ON THE FORECAST OF PERIODIC EVENTS FROM INCOMPLETE RECORDS

by

Naunihal Singh and K. L. Ahuja

Defence Science Laboratory, Delhi

ABSTRACT

In practice several natural phenomena like wind directions, death rates, egg production etc. follow circular normal distribution. In this paper the problem of estimating the parameters of this distribution from incomplete records has been worked out by the method of maximum likelihood. Second order partial derivatives are given to determine the standard error of the estimates. A numerical example is added to illustrate the practical application of the method.

Introduction

The problem of measuring parameters like amplitude, period and direction of certain natural phenomena following cyclic variations has long been recognised especially in cases where the record of events is not completely known. As for illustration, it can be said that most of the economic phenomena in nature behave in cyclic fashion. It is indeed a complicated matter to find out the nature of a particular phenomenon it follows, due to imposition and super-imposition of various factors. Though number of methods have been developed, the question still remains open in view of the complexities involved partly in the phenomenon and partly due to the incompleteness of the methods themselves.

E. J. Gumbel[1] has discussed a specialised distribution function, called the circular normal distribution which is essentially due to Mises[2]. Under this distribution, the chance of generating false cycles is mitigated by specifying the length of the period which is established by some statistical considerations. The distribution function possesses the properties of linear normal distribution under certian conditions and has wide practical applications. Phenomena as wind direction, storms, deaths, production of components by automatic system, rainfall etc., follow this type of pattern. Circular normal distribution is important and useful also for making predictions of the trend from the available information. As for an instance, let us suppose that we have recorded the dates of the occurrence of storms in a month for the last ten years. It so happens that due to certain climatic difficulties or instrumental defects, it could not be possible to record in the dates for few consecutive months. The data as such is not complete and such situations are generally encountered in our day to day planning. On the basis of the incomplete information, we are interested in to find the period of the maximum occurrence with a certain amount of dispersion, in the past and future provided the conditions do not change abnormally. Thus if the period of storms is predicted with a certain degree of accuracy, it would be of great help in directing the navigations of aeroplanes and ships. Similarly by noting the behaviour of wind direction, it may be possible to detect the radioactive fallout over a particular locality.

The distribution under consideration has two parameters, one giving the mean direction and the other the concentration of observations. In this paper we have discussed the method of estimation of parameters from incomplete records. The incomplete records studied in this paper arise under two situations.

(1) Occurrences are not recorded for the later part of the year for a number of years.

(2) Total number of occurrences are noted but the actual dates are not recorded.

The two situations considered above are analogous to truncation and censoring respectively for linear distribution. The problem of estimation of parameters for these situations has been considered extensively by Cohen[3], Singh[4] and others. 2nd order partial derivatives are obtained to determine the standard error of the estimates. A numerical example is added to illustrate the practical utility of the method.

2. π—Diagram of a Circular Distribution

Let N points or observations lie in a manner on the circumference of a circle of unit radius and let a_i $(i = 1, 2, 3 \ldots\ldots\ldots\ldots N)$ be the angles corresponding to each observation. To characterize such circular observations, their vector mean has been used and this is calculated by introducing ractangular coordinates. $x_i = \cos a_i$, $y_i = \sin a_i$

then

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \qquad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \quad \ldots \quad (2\cdot1)$$

defines the vector mean.

It is, however, convenient to express the vector mean in terms of polar coordinates. Let $\bar{a}$ and a_o be the solution of $\bar{x} = \bar{a} \cos a_o$, $\bar{y} = \bar{a} \sin a_o$. It is obvious that this solution is unique unless $\bar{x} = \bar{y} = o$ and is given by

$$\bar{a} = \left\{ (\bar{x})^2 + (\bar{y})^2 \right\}^{\frac{1}{2}} \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (2\cdot2)$$

and

$$a_o = \tan^{-1} \frac{\bar{y}}{\bar{x}} \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (2\cdot3)$$

where the quadrant in which α_o is to lie is to be determined by the signs of $\bar{x}$ and $\bar{y}$. The vector strength $\bar{a}$ at the mean direction α_o is written with the help of (2·1) as

$$\bar{a} = \frac{1}{N} \sqrt{\left(\sum_{i=1}^{N} \cos \alpha_i \right)^2 + \left(\sum_{i=1}^{N} \sin \alpha_i \right)^2} \quad \ldots (2\cdot4)$$

The parameter α_o is analogous to the mean of a linear variate.

The observations may be given individually or in the grouped form. In the individual case, the calculation of sine and cosine for each observation will give $\bar{a}$ and α_o . In the grouped case, let p_i denote the frequencies attributed to certain intervals, say months, of thirty days each. Then one day corresponds to one degree on the π—diagram. We attribute frequencies to be spread uniformally over each month, then α_o will denote the direction of the maximum frequency on a day. The observations may be arranged according to the following scheme.

Scheme for the circular distribution over a year.

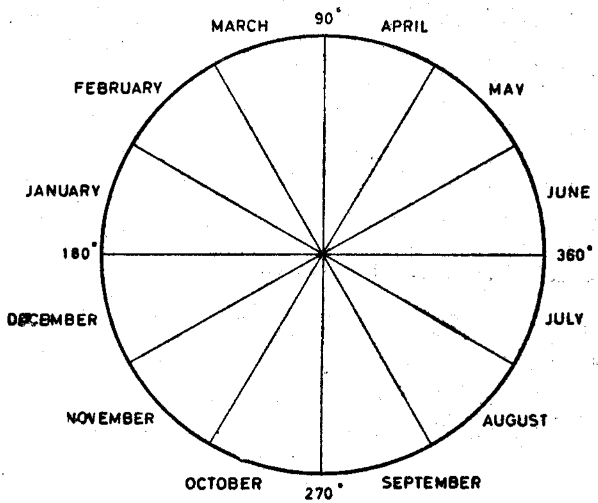


Fig. 1

In the tabular form it can be represented as

Angle	Month	Frequency	Angle	Month	Frequency
0—30°	June	p_1	180°—210°	Dec.	p_7
30°—60°	May	p_2	210°—240°	Nov.	p_8
60°—90°	Apr.	p_3	240°—270°	Oct.	p_9
90°—120°	Mar.	p_4	270°—300°	Sep.	p_{10}
120°—150°	Feb.	p_5	300°—330°	Aug.	p_{11}
150°—180°	Jan.	p_6	330°—360°	Jul.	p_{12}

The mean direction can be obtained from the following trigonometric rule

$$\sum_{i=1}^{N} \cos\alpha_i = \text{July} - \text{Jan} + 0\cdot5\,(\text{May} - \text{March} - \text{Nov} + \text{Sept})$$

$$+ 0.86603\,(\text{June} - \text{Feb} - \text{Dec} + \text{Aug}) \quad \ldots \quad \ldots \quad (2.5)$$

$$\sum_{i=1}^{N} \sin\alpha_i = \text{April} - \text{Oct} + 0.5\,(\text{June} + \text{Feb} - \text{Dec} - \text{Aug})$$

$$+ 0.86603\,(\text{May} + \text{March} - \text{Nov} - \text{Sept}) \quad \ldots \quad \ldots \quad (2.6)$$

3. Circular Normal Distribution

The density function of the circular normal distribution as defined by Gumbel[1] is given by

$$f(\alpha = \frac{1}{2\pi I_o(k)} e^{k\cos(\alpha-\alpha_o)} \quad o \leqslant \alpha \leqslant 2\pi \quad \ldots \quad \ldots \quad \ldots \quad (3.1)$$

where α_o and k have their usual meaning and $Io(k)$ is the modified Bessel Function of the first kind and pure imaginary argument. For $k = o$ (3·1) reduces to the uniform circular distribution and for large values of k to N (0·1) and the distribution assumes the form

$$f(\beta) = \frac{1}{\sqrt{2\pi}} e^{-\beta^2/2} \quad , \quad -\infty \leqslant \beta \leqslant \infty \quad \ldots \quad (3\cdot2)$$

where

$$\beta = \sqrt{k}\,(\alpha - \alpha_o)$$

4. Maximum Likelihood Estimation

Situation I (Occurrences are not recorded for the later part of a year for number of years.)

Let α_i (i=1, 2, 3.......... n) be the angle corresponding to the i th occurrence. The likelihood function of such n consecutive occurrences is given by

$$L_1 = \prod_{i=1}^{n} \lambda(\alpha) \quad 0 \leqslant \alpha \leqslant \alpha_n \quad \ldots \quad \ldots \quad \ldots \quad (4\cdot1)$$

where

$$\lambda(\alpha) = c\,f(\alpha)$$

and $$\frac{1}{c} = \frac{1}{n!} \int_{o}^{\alpha_n} f(\alpha)\, d\,\alpha$$

Taking log of (4·1) and differentiating w.r.t. a_o and k we get

$$\frac{\lambda}{\partial a_o} log L_1 = k \sum_{i=1}^{n} sin\ (a - a_o) + \frac{n}{1 - F\ (a_n, a_o)} \left(f(a_n) - \Psi\ a_o) \right)$$

and

$$\frac{\partial}{\partial k} log L_1 = \sum_{i=1}^{n} cos\ (a - a_o) + \frac{n}{1 - F\ (a_n, a_o)} \left(\beta\ (a) - \frac{I'_o\ (k)}{I_o\ (k)} F\ (a_n, a_o) \right) - n \frac{I'_o\ (k)}{I_o\ (k)} \quad (4\cdot2)$$

where

$$\Psi\ (a_o) = \frac{1}{2\pi I_o(k)}\ e^{k cos a_o}$$

$$F\ (a_n, a_o) = \frac{1}{2\pi I_o(k)} \int_{\alpha n}^{2\pi} e^{k cos(a - a_o)}\ da$$

$$\beta\ (a) = \frac{1}{2\pi I_o\ (k)} \sum_{J=1}^{\infty} \frac{k^{J-1}}{(j-1)!} \int_{\alpha n}^{2\pi} cos^J (a - a_o)\ da$$

and $I_o^s\ (k)$ is the sth derivative of $I_o(k)$ $(s = {}', {}'')$

On equating to zero, equations (4·2) can be written as

$$(1 + a^2)\ cos^2 a_o + 2aB cos a_o + B^2 - 1 = 0$$

and

$$\frac{I'_o(k)}{I_o(k)} = \beta(a) + \frac{1 - F(a_n, a_o)}{n} \left(cos a_o \sum_{i=1}^{n} cos a_i + sin a_o \sum_{i=1}^{n} sin a_i \right) \quad (4\cdot3)$$

where

$$a = \frac{\sum_{i=1}^{n} sin a_i}{\sum_{i=1}^{n} cos a_i} \text{ and } B = \frac{n}{1 - F\ (a_n, a_o)} \times \frac{f\ (a_n) - \Psi\ (a_o)}{k \sum_{i=1}^{n} cos a_i}$$

The parameters can be estimated from (4·3) with the aid of tabels (2, 3, 4) given in Ref. [1], the graph and the Appendices I, II given at the end of the paper. Since the equations are not explicit, the solution may be arrived at by successive approximation. The procedure has been illustrated in section (6).

Situation II (Total number of occurrences are noted but the actual dates are not recorded).

The likelihood function under this situation is written as

$$L_2 = A \prod_{i=1}^{n} f\ (a_i)\ [\ F\ (a_n, a_o)\ ]^{N-n} \quad .. \quad .. \quad .. \quad (4\cdot4)$$

$$A = N\ !/n!\ (N - n)!$$

using the same procedure as above, we have

$$\left.\begin{aligned} \frac{\partial}{\partial \alpha_o} \log L_2 &= k\cos\alpha_o \sum_{i=1}^{n} \sin\alpha_i - k\sin\alpha_o \sum_{i=1}^{n} \cos\alpha_i + (N-n)(X-Z) \\ \text{and} \quad \frac{\partial}{\partial k} \log L_2 &= \sum_{i=1}^{n} \cos(\alpha_i - \alpha_o) + (N-n)\left[\frac{\beta(\alpha)}{F(\alpha_n, \alpha_o)} - \frac{I'_o(k)}{I_o(k)}\right] \\ &\quad - n\frac{I'_o(k)}{I_o(k)} \end{aligned}\right\} \quad (4\cdot5)$$

$$\text{where } X = \frac{f(\alpha_n)}{F(\alpha_n, \alpha_o)}, \quad Z = \frac{\psi(\alpha_o)}{F(\alpha_n, \alpha_o)}$$

On equating to zero, equations (4·5) can be written as

$$\left.\begin{aligned} &(1 + a^2)\cos^2\alpha_o + 2ab\cos\alpha_o + b^2 - 1 = o \\ \text{and} \quad &\frac{I'_o(k)}{I_o(k)} = \frac{(1-p)\beta(\alpha)}{F(\alpha_n, \alpha_o)} + \frac{1}{N}\left[\cos\alpha_o \sum_{i=1}^{n} \cos\alpha_i + \sin\alpha_o \sum_{i=1}^{n} \sin\alpha_i\right] \end{aligned}\right\} (4\cdot6)$$

$$\text{where} \quad b = \frac{(N-n)(X-Z)}{k\sum_{i=1}^{n} \cos\alpha_i}, \quad p = \frac{n}{N}$$

The equations given at (4·6) are also not explicit and the solution is obtained by iteration process. A numerical example for this case is also illustrated in Section (6).

5. Precision of Estimates

To determine the asymptotic variances and covariances of the estimates, it is essential to construct an information matrix for which the 2nd order partial derivatives of the likelihood functions L_1 and L_2 are given below:

SITUATION I

$$\frac{\partial^2}{\partial \alpha_o^2}\log L_1 = -k\cos\alpha_o \sum_{i=1}^{n} \cos\alpha_i - k\sin\alpha_o \sum_{i=1}^{n} \sin\alpha_i + \frac{nk}{1-F(\alpha_n, \alpha_o)} \left(\sin(\alpha_n - \alpha_o) f(\alpha_n) + \sin\alpha_o \psi(\alpha_o)\right) + \frac{n}{(1-F(\alpha_n, \alpha_o))^2}\left\{f(\alpha_n) - \psi(\alpha_o)\right\}^2 \quad (5\cdot1)$$

$$\frac{\partial^2}{\partial k^2}\log L_1 = \frac{n}{1-F(\alpha_n, \alpha_o)}\left[\phi(\alpha) - 2\frac{I'_o(k)}{I_o(k)}\beta(\alpha) + (1 + F(\alpha_n, \alpha_o))\frac{I_o'^2(k)}{I_o^2(k)} - \frac{I_o''(k)}{I_o(k)}\right] + \frac{n}{(1-F(\alpha_n, \alpha_o))^2}\left(\beta(\alpha) - \frac{I_o'(k)}{I_o(k)}F(\alpha_n, \alpha_o)\right)^2 \quad (5\cdot2)$$

$$\frac{\partial^2}{\partial \alpha_o \partial k}\log L_1 = \cos\alpha_o \sum_{i=1}^{n} \sin\alpha_i - \sin\alpha_o \sum_{i=1}^{n} \cos\alpha_i + \frac{n}{(1-F(\alpha_n, \alpha_o))^2} \left(\beta(\alpha) - \frac{I'_o(k)}{I_o(k)}\right)\left(f(\alpha_n) - \psi(\alpha_o)\right) + \frac{n}{1-F(\alpha_n, \alpha_o)} \left[\cos(\alpha_n - \alpha_o) f(\alpha_n) - \cos\alpha_o \psi(\alpha_o)\right] \quad (5\cdot3)$$

where

$$\phi(\alpha) = \frac{1}{2\pi I_o(k)} \sum_{j=1}^{\infty} \frac{k^j}{j!} \int_{\alpha_n}^{2\pi} \cos^{j+2} (\alpha - \alpha_o)\, d\alpha$$

SITUATION II

$$\frac{\partial^2}{\partial \alpha_o^2} \log L_2 = - k\cos\alpha_o \sum_{i=1}^{n} \cos\alpha_i - k\sin\alpha_o \sum_{i=1}^{n} \sin\alpha_i + k(N-n) \left[X \sin(\alpha_n - \alpha_o) + Z \sin\alpha_o \right] - (N-n)(X-Z)^2 \quad (5\cdot4)$$

$$\frac{\partial^2}{\partial k^2} \log L_2 = (N-n) \left[\frac{\phi(\alpha)}{F(\alpha_n, \alpha_o)} - \left(\frac{\beta(\alpha)}{F(\alpha_n, \alpha_o)} \right)^2 \right] + N \left(\frac{I'^2_o(k)}{I_o^2(k)} - \frac{I_o''(k)}{I_o(k)} \right) \quad (5\cdot5)$$

$$\frac{\partial^2}{\partial \alpha_o \partial k} \log L_2 = \cos\alpha_o \sum_{i=1}^{n} \sin\alpha_i - \sin\alpha_o \sum_{i=1}^{n} \cos\alpha_i + (N-n) \left[X\cos(\alpha_n - \alpha_o) - Z\cos\alpha_o - \frac{\beta(\alpha)}{F(\alpha_n, \alpha_o)} (X - Z) \right] \quad (5\cdot6)$$

The term $\varphi(\alpha)$ can be evaluated directly for the desired values of k and α_n.

6. Numerical Example

If the data under consideration is distributed over a year, the observations may be grouped into monthly periods. Since the lengths of months vary from 28 to 31 days, it is essential to eliminate the influence of inequalities of months by multiplying the observed frequencies for January, March, May, July, August, October and December by 0·96774 and for February by 1·07143. The frequencies thus obtained are again to be adjusted by multiplying them by the factor

$$Q = \frac{\sum_{i=1}^{12} p_i}{\sum_{i=1}^{12} p'_i}$$ so as to keep the total frequencies same, p_i being the observed frequencies and p_i, the adjusted ones.

The example considered in the paper is taken from Ref. [5] and concerns with the death rates recorded in U.S.A. from Sept. 1946 to August 1951. Let us assume that it could not be possible to record in the death rates for the months of July and August for some years. The frequency distribution of the death rates is as follows:—

TABLE II

Death Rates U.S.A. Sept. 1946—August 1951

1 Month	2 Obs	3 First adj.	4 Sec. adj.	1 Month	2 Obs	3 First adj.	4 Sec. Adj.
Feb.	10·66	11·42	11·64	Mar.	10·80	10·45	10·05
Jan.	10·50	10·16	10·35	Apr.	10·24	..	10·24
Dec.	10·42	10·08	10·27	May	9·60	9·29	9·46
Nov.	9·70	..	9·70	June	9·42	..	9·42
Oct.	9·42	9·12	9·29				
Sept.	9·08	..	9·08			$n = 99{\cdot}84$	

π—diagram of death rates

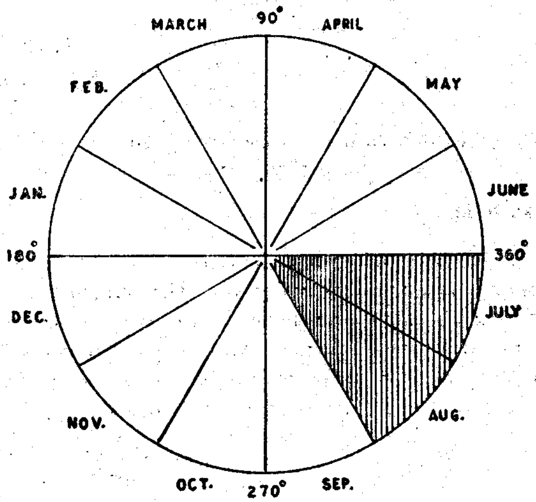


FIG. 2

where shaded area denotes the period for which the data are not recorded.

A rough estimate of the mean direction a_o is given by

$$tan\ a_o = \frac{\sum_{i=1}^{n} sin a_i}{\sum_{i=1}^{n} cos a_i} \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (6{\cdot}1)$$

where

$$\sum_{i=1}^{n} \cos a_i = -\text{Jan} + 0\cdot5\,(\text{May} - \text{March} + \text{Sept} - \text{Nov}) + 0\cdot86603\,(\text{June} - \text{Dec} - \text{Feb})$$

$$\sum_{i=1}^{n} \sin a_i = \text{April} - \text{Oct} + 0\cdot5\,(\text{June} - \text{Dec} + \text{Feb}) + 0\cdot86603\,(\text{May} + \text{March} - \text{Sep} - \text{Nov})$$

and that of $\bar{a}$ is given by

$$\bar{a} = \frac{1}{n}\sqrt{\left(\sum_{i=1}^{n} \cos a_i\right)^2 + \left(\sum_{i=1}^{n} \sin a_i\right)^2} \quad \cdots \quad (6\cdot2)$$

The value of k for a given value of $\bar{a}$ is obtained from Table-II Ref. [1].

Thus the values of α_o and k are as follows:—

$\alpha_o = 161°$ and $k - 0\cdot477$

SITUATION I

On substituting the values of k and α_o in the first equation of (4·3), we get an estimate of α_o equal to 167°. Utilising this set of values of α_o and k in the second equation of (4·3), we get the value of k as 0·373. The graph gives the value of k as 0·810. By using the iteration process, the values of k and α_o converge to the values 0·980 and 161° respectively. Variances and covariances of the estimates have also been determined and given below.

Var $(\alpha_o) = 0\cdot07934$ Var $(k) = 0\cdot00606$ Cov $(\alpha_o, k) = 0\cdot00011$. Therefore the confidence limits for α_o and k are given by $161° \pm 0\cdot282°$ and $0\cdot880 \pm 0\cdot078$ respectively.

Interpretation of Results

This means that the maximum occurrence of deaths will take place during the period corresponding to 160·718° to 161·282° i.e. near about 19th and 20th January. Therefore, under the assumption that conditions from year to year remain same, the necessary steps may be taken to provide more medical assistance during this period. Similarly; results may be obtained under situation II and an interpretation may be given accordingly.

Acknowledgement

Authors are grateful to Dr. R. S. Varma, Director, Defence Science Laboratory for permission to publish the paper and to Dr. P. V. K. Iyer, Assistant Director, D. S. L. for help and encouragement.

References

1. Gumbel, E. J., *et al*, *J. Amer. Stat. Asso.*, **48**, 131, 1953.
2. Mises, R. V., "*Physikalische Zeitschrift*", **19**, 490, 1918.
3. Cohen, A. C., *Ann. Math. Stat.*, **21**, 557, 1950.
4. Singh, Naunihal, *J. Roy, Stat. Soc.*, **22**(2), 307, 1960.
5. Gumbel, E. J., *J. Amer. Stat. Asso.*, **49**, 267, 1954.

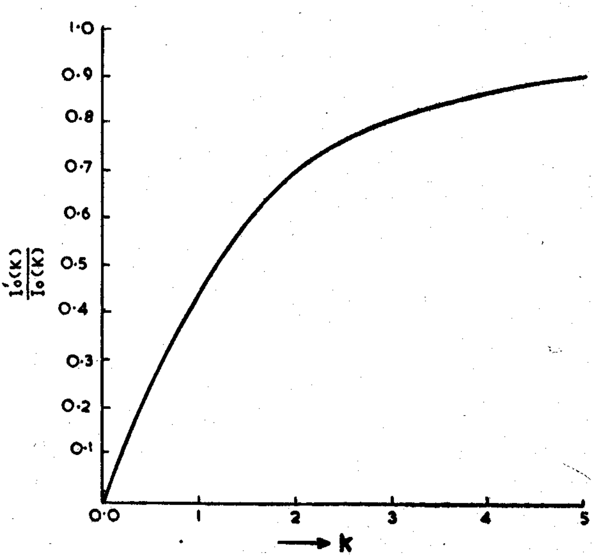
Appendix I

$$\textit{Values of } \beta(a) = \frac{1}{2\pi I_o(k)} \sum_{j=1}^{\infty} \frac{k^{j-1}}{(j-1)!} \int_{a_n}^{2\pi} \cos^{j}(a - a_o)\, aa \textit{ for}$$

Different values of a_o *and* k.

k \ α_o	30°	45°	60°	90°	300°
0·1	—0·03741	—0·07301	—0·09983	—0·12183	0·14977
0·2	0·00410	—0·03415	—0·06246	—0·08505	0·16196
0·3	0·04464	0·00374	—0·02599	—0·04908	0·17432
0·4	0·08394	0·04042	0·00941	—0·01414	0·18676
0·5	0·12178	0·07565	0·04332	0·01954	0·19917
0·6	0·15785	0·10925	0·07577	0·05179	0·21149
0·7	0·19068	0·14106	0·10655	0·08248	0·22364
1·0	0·28116	0·22502	0·18808	0·16431	0·25831
2·0	0·46052	0·38952	0·35145	0·33258	0·34632
3·0	0·52048	0·44658	0·41365	0·36829	0·39633

Appendix II



THE RATIO $\frac{I'_o(K)}{I_o(K)}$ IS PLOTTED AGAINST DIFFERENT VALUES OF K