is the number of trials. The fit of the Binomial distribution has been teated by the Chi-square test. The results are given in Table 3. The fit is seen to be excellent. The Binomial fit appears to be much better than the Poisson fit. On the basis of the Binomial model, the probabilities of 1, 2, 3 occasions of the calamitous behaviour of the monsoon in a five-year period are 0.379, 0.125 and 0.021 respectively.

The parameters of the Poisson and the Binomial distribution have to be obtained from data. In view of this the stability of the probabilities obtained on the basis of these models would depend on the stability of the parameter determined from the data sample. The mean probability of an occasion of the calamitous behaviour of the monsoon for the whole period is about 0.14, whereas, the values for the first and second half of the total period are 0.17 and 0.11 respectively. This variation in mean over a period of the order of 50 years is perhaps not large.

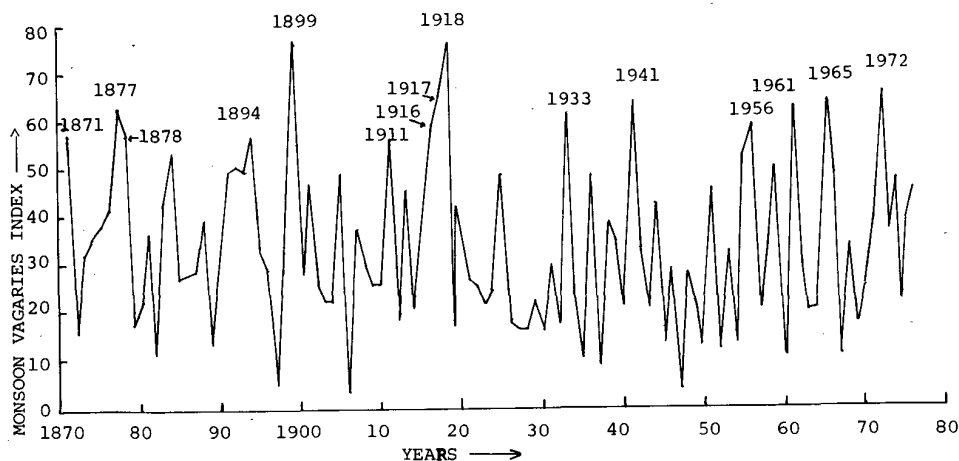


Fig. 2. Monsoon Vagaries Index (MVI) for India (1871-1978).

CONCLUDING REMARKS

The calamitous behaviour of the monsoon appears to be random, and the number of the calamitous behaviour in a five-year period is distributed according to the Binomial law. The Poisson distribution also shows a very good fit.

ACKNOWLEDGEMENT

The authors convey their thanks to the Director for the facilities to pursue this study and to Miss C.P. Ghosh for typing the manuscript.

REFERENCES

- Mooley, D.A.,1975. "Vagaries of the Indian summer monsoon during the last ten years", Vayu Mandal, Vol. 5, Nos. 2 and 3, pp. 65-66.
- Mooley, D.A.,1976. "Worst summer monsoon failures over the Asiatic monsoon area", Proceedings of the symposium on Droughts in Asiatic Monsoon Area held at Pune during 14-16 Dec. 1972, Indian National Science Academy, pp. 34-43.
- Pandras, L.A.,1976. "Droughts and floods in India and some other countries near and far from India", Proceedings of the symposium on Droughts in Asiatic Monsoon Area held at Pune during 14-16 Dec. 1972, Indian National Science Academy, 91-101 pp.

- Owen, D.B.,1962. Hand book of Statistical Tables. Addison-Wesley Publishing Company, London, 580 pp.
- World Meteorological Organisation.,1966a. Climatic Change. WMO Tech. Note No. 79, WMO No. 195, TP-100 80 pp. Geneva.
- World Meteorological Organisation.,1966b. Some methods in climatological analysis. WMO Tech. Note No. 81 WMO - 199 TP-103 pp. 31, Geneva.

PROBLEMS IN STATISTICAL CLIMATOLOGY

— Concluding Remarks of the Conference —

1. INTRODUCTION

Climatology derives most of its basic concepts from an appreciation of various series of meteorological observation taken over extended periods of time, as written by Crowe(1971). At first step, the average, arithmetic mean, can be calculated to reduce a mass of data obtained during the periods. It has been considered that this mean values show the general conditions of the atmosphere. Hence, climate has been defined as an average state of the atmosphere. This may be the most typical definition in the last century, which was a golden era of climatography based on the description of the observed meteorological materials (Leighiy,1949). Thus the statistical method in research influenced on the concept or definition of Climate.

In the present paper, problems in statistical climatology are reviewed, summarizing the results obtained previously and presented at the Symposium.

2. STATISTICS AND CLIMATOLOGY

a) Statistical climatology or climatological statistics

Climatology was born as a son of statistics of meteorological data. Climatology is therefore concerned with collecting and processing meteorological data, summarizing meteorological information, estimating parameters, and discovering climatological empirical laws. One of its application from is a statistical long-range forecast, based upon a systematic statistical examination of the past behavior of climatic elements.

Strictly speaking climatological statistics differs of course from statistical climatology: goal of the former is statistics and the latter climatology. However, the words have been used not always in proper meaning, because it is not so easy to separate them clearly. For instance, Crowe(1971) dealt with the statistics of description under the title of "climatological statistics", as applied to monthly data for temperature and precipitation. Actually, the results of statistics show climatological implications.

b) History of statistical climatology

Eventhough the statistical procedure has long been developed along with the history of climatology since the last century, Conrad(1944) summarized the sta-

3. PROBLEMS IN STATISTICAL CLIMATOLOGY

Problems in statistical climatology can be summarized as the following eight topics. These were drawn from the results of discussion between the Organizing Committee and Dr. R. Sneyers, chairman of the Symposium.

Namely; (i) Time series and assessment of randomness: Applied problems of this item may be homogeneity of series and climatic change.

(ii) Theoretical distributions: Single values, extreme values, continuous or discrete variates, and Markov chains are included in this items. Applied problems are for instance statistical prediction and simple random climatic models.

(iii) Joint (multivariate) distributions: Continuous or discrete variates, estimation when one margin is known, multivariate analysis, and factor analysis are the problems. Their applications are statistical prediction, simple random climatic models, and statistical description.

(iv) Statistical quality control: Applications of this item are outliers in series of observations, and quality of predictions (numerical, dynamic, etc.).

(v) Stochastic models of meteorological fields: Applications are estimation for lacking points or optimal density of networks

(vi) Discriminant analysis: Examples of application are climatic classifications and weather type classifications.

(vii) Stochastic models and autoregressive models: This is applicable to climatic models or stochastic dynamic prediction.

(viii) Circular distributions: Harmonic analysis, spectral analysis, cross spectral analysis and test of significance are included. Application is climatic models.

At the Symposium more than fourteen papers were concerned with the topics of item (ii) mentioned above. The topics of item (i) were mentioned by more than four papers, while item (iii) by six papers. Contrary to expectation, items (iv) (vi) (vii) and (viii) were taken up only by two or three papers respectively.

4. FUTURE PROBLEMS

a) Thema to be studied

As have been shown by number of the papers presented at the Symposium and also considering the social needs, the thema to be studied in future can be summarized as follows:

- (i) Climatic variation or change and prediction of future climate,
- (ii) Estimation future population, food or energy in relation to the climatic conditions,
- (iii) Statistical test of randomness and theoretical distributions of climatic data for each elements in each climatic regions.

tistical methods in climatology first in the middle of this century. Revised his book (Conrad and Pollak, 1950) contributed for advancement of climatological research. A comprehensive textbook by Brooks and Garruthers (1953), which is one of the best reference book at the same time, appeared beginning of the second half of this century. In this handbook, formulation of significance tests, analysis of variance, periodogram analysis, correlogram and other methods were introduced based on the greatest developments in statistics in the first half of the twentieth century. Since statistical analysis applies to samples from populations of data, the sequences of climatological data must be defined so as to be samples from populations (Oliver, 1973). From this viewpoint, the book was the first mile stone of the statistical climatology, which should be studied by sample theories or stochastics.

In China, a textbook on climatic statistics was written by Yao (1965) on the basis of classical statistics.

Godske (1966a) presented a comprehensive paper on statistical meteorology at the WMO Inter-Regional Seminar on Statistical Analysis and Prognosis in Meteorology, Paris, in October, 1962. In this paper he reviewed the problems of information in meteorology including the scale of information, the flow of information in classical climatology, in numerical field prognosis, in routine weather forecasting, in studies of the representativeness of meteorological stations and in synoptic climatology. After describing the statistical methods in climatology in detail, he proposed a definition of climate and climatology: Climatology is the science of the multivariate distributions of meteorological elements with time and space (Godske, 1966b). This definition based on the statistical research methods must be one of the most important results in the history of climatology after the war.

Standard textbooks on climatology treat the practical, simple statistics in most cases. They were concerned mainly with homogeneous data, average, deviation, frequency curve, probable error, mode, correlation coefficient, harmonic analysis etc. (Fukui, 1938). Landsberg (1947; 1958) mentioned also the statistical method for climatological materials. A textbook on climatology in Soviet Union, devoted one seventh of total pages to practical method of climatological statistics (Kostin and Pokrovskaya, 1953). More complete description was appeared in a textbook by Alissow et al. (1956): homogeneity and inhomogeneity and reduction of the records were the main topics in climatological statistics and the thorough description of the statistics of each climatological elements were given as a chapter of statistical climatology. It must be pointed out, however, that the textbooks in this stage have not yet mentioned the problems such as significance test, analysis of variance, factor analysis and small sample theories.

As far as I know, only the textbook written by Suzuki (1968) treats such newly developed fields of statistics systematically.

There were many symposiums, seminars and commissions which set the theme on

statistical climatology. Besides, there published so many research papers from the standpoint of statistical climatology. Historical development of this field should be reconsidered by reviewing them in detail in future.

C) Development of statistical climatology in Japan

Development of the statistical climatology in Japan will be first mentioned briefly. After World war II, the long-range weather forecasting started by the Statistical method. M. Ogawara published the results first in 1948. Also, climatic fluctuations were studied extensively by computing correlation coefficients by T. Yamamoto and others in the second half of 1940's.

1949 may be one of the monumental year in the history of statistical climatology in Japan. Kisho-tokei konwakai, a study group for meteorological statistics, was established in this year and began to publish a journal "Meteorology and statistics" regularly. This journal written in Japanese was distributed only among the researchers in Japan, but it was early enough, if we consider that the first volume of "Tellus" was published in Sweden in the same year and also "Archiv für Meteorologie, Geophysik, und Bioklimatologie" in Austria.

Takahashi published a book entitled "Meteorological statistics" in 1944. The confusion of the research environment at the period of the end of War made it impossible to distribute this book to us. Based on the radio-lecture, Takahashi(1952) compiled a popular booklet on weather and statistics, in which he dealt with the significance of climatology as a science of mean value. He pointed out the importance of frequency distribution as well as deviation, secular variations, extreme values and periodogram analysis as a climatological presentation.

The small sample statistics were introduced in Japan during 1940's. The study group mentioned above studied it eagerly. M.Masuyama, M.Ogawara, E.Suzuki, M.Hirose, T.Ozawa, T.Fujita, K.Tomatsu and many other researchers mainly in the Meteorological Research Institute played an important role in the group. The most brilliant age for the development was occurred in 1950's.

Watanabe(1958) published a book entitled "Modern method of meteorological research", in which he introduced many practical methods of climatological statistics for each climatic elements. Kunisawa and Suzuki(1961) published a textbook on the statistical practice with a wealth of examples of meteorological or climatological materials.

Suzuki(1968) wrote a book entitled "statistical meteorology", which aimed to make clear the meteorological and climatological phenomena by the statistical method of analysis, considering the systematization of the statistical methods. It must be pointed out that this is the first comprehensive book dealt with the boundary region between statistics and climatology in Japan and, at the same time, written on the ground of modern statistics.

- (iv) Stochastic models of meteorological fields in the respective global, regional regional local and micro-scales, and
- (v) Climatic models.

b) International project studies or joint studies

Concerning the data used in statistical climatology, international cooperation must be first needed. As the ICSU (International Council of Scientific Unions) Panel on World Data Centres has reaffirmed, for instance, the World Data Centres exist for the benefit of the world-wide community of scientists. The resolution to the publication of climatic data said that "it is most important that climatological research workers should be able to obtain by some convenient means the meteorological data needed for their work (ICSU, 1979).

CODATA (Commission on Data for Science and Technology), one of the ICSU body, is working also on the data related to statistical climatology. Tomlinson (1979) reported interdisciplinary cooperation and technical exchange in handling of space- and time-varying data. In the field of climatology, WMO (World Meteorological Organization) plays an important role, but it was pointed out by CODATA that natural variability, likelihood of events such as droughts or floods, and changes in climatic means and variability with are the problems concerning the climate for the data to answer. Further, the problems presented are: Cressman analyses, polynomial fitting, eigenfunction expansions, optimal analyses, variational methods, special methods to preserve gradients, and filtering and smoothing.

Recently, World Climate Programme (WCP) has started with the four components: Climatic Data Programme (CDP), Climatic Applications Programme (CAP), Climatic Impact Study Programme (CIP), and Climatic Change and variability Research Programme (CRP). In these programmes, it is needless to say that the statistical approach is the most important. For instance, research elements in the last programme mentioned above are (i) climate model development, (ii) climate predictability, (iii) climate sensitivity, (iv) climatologically significant processes, (v) climate diagnostics, and (vi) climate data requirements.

In such cooperation with the international projects or joint studies, the statistical climatology will make progress intensively. As has been mentioned, the statistical climatology is an interdisciplinary science between statistics and climatology. It is hoped therefore that the statistical societies and the meteorological societies will cooperate in arranging meetings like the present Symposium in near future again.

REFERENCES

Alissow, B.P., Drosow, O.A. and Rubinstein, E.S., 1956. Lehrbuch der Klimatologie.

- Deutscher Verl. d. Wissenschaften, Berlin, 536pp.
- Brooks, C.E.P. and Carruthers, N., 1953. Handbook of Statistical Methods in Meteorology. Met. Office 538, Her Majesty's Stationary Office, London, 412pp.
- Conrad, V., 1944. Method in Climatology. Harvard Univ. P., Cambridge.
- Conrad, V. and Pollak, L.W., 1950. Methods in Climatology. Harvard Univ. P., Cambridge, 459pp.
- Crowe, P.R., 1971. Concepts in Climatology. Longman, London, 589pp.
- Fukui, E., 1938. Climatology. Kokinshoin, Tokyo, 566pp.
- Godske, C.L., 1966a. Methods of statistics and some applications to climatology. WMO, Tech. Note 71:8-86.
- Godske, C.L., 1966b. Statistical approach to climatology. Arch. Met. Geoph. Biokl., B, 14:269-279.
- ICSU, Panel on World Data Centres, 1979. Fourth consolidated guide to international data exchange through the World Data Centres, 113pp.
- Kunisawa, K. and Suzuki, E., 1961. Practice in Statistics. Seirinshoin, Tokyo, 251pp. (J).
- Kostin, S.I. and Pokrovskaya, T.W., 1953. Klimatologiya. Gidrometeorologicheskoe Izdatelstwo, Leningrad, 427pp.
- Landsberg, H.E., 1947; 1958. Physical Climatology. Rev. ed. 283PP.; 2nd ed., Gray Printing Co., Doboys, Pa., 446pp.
- Leighly, J., 1949. Climatology since the year 1800. Trans. Amer. Geoph. Union 30:658-672.
- Oliver, J.E., 1973. Climate and Man's Environment. John Wiley, New York, 517pp.
- Suzuki, E., 1968. Statistical Meteorology. Chijinshokan Co., Tokyo, 314pp. (J)
- Takahashi, K., 1944. Meteorological Statistics. Kawade Shobo, Tokyo, 112pp. (J)
- Takahashi, K., 1952. Weather and Statistics. Hobunkan, Tokyo, 111pp. (J)
- Tomlinson, R.F., 1979. Interdisciplinary cooperation and technical exchange in handling of space- and time-varying data. In: Proc. 6-th Intern. CODATA Conf., Pergamon P., Oxford, 311-319.
- Watanabe, T., 1958. Modern Method of Meteorological Research. Gihodo, Tokyo, 302pp. (J)
- Yao, Ch., 1965. Climatic Statistics. Science Publ., Peking, 246pp. (in Chinese).

M.M. Yoshino
Cochairman