SYMPOSIA

- Computer Intensive Techniques in Agricultural Research
- Economic Reforms in Agriculture Sector - A Statistical Assessment

PROCEEDINGS

Edited by

PREM NARAIN
R.K. PANDEY

INDIAN SOCIETY OF AGRICULTURAL STATISTICS

Golden Jubilee Conference
19-21 December 1996

Organiser

INDIAN AGRICULTURAL STATISTICS RESEARCH INSTITUTE
Library Avenue, New Delhi-110012

Sponsored by

NATIONAL ACADEMY OF AGRICULTURAL SCIENCES
Avenue II, IARI Campus, New Delhi-110012

CONTENTS

Analysis of Directional Data in Agricultural Research Using DDSTAP

Ashis SenGupta
Indian Statistical Institute, Calcutta

ABSTRACT

Data on orientations, displacements, angular measurements, directions, periodic phenomena, measuremental errors, etc. are often cast in the arena of circular or directional data. Usual statistics for linear data are generally not even meaningful for directional data. Statistical analyses of such data call for new techniques. This is illustrated through "data" from a real-life experiment on the onslaught of insect vectors laying waste huge areas of agricultural fields. Inferences on the expected arrival time, the direction of flight, adaptive modelling of flight directions and flight times after preliminary explanatory data analyses, etc. together with the determination of the predictive density are only just a few practical problems that beg applications of analysis of directional data. The interplay of time, temperature, humidity, fertility gradients, insect-density, agricultural production, etc. call for circular-linear regression using probability models which are enhanced through the principle of maximum entropy. The above analyses demand extensive computer work. However, many of these problems may be solved through DDSTAP, a statistical software package for analyses of directional data, being currently developed by this author.

Key words: Circular normal distribution, DDSTAP, Directional data, Windborne insect vectors.

1. *Introduction*

Often in agricultural research one is faced with the analysis of data relating to orientations, displacements, angular measurements, directions, periodic phenomena, measuremental errors, etc. Such data are usually referred to as Directional Data (DD) in general, and for the special case of such observations that can be mapped to $[0, 2\pi)$ as circular data. It is of interest then, e.g., to obtain initial representations of the data for facilitating preliminary exploratory data analysis; estimate and conduct formal statistical tests for establishing postulated preferred times and directional onsets of and/or for establishing comparisons between different species of insect vectors laying waste vast areas of agricultural fields; estimate and infer on the change in the preferred direction or time of such an onset; determine the relationship of interplay between time, temperature, humidity and insect-densities for varieties of insects over crops; predict, say wind or storm directions, or the direction or time of the next onset of insect vectors, etc.

Analysis of and procedures for statistical inference for DD differ considerably from methods employed for the usual linear data. For example, the sample average of the directions and the corresponding sample variance are not only misleading but even meaningless for representing measures of central tendency and scatter, respectively, for DD. New probability distributions need to be enhanced as defined on the circle and sphere, i.e. on $S^{(p-1)}$ in contrast to the commonly known usual distributions on R and R^p. A general introduction to statistical inference for DD is available in Mardia [4]. Here we exhibit some recent as well as some new techniques we have developed for the analysis of DD. An initial summarization of data may be graphically through Rose diagram, Cuboidogram (SenGupta [6]), etc. Lack of preference of any direction translates to randomness, isotropy or simply, uniformity on the circle. A test for circular uniformity is thus the very first step to model such data. We present below such a test (SenGupta and Chang [7]) which is also a robust invariant optimal test against a very rich family of circular distributions, e.g., the family of wrapped stable (WS) distributions. This family encompasses *all* unimodal symmetric circular distributions that one would need to consider to model such data (without outliers) in practice, in general. Once CU is rejected, we (SenGupta and Chang [7]) choose the appropriate distribution from the WS family consisting of an infinite number of distributions by an adaptive search, using the computer, based on the criterion of p-value with a recently enhanced circular Goodness-of-Fit test. We note that the circular normal (CN) distribution is matched for all practical purposes by the wrapped normal (WN) distribution of the WS family. Estimators of model parameters are then given. Confining to the CN distribution we next present various (Mardia [4]; SenGupta and Jammalamadaka [8] - SJ henceforth) one sample tests for the mean direction and the concentration parameter. These can be extended to multisample tests also, e.g. the ANOMED, the multi-sample test for homogeneity of mean directions, which is parallel to the ANOVA for linear data. The concept of, and some models (Johnson and Wehrly [3]) for circular-linear regression are then introduced. Predictive density (Jammalamadaka and SenGupta [2]-JS henceforth) for the next observation is obtained for an underlying CN distribution and it is noted that this density is quite robust even against the entire corresponding WS family.

Most of the methods presented above demand extensive computer work for their implementations and use in practice for real-life problems. Only recently are such methodologies emerging with the advent of computers. Currently a new package for analysis of DD, "DDSTAP", (SenGupta [6]) is being developed by this author at Indian Statistical Institute which incorporates several newly derived techniques, both an exploratory data analysis and on formal statistical inference. It is hoped that this package will be equipped to tackle many of the above commonly encountered statistical problems with DD in agriculture.

2. *A Real-Life Data Set*

We now describe briefly a real-life experiment in agricultural research and establish the need and usefulness of application of analysis of DD specifically for this experiment. The ideas introduced easily apply to many agricultural problems in general.

It is well-known that many economically devastating diseases of crops caused by viruses and tiny organisms are infected by insect vectors. These small insects are capable of windborne migrations over range of hundreds of kilometers and hence their movements have significant bearings on the epidemiology of disease outbreaks. Estimation and prediction of the movements of these insects and their incidence-distributions directly or even using pertinent auxiliary variables, such as meteorological data on atmospheric structure above the experimental site, would be quite rewarding for both policy-formation and economic considerations.

An Indo-UK joint project was recently undertaken on this aspect by Riley *et al* [5]. They conducted a preliminary investigation of the windborne movements of insect vectors of crop diseases in Haringhata, West Bengal, in northeast India. Data collected included information on timing and direction of migration. Riley *et al* observe: "The ultimate objective is to develop methods for understanding and forecasting disease outbreaks, so that control strategies may be devised."

Here we will be confining ourselves to the data an aerial trapping of insect groups by a net attached to the tethering line of kytoon (an aero-dynamically-shaped balloon). Further details may be obtained from Riley *et al.* From the distribution of the different insect taxa caught and identified in the aerial net, it was seen that the mirid bug *Cyrotorhinus lividipennis* (C. *liv.*, henceforth) was the commonest insect caught, accounting for approxly. 20 percent of the total. We thus choose here C. *liv.* for the sake of illustration of our methods. Table 1 exhibits the data on the incidence-distribution over the period of 24 hours. Here, analysis of the incidence-distribution over the flight directions (which are directly directional variables) observed is quite important too. Unfortunately, such data have not been reported in detail in Riley *et al* [5].

In tune with the objective of the study, we feel that the following analyses are needed and will be quite enlightening. We see to exploit whether the migration is random, that is, whether onslaught of C. *liv.* is uniform over time. If not, a theoretical probability distribution should be fitted. Inferences on the parameters of such a distribution then naturally follow. Regression analysis can reveal interplay between variables, biological, atmospheric, etc. encountered. Predictions of the most probable time and that of the direction of the next onslaught are of prime usefulness. Based on such analyses, we hope that optimal collection of crops, evasive or even deterring techniques, say through aerial spraying in predicted times and/or directions, can be adopted to protect crops and agricultural fields from a devastating onslaught.

3. *Graphical Representations*

Graphical representations can be quite helpful for Exploratory Data Analysis and preliminary summarization of the data. For example, Rose diagram, Circular histogram and the Cuboidogram recently introduced by the author (see e.g. SenGupta [6]) may serve to explore initially the symmetry, skewness and other features of circular data. A plot of the fitted curve - the suggested theoretical probability distribution, on say, the Circular histogram and on the best-fitting standard CN distribution, can establish the closeness of the fit and the need for alternative new distributions, e.g. those belonging to the WS family, respectively. These are achieved through DDSTAP.

4. *The Models*

The most commonly used circular distribution is the circular normal (CN) or the von Mises distribution, CN (μ_0, κ), with the p.d.f.,

$$f(\theta;\mu_0,\kappa) = [2\pi I_0(\kappa)]^{-1} \exp\{\kappa\cos(\theta-\mu_0)\}, 0<\theta, \mu_0 \le 2\pi, \kappa>0 \quad (1)$$

We consider the WS family consisting of the densities $f(\theta;\eta,a,\mu_0)$ to model directional data,

$$f(\theta;\eta,a,\mu_0) = (2\pi)^{-1}\left[1+2\sum_{p=1}^{\infty}\eta^{p^a}\cos p(\theta-\mu_0)\right] \quad (2)$$

$$0\le\theta,\ \mu_0<2\pi;\ 0\le a\le 2; 0\le\eta\le 1$$

$a = 2$ gives Wrapped Normal (WN), $a = 1$ Wrapped Cauchy (WC), $\eta = 0$ gives Circular Uniform (CU). Retaining terms till only $p = 1$ gives Cardioid.

This family is obtained by "Wrapping" (mod 2π) the family of symmetric unimodal stable distributions. It gives a rich coverage of kurtosis in the data, not within the capture of CN distribution. Further, it is also envisaged as the "omnibus" family which can be used to model any circular data corresponding to a symmetric unimodal circular distribution that one may ever usually encounter in practice.

We now consider several formal statistical inference problems that are often needed in agricultural research involving DD.

Let $\theta_1, \ldots, \theta_n$ be a random sample from the underlying population. Let $S=\sum_{i=1}^{n}\sin\theta_i$, $C=\sum_{i=1}^{n}\cos\theta_i$, and $\bar{R}_p=\sum_{i=1}^{n}\cos p(\theta_i-\bar{\theta})$, $\bar{R}_1\equiv\bar{R}$, where $\bar{\theta}$ is the sample mean direction given by $\tan\bar{\theta} = S/C$.

5. *Estimation*

MLEs of μ and κ of the CN distribution are solutions to

$$\hat{\mu} = \bar{\theta},\ A(\hat{\kappa}) = \bar{R},\ A(x) = I_1(x)/I_0(x) \quad (3)$$

where $I_r(.)$ is the modified Bessel function of the r-th order. The above MLEs coincide with the corresponding estimators obtained by the method of trigonometric moments.

The MLEs for the three parameters of any member of the WS family are not available in any useful closed form. However, here we note that the method of trigonometric moments does conveniently yield simple estimators of these parameters. By exploiting the Fourier series representation of any circular distribution and the added property of symmetry assumed for our WS family, these estimators are given by

$$\hat{\mu} = \bar{\theta},\ \hat{\eta} = \bar{R},\ \hat{a} = \ln[\ln \bar{R}_2/\ln \bar{R}]/\ln 2 \tag{4}$$

It is interesting to note that the moment estimators of μ and η do not depend on that for a. One may thus explore alternative methods of estimation of a, retaining the simple moment estimators of μ and η for use as needed. For example, one may estimate (adaptively) a as the value of the parameter of that distribution of the WS family which gives the best fit to the data as established by a Goodness-of-Fit test. The above estimators, including $\hat{\kappa}$ and the adaptive estimator of a, which require iterative techniques for their computations, may be obtained using DDSTAP.

6. *Test for Randomness or Uniformity*

For testing CU against WS family, a robust, optimal invariant test when all parameters are unknown is given by

$$\omega : R^2 = S^2 + C^2 > K \tag{5}$$

and when μ_0 known, is given by

$$\omega : C^2 > K_0 \tag{6}$$

The exact cutoff points K and K_0 are obtained through numerical integration involving standard and modified Bessel functions or through extensive simulations. DDSTAP incorporates these cutoff points to return the result of the tests.

7. *Goodness-of-Fit Tests*

For testing for the appropriateness of CN distribution or for obtaining the best WS member to model the data, we use the *p*-value of the corresponding circular Goodness-of-Fit tests as described below.

Watson's U^2-statistic is an adaptation of the well-known Cramer-von Mises statistic W^2 to make it independent of the choice of the origin for θ around the circle. Let F be a completely specified distribution function and F_n the empirical distribution function based on n observations. Then,

$$U^2 = n \int_0^{2\pi} [F_n(\theta) - F(\theta) - \int_0^{2\pi} \{ F_n(\phi) - F(\phi) \} dF(\phi)]^2 dF(\theta) \qquad (7)$$

Stephens has established that for $n \geq 8$, the percentage points of the modified statistic, U^{*2},

$$U^{*2} = (U^2 - .1/n + .1/n^2)(1.0 + 0.8/n) \qquad (8)$$

can be referred to his tables, see e.g. Table 7.3 in Mardia, pg. 182.

However when the n observations are available only in a grouped form, grouping corrections need to be incorporated. Such corrections were not available and only very recently, Brown suggested the following U_G^2 as the required modified U^2-statistic. Using the results of a Brownian bridge, he also justified that U_G^2 can be referred to the same table of percentage points for U^2 as usual.

Let the data be grouped into s periods, O_i is observed frequency and $E_i = n\Delta_i$, $\Delta_i \geq 0$, the expected frequency under F, $i = 1, \ldots, s$. Then,

$$U_G^2 = \sum_{i=1}^{s} \frac{\Delta_i^2}{6}\left(1 - \frac{\Delta_i}{2}\right) + \frac{1}{12N} \sum_{i=1}^{s} \Delta_i (O_i - E_i)^2$$

$$+ \frac{1}{n}\left\{ \sum_{i=1} \Delta_i Y_i^2 - \left(\sum_{i=1}^{s} \Delta_i Y_i \right)^2 \right\} \qquad (9)$$

where $Y_i = \sum_{j \leq i-1} (O_j - E_j) + (O_i - E_i)/2$. Theoretically, the computations for the E_i s, or equivalently, Δ_is, with a WS distribution involve integrating an infinite sum term by term. However, in practice, it is known, that $f(\theta; .)$ is a rapidly convergent series and it sufficed to retain at most the first five terms (three terms for WN, see e.g. Mardia [4], Pg. 56) for computing the density and atmost twenty terms for computing $\int_0^{2\pi} F(\theta)\, d\theta$ corresponding to the WS distribution.

8. *Tests for Mean Direction*

We consider here CN populations only. Tests for the mean directions in WS distributions may also be constructed based on their moment estimators given above.

Recall that in general the CN population is a REF, but with κ known it becomes a member of (1, 2) CEF.

8.1 Case of κ Known

Consider testing $H_0 : \mu = 0$ against $H_1 : \mu > 0$. Let κ_0 be the known value of κ. Here one may use the usual LRT. However, no small sample optimal properties of this LRT is known. The approach of using ancillarity here leads to an optimal test, but this is conditional on the ancillary R and seems not to be popular with the users. We present now an unconditional optimal, LMP, test for the situation. Direct differentiation yields,

$$\omega : S > K, \text{ where } K \text{ is the cut-off point} \quad (10)$$

For testing against two-sided alternative, we enhance the LMPU (Unbiased) test. Exploiting the symmetry of CN and the principle of reflection, this yields an elegant test given by,

$$\omega : n\kappa_0 \, \bar{S}^2 - \bar{C} > K^*, \text{ where } K^* \text{ is the cut-off point} \quad (11)$$

Both the above tests are for all sample sizes, locally most powerful (unbiased) and the first one has monotone increasing power in $(0, \pi)$. These are also consistent tests. Note that by virtue of the multivariate CLT, $(\bar{C}, \bar{S})$ follows a bivariate normal distribution. Hence with large samples, for any given level of significance, K and K^* can be easily obtained. However, for small samples, their determination requires non-trivial computations and calls for computer programs. DDSTAP performs the above optimal simple unconditional LMP and LMPU tests proposed by SJ and SenGupta and Chang [7] for all sample sizes-both for large as well as for small samples.

8.2 Case of κ Unknown

When κ is unknown, the principle of similarity or meaningful invariance does not lead to any reduction and hence no unconditional useful test is available. One approach, a very restrictive one, may be to use a conditional test (Mardia [4] pg. 143). However, it has been shown in SJ that in this case an unconditional asymptotically optimal test, e.g. Neyman's C_α test can be derived. Let $\phi = \ln f(\theta, \kappa)$. Then at $\mu = 0$, $\phi_\mu = \kappa \sin\theta$, $\phi_\kappa = \cos\theta - A(\kappa)$. Assume $\kappa < K_0 < \infty$. Then straightforward computations establish that all the conditions for ϕ_μ and ϕ_κ to be Crámer functions are satisfied.

Consider testing $H_0 : \mu = 0$ against $H_1 : \mu > 0$. The C_α- test has been shown to reduce to the simple form,

$$\omega : Z_n = \sqrt{\kappa}\, S/(nA\,(\hat{\kappa})^{1/2}) > \tau_\alpha \tag{12}$$

where $\hat{\kappa}$ is the MLE of κ and τ_α is the upper 100α- percent point of the standard normal distribution.

For testing against two-sided alternatives, a test can be based on the same test statistic as above, with obvious modification of the critical region. Neyman, in general, has shown that this also yields an asymptotically locally most powerful test.

9. *Tests for the Concentration Parameter*

Observe that when μ is known, CN distribution reduces to a one-parameter REF with the canonical parameter coinciding with the natural parameter of interest κ. When μ is unknown, we are back to the two parameter REF. Further, as shown above in the context of testing circular uniformity, μ may be treated as a 'location' parameter and principle of invariance may be exploited when it appears as a nuisance parameter..

Consider first the case when μ is known, and wnlog taken to be 0. We want to test $H_0 : \kappa = \kappa_0$ against one and two sided alternatives. Recall that $\kappa_0 = 0$ is equivalent to testing for uniformity. Note that then the corresponding optimal UMP and UMPU tests are simply based on $\bar{C}$.

Next let μ be unknown. Note that R is a maximal invariant statistic here. It is easy to show that the pdf of R, though not a member of the exponential family, has the monotone likelihood ratio property wrt κ. It then follows that the UMPI (Invariant) test has the critical region,

$$\omega : \; R > R_\alpha, \; R_\alpha \text{ being the cutoff point} \tag{13}$$

For two-sided alternatives, the UMPUI test is based on R also, with a two-sided critical region.

The exact distribution of R is complicated. For $\kappa > 2$, and for large samples, Stephen has given a simple approximation,

$$2\gamma(n-R)\chi^2_{n-1},\ \gamma^{-1} = \kappa^{-1} + 3/8\kappa^{-2} \tag{14}$$

For small samples and exact cutoff points, in general, one needs to call upon the computer programs.

10. Regression

As mentioned earlier, it is quite common in the context of regression analysis in agricultural studies to encounter situations where the response variable is linear but the explanatory variables may not be linear only but circular also. The following results (Johnson and Wehrly [3]) may be quite useful in that context.

Theorem 1 : Let Θ and X have the joint density function

$$f(\theta, x) = c.\exp\left\{-\frac{1}{2}X^T\,\Sigma^{-1}x + \lambda^T\Sigma^{-1}x + a(\theta)^T\Sigma^{-1}x\right\} \tag{15}$$

where c is a constant of integration, $a(\theta)^T = (a_1(\theta), \ldots, a_q(\theta))$

$$a_i(\theta) = \sum_{j=1}^{p}\ \sum_{k=1}^{n} a_{ijk}\cos[k(\theta_j - \mu_{ijk})]$$

$$= \sum_{j=1}^{p}\ \sum_{k=1}^{n}[\alpha_{ijk}\cos(k\,\theta_j) + \beta_{ijk}\sin(k\theta_j)]$$

$$i = 1, \ldots, q \tag{16}$$

$x \in \mathbb{R}^q$, $\theta \in [0, 2\pi)^p$ and Σ^{-1} is positive definite. Then $f(\theta, x)$ maximizes the entropy of multivariate angular-linear distributions subject to $E[XX^T]$, $E(X)$, and $E[X \otimes H(\Theta)]$, where $\otimes$ is the Kronecker product, taking specified values consistent with expectation with respect to the distribution (15).

Remark 1: The conditional distribution of X given $\Theta = \theta$ is q-dimensional multivariate normal with mean $\lambda + a(\theta)$ and convariance matrix Σ.

The density (15) also provides a means for predicting X_1 from X_2 and Θ by writing $X = \left(X_1^T \mid X_2^T\right)^T$ and looking at $f(x_1 \mid x_2, \theta)$. Results for the conditional distribution of X_1 given x_2 and θ follow from the usual results for the conditioning of one multivariate normal vector on another. If we partition Σ, λ and $a(\theta)$ correspondingly, we find that the distribution of $X_1 = (X_1, \ldots, X_r)^T$ given x_2 and θ is the r-dimensional normal distribution with mean $\lambda_1 + \Sigma_{12}\Sigma_{22}^{-1}\,[x_2 - (\lambda_2 + a_2(\theta))]$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\,\Sigma_{21}$.

Note that each component X_i, $i = 1, \ldots, r$ of X_1 has a variance not depending on the conditioning variables and a mean of the form

$$v_0 + \sum_{i=r+1}^{q} v_i x_i + \sum_{i=r+1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{n} [\gamma_{ijk} \cos(k\theta_j) + \delta_{ijk} \sin(k\theta_j)] \quad (17)$$

Consequently, this model leads to a natural method of predicting a vector X from a vector θ of directions. Here we fit the best Fourier series of n-th degree in the individual θ's to the individual X's. As shown, some X_i's may also be included into the set of predictor variables.

Situations where response variables are circular with explanatory variables being circular and/or linear are being dealt with recently. We expect to incorporate these also in DDSTAP.

11. *Prediction*

Most statisticians are of the opinion that the singlemost aim of statistics is to predict the nature of the future occurrence of the event under study. It is even more so for Directional Data, since their applications are in a variety of real-life problems ranging from bio and circadian rhythms to onset times of certain diseases, from directional propagations of tornadoes and twisters to angular displacements of magnetic fields, etc.

It has been demonstrated in JS through exact numerical computations and simulations that the predictive density derived under the assumption of a CN population, not only tends to a CN population with more stability (higher concentration parameter), but also is robust against the symmetric unimodal wrapped stable family of distributions-possibly the "omnibus" family for unimodal symmetric circular distributions. The following theorem presents the results for the CN distribution.

Theorem 2 : 1. The predictive density of θ_{n+1} is given by

$$g(\theta_{n+1} \mid \theta_1, \ldots, \theta_n) = \phi_n\,(R_n^2)/\phi_{n+1}\,(R_{n+1}^2) \quad (18)$$

where

$$\phi_n (R^2) = \int_0^\infty u J_0 (Ru) J_0^n (u)\, du \tag{19}$$

$J_0(x)$ being the standard Bessel function of zero order.

2. $g(.)$ above is symmetric and unimodal with its mode at $\bar{\theta}_0$, the MLE of μ_0 based on the past data of n observations, $\theta_1, \theta_2, \ldots, \theta_n$.

3. $g(.)$ is proportional to $f(\theta_{n+1}; \bar{\theta}_0, 2\bar{R})$ for large n.

Similar results as above can be derived for l-angular and l-modal CN distributions on the circle as well as the Langevin distribution on the sphere.

12. *Statistical Package: DDSTAP*

The software DDSTAP (SenGupta [6]) is a menu-base statistical package being currently developed by this author. It gives both graphical representations as well as formal statistical inferential results on the analysis of DD. Several recent techniques, both graphical - Cuboidogram, Changeogram, etc., as well as analytical-adaptive modeling in the presence of outliers, robust predictive HPD, etc., are also incorporated in this package. In several cases, exact or small-sample analyses are provided in addition to the usual large-sample approximate results. It is envisaged that data on spheres, in addition to those on circles, may be analyzed through DDSTAP.

13. *Data Analysis*

We illustrate the need and use of analysis of DD in agricultural research through the real-life data set described in Section 2 above. We consider the data on the incidence-distribution of C. *liv.* over the period of 24 hours as reported in Table 4 of Riley *et al* [5]. These are reproduced in columns 1 and 3 of our Table 1. Though such daily time is not directly an angular or directional variable, the inherent periodicity lends itself to be treated as a circular random variable by identifying 360 degrees with the 24 hour period. Such conversion is shown in column 2 of our Table 1. The analyses presented here were done by exploiting the package DDSTAP.

Table 1 : Mean aerial densities (number per 104 cubic metres) of C. *liv.* caught in the aerial net at different periods of the day or night, between 14-23 November, 1992.

Sample period (hrs.)		Degrees	Mean
Dusk	(16.30-18.30)	0-30	11.37
First part of night A	(18.30-21.30)	30-75	11.39
First part of night B	(21.30-00.00)	75-112.50	4.65
Second part of night	(00.00-04.30)	112.50-180	11.38
Dawn	(04.30-06.30)	180-210	10.18
Day	(06.30-16.30)	210-360	0.69

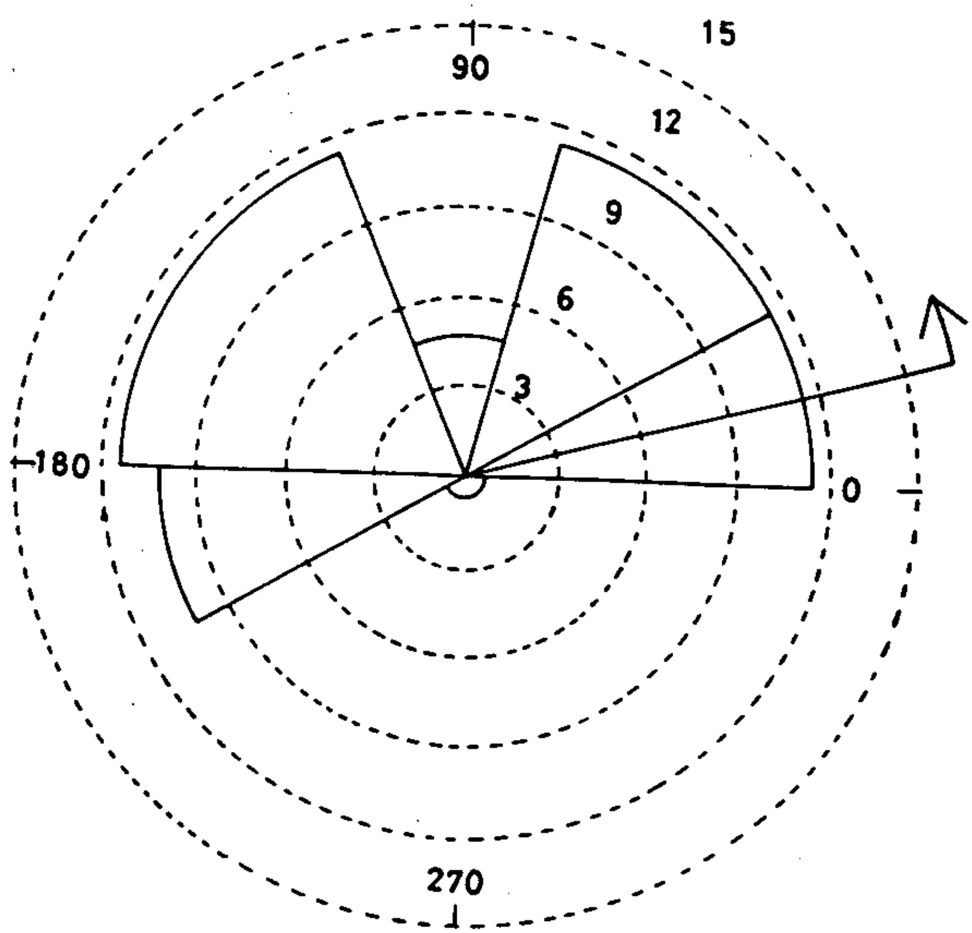

Fig. 1, Rose diagram for *Cyr. liv.* data

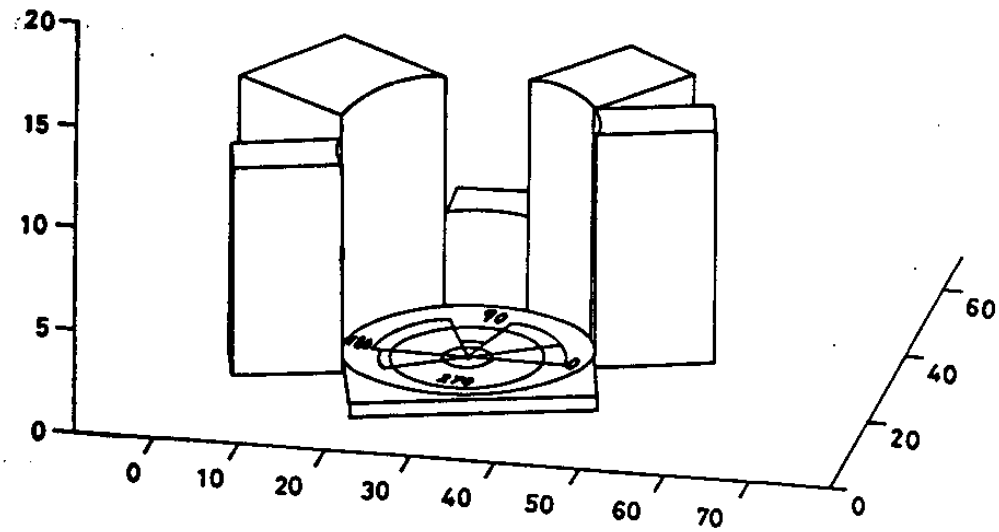

Fig. 2. Cuboidogram for C. *liv.* data

Initial graphical representations of the data were called for to reveal some characteristics, such as skewness, kurtosis, multimodality, etc. of the underlying population distribution. For example, Figures 1 and 2 present the Rose diagram and the Cuboidogram for C. *liv.* data. These indicate that the distribution is not uniform but rather a symmetric circular distribution be fitted to the data. We treat the sample size as large and thus take $\hat{\mu}$ as the true mean direction μ. The data is then transformed to the standardized data, $\theta^* = \theta - \mu$. This is shown in column 3 of Table 2. We then proceed to analyze this transformed data using the fact that the mean direction for the underlying distribution may now be taken as known to be equal to zero degrees. This facilitates computations. However, if the sample size is not to be considered large, we could start with original data set and perform all our usual tests corresponding to the ones conducted below by merely calling from DDSTAP their appropriate versions for unknown μ. The results on the formal tests for circular uniformity and that for the Goodness-of-Fit for the circular normal distribution are presented in Table 2. The CU distribution is rejected as expected, while the CN is seen to give an acceptable good fit to the data. Figure 3 exhibits the fitted CN distribution.

Having decided on the CN distribution from above, one may then proceed with the various aspects of statistical inference. We note that the estimated mean direction, $\hat{\mu}$ is 94.30 degrees which translates to

Table 2 : Inference with CN Model for *C. liv.* Data

no. of classes = 6

$\mu = 94.301228$ $\quad r^2 = 388.927182$

$r^{*2} = 386.564518$ $\quad \rho = 0.397125$

$\kappa = 0.865111$

Test for Circular Uniformity:

$\sqrt{2n}\,C^* = 3.957727$

CU rejected at $\alpha = .05$

Goodness-of-fit Test for Circular Normality:

$U_g^2 = 0.159843$ $\quad U_g^{*2} = 0.141998$

CN not rejected at $\alpha = .05$

Table 2A: Goodness-of-Fit Table with Fitted CN Distribution

θ-class		θ^*	O_i	E_i
0.00	30.00	280.698772	11.370000	4.086678
30.00	75.00	318.198772	11.390000	9.802765
75.00	112.50	359.448772	4.650000	10.106959
112.50	180.00	51.948772	11.380000	13.185099
180.00	210.00	100.698772	10.180000	2.975359
210.00	360.00	190.698772	0.690000	9.457993
Total Frequency			49.660000	49.614855

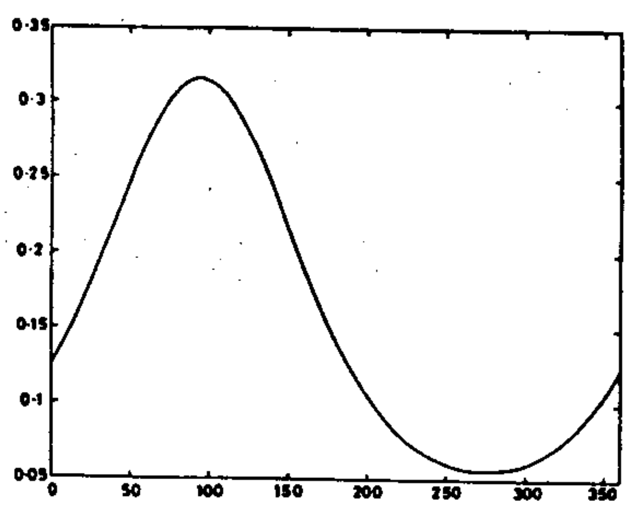

Fig. 3. Fitted CN curve for *Cur. liv.* data

22.47 hrs. The concentration parameter is estimated as $\hat{\kappa} = 0.8651$. The agricultural scientists may have reasons to hypothesize a particular time, or direction, say μ_0, as the time most preferred by *C. liv.* for their onslaughts. One can then use the results from Section 8 and DDSTAP to formally test that hypothesis. Similarly, regression "lines" may be fitted to the circular response variable, time or direction, when information on explanatory variables are made available. Further, from Section 11 we note, e.g., that the most likely time (predicted) for the onslaught by the next *C. liv.* is 22.47 hrs.

Remarks 2: It is possible to envisage a better fit to the data through some member of the WS family. DDSTAP does identify such a distribution which in general need not be CN. It has been however seen here the WN gives the best fit to the above data. Since it is well known that CN and WN are identical for the purposes of statistical inference in practice, unless the sample size is quite large, it suffices here to proceed with the CN distribution as the underlying model.

ACKNOWLEDGEMENTS

The author is thankful to Professor D.D. Majumdar, Professor Emeritus, ISI and CSIR, for making available to him a copy of Riley *et al* [5] with encouragements and to Professor Prem Narain, IARI, for his encouragement to work on this problem.

REFERENCES

[1] Chang, H., 1991. Some optimal tests in directional data. Ph.D. Dissertation. Dept. of Statistics and AP, University of California, Santa Barbara.

[2] Jammalamadaka, S.R. and SenGupta, A., 1997. Predictive inference for directional data. Tech. Rep. No. 340, Dept. of Statistics & AP, University of California, Santa Barbara.

[3] Johnson, R.A. and Wehrly, T.E., 1978. Some angular-linear distributions and related regressions models. *J. Amer. Statist. Assoc.*, 73, 602-606.

[4] Mardia, K.V., 1972. *Statistics of directional data*. Academic Press.

[5] Riley, J.R., Reynolds, D.R., Mukhopadhyay, S., Ghosh, M.R., Chowdhuri, M.M., Sarkar, T.K., Nath, P.S., Sarkar, S., Das, B.K., De, B.K., Satpathi, C.R., Guha, M.K., Dutta Majumdar, D., Das, J. and De, A.K., 1993. A preliminary investigation of the windborne movements of insect vectors

of crop diseases in Northeast India. *NRI Project A0166*, Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta.

[6] SenGupta, A., 1997. DDSTAP- *A statistical software package for the analysis of directional data.* (Under preparation) Applied Statistics Division, Indian Statistical Institute, Calcutta.

[7] SenGupta, A. and Chang, H., 1996. A robust optimal test for uniformity for directional data against wrapped stable mixture family. Submitted for publication.

[8] SenGupta, A. and Jammalamadaka, S.R., 1997. Geometry and optimal unconditional tests for the mean direction of the von Mises distribution. Resubmitted to *Ann. Instt. Statist. Math.*