Modelling wind direction data of Langkawi Island during Southwest monsoon in 2019 to 2020 using bivariate linear functional relationship model with von Mises distribution

Nurkhairany Amyra Mokhtar¹*, Yong Zulina Zubairi², Abdul Ghapor Hussin³, Basri Badyalina¹, Ahmad Faiz Ghazali¹, Fatin Farazh Ya’acob⁴, Palaniappan Shamala¹, Lee Chang Kerk¹

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Cawangan Johor Kampus Segamat, 85000 Segamat, Johor, Malaysia.
²Centre for Foundation Studies in Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia.
³Faculty of Defence Sciences and Technology, National Defence University of Malaysia, Kem Sungai Besi, 57000 Kuala Lumpur, Malaysia.
⁴Faculty of Business and Management, Universiti Teknologi MARA Cawangan Johor Kampus Segamat, 85000 Segamat, Johor, Malaysia.

*E-mail: nurkhairany@uitm.edu.my

Abstract. The weather in Malaysia is characterised by two monsoons, namely, the southwest monsoon from May to September, and the northeast monsoon from November to March. Wind direction is essential in observing the weather patterns and global climate. In this study, our interest is on investigating the relationship of the wind direction data of Langkawi Island in Malaysia during the southwest monsoon from year 2019 to 2020. It is essential to highlight that wind direction data that is circular and this requires different statistical techniques from the techniques that are used to analyse linear data. In this paper, we model the relationship of wind direction data by using the bivariate functional relationship model with von Mises distribution. The magnificence of this model is that the existence of error terms in all variables is considered. When modelling the data, outliers of the wind direction data are identified by using the covratio method that considers row deletion. The covariance matrix of the parameter estimates of the data is obtained by using the Fisher information matrix. Also, the result is supported by the Q-Q plots of the von Mises that indicate the goodness-of-fit of the wind direction data to the von Mises distribution. Then, maximum likelihood estimation is used in obtaining the parameter estimates of the data and hence, the model of the wind direction data is attained. The implications of this study provides an improved comprehension of the behaviour of wind direction and may be used for the prediction of wind energy in future. Keywords: circular data, wind direction, functional relationship model, big data, statistical modelling
1. Introduction

Dealing with circular data is inherently different from dealing with linear data due to the periodic nature of circular data that is presented in the form of degrees or radians. Measurements on the circle, from 0° to 360°, or from 0 radian to 2π radians reflect the same direction, while for linear scale they would be located at opposite ends. For this purpose, circular knowledge requires unique methods of analysis [1].

Circular data occurs in many areas such as astronomy, geology, meteorology, physics, biology, image analysis and medical applications. In physics, for example, angular data is collected in bubble chamber experiments, where points representing events are observed through a circular window [2]. In studying the phenomenon of pole-reversal, geologists may be interested in the path of the magnetic pole of the earth and research the flow of rivers. In medical application, the angles of knee flexion are tested to evaluate the rehabilitation of orthopaedic patients [3]. In political science, cyclical timing attacks of domestic terrorism events were studied [4].

When dealing with data, the outlier would appear as anomalies. The presence of outliers may affect the fitted model, which makes scientific conclusions should be paid attention [5]. Thus, this study considers in identifying the presence of an outlier if there is any.

In this study, our interest is to propose a statistical model of wind direction. The wind direction data of Langkawi Island was obtained from Malaysia Meteorological Department. Langkawi Island is known as a UNESCO Global Geopark and it plays a crucial role in the development of heritage tourism [6]. The wind direction data of Langkawi island was recorded at the latitude of 6°20’N with the longitude of 99°E’ for its maximum daily direction at the elevation of 6.4 meters. Figure 1 shows the location of Langkawi Island in Malaysia.

![Figure 1. Location of Langkawi Island in Malaysia](image-url)

In this paper, Section 2 describes the von Mises distribution that is used to analyse the data. Section 3 discusses on bivariate linear functional relationship model that is fitted to model the data. Section 4 describes the parameter estimation used in modelling the data meanwhile Section 5 describes the outlier checking for the data. The results of the study are explained in Section 6 and then concluded in Section 7.

2. Von Mises Distribution

The von Mises distribution is widely used to model statistically circular data. It consists of two parameters namely mean direction and concentration parameters. When the concentration parameter is at its minimum value of zero, the distribution is also a uniform distribution over the unit circle [7].
Use of the model is aplenty in the literature. In 2006, the von Mises distribution was used to determine the strength of minerals, namely biotite and feldspar in thin sections prepared parallel to the magnetic foliation plane [8]. In 2010, the von Mises distribution was applied to study wind direction for weather prediction over the Pacific Northwest [9]. Besides, von Mises distribution was used to describe the orientations of fault plane solutions before, during and after seismic swarm earthquakes in 2004 to 2008 [10]. In imaging, this distribution was applied in image analysis assessment [11].

The probability distribution function of the Von Mises distribution is given by

\[ g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\theta - \mu)) \]  

where \( I_0(\kappa) \) is the modified Bessel function of the first kind and order zero, which can be defined by

\[ I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \exp(\kappa \cos \theta) \, d\theta \]  

where \( \mu \) is the mean direction and \( \kappa \) is the concentration parameter for \( 0 \leq \theta < 2\pi \) and \( \kappa > 0 \).

3. **Bivariate Linear Functional Relationship Model**

The functional relationship model is a type of error-in-variables models (EIVM). The EIVM differs from the ordinary or classical linear regression model in that the variables are masked by measurements error and are not observed directly [12]. In reality, the existence of measurement errors arises into the observations in which by ignoring these measurement errors may directly affect the desirable criteria of the estimators [13].

EIVM is said to be the most statistically correct technique for reactivity ratio estimation as it considers the presence of error in all variables [14]. It is the other type of regression model that considers the measurement errors for both \( x \) and \( y \) variables. In contradictory, in ordinary linear regression analysis, the explanatory variable \( x \) is assumed to be fixed and measured without error, and the \( y \)-variable is considered with the error term where it is identified as the response variable [15]. However, there is no difference between ‘explanatory’ and ‘response’ variables in EIVM [16].

EIVM consists of three types which are functional, structural and ultrastructural. In the functional relationship model of EIVM, the underlying variables are fixed [15]. Meanwhile, in the structural relationship model, the variables are random whereas in the ultrastructural relationship model, the variables are the synthesis of linear functional and structural relationship model. In EIVM, a nested-iterative loop is advised to be used to obtain the estimates of the variables and parameters [14].

In a linear functional relationship model, there are two variables that are linearly related to estimate a relationship between them. Unfortunately, observations on both quantities are subject to error [17]. The fitting of a linear relationship with errors in the continuous linear variables or error-in-variables model (EIVM) had been explored since the 19th century when Adcock investigated estimation properties under restrictive assumptions in ordinary linear regression models [18]. In 2015, this bivariate linear functional relationship model is used to study the relationship of wind direction data collected from the Holderness coastline situated at the Humberside coast of the North Sea, United Kingdom [19].

4. **Parameter Estimation**

In this paper, the wind direction data is modelled by using a linear functional relationship model for circular data. This model is used because we would like to model the relationship of the wind direction data between the two years, 2019 and 2020. The linear functional relationship model for circular data is given by

\[ Y = \alpha + X \mod 2\pi \]  

in which both \( X \) and \( Y \) variables are considered with the random errors \( \delta_i \) and \( \varepsilon_i \), respectively, where \( X_i = x_i + \delta_i \) and \( Y_i = y_i + \varepsilon_i \).
With the assumption of equal error concentration that has been discussed by Mokhtar, the errors are distributed with von Mises distribution of $\delta_i \sim VM(0, \kappa)$ and $\varepsilon_i \sim VM(0, \kappa)$ [20]. The parameter estimation is derived by using the method of maximum likelihood. The log-likelihood function of the Von Mises distribution is given by

$$
\log L (\alpha, \kappa, X; x, y) = -2n \log 2 \pi - 2n \log I_0 (\kappa) + \kappa \sum_{i=1}^{n} \cos(x_i - X_i) + \kappa \sum_{i=1}^{n} \cos(y_i - \alpha - X_i)
$$

for $0 \leq x < 2\pi, 0 \leq y < 2\pi$ and $\kappa > 0$ where $I_0 (\kappa)$ is the modified Bessel function of the first kind and order zero, which can be defined by:

$$
I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} e^{\kappa \cos \theta} d\theta
$$

where $\kappa$ is the concentration parameter of the error term.

The estimation of the variable $X_i$ is obtained iteratively given by

$$
\hat{X}_{1i} \approx \hat{X}_0 + \frac{\sin(x_i - \hat{X}_0) + \sin(y_i - \hat{\alpha} - \hat{X}_0)}{\cos(x_i - \hat{X}_0) + \cos(y_i - \hat{\alpha} - \hat{X}_0)}
$$

and the estimation of the rotation parameter is given by

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

where $S = \sum_{i=1}^{n} \sin(y_i - \hat{X}_i)$ and $C = \sum_{i=1}^{n} \cos(y_i - \hat{X}_i)$.

The concentration parameter $\kappa$ for the case of equal error concentration is estimated using the approximation given by [21]:

$$
A_1^{-1}(w) = \begin{cases} 
2w + w^3 + \frac{5}{6}w^5 & \text{when } w < 0.53 \\
-0.4 + 1.39w + \frac{0.43}{(1 - w)} & \text{when } 0.53 \leq w < 0.85 \\
1 & \text{when } w \geq 0.85
\end{cases}
$$

Thus, the estimation is given by

$$
\hat{\kappa} = A_1^{-1}(w)
$$

where $w = \frac{1}{n} \left( \sum_{i=1}^{n} \cos(x_i - \hat{X}_i) + \sum_{i=1}^{n} \cos(y_i - \hat{\alpha} - \hat{X}_0) \right)$.

Thus, it is

$$
\hat{\kappa} = A_1^{-1} \left( \frac{1}{n} \left( \sum_{i=1}^{n} \cos(x_i - \hat{X}_i) + \sum_{i=1}^{n} \cos(y_i - \alpha - \hat{X}_i) \right) \right)
$$

It is worthwhile to note that for circular case, the estimation of a concentration parameter (whose inverse is equivalent to the variance for linear data) is to be corrected by dividing it by 2 [22]. Hence, the estimate becomes $\hat{\kappa} = \frac{\hat{\kappa}}{2}$.
5. Identifying Outlier

Outliers refer to those data instances that make it more difficult to match the desired model [23]. Thus, in this section, the presence of an outlier is identified through the method of covratio. The covratio method is studied to detect the presence of outlier for circular data [23].

The covratio equation is developed from the Fisher Information matrix of the parameter estimates and is defined by

$$\text{COVRATIO}_{(-i)} = \frac{|\text{COV}|}{|\text{COV}_{(-i)}|}$$

where $|\text{COV}|$ is the determinant of the covariance matrix of the parameter estimates is given by

$$|\text{COV}| = \frac{1}{n^2k[A'_1(k)]^2}$$

and $|\text{COV}_{(-i)}|$ is the determinant of the covariance matrix for the reduced data set by excluding the $i$-th row.

The cut-off equation for outlier detection in bivariate functional relationship model for circular data with $y = 3.7586n^{-0.71}$ [24]. Since the sample size used for this study is $n = 151$, thus the cut-off point becomes $y = 3.7586n^{-0.71} = 0.170382$. This cut-off equation is used to detect the outlier for the wind direction data in this study.

6. Results

Preliminarily, we carried out the univariate analysis for the wind direction data during the southwest monsoon and described graphically with the rose diagram for the years 2019 and 2020. Figures 2 and 3 show the rose diagrams of the wind direction data of Langkawi Island in 2019 and 2020, respectively.

From the rose diagrams, we can see that the patterns of the data for both years are different. Thus, in this paper, we would like to investigate the relationship of the data for both years and describe it in the form of a bivariate functional relationship model for circular data.

The Q-Q plots of the von Mises distribution for the wind direction data of both years 2019 and 2020 are shown in Figures 4 and 5, respectively. The Q-Q plots indicate the goodness-of-fit of the data in both years to the von Mises distribution.
The next step is to figure out if any outlier presents in the data. Figure 6 describes the value of \( \text{COVRAATIO}_{(\sim t)} \) for the wind direction data in this study and it shows that none of the values is greater than the cut-off equation of \( y = 0.170382 \). Thus, none of the data is classified as the outlier. Therefore, the statistical model for the wind direction data of Langkawi Island is proposed without having to eliminate any of the data.

After that, the wind direction data is fitted to the bivariate functional relationship model described in Section 3 and 4. The parameter estimate values of the data are shown in Table 1.

| Detail                        | Value   |
|-------------------------------|---------|
| Rotation parameter, \( \hat{\alpha} \) | 6.00350 |
| Variance of \( \hat{\alpha} \)    | 0.01887 |
| Concentration parameter, \( \hat{\kappa} \) | 1.29619 |
| Variance of \( \hat{\kappa} \)    | 0.01146 |

From Table 1, we can say that the model proposed for wind direction data of Langkawi Island during southwest monsoon season in 2019 and 2020 is \( Y = 6.00350 + X \ (\text{mod} \ 2\pi) \) with a small error
concentration parameter of 1.29619. The parameter estimates show that the rotation parameter is 6.00350 which is very near to 2\pi radians. The variance of \( \hat{\alpha} \) and \( \hat{\kappa} \) are small given by 0.01887 and 0.01146, respectively which indicate that the values estimated are consistent and less dispersed.

7. Conclusion
To conclude, this paper discusses on modelling wind direction data of Langkawi Island during southwest monsoon season in 2019 and 2020 by using an error-in-variables model. The bivariate data is considered with the presence of error terms. The existence of outlier in the data is checked through the \textit{covratio} statistic and it is shown that there is no outlier present in the data. The von Mises distribution is considered to estimate the parameters of the wind direction data through maximum likelihood estimation. From this model, we can say that the rotation parameter is very near to 2\pi radians and the value of the concentration parameter of the error terms is small and less concentrated. This model may be used in predicting the wind direction of Langkawi Island during the southwest monsoon and thus helps in managing the outdoor activities with the concern of safety and weather. Further studies may apply this model in studying the wind direction of any other places.

Acknowledgements
The authors gratefully acknowledge the financial support by Universiti Teknologi MARA Cawangan Johor under grant Bestari 600-UiTMJC (PJIA. 5/2) for supporting this study.

References
[1] Cremers J and Klugkist I 2018 One Direction? A Tutorial for Circular Data Analysis Using R With Examples in Cognitive Psychology \textit{Frontiers in Psychology} 9 Article 2040 pp 1-13
[2] Mardia K V 1972 \textit{Statistics of Directional Data} (New York: Academic Press London)
[3] Jammalamadaka SR and Sengupta A 2001 \textit{Topics in Circular Statistics} (Singapore: World Scientific Publishing)
[4] Gill J and Hangatner D 2010 Circular Data in Political Science and How to Handle it \textit{Political Analysis} 18 pp 316-336
[5] Altman N and Krzywinski M 2016 Analyzing outliers: influential or nuisance? \textit{Nature Methods} 13 pp 281–282.
[6] Azman N, Halim S A, Liu O P and Komoo I 2011 The Langkawi Global Geopark: Local community’s perspectives on public education \textit{International Journal of Heritage Studies} 17(3) pp 261-279
[7] Lark R M, Clifford D and Waters C N 2014 Modelling complex geological circular data with the projected normal distribution and mixtures of von Mises distributions \textit{Solid Earth} 5 pp 631–639.
[8] Sen K and Mamtani M A 2006 Magnetic fabric, shape preferred orientation and regional strain in granitic rocks \textit{Journal of Structural Geology} 28 pp 1870–1882
[9] Bao L, Gneiting T, Grimit E P, Guttorp P and Raftery A E 2010 Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction \textit{American Meteorological Society} 138 pp 1811-1821
[10] O’Brien J F, Roman D C, Dixon J P, Power J A and Arnold R 2012 Multiple causes for non-eruptive seismic swarms at Mt. Martin, Katmai Volcanic Cluster, Alaska (2004–2008) \textit{Journal of Volcanology and Geothermal Research} 229 pp 13–22
[11] Gabarda S and Cristóbal G 2012 No-reference image quality assessment through the von Mises distribution \textit{Journal of the Optical Society of America A} 29 (10) pp 2058-2066
[12] Hussin A G 2004 Numerical Comparisons for Various Estimators of Slope Parameters for Unreplicated Linear Functional Model \textit{Matematika} 20(1) pp 19-30
[13] Mamun A S M A, Hussin A G, Zubairi Y Z, Imon A H M and Rana S 2017 Small-Sample confidence interval for the slope of linear structural relationship model \textit{Electronic Journal of Applied Statistical Analysis} 10(2) pp 374-383
[14] Scott A J and Penlidis A 2018 Computational Package for Copolymerization Reactivity Ratio Estimation: Improved Access to the Error-in-Variables-Model Processes 6(1) pp 8

[15] Rousseeuw P J and Leroy A M 2005 Robust regression and outlier detection Vol. 589 (New York: John Wiley & Sons)

[16] Hassan S F, Hussin A G and Zubairi Y Z 2010 Estimation of Functional Relationship Model for Circular Variables and Its Application in Measurement Problem Chiang Mai Journal of Science 37(2) pp 195-205

[17] Lindley D V and El-Sayyad G M 1968 The Bayesian Estimation of a Linear Functional Relationships Journal of the Royal Statistical Society: Series B 30 (1) pp 190-202

[18] Adcock R J 1878 A Problem in Least Squares Analyst 5(2) pp 53-54

[19] Mokhtar N A, Zubairi Y Z and Hussin A G 2015 A simple linear functional relationship model for circular variables and its application. In Proceedings of the 9th International Conference on Renewable Energy Sources (RES ‘15), Kuala Lumpur, Malaysia pp 57-63

[20] Mokhtar N A 2016 Simultaneous Functional Relationship Model and Outlier Detection using Clustering Technique for Circular Variables MSc Thesis National Defence University of Malaysia.

[21] Fisher N I 1993 Statistical Analysis of Circular Data (New York: Cambridge University Press)

[22] Caires S and Wyatt L R 2003 A Linear Functional Relationship Model for Circular Data with an Application to the Assessment of Ocean Wave Measurement American Statistical Association and the Internal Biometric Society Journal of Agricultural, Biological and Environmental Statistics 8 (2) pp 153-169

[23] Boukerche A, Zheng L and Alfandi O 2020 Outlier Detection: Methods, Models, and Classification ACM Computing Surveys 53 (3) Article 55 pp 1-37

[24] Mokhtar N A, Zulina Y Z, Hussin A G and Moslim N H 2019 An Outlier Detection Method for Circular Linear Functional Relationship Model Using Covratio Statistics Malaysian Journal of Science 38 (Special Issue 2) pp 46 – 54