Computer and Information Science
www.ccsenet.org/journal.html
Vol. 1, No. 4
November 2008

An Alternative Analysis of Two Circular Variables via Graphical Representation: An Application to the Malaysian Wind Data

Yong Zulina Zubairi (Corresponding author)
Centre for Foundation Studies in Science
University of Malaya
50603 Kuala Lumpur, Malaysia
Tel: 6-03-7967-5971 E-mail:yzulina@um.edu.my
Fakhrulrozi Hussain
Institutes of Postgraduates Studies
University of Malaya
50603 Kuala Lumpur, Malaysia
Tel: 6-03-7967-4263 E-mail:fakhrul@perdana.um.edu.my
Abdul Ghapor Hussin
Centre for Foundation Studies in Science
University of Malaya
50603 Kuala Lumpur, Malaysia
Tel: 6-03-7967-5821 E-mail:ghapor@um.edu.my

Abstract

The relationship between variables is vital in data analysis. The scatter plot, for instance, gives an easy preliminary exploratory analysis for finding relationship between two variables, if any. Statistical method such as correlation and linear relationship are standard tools in most statistical packages. For circular variables that take value on the circumference of a circle, the analysis however is different from those of the Euclidean type variables because circumference is a bounded closed space. Unlike linear variable, standard statistical packages for circular variables are limited. This paper proposes a graphical representation of two circular variables as a preliminary analysis using the MATLAB environment. A plot called Spoke plot is developed to visually display relationship between two circular variables and linear correlation. As an illustration, the Malaysian wind data is used in the analysis. This new type of representation promises an alternative approach in the preliminary analysis of circular data.

Keywords: Spoke plot, Wind data, Circular, Correlation, Linear Relationship

1. Introduction

Circular variables take values on the circumference of a circle and thus are bounded in a closed space. Unlike the usual Euclidean type variables, the analysis of directional data is different. Further readings in circular variables can be found in Fisher (1993) and Mardia (1999, 1972). A number of environmental data are circular in nature such as wave, wind direction, compass bearing, clock and others. Statistical softwares devoted to the analysis of circular variable are limited. At present, there are only two softwares in the market namely AXIS (Handerson et el., 2002) and ORIANA (Oriana Software, 2007), that offer some statistical analysis of circular variable on a window-based environment. Although these packages offer a range of graphical and analytical techniques required for statistical analysis of circular data, they have some limitations. For example, further analysis such as statistical inference, analyzing grouped data sets, circular plots of the corresponding probability density functions as well as regression and correlation of circular data are not available in the existing packages. In this study, an alternative diagrammatical representation of correlation and linear relationship analysis between two circular variables are developed. By using Matlab (Matlab Software, 2004), the programs generate graphical and calculated outputs of the analysis with a window-based environment. By interfacing with the existing softwares, this analysis could enhance the exploratory analysis of circular variable with respect to software development.

2. Theoretical formulation of the analysis of circular data

2.1 Circular data

Data on the angular displacements, directional propagations and in general periodic occurrence can be casted within the framework of directional or circular data. In other words, circular data is one which takes values on the circumference of a circle, i.e. they are angles in the range of $(0,2\pi)$ radians or $(0^0,360^0)$. To analyze this type of data, we must use techniques differing from those of the usual Euclidean type variables because the circumference is a bounded closed space, for which the concept of origin is arbitrary or undefined. Thus, the techniques that have been used for continuous linear data cannot be applied for circular data. Furthermore, continuous linear data are realized on the straight line or real line which may be analyzed straightforwardly by usual statistical techniques.

2.2 Analysis between two circular variables

As mentioned earlier, unlike linear variables, standard statistical package for analysing circular variables is limited (Friendly, 2002) and evidence on the utilization of such packages in published works is also few. For example Hussin et el. (2006) evaluated the performance of AXIS by running several exploratory analysis of the Malaysian Wind data. This paper will focus on the analysis of two circular variables. The correlation measure and linear relationship will be used in the development of the preliminary analysis of two circular variables in addition to a graphical representation.

2.3 Correlation between circular variables

Correlation or also known as a measure of a correlation coefficient indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. When the data are linear there are several coefficients, measuring the degree of correlation, adapted to the nature of data.

A similar measure of association between two circular variables is not well known. Given n pairs of circular data $(\theta_1,\varphi_1),...,(\theta_n,\varphi_n)$, where $0 \le \theta_i,\varphi_i < 2\pi$ of circular variables θ and φ, the circular correlation coefficient given by Fisher is defined by

$$\hat{\rho}_T = \frac{\sum\limits_{1\le i\le j\le n}\sin(\theta_i-\theta_j)\sin(\varphi_i-\varphi_j)}{\sqrt{\sum\limits_{1\le i\le j\le n}\sin^2(\theta_i-\theta_j)\sum\limits_{1\le i\le j\le n}\sin^2(\varphi_i-\varphi_j)}} \tag{1}$$

Alternatively, one can transform equation (1) to

$$\hat{\rho}_T = \frac{4(AB-CD)}{\sqrt{(n^2-E^2-F^2)(n^2-G^2-H^2)}} \tag{2}$$

where

$A=\sum(\cos\theta_i\cos\varphi_i)$ $B=\sum(\sin\theta_i\sin\varphi_i)$

$C=\sum(\cos\theta_i\sin\varphi_i)$ $D=\sum(\sin\theta_i\cos\varphi_i)$

$E=\sum(\cos^2\theta_i)$ $F=\sum(\cos^2\theta_i)$

$G=\sum(\cos^2\varphi_i)$ $H=\sum(\sin^2\varphi_i)$

2.4 Linear Association between two circular variables

The regression model when both variables are circular produces very interesting form. Hussin (2007) present the hypothesis testing of parameters for ordinary linear circular regression model assuming the circular random error distributed as von Mises distribution. The model is given by

$$\varphi = \alpha + \beta\theta + \varepsilon \pmod{2\pi} \tag{3}$$

where ε is a circular random error having a von Mises distribution with mean circular 0, and concentration parameter κ, which can be written as $\varepsilon \sim VM(0,\kappa)$. We can estimate α and β by maximum likelihood estimation. Based on von Mises density function, the log likelihood function for model (3) is given by

$$\log L(\alpha,\beta,\kappa;\theta_1,...,\theta_n,\varphi_1,...,\varphi_n) = -n\log(2\pi) - n\log I_0(\kappa) + \kappa\sum\cos(\varphi_i-\alpha-\beta\theta_i) \tag{4}$$

By differentiating log L with respect to α,β and κ the estimates of α and β namely $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\kappa}$ are given by

$$\hat{\alpha} = \begin{cases} \tan^{-1}\left(\dfrac{S}{C}\right), & S > 0, C > 0 \\ \tan^{-1}\left(\dfrac{S}{C}\right) + \pi, & C < 0 \\ \tan^{-1}\left(\dfrac{S}{C}\right) + 2\pi, & S < 0, C > 0 \end{cases} \tag{5}$$

where $S = \sum \sin(\varphi_i - \hat{\beta}_{i-1}\theta_i)$ and $C = \sum \cos(\varphi_i - \hat{\beta}_{i-1}\theta_i)$ and

$$\hat{\beta}_i \approx \hat{\beta}_{i-1} + \frac{\sum \theta_i \sin(\varphi_i - \hat{\alpha} - \hat{\beta}_{i-1}\theta_i)}{\sum \theta_i^2 \cos(\varphi_i - \hat{\alpha} - \hat{\beta}_{i-1}\theta_i)} \tag{6}$$

respectively. Further, the maximum likelihood estimate of $\kappa$is

$$\hat{\kappa} = A^{-1}\left(\frac{1}{n}\sum \cos(\varphi_i - \hat{\alpha} - \hat{\beta}\theta_i)\right)\cdot \tag{7}$$

The approximation given by Dobson for the function A (ratio of the modified Bessel functions for the first kind of order one, and the first kind of order zero) for the von Mises concentration parameter, κ is

$$A^{-1}(w) \approx \frac{9 - 8w + 3w^2}{8(1 - w)}. \tag{8}$$

These expressions may be solved iteratively given some suitable "initial guesses" at the estimate. The estimated $\hat{\beta}$ is obtained by iterative procedure at some predetermined stopping rules.

More often than not, an investigation on the relationship between two circular variables is required. As mentioned earlier, the strength of the correlation between two circular variables can be numerically computed using the correlation coefficient $\hat{\rho}_T$ as shown in (1). The $\hat{\alpha}$ and $\hat{\beta}$ for linear association between two circular variables can be calculated as shown in (5) and (6). In this analysis, however, the von Mises concentration parameter, κ is not include in the calculation.

2.5 Diagrammatical representation of two circular variables

Many statistical tools exist for analyzing their structure, but, surprisingly, there are few techniques for exploratory visual analyses, and for depicting the patterns of relations among variables (Friendly, 2002). The concepts of "one picture is worth a thousand words" have been used by some researchers such as Linden (2005) who developed visual display in the disease management program, Vinnakota (1988) designs charts to understand composite beam design problem and Friendly (2001) wrote macro programs for graphical analysis to reveal features of categorical data that are not apparent in traditional numerical summaries.

In most exploratory data analysis of two variables, visual representation of the correlation can provide better understanding of the association between the two variables. In the analysis of two circular variables, a diagrammatical representation to describe the relationship is developed and known as the Spoke plot. The Spoke plot consists of inner and outer rings in which lines are used to connect the pair of points (θ_i, φ_i,). Together with the Spoke plot, the program calculates correlation and parameters of the linear association between two variables.

3. Source of data

In this study, the Malaysian wind data that has been obtained from the Malaysian Meteorological Services Department is used. The data was collected daily and measured by anemometer at several airport locations throughout Malaysia over period of one year in 2005 and at a telecommunication tower in Seberang Jaya in April 2002.

4. Result and findings

A call function is developed using MATLAB version 7.0, in the analysis of two directional data. After running the call function "spokecorrelation" for the data, an output window is generated that gives the:

i. calculated correlation value of two circular variables.

ii. calculated linear association measure of two circular variables.

iii. relationship of two circular variables using Spoke plot.

Thus one can easily make comparison of the three analyses in one output window. For illustration, we run an analysis for three datasets of wind data. They are the:

i. wind direction data at maximum speed in January 2005 between KLIA and Ipoh, with the objective to compare two sets of wind data at two different locations.

ii. wind direction data in January between KLIA (time = 0000, pressure = 1000hpa) and KLIA (time = 0000, pressure = 500hpa), with the objective to compare two sets of wind data at different pressures.

iii. wind direction data recorded at Telecommunication tower, Seberang Jaya in April 2002 between the heights 45.72m and 75.28m, with the objective to compare two sets of wind data at different level.

The results are shown in Figure 1, Figure 2 and Figure 3, respectively.

5. Conclusion

The need of software development in the analysis of two circular variables is necessary. This is because a variety of environmental data are circular in nature such as wave, wind direction, frequency, compass bearing, clock and others. In this study, a program that evaluates statistical functions as well as diagrammatical representation is presented. Using MATLAB, the output window gives all the three analyses namely Spoke plot, correlation and linear association in one diagram. This research could be improved by developing further the Graphical User Interface (GUI) into the packages for the ease of application by the user.

References

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge University Press.

Friendly, M. (2001). A Reader's Guide to Visualizing Categorical Data, *SAS SUGI proceedings*, 26(173).

Friendly, M. (2002). Corrgrams: Exploratory display for correlation matrices, *The American Statistician,* 56(4), 316-324.

Handerson, P.A., Seaby, R. M. H.(2002), *Axis Software, version 1.1,* Pisces conservation Ltd.

Hussin, A.G., Jalaludin, J.F., and Mohamed, I. (2006). Analysis of Malaysian Wind Direction Data using AXIS, *Journal of Applied Science Research*, 11,1019-1021.

Hussin, A.G. (2007). Hypothesis Testing of Parameters for Ordinary Linear Circular Regression. *Journal of Applied Sciences Research*, 3,185-188.

Linden, A. and Roberts, N. (2005). Using Visual Displays as a Tool to Demonstrate Disease Management Program Effectiveness. *Journal of Disease Management*, 5, 301-310.

Mardia, K. V. (1972). *Statistics of directional data*. Academic Press Inc.

Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics*. Academic Press Inc.

Matlab Software (2004), *Version 7.0.0.19920(R14)*, The Mathworks Inc.

Oriana Software (2007), *Version* 2.02(e), Kovach Computing services.

Vinnakota, S., Foley, C. and Vinnakota, M., (1988). Design of Partially or Fully Composite Beams, with Ribbed Metal Deck, Using LRFD Specifications. *Engineering Journal Second Quarter*, 60-78.

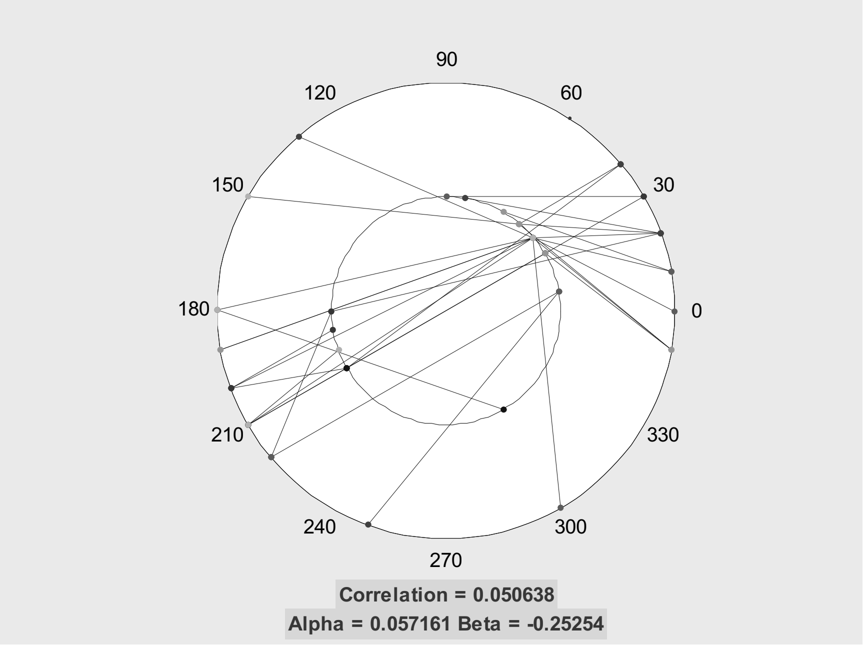

Figure 1. Spoke Plot of wind direction data at maximum speed

in January 2005 between KLIA and Ipoh.

From the Spoke plot in Figure 1, it can be seen that a number of lines crossing the inner ring implies that there is no correlation between the variables. To support the finding, the calculated correlation value of Equation (1) is 0.0506 which indicates no correlation. The linear association value also shows the absence of one to one linear relationship between the two circular variables.

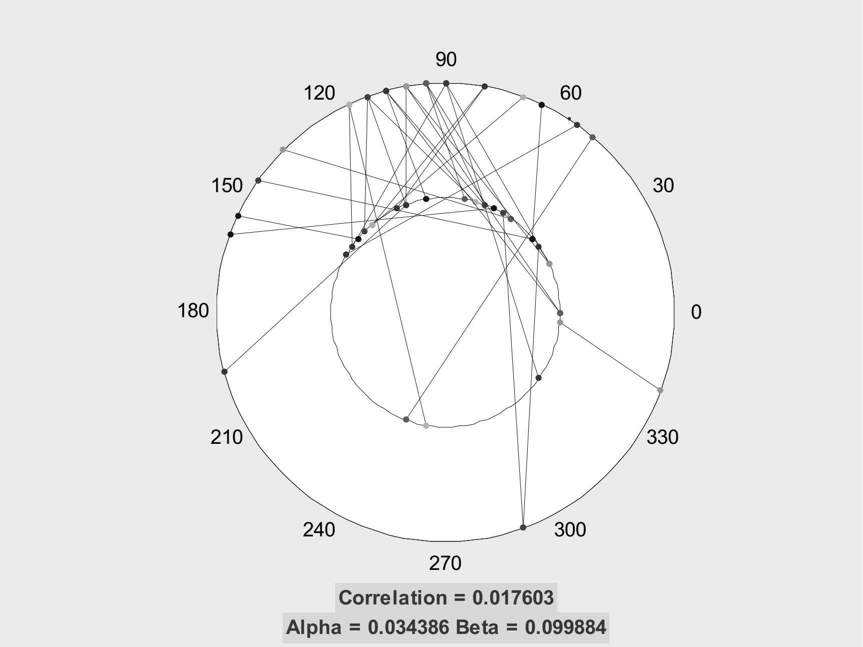

Figure 2. Spoke Plot of Wind direction data in January between KLIA

(time = 0000, pressure = 1000hpa) and KLIA (time = 0000, pressure = 500hpa).

From the Spoke plot in Figure 2, it can be seen again a number of lines crosses the inner ring indicates no correlation between the variables. To support the finding, the calculated correlation value is 0.0176 which implies no correlation. On the linear association, there seems to be evidence of a small one to one relationship.

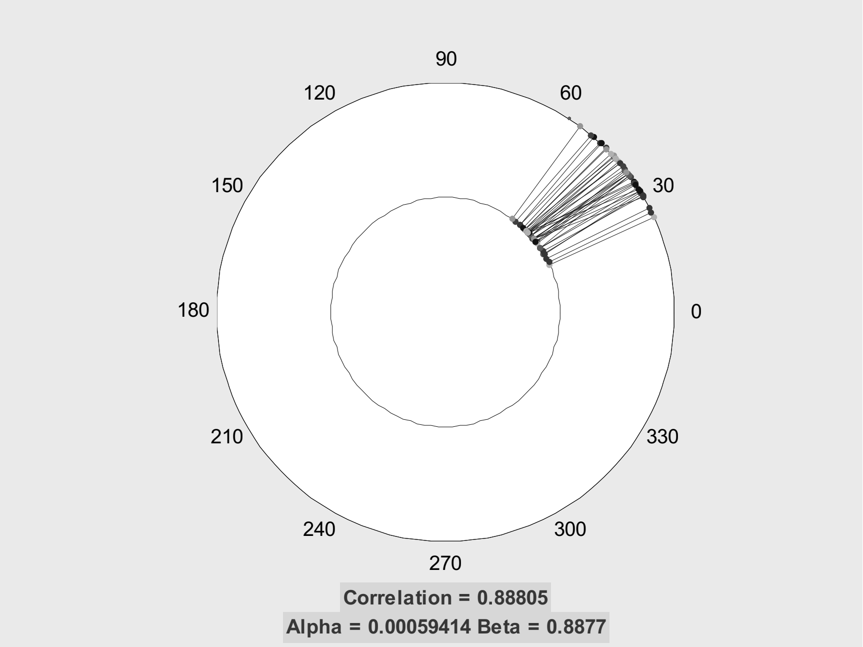

Figure 3. Spoke Plot of Wind direction data recorded at Telecommunication tower,

Seberang Jaya in April 2002 between height 45.72m and height 75.28m.

From the Spoke plot in Figure 3, it can be seen that none of the line crosses the inner ring and this suggests the presence of a strong correlation between the two variables. To support the finding, the calculated correlation value of Equation (1) is 0.8880 which indicates a strong correlation. On the linear association, there seems to be some evidence of a strong one to one relationship.