Comparison of estimation methods of the joint density of a circular and linear variable

Sahana Bhattacharjee1, Kishore Kumar Das2

*1,2Department of Statistics, Gauhati University, Guwahati-781014, India*

*Abstract:* A detailed account on environment and its related problems often requires a study on variables, some of which may be linear and others, circular in nature. Specialized techniques are required to analyze circular variables and also, the inter-relationship between circular and linear variable. In this paper, estimation methods of joint density of a circular and linear variable have been compared with the aid of three real life examples. Therein, we have computed the maximum likelihood estimates of the l-modal circular normal distribution and used the Monte Carlo approximation method to approximate its distribution function. Finally, the theory of Copula function has been put to use for exploring the association between the two variables. The study reveals the superiority of one of the methods over the others, as exemplified by all the three data sets.

*Key words: l-modal circular normal distribution, Maximum Likelihood Estimation, Monte Carlo Approximation, Comparison of Circular-linear joint density estimation methods, Copula function.*

1. **Introduction, motivation and objective**

**1.1 Background of the Study**

In this age of growing population and rapid advancement in technology, the reserves of fossil fuels are depleting at a very fast rate. This calls for the search for alternative natural resources of energy, which are also environment-friendly [13]. Wind energy is one such ever-abundant natural resource which is clean and pollution-free. The assessment of the wind characteristics and their inter-relationships is necessary for the proper installation of wind turbines for tapping of wind energy.

Further, several studies ([14], [18], [21]) have shown that the wind speed and direction and their joint effect influenced the growth of sewage-associated bacteria and sediment re-suspension. These call for the analysis of inter-relationship between wind speed and direction of a place, which will play a significant role in maintaining the water quality of water bodies located in its vicinity. It needs mention here that wind direction is a circular variable which arises in terms of angles and is measured in degrees whereas wind speed is a linear random variable. Similarly, [9] have carried out a case study to analyze the effect that wind direction has on ozone levels, of which the former is circular and the latter is linear. This calls for the incorporation of specialized techniques for jointly analyzing a circular and linear variable

Several correlation measures and regression models have been proposed in literature to study the circular-linear and linear-circular relationship. [22] have demonstrated the use of a circular-linear rank correlation coefficient for the association between wind direction and pollution concentration. [11] have proposed parametric models for the joint distribution of linear and angular random variables, based on the maximum entropy principle and by the specification of marginal distributions. [9] have considered a regression model to estimate linear response from the sine and cosine components of the angular variable and used it to explore the effect of wind direction on ozone levels. [15] has proposed a bivariate circular-linear correlation coefficient to assess the relation between a circular and linear variable and used it to explore the relationship between wind direction and ozone levels. [5] has constructed circular distributions based on non-negative trigonometric sums and later on, the same author in 2007 [6], used them to construct circular-linear and circular-circular density and applied them to real-life data sets. The advantage of using circular distributions based on non-negative trigonometric sums is that it allows for the incorporation of the multimodality or skewness present in the data, unlike the classical circular models like the von-Mises distribution. [8] have constructed circular-linear densities wherein the estimation of the marginals and joining densities were done using both parametric and non-parametric kernel density method. The following section throws some light on circular and linear random variable and circular-linear joint distribution.

**1.2 Circular-Linear Random Variable and Circular-Linear Distribution**

A circular random variable [9] is the one with support on the unit circle. After having chosen the initial direction and orientation of the unit circle (clockwise or anticlockwise), the distribution function of the variable, say, is defined as a function on the whole real line [16] as

If *F* is absolutely continuous, then it has a probability density function *f* such that

Further, f integrates to unity over its support, is non-negative and satisfies the periodicity condition for all and for all integers.

A linear random variable, say X, is a continuous r.v having support on the whole real line or an interval on it [8].

A circular-linear random variable takes values in the cylinder or in a subset of it and the circular-linear density of say satisfies the periodicity condition in the circular argument, i.e.

**1.3 Objective of the Study**

This paper aims to compare methods of estimation of joint density of a circular and a linear variable with the help of three real life data sets, two datasets being from the meteorological field and one being from the biological field. Therein, we have performed the maximum likelihood estimation of the best fitting l-modal circular normal density to the wind direction data of a particular data set and the distribution function of the same has been approximated using Monte Carlo method. Also, the theory of Copula function has been used to assess the dependence between the two variables.

**1.4 Data Source**

To fulfill the objectives of this paper, the first data set (Data set-I) considered is the daily data on wind direction and speed, measured at 08:30 a.m. for the years 2012 and 2013,which has been collected from the Regional Meteorological Center (RMC), Guwahati, Assam, India for the Meteorological station Mohanbari (Dibrugarh), belonging to the East Zone of Assam. The wind direction is measured in degrees as per 16 Compass whereas the wind speed has been reported in Kilometer/hour. The time period for the Monsoon season has been fixed according to the RMC norms. The second data set (Data set-II) on wind direction and ozone level taken at 06:00 a.m. at four-day intervals between April 18th and June 29th, 1975 at a weather station in Milwaukee, has been taken from [7, Appendix B.18] who procured it from [12]. The third data set (Data set-III), that on the distance and direction moved by small blue periwinkles, when they had been shifted down shore from the height at which they normally live, has been procured from [7, Appendix B.20] who reproduced it from [23].

1. **Methodology**
	1. **Estimation of Joint Density of Linear and Circular variable**

Here, three methods of estimation of circular-linear joint density have been considered initially as described below:

METHOD I:

[11] have proposed a method for obtaining circular-linear density with specified marginals of the linear and circular variable. Denoting the circular and linear marginal densities by and and their distribution functions by and respectively, the density for the circular-linear distribution is

(1)

where , the joining density, is circular in nature.

An estimator of can be obtained by estimation of the marginal and joining densities.

The first method consists in estimating the joint density using the formula in equation (1) with the help of the algorithm developed by [8]. Here, the marginals are estimated using the Maximum

Likelihood Method and the joining density has been estimated using Circular Kernel Density Method. This comprises of the ‘Mixed approach’ of estimation of the joint density of and X as both parametric and non-parametric methods of estimation have been incorporated.

METHOD II:

The second method would consist in estimating the joint density of the circular-linear variable from the joint density of two linear variables, sayby wrapping one of these variables around a unit circle, say X, using the relation [11].

But since in all the three real-life data sets that have been considered for this study, one variable is circular and another is linear in nature, this method of estimation is not applicable to any of them. The application of this estimation method requires the data to be such that both the variables are linear in nature.

METHOD III:

Another method of obtaining the circular-linear joint density can be found in [19]. The authors started with the assumption that the linear and the circular variable is independently distributed. Finally, their joint density is estimated by multiplying their assumed marginal distributions. They used the “Independent von-Mises Gaussian” distribution to model the color image segmentation in the LCH color space, wherein the circular variable (hue) follows von-Mises distribution and the linear variables (chroma and lightness) follow Gaussian distribution.

Since this estimation method requires the circular and linear variables to be independently distributed, one has to first perform a test of independence on these variables. If the hypothesis of independence is accepted, the next step would be to estimate the marginal densities of the variables and multiply their marginals to finally obtain the circular-linear joint density.

Hence, it can be seen that method II of estimation being ruled out due to non-applicability to either of the three data sets, our study would carry out a comparison between only the methods I and III through the three real-life data sets.

* 1. **Test of Uniformity using Copula**

It can be seen from equation (1) that the joint density can be expressed in terms of copula function, which makes it convenient to deal with them computationally. Copula functions are multivariate distribution functions with uniform marginals. Copula functions have got varied applicability. [24] had modeled the dependence between the bivariate binary variables by using a copula function and compared the proposed model with the odds ratio and tetrachoric model. The application of copula theory in finance has been discussed by [3]. [17] showed the use of conditional copula in assessing the Value at Risk of portfolio with two assets.

The Sklar's theorem in the theory of copula functions [20] states in the bivariate case that if is a joint distribution function with marginal distributions and respectively, then there exists a copula such that

If the marginal r.v's are absolutely continuous, Sklar's theorem can be interpreted in terms of the corresponding densities. Suppose we denote the copula density by . Then we can re-write equation (2) as

We denote the circular-linear copula density of and X by and it can be seen from the equations (1) and (3) that copula can be linked with the circular joining density as

hence, the circular-linear density in (1) can be written in terms of copula density as

 (4)

if X and Y are independent and then we have

thus, if and *X* are independent, we must have

Hence, if the populations to which the and X samples belong, are independent, the joint sample of and X must belong to the uniform distribution. This result serves as a means to test if and X is independently distributed. The uniformity of the joining density of and X, which is circular, can be assessed using any one of uniformity tests among the several uniformity tests available in the literature. This independence relation between the variables can further be verified using the linear-circular correlation coefficient developed by [15].

This clearly shows one advantage of the method-I over the other two methods, for it not only provides us with the joint density of the circular and linear variable, but also let us assess the association between them through it.

**2.3 Estimation of Best Fitting Circular Marginal density and Goodness of Fit Test for Circular Distribution**

In this section, we discuss the best fitting circular marginal densities for each of the three data sets considered for the study along with the goodness-of-fit test used.

Data set-I:

Figure (1) displays the histogram of the wind direction data measured at 08:30 a.m. during the Monsoon season for Dibrugarh meteorological station. From the histogram iit is clear that the

data has three equidistant modes. So, we chose to fit the l-modal circular normal density to this data, where l represents the number of modes in the data.

**2.3.1 *l*-modal Circular Normal Density**

The p.d.f of the *l*-modal circular normal distribution is given by

where is the modified Bessel function of the first kind and order zero; and are the mean direction and concentration parameters of the population respectively and the parameter l represents the number of modes of the distribution [10].

In the following section, the maximum likelihood estimates of the parameters of this distribution have been obtained.

**Maximum Likelihood Estimation of the Parameters of *l*-modal Circular Normal Distribution**

Let be a random sample from. Then the likelihood function is given by

Although l is a parameter and it stands for the number of modes of the distribution, we draw the histogram (density) of the data under consideration and find out an estimate of l graphically.

Taking this value of l, we obtain the m.l.e of the parameters and of the distribution, which are as shown below:

the likelihood equations are:

 (5)

and

 (6)

solving equation (5), we get the m.l.e of as

 (7)

and solving equation (6) yields the *m.l.e* of as

 (8)

where is the quadrant-specific inverse and is the inverse function of the ratio of the first and zeroth order Bessel functions of the first kind.

**Distribution Function of the *l*-modal Circular Normal Distribution**

The distribution function of the l-modal circular normal distribution is given by

(9)

it can be seen from the above equation that the integral that appears on the right hand side is complex in nature and cannot be evaluated using plain integration. We, thus take resort to the Monte Carlo approximation of the above integral.

**Monte Carlo Approximation of Integrals**

The Monte Carlo method consists in expressing the integral as the expectation of a random variable and approximating the expectation by generating random variables from the distribution of the r.v.

The equation (9) can be re-written as

where and .

Thus, t can be thought of as being distributed as a uniform random variable in the range so that becomes its p.d.f.

Let be a random sample from the distribution. Then we have the empirical approximation of as

By virtue of SLLN,

i.e. the approximation is convergent with probability 1 [1].

The standard error of the Monte Carlo estimate of (NYU Courant Computer Science, <http://www.cs.nyu.edu/courses/fall06/G22.2112-001/MonteCarlo.pdf>, online resource) is given by

where

Data set-II:

The data on wind direction as mentioned in section (1.4) follows von-Mises distribution and an account of this distribution is given below:

**2.3.2 von-Mises distribution**

The p.d.f of the von Mises distribution is given by

 (10)

where is the modified Bessel function of the first kind and order zero; and are the mean direction and concentration parameters of the population respectively. It is also referred to as the Circular Normal distribution.

The c.d.f of the von Mises distribution [10] is given by

 (11)

the parameter estimation has been done using the maximum likelihood method.

Data set-III:

The data on direction towards which periwinkles move as mentioned in section (1.4) follows Wrapped Cauchy distribution and a brief description of the distribution is given below:

**2.3.3 Wrapped Cauchy distribution**

The p.d.f. of the Wrapped Cauchy distribution is given by

 (12)

where and are the parameters of the distribution.

The c.d.f of this distribution [16] is given by

 (13)

The maximum likelihood method has been used to estimate the parameters of the distribution.

**2.3.4 Goodness of Fit Test for Circular Distribution**

By means of probability integral transformation, we can transform any continuously distributed r.v into a Uniformly distributed r.v. Analogous to the linear case, the probability integral transformation of a circular distribution corresponds to the transformation of the circle which sends the circular r.v to , where F is the c.d.f of the circular distribution of whose, orientation and initial direction have been chosen [16].

If *F* is continuous, then the transformed r.v

is distributed uniformly on the circle and any test of uniformity applied to *U* gives us a goodness-of-fit test. We consider the hypothesis to be tested as the data has come from a given distribution on the circle with distribution function We then fit the distribution to the data and get the fitted c.d.f The goodness-of-fit of the fitted distribution to the data is then assessed by checking if the transformed observations can be thought of as belonging to the uniform distribution.

In this paper, we have made use of the Watson's U2 Test for circular uniformity [16] for all the three data sets, which is the circular analogue of the Cramr-von Mises statistic W2 that arises on the real line.

**2.4 Estimation of Best Fitting Linear Marginal Density and Goodness of Fit Test**

In the following sections, we describe the best fitting linear marginal densities to each of the three data sets and the goodness of fit test used to justify their fitting.

Data set-I:

For the data on wind speed as mentioned in section (1.4), the Generalized Extreme Value distribution comes out to be the best fit. A brief account on this distribution is given below:

**2.4.1 Generalized Extreme Value Distribution**

The Generalized Extreme Value (GeV) distribution is a family of continuous probability distributions developed within extreme value theory to combine the Gumbel, Frchet and Weibull families, which is also known as type I, II and III extreme value distributions ([2], [25]). By the extreme value theorem the GEV distribution is the only possible limit distribution of properly normalized maxima of a sequence of independent and identically distributed random variables. It is widely used to model extreme events in insurance and finance. Here, we have used the GeV distribution to model the wind speed for the selected study area.

The p.d.f of the Generalized Extreme Value Distribution is given by

where are the ‘location’, ‘scale’ and ‘shape’ parameters respectively. The c.d.f of Generalized extreme value distribution is given by

The parameters of this distribution have been estimated using the Maximum Likelihood Method of estimation.

Data set-II:

The data on ozone level as mentioned in section (1.4) has been found to be best fitted by the Normal distribution. An account of the distribution is given below:

**2.4.2 Normal Distribution**

The p.d.f of the Normal distribution is given by

 (16)

The c.d.f of the distribution is given by

 (17)

the parameter estimation has been done using the maximum likelihood method.

Data set-III:

The data on distance moved by the periwinkles as mentioned in section (1.4) has been found to be best fitted by the Normal distribution, description of which is given above.

* + 1. **Goodness of Fit Test for the linear distributions**

The Chi-square Goodness-of-fit test and Kolmogorov-Smirnov one sample test has been used to assess the goodness of fit of the Gev distribution to the data on wind speed. The assessment of the goodness of fit of the normal distribution to the data on linear variable in data set II and III has been carried out using the Shapiro-Wilk normality test.

1. **Analysis and Results**

In the following sections, under each of the two applicable estimation methods of the circular-linear joint density, we show the estimation of the parameters, perform the goodness of fit of the marginal and joint densities for all the three data sets and finally, discuss the results.

**3.1 Estimation of the parameters, goodness of fit tests of the marginal and joint densities for the Data set-I:**

**3.1.1 Maximum Likelihood Estimation of Parameters of Wind Direction Density and Wind Speed Density**



Figure 1: Histogram of the wind direction data measured at 08:30 a.m. during the Monsoon season for Dibrugarh meteorological station

From the histogram of the given data presented in figure (1), it can be observed that the number of modes is 3. So, l = 3. Substituting the values of l and in (7) and (8), we get the maximum likelihood estimates of and (both measured in radians) as

The maximum likelihood estimates of the parameters of the Generalized Extreme Value distribution have been found to be

**3.1.2 Approximation of Distribution Function of *l*-Modal Circular Normal Density using Monte Carlo Method**

For approximating we draw samples of size Table (1) enlists the Monte Carlo approximations of and the corresponding standard errors for all the distinct values of appearing in the sample, which belong to the range and for different values of n. It can be seen that the standard error of the estimate decreases with increase in the sample size.

Table 1: Monte Carlo approximations of and the corresponding standard errors for different values of and *n*

|  |
| --- |
|  0.3491 0.8727 2.4435 3.1416 4.0143  |
|  *Se* *Se* *Se* *Se* *Se* |
| 100 0.0562 0.0230 0.1398 0.0347 0.3896 0.0488 0.5007 0.0500 0.6386 0.0481 |

 1000 0.0562 0.0073 0.1398 0.0110 0.3896 0.0154 0.5007 0.0158 0.6386 0.0152

 10000 0.0562 0.0023 0.1398 0.0035 0.3896 0.0049 0.5007 0.0050 0.6386 0.0048

 100000 0.0562 0.0007 0.1398 0.0011 0.3896 0.0015 0.5007 0.0016 0.6386 0.0015

|  |
| --- |
|  |

 4.3633 4.7124 5.5851 5.9341 6.2832

|  |
| --- |
|  |

 *Se* *Se* *Se* *Se* *Se*

|  |
| --- |
|  |

 100 0.6948 0.0461 0.7509 0.0347 0.8888 0.0315 0.9441 0.0232 1.0000 0.0027

 1000 0.6948 0.0146 0.7509 0.0110 0.8889 0.0030 0.9441 0.0073 1.0000 0.0008

 10000 0.6948 0.0046 0.7509 0.0035 0.8889 0.0010 0.9441 0.0023 1.0000 0.0008

 100000 0.6948 0.0015 0.7509 0.0011 0.8889 0.0001 0.9441 0.0007 1.0000 0.0003

|  |
| --- |
|  |

**3.1.3 Goodness-of-Fit Test for Wind Direction Density and Wind Speed Density**

For wind direction density, we frame our null hypothesis as

H01: The data has come from *l*-modal circular normal distribution

Using the goodness-of-fit test based on the Watson's *U2* Test for circular uniformity, we get the value of the test statistic as.

Whereas the critical value at 1% level of significance has come out to be 0.267. The *p-*value of the test is 0.056 > 0.05. Hence, we accept our null hypothesis H01 and conclude that possibly, the data has come from the *l*-modal circular normal distribution.

Figure (2) shows the e.c.d.f plot for the wind direction data measured at 08:30 a.m. during the Monsoon season for Dibrugarh meteorological station. The plot clearly shows that the *l*-modal circular normal distribution is a good fit to the data on wind direction.

For wind speed density, we frame our null hypothesis to be tested as

H02: The data has come from Generalized Extreme Value population

The Chi-square test statistic and the Kolmogorov-Smirnov test statistic calculated from the data have been found to be

Whereas the critical value of these statistics at 5% level of significance are 7.815 and 0.133 respectively. The p-value for chi-square test is 0.92465623 > 0.05 and that for the Kolmogorov-Smirnov test is 0.999 > 0.05. Hence, we accept H02 and conclude that possibly, the data has come from Generalized Extreme Value distribution.

Figure (3) shows the e.c.d.f plot for the wind speed data measured at 08:30 a.m. during the Monsoon season for Dibrugarh meteorological station. The plot clearly shows that the Generalized Extreme Value distribution is a good fit to the data on wind speed.



Figure 2: e.c.d.f plot for the wind direction data measured at 08:30 a.m. during the Monsoon season for Dibrugarh meteorological station.



Figure 3: e.c.d.f plot for the wind speed data measured at 08:30 a.m. during the Monsoon season for Dibrugarh meteorological station

**3.1.4 Estimation of the Joint Density of Wind Speed and Wind Direction**

Method-I:

We see that the l-modal circular normal density and the Generalized Extreme Value density have been fitted to the wind direction and wind speed respectively and both have come out to be good fit to the respective data sets. The joining circular density has been estimated using Circular Kernel density method with Gaussian Kernel and least-squares cross-validatory bandwidth.

For testing the uniformity of the joint sample of and X, we apply the Watson's U2 circular uniformity test to it. The value of the test statistic has been found to be

Whereas the critical value at 5% level of significance is 0.187. The *p-*value of the test is 0.00 < 0.05. Hence, we reject the hypothesis of uniformity and conclude that the joining circular density is not uniform. This leads us to the rejection of the hypothesis of independence of and *X*. Also, it can be seen that the value of the linear-circular correlation coefficient for this sample is 0.05 and the *p*-value for the test of independence based on this coefficient is 0.017 < 0.05 [4].

This shows that there is a relation between the wind speed and direction measured at 08:30 a.m. during Monsoon season for Dibrugarh meteorological station. Further, figure (4) shows the histogram of the joint sample of and X. It is clear from this figure that the sample has not come from a uniform distribution.



Figure 4: Histogram of the joint sample of wind speed and direction measured at 08:30 a.m. during the Monsoon season for Dibrugarh meteorological station

Method-III:

We first perform the test of independence of wind direction and speed, which is a pre-requisite to the application of this method. As reported in section (3.1.4), under Method-I, the p-value for the test of independence based on this coefficient is 0.017 < 0.05. This means the hypothesis of independence is rejected and so, we cannot apply the third method of estimation of the circular-linear joint density. Figure (5) displays the joint density plot of wind direction and speed.



Figure 5: Joint density plot of wind speed and direction measured at 08:30 a.m. during the Monsoon season for Dibrugarh meteorological station

**3.2 Estimation of the parameters, goodness of fit tests of the marginal and joint densities for the Data set-II:**

**3.2.1 Maximum Likelihood Estimation of Parameters of Wind Direction Density and Ozone Level Density**

The maximum likelihood estimates of and (both measured in radians) of von Mises distribution are

The maximum likelihood estimates of parameters and of the Normal distribution are

**3.2.2 Goodness-of-Fit Test for Wind Direction Density and Ozone Level Density**

For Wind Direction density, the hypothesis to be tested is

H03: The data has come from von Mises distribution

Using the goodness-of-fit test based on the Watson's U2 Test for circular uniformity, we get the value of the test statistic as.

Whereas the critical value at 1% level of significance has come out to be 0.079. The p-value of the test is > 0.05. Hence, we accept our null hypothesis H03 and conclude that possibly, the data has come from the von Mises distribution.

Figure (6) shows the P-P plot of the von-Mises distribution fitted to the wind direction data measured at 06:00 a.m. at four-day intervals at a weather station in Milwaukee. It clearly shows that von Mises distribution is a good fit to the data.

For Ozone Level density, we frame our null hypothesis to be tested as

H04: The data has come from Normal distribution

The Shapiro-Wilk Normality test statistic based on the data has been found to be

The p-value for the test is 0.8605 > 0.05. We thus, accept H04 and conclude that possibly, the data has come from Normal distribution.

Figure (7) displays the Q-Q plot of the Normal distribution fitted to the ozone level data measured at 06:00 a.m. at four-day intervals at a weather station in Milwaukee. It is evident from the graph that the Normal distribution is a good fit to the data.



Figure 6: von Mises P-P plot of wind direction measurement taken at 06:00 a.m. at four-day intervals at a weather station in Milwaukee



Figure 7: Normal Q-Q plot of ozone level measurement taken at 06:00 a.m. at four-day intervals at a weather station in Milwaukee

**3.2.3 Estimation of the Joint Density of Wind Speed and Ozone Level**

Method-I:

The von Mises distribution and the Normal distribution have been found to be the best fit to the data on wind direction and ozone level respectively. The Circular Kernel density method with Gaussian Kernel and least-squares cross-validatory bandwidth has been used to estimate the joining circular density .

The value of the Watson's *U2* circular uniformity test statistic for testing the uniformity of the joint sample of and X has been found to be whereas the critical value at 5% level of significance is 0.187. The *p-*value of the test is 0.00 < 0.05. Hence, we reject the hypothesis of uniformity and conclude that the joining circular density is not uniform. This leads us to the rejection of the hypothesis of independence of and *X*. This is further verified by the observation that the value of the linear-circular correlation coefficient for this sample is 0.522 and the *p*-value for the test of independence based on this coefficient is 0.003 < 0.05. Also, the histogram of the joint sample of and *X* clearly presented in figure (8) shows that it does not belong to the uniform distribution.

Method-III:

Since the application of the third method requires the independence of the linear and the circular variable, we first carry out the test of their independence. As reported in section (3.2.3), under Method-I, the p-value for the test of independence based on this coefficient is 0.003 < 0.05. This indicates the rejection of the hypothesis of independence and so, third method of estimation of the joint density of wind direction and ozone level is not applicable in this case also.

Figure (9) displays the joint density plot of wind direction and ozone level.



Figure 8: Histogram of the joint sample of wind speed and ozone level measured at 06:00 a.m. at four-day intervals at a weather station in Milwaukee

Figure 9: Joint density plot of wind speed and ozone level taken at 06:00 a.m. at four-day intervals at a weather station in Milwaukee

**3.3 Estimation of the parameters, goodness of fit tests of the marginal and joint densities for the Data set-III:**

**3.3.1 Maximum Likelihood Estimation of Parameters of Direction Density and Distance Density**

The maximum likelihood estimates of parameters and (both measured in radians) of the Wrapped Cauchy distribution have been found to be

The maximum likelihood estimates of the parameters and of the normal distribution are

**3.3.2 Goodness-of-Fit Test for Direction Density and Distance Density**

For the direction density, the hypothesis to be tested is

H05: The data has come from Wrapped Cauchy distribution

Using the goodness-of-fit test based on the Watson's U2Test for circular uniformity, we get the value of the test statistic as whereas the critical value at 5% level of significance has come out to be 0.187. The *p-*value of the test is > 0.05. Hence, we accept our null hypothesis H05 and conclude that possibly, the data has come from the Wrapped Cauchy distribution.

Figure (10) depicts the e.c.d.f plot of the Wrapped Cauchy (WC) distribution fitted to the data on direction towards which moved by periwinkles and it is evident from the graph that the WC distribution is a good fit to the data.

For the distance density, the null hypothesis to be tested is

H06: The data has come from Normal distribution

The Shapiro-Wilk Normality test statistic based on the data has been found to be . The *p*-value for the test is 0.2652 > 0.05. We thus, accept H06 and conclude that possibly, the data has come from Normal distribution.

Figure (11) shows the Q-Q plot of the Normal distribution fitted to the data on distance moved by periwinkles. It is clear from the graph that the Normal distribution is a good fit to the given data.



Figure 10: e.c.d.f plot for the data on direction towards which moved by the periwinkles on moving them down shore from the height at which they normally live



Figure 11: Normal Q-Q plot for the data on distance moved by the periwinkles on moving them down shore from the height at which they normally live

**3.3.3 Estimation of the Joint Density of Direction and Distance Density**

Method-I:

The Wrapped Cauchy distribution and the Normal distribution have been found to be the best fit to the data on direction towards which moved and the distance moved, respectively. The joining circular density has been estimated using Circular Kernel density method with Gaussian Kernel and least-squares cross-validatory bandwidth.

The value of the Watson's *U2* circular uniformity test statistic for testing the uniformity of the joint sample of and X has been found to be

Whereas the critical value at 5% level of significance is 0.187. The *p-*value of the test is 0.00 < 0.05. Hence, we reject the hypothesis of uniformity and conclude that the joining circular density is not uniform. This leads us to the rejection of the hypothesis of independence of and *X*. It can further be verified by observing that the value of the linear-circular correlation coefficient for this sample is 0.2935 and the *p*-value for the test of independence based on this coefficient is 0.008 < 0.05.

Figure (12) which presents the histogram of the joint sample of direction and distance moved by the periwinkles also justifies the claim that the population which the sample comes from is not uniformly distributed.



Figure 12: Histogram of the joint sample of direction and distance moved by the periwinkles on moving them down shore from the height at which they normally live

Method-III:

Since the application of the third method requires the independence of the linear and the circular variable, we first carry out the test of their independence. As reported in section (3.3.3), under Method-I, the *p*-value for the test of independence based on this coefficient is 0.008 < 0.05. This indicates the rejection of the hypothesis of independence of the direction and distance moved and so, third method of estimation of their joint density is not applicable in case of the data set-III,Figure (13) shows the joint density plot of the direction and distance moved.

Figure 13: Joint density plot of the direction and distance moved by the periwinkles on moving them down shore from the height at which they normally live

**7. Summary and Discussion**

This paper mainly attempts to compare the methods of estimation of joint density of a circular and a linear variable using three real life data sets. One of the data sets consist in observations on wind direction and speed; the second data set contains observations on wind direction and ozone level and the third data set comprises of observations on direction towards which moved and distance moved by periwinkles. Both the methods of estimation of the joint density of a circular and a linear variable call for the estimation of their marginal densities. So, their estimation has been done and the goodness-of-fit tests have also been performed. Therein, under the first data set, we have carried out the maximum likelihood estimation of the parameters of the *l*-modal circular normal density and approximation of its distribution function with the help of Monte Carlo method, which comes out to be a good fit to the data on wind direction. Furthermore, the first method of joint density estimation (Method-I) enables one to use the theory of Copula function to explore the relationship between the linear and circular variable. The test of uniformity of the circular joining density based on Copula function, in case of all the three data sets reveal that there is an association between the circular and linear variable under consideration. The result has been verified by tallying it with the value of the linear-circular correlation coefficient in each case. Method-III could not be applied to any of the data sets because the linear and circular variable is independent in neither of these cases, which is the necessary for the application of this method. As far as the Method-II is concerned, it has been ruled out at the very outset because the form in which the data arise doesn’t allow the application of this method.

Thus, in the end, it can be concluded that the first method viz. Method-I of the circular-linear joint density estimation is superior as compared to the method-III (and also, method-II) because it does not require any condition to hold good for its application and also, the circular joining density via the theory of Copula function can be used to assess the dependence between the variables, without having to calculate a correlation coefficient separately.

1. **Acknowledgements**

The authors would like to thank the Department of Science and Technology, New Delhi, for providing financial assistance to Miss Sahana Bhattacharjee through the Innovation in Science Pursuit for Inspired Research (INSPIRE) programme. Also, thanks are due to the anonymous reviewers whose valuable suggestions have led to the enhancement of the quality over the earlier version of the manuscript.

**References**

[1] Caflisch, R.E. (1998). Monte Carlo and Quasi-Monte Carlo Methods. *Acta Numerica*, **7**, 1-49.

[2] Cheng, E. and Yeung, C. (2002). Generalized extreme gust wind speeds distributions. *J Wind Eng Ind Aerodyn*, **90**:1657–1669.

[3] Chiou, S.C. and Tsay, R.S. (2008). A Copula-based Approach to Option Pricing and Risk Assessment. *Journal of Data Science,* **6**: 273-301.

[4] Das, K.K. and Bhattacharjee, S. (2016). The inter-relationship between wind direction and wind speed for 5 meteorological stations across Assam, India: A case study. *International Journal of Advanced Research,* **4(5),** 591-603.

[5] Fernández-Durán, J.J. (2004). Circular distributions based on nonnegative trigonometric sums. *Biometrics*, **60**, 499-503.

[6] Fernández-Durán, J.J. (2007). Models for Circular-Linear and Circular-Circular Data Constructed from Circular Distributions Based on Nonnegative Trigonometric Sums. *Biometrics*, **63**, 575-589.

[7] Fisher, N.I. (1993). *Statistical Analysis of Circular Data.* Cambridge University Press, Cambridge.

[8] Garcia-Portugués, E., Crujeiras, R.M. and González-Manteiga, W. (2011). Exploring Wind Direction and SO2 concentration by circular-linear density estimation. *Reports in Statistics and Operations Research*, Universidade De Santiago De Compostela, Departamento De Estatstica E Investigacin Operativa, 11-01.

[9] Jammalamadaka, S.R. and Lund, U.J. (2006). The effect of wind direction on ozone levels: a case study. *Environmental and Ecological Statistics*, **13**, 287- 298.

[10] Jammalamadaka, S.R. and Sengupta, A. (2001). *Topics in Circular Statistics*. World Scientific Publishing Co. Pte. Ltd., Singapore.

[11] Johnson, R.A. and Wehrly, T.E. (1978). Some Angular-Linear Distributions and Related Regression Models. *Journal of the American Statistical Association*, **73**, 602-606.

[12] Johnson, R.A. and Wehrly, T.E. (1977). Measures and models for angular correlation and angular-linear regression. *J. R. Statist. Soc.* **B 39,** 22-229.

[13] Kollu, R., Rayapud, S., Narasimham, S., and Pakkurthi, K. (2012). Mixture probability distribution functions to model wind speed distributions. *International Journal of Energy and Environmental Engineering*, **3:27,** 1–10

[14] Krstulovic, N. and Solic, M. (1991). Spatial distribution of faecal pollution indicators in the kastela bay under different meteorolgical conditions. *Acta Adriatica*, **32**, 827-837.

[15] Mardia, K.V. (1976). Linear-circular correlation coefficients and rhymometry. *Biometrika*, **63**, 403-405.

[16] Mardia, K.V. and Jupp, P.E. (2000). *Directional Statistics*. John Wiley & Sons Ltd, Chichester.

[17] Palaro, H.P. and Hotta, L.K. (2006). Using Conditional Copula to Estimate Value at Risk. *Journal of Data Science,* **4**: 93-115.

[18] Roslev, P., Bastholm, S., and Iversen, N. (2008). Relationship between fecal indicators in sediment and recreational waters in a danish estuary. *Water, Air & Soil Pollution*, **194,** 13-21.

[19] Roy, A., Parui, S.K. and Roy, U. (2012). A Mixture Model of Circular-Linear Distributions for

Color Image Segmentation. *International Journal of Computer Applications,* **58(9),** 6-11.

[20] Sklar, A. (1959). Fonctions de répartition *n* dimensions et leurs marges. *Publications de l’Institut de Statistique de L’Université de Paris*, **8**, 229–231

[21] Smith, P., Carroll, C., Wilkins, B., Johnson, P., Nic Ganhainn, S., and Smith, L.P. (1999). The effect of wind speed and direction on the distribution of sewage- associated bacteria. *Letters in Applied Microbiology*, **28**, 184-18

 [22] Somerville, M. C., Mukherjee, S., Fox, D. L. and Stevens, R.K. (1994). Statistical approaches in sector analyses for assessing local source impacts. *Atmospheric Environment*, **28**, 3483-3493.

[23] Underwood, A.J. and Chapman, M.G. (1992). Experiments on topographic influences on density and dispersion of *Littorina unifasciata* in New South Wales In P*roceedings of the Third International Symposium of Littorinid Biology.* The Malacological Society of London.

[24] Li, X., Li, L. and Fang, R. (2014). Copula-based Logistic Regression Models for Bivariate Binary Responses. *Journal of Data Science,* **12**: 461-476.

[25] Ying, A. and Pandey, M. (2007). The rth largest order statistics model for extreme wind speed estimation. *J Wind Eng Ind Aerodyn*, **95(3)**:165–182.

Sahana Bhattacharjee

Department of Statistics

Gauhati University

Guwahati, Assam, India

Phone no.: +9195086-60795

Email: sahana.bhattacharjee@hotmail.com

Kishore Kumar Das

Department of Statistics

Gauhati University

Guwahati, Assam, India

Email: daskkishore@gmail.com