



## Analysis of Malaysian Wind Direction Data Using ORIANA

Siti Fatimah Hassan (Corresponding author)

Centre for Foundation Studies in Science

University of Malaya, 50603 Kuala Lumpur, Malaysia

Tel: 6-013-2508490 E-mail: s\_fatimahh@yahoo.com

Abdul Ghapor Hussin

Centre for Foundation Studies in Science

University of Malaya, 50603 Kuala Lumpur, Malaysia

Tel: 6-03-7967-5996 E-mail: ghapor@um.edu.my

Yong Zulina Zubairi

Centre for Foundation Studies in Science

University of Malaya, 50603 Kuala Lumpur, Malaysia

Tel: 6-03-7967-3241 E-mail: yzulina@um.edu.my

### Abstract

Of late, analysis of circular variables or directional data have gained much attention as they describe most of the environmental phenomena such as waves, wind gust, tornados and others. Unlike linear data, the availability of statistical software dedicated to analyze of directional data is scarce. Furthermore, the analyses are limited to descriptive summary, point estimation and comparison of means. This could partly due to the difficulty in statistical analysis of circular data because of disparate topologies between circle and straight line. For example, if the angles are recorded in the range  $[-\pi, \pi]$  radian or  $(0^\circ, 360^\circ]$ , then the direction close to the opposite end-points are near neighbours in a metric if we refer to the topology of circle, but maximally distant in linear metric. Thus, the “distance” between  $350$  and  $15$  angular degrees is more commonly thought as  $25^\circ$  opposed to the  $335^\circ$  as a standard calculation. In this paper, we describe the analysis of Malaysian wind direction data using the newly improved statistical software, ORIANA designed to analyze circular data. Exploratory data analysis (EDA) based on descriptive statistics, graphical display of the data and comparison of samples are discussed.

**Keywords:** Directional data, Circular data, ORIANA software, Wind direction

### 1. Introduction

We come across directional or circular data almost everywhere in applied science. They are widely used in biology, geography, astronomy, meteorology, medicine and many other areas. Data measured in the form of angles or two dimensional orientations is unlike the linear data and it cannot be treated in the same way as linear data. Furthermore, most of the methods used in statistical analysis of linear data cannot be applied in circular data due to different topology between circle and straight line. Thus, the need of analyzing directional or circular data is really indispensable. For the past few years, many new techniques for analyzing circular data have been developed, for example see Fisher (1993), Mardia (2000) and Jammalamadaka (2001). However, these are often computationally intensive and to find the suitable friendly software is always an issue. Researchers in the area of directional statistics are looking for the friendly software like SPLUS and MATLAB that have been widely used for linear data.

### 2. The ORIANA Software

ORIANA, the statistical software dedicated to the analysis of circular variables, was first introduced on 31<sup>st</sup> December 2003 and was further upgraded with the latest version issued on 3<sup>rd</sup> May 2007. In view of the scarcity of commercialize statistical software in the market, the software can be useful for students, researcher, scientists and also medical professional in areas which involve in analyzing circular data. It has a window-based environment with several options in the drop down menu. There are five windows in ORIANA including the main window, status window, data editor, results window, graph windows and notepad. The tutorial guide also provided in ORIANA in order to help users

become familiar with this software. It offers various graphical and analytical analysis as well as to calculate a variety of the special types of statistics involving the data measured in degrees, time of day or other circular form. Some of the features that are available in ORIANA include summary statistics for each sample, one sample test, multi sample test and also correlation between samples. As an illustration, the statistical analysis of Malaysian wind direction data is performed using this software.

### 3. Malaysian Wind Direction Data

In this study, the data was obtained from Malaysian Meteorological Services Department and it consists of the observations of surface wind at Kuala Lumpur International Airport (KLIA) station. The data was recorded daily in 2005 at time 1200 and was measured by anemometer at location 16.3 m above ground level, latitude 02°44' N and longitude 101°33' E. Two sets of data have been considered in this study known as Northeast monsoon data (January and February 2005) and Southwest monsoon data (July and August 2005).

### 4. Descriptive Statistics of Malaysian Wind Direction Data

#### 4.1 Circular plot

In most exploratory data analysis, graphical representations are often used to summarize the data. ORIANA offers wide range of representation which included rose diagram, circular histogram, raw data plot, arrow data plot and linear histogram. However, the most commonly used types of graph for circular data is the rose diagram. The rose diagram is a histogram displayed in a circle, similarly to the pie chart for linear data. However, each sector represents the frequency or number of observations which falls in the range of angles. The concentric circles show the frequency of the observations for each angular value. Fig. 1 and Fig. 2 show the rose diagram of wind direction for Northeast and Southwest data set respectively. Also shown in the plots are the mean direction of the observation which denoted by thick line running from the center of the diagram to the outer edge and is given by 50° and 240° for Northeast and Southwest data set respectively together with their confidence interval which can be set either at 95% or 99% level.

Another diagrammatical representation available in ORIANA is the circular histogram. The circular histogram as shown in Fig. 3 and Fig. 4 are quite similar to the rose diagram, except that the wedge-shaped in rose diagram become parallel-sided bar that show the number of observations within that class range.

Alternatively, raw data plots are shown in Fig. 5 and Fig. 6 can be used to view the spread of the data and describe the distribution of the data.

Another graph shown in Fig. 7 and Fig. 8 are the arrow graphs and these graphs are similar to rose diagram, but instead of wedge-shaped sector they have arrows to indicate the number of observations for each class.

#### 4.2 Basic statistics for each sample

To complement the graphical analysis, summary statistics are often deployed in the EDA. Similar to any window-based environment, the summary statistics in ORIANA can be obtained from the drop down menu. In other words, from the Analyses menu, drop down to choose Stats... and further to the Statistics dialog box, in which one can choose several summary statistics and statistical tests. For this study, the summary statistics are given as in Table 1.

The summary statistics as shown in Table 1 depicts the following:

- The numbers of observations for this study are 59 and 62 observations for Northeast and Southwest respectively.
- The mean direction ( $\mu$ ) is the direction of the resultant vectors with given corresponding angles and is defined by

$$\mu = \begin{cases} \tan^{-1}\left(\frac{S}{C}\right), & S > 0, C > 0 \\ \tan^{-1}\left(\frac{S}{C}\right) + \pi, & C < 0 \\ \tan^{-1}\left(\frac{S}{C}\right) + 2\pi, & S < 0, C > 0 \end{cases}$$

where  $S = \sum_{i=1}^n \sin(\theta_i)$  and  $C = \sum_{i=1}^n \cos(\theta_i)$ . For Northeast data, the mean angle is 52.189°, while for Southwest data the mean angle is 247.019°.

- Length of mean vector is defined by  $r = \frac{1}{n} \sqrt{C^2 + S^2}$ . It is the length of the directions of the resultant vectors at given angles and the range is between 0 and 1. The larger value of  $r$  implies that the observations are closely clustered around the mean. For this data set, the value of  $r$  is 0.828 and 0.566 for Northeast and Southwest data set respectively. The value of  $r$  for Northeast is higher than Southwest and it is closer to 1 meaning that the observations for Northeast are clustered closely around the mean as compared to Southwest.

- Median is defined as a direction that divides the data into two equal size groups and in ORIANA the median may be calculated by minimizing the function of  $d(\theta) = \pi - \frac{1}{n} \sum |\pi - |\theta_i - \theta||$ . As opposed to linear data, where we can easily find the midpoint between observations as we arranged the data in increasing order. However, for circular data it is rather complicated since the data are in closed curve and can always be rotated around the circle. This has caused a problem on how to choose an appropriate axis on the circular scale. For the given data set, the calculated median is  $60^\circ$  for Northeast data set and  $245^\circ$  for Southwest data set.
- Concentration is denoted by  $\kappa$  and is a parameter that related to von Mises distribution. It is also related to the length of mean vector and the value of  $\kappa$  given by ORIANA is the maximum likelihood estimation of population concentration. A large value of  $r$  will imply the large value of  $\kappa$ . For this data set, the concentrations are 3.246 and 1.377 for Northeast and Southwest data set respectively.
- Circular variance is the measures of dispersion of circular data. Variance is related to length of mean vector and calculated by using  $V = 1 - r$ . The values of circular variance for the Northeast and Southwest data set are 0.172 and 0.434 respectively.
- Circular standard deviation is an analogue to linear counterpart but it is calculated in a much different way and is given by  $S = \sqrt{(-2 \ln(r))}$ . The values are  $35.236^\circ$  and  $61.137^\circ$  for Northeast and Southwest data set respectively.
- Standard error of mean are calculated based on the length of mean vector and concentration parameters as well as assuming that the data set follows a von Mises distribution. For the given data set, the standard error of mean is  $4.551^\circ$  and  $8.242^\circ$  for Northeast and Southwest data set respectively.
- Finally, the 95% and 99% confidence intervals are derived from the standard error as for the normal distribution. It is defined as 95% or 99% probability that the true mean vector is greater and less than this value. At 95% confidence interval, the values are between  $43.268^\circ$  and  $61.11^\circ$  for Northeast data set whereas  $230.861^\circ$  and  $263.177^\circ$  for Southwest data set. While at 99% confidence interval, the value is between  $40.466^\circ$  and  $63.913^\circ$  for Northeast data set whereas  $225.785^\circ$  and  $268.252^\circ$  for Southwest data set.

#### 4.3 Testing for uniformity

The samples can be tested whether they are uniformly distributed or otherwise, that is to test whether all directions are equal likely. By using one sample test such as Rayleigh's Uniformity Test where the Z value is calculated simply as  $Z = nr^2$ , with  $n$  is the number of observations and  $r$  is the length of the mean vector. A longer mean vector will give larger value of Z and greater concentration of the data around the mean. Thus, the likelihood of the data being uniformly distributed is less. From the test statistics, it gives a Z value of 40.42 with p-value  $< 0.001$  for Northeast data set and 19.857 with p-value  $< 0.001$  for Southwest data set. Hence at 5% significant level, we reject the null hypothesis that the samples were uniformly distributed.

#### 4.4 Comparison between samples

Watson-William F-Test can be used to compare two or more samples to determine if their mean angles differ significantly. The F statistic is the same as Fisher's variance ratio statistic which is commonly used in linear statistics. This test assumes that the two samples are independent and drawn at random from a population with a von Mises distribution. It also assumes that the concentrations of the two samples are similar and that they are sufficiently large, normally greater than 2. For this data set, the test statistics give the value of 254.151 and the probability value associated with the null hypothesis is less than 0.001. Hence, null hypothesis can be rejected at 5% significant level implying that the mean directions for the two monsoons are not the same.

### 5. Conclusion

We found that ORIANA is user friendly software, and proves to be a useful package in analyzing circular data. For EDA, it offers a wide selection of graphical display and descriptive statistics which deemed sufficient in statistical analysis of circular data. Nevertheless it has some limitations. For example, other features that is available for linear data set such as measure of skewness and kurtosis. Further analysis such as statistical inference, probability distribution function and predictive modeling such as regression are also not available in the statistical package. Thus, there is much to be done in developing a more comprehensive analysis of circular data which can be incorporated into the statistical package.

### References

- Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge University Press.
- Jammalamadaka, S.R. & SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific Publishing Co. Pte. Ltd.
- Mardia, K. V. & Jupp, P.E. (2000). *Directional Statistics*. John Wiley, Chichester.
- ORIANA Software Version 2.02e

Table 1. Summary statistics for each sample

Variable	Northeast (Jan & Feb)	Southwest (July & Aug)
Data Type	Angles	Angles
Number of Observations	59	62
Data Grouped?	No	No
Group Width (& Number of Groups)		
Mean Vector ( $\mu$ )	52.189°	247.019°
Length of Mean Vector ( $r$ )	0.828	0.566
Median	60°	245°
Concentration	3.246	1.377
Circular Variance	0.172	0.434
Circular Standard Deviation	35.236°	61.137°
Standard Error of Mean	4.551°	8.242°
95% Confidence Interval (-/+ ) for $\mu$	43.268° 61.11°	230.861° 263.177°
99% Confidence Interval (-/+ ) for $\mu$	40.466° 63.913°	225.785° 268.252°

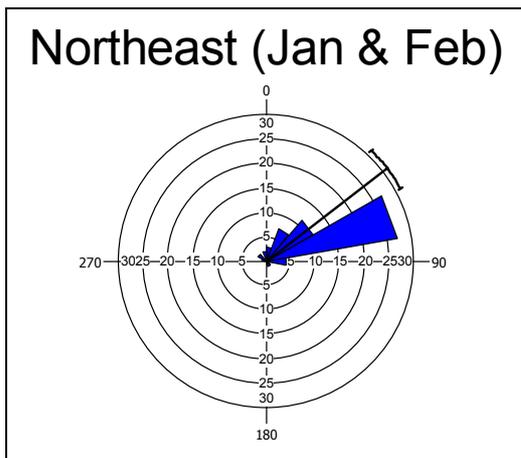


Figure 1. Rose diagram of Northeast data set.

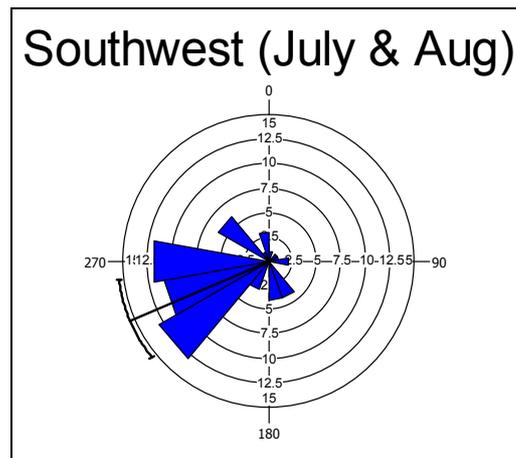


Figure 2. Rose diagram of Southwest data set.

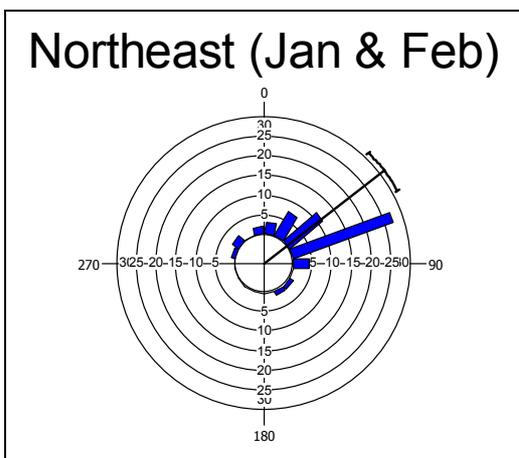


Figure 3. Circular histogram of Northeast data set.

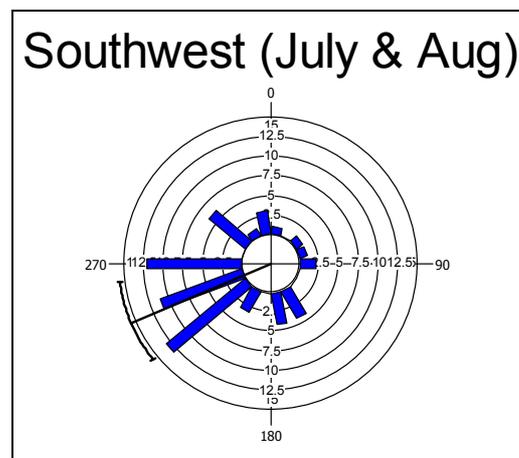


Figure 4. Circular histogram of Southwest data set.

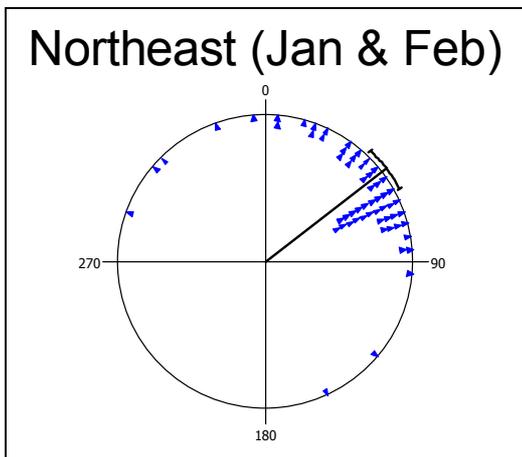


Figure 5. Raw data plot of Northeast data set.

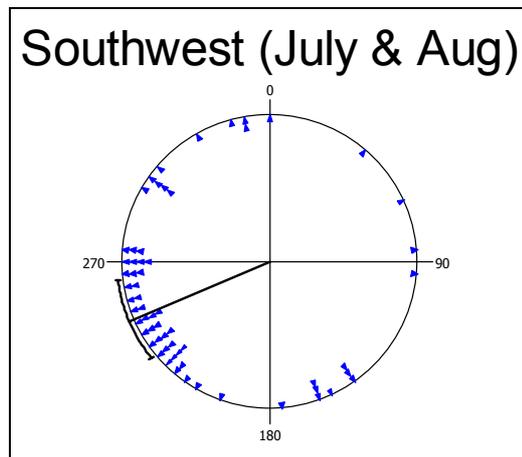


Figure 6. Raw data plot of Southwest data set.

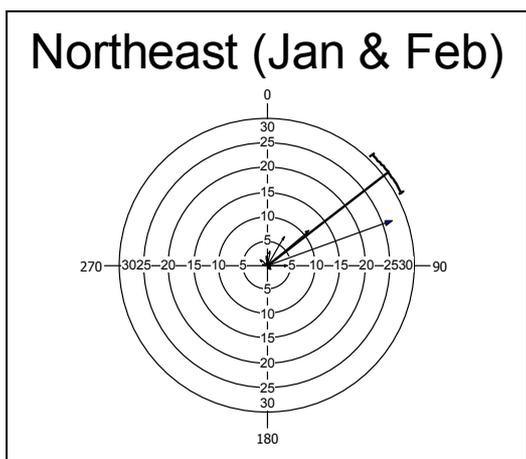


Figure 7. Arrow graph of Northeast data set.

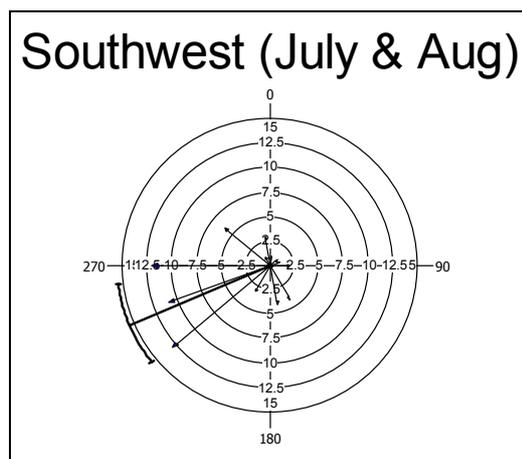


Figure 8. Arrow graph of Southwest data set.