Application of mixtures of von Mises-Fisher model to investigate the statistical characteristics of the wind direction data

NURULKAMAL MASSERAN [1, 2], AHMAD MAHIR RAZALI [1, 2],
KAMARULZAMAN IBRAHIM [1, 2]
[1]School of Mathematical Sciences,
Faculty of Science and Technology,
[2]Centre for Modeling and Data Analysis (DELTA)
Faculty of Science and Technology,
Universiti Kebangsaan Malaysia,
43600 UKM Bangi, Selangor,
MALAYSIA
kamalmsn@ukm.edu.my, mahir@ukm.edu.my, kamarulz@ukm.edu.my

Abstract: - Wind direction has a substantial effect on the environment and human lives. Wind direction influences the dispersion of particulate matter in the air and affects the construction of engineering structures, such as towers, bridges, and tall buildings. In fact, knowledge of the wind direction and wind speed can be used to obtain information about the energy potential. This study investigates the characteristics of the wind regime involving the wind direction in Kudat, Malaysia using a mixture of von Mises-Fisher model (mvMF). The suitability of each mvMF was judged based on a graphical representation and goodness-of-fit statistics. In addition, the best-fit mvMF model was compared with the circular distribution based on nonnegative trigonometric sums to determine the best model. The results found that the mvMF model with $H \geq 2$ components is the best model. Additionally, the circular density plots of the suitable model clearly show the dominant wind directions in the Kudat region.

Key-Words: - Circular distribution based on nonnegative trigonometric sums, directional statistics; directional distribution, mixture of von Mises-Fisher distributions, wind direction modelling.

1 Introduction

Wind direction is the direction from which the wind is blowing. It is expressed in terms of degrees measured clockwise from geographical north, which can be represented as an angle measured from a point chosen as the "zero direction". The starting point and rotation from this point, regardless of whether it is clockwise or anticlockwise, are taken as positive values. Observations using these two dimensions are also called circular or directional data [1]. In practice, the wind direction is an important feature that should be considered in building wind turbines and in structural and environmental design analysis [2, 3, 4]. Wind direction has also been recognised as an important aspect for the evaluation of wind energy because wind direction data can complement wind speed data to yield information about the energy potential

Wind direction is a type of directional data. Thus, it has unique characteristics that are different from standard linear or real-line data sets. Such distinctive features have made directional statistics analysis substantially different from linear analysis [5, 6]. Let θ be a random variable that measures the directional data that take values in the range 0° to 360° or 0 to 2π. An analysis of θ would depend on the selection of the starting point as the "zero-direction" and the sense of rotation, i.e., clockwise or anti-clockwise. For example, in Figure 1, if the zero direction is due east, corresponding to anti-clockwise rotation, the data will take the value of 60°, whereas if the zero direction is due north, corresponding to clockwise rotation, the data will take the value of 30°. However, the "beginning" are always coincides with the "end", i.e., 0°-360°, and the measurement is also periodic, with θ being the same as $\theta + p \times 2\pi$ for any integer p [1].

In addition, directional data that take values of $\theta = 0°$ to 360° or $\theta = 0$ to 2π are commonly termed polar coordinate data with magnitude = 1, namely, (1, θ). On the other hand, the directional data can be transformed into rectangular coordinate form, (*X, Y*), through $x = \cos\theta$ and $y = \sin\theta$ for every θ. Figure 2 shows the directional data in terms of both polar coordinates (1, θ) and rectangular coordinate (*X, Y*). There are many other

unique features of directional data; for further reference, see [1, 7, 8]

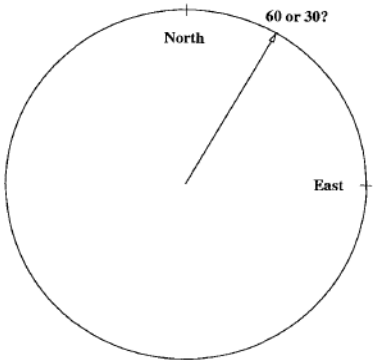

Fig 1. The observed directional data depend on choice of origin and the sense of rotation [1]

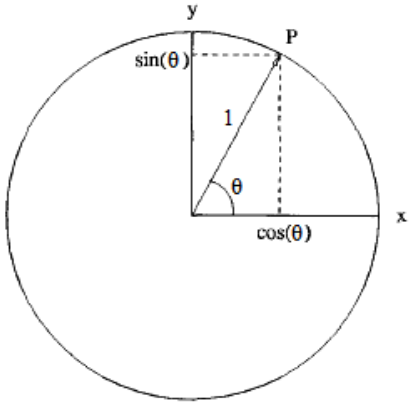

Fig 2. Relationship between polar coordinate data $(1, \theta)$ and rectangular coordinate data (X, Y) [1]

Based on the features of directional data described above, it is clear that frequently used statistical analyses cannot be used indiscriminately when analysing directional data. In this study, we describe some of the characteristics of the wind direction in the region of Kudat, Malaysia, by analysing directional data to gain some insight regarding the behaviour of the wind regime.

2 Study area and data

Sabah is a state of Malaysia, located in the northern section of the island of Borneo. It is the second largest state in the country after Sarawak, which it borders to the southwest. Sabah is relatively wet (annual precipitation exceeding 200 mm) due to the tail effect of typhoons, which frequently traverse the Philippine islands across the South China Sea. It is worth mentioning that from April to November each year, when typhoons frequently develop over the west Pacific and move westward across the Philippines, the south-westerly winds over the northwest coast of Sabah may reach speeds of 10.30 m/s or more [9, 10]

Fig 3. Map of Sabah state

The data used in this study were obtained from the Malaysian Meteorological Department. In this study, hourly wind direction data from 1 January 2007, to 30 November 2009 were used. Wind direction data are circular because they are recorded in terms of degrees, from 0° to 360°. However, for modelling, data transformation into radian units can be performed easily. Apart from that, the missing data has been estimated by using the method of single imputation [11].

3 Descriptive statistics

Before a detailed analysis is conducted, it is important to evaluate the descriptive statistics to obtain some preliminary information about the data. As mentioned above, directional data have many features that differ from those of standard linear data sets. For example, the arithmetic mean, which is commonly used for linear data, cannot be used as a measure of the centre of the directional data. The sample variance s^2, which depends on the sample mean, also suffers from the same problem. Thus, we need an alternative measure of centre and dispersion when dealing with directional data [1]. Let $\theta_1, \theta_2 \ldots, \theta_n$ be a set of directional data; the mean direction can then be calculated as

$$\bar{\theta} = \begin{cases} \tan^{-1}\left(S/C\right) & , \text{ if } C > 0, S \geq 0 \\ \pi/2, & , \text{ if } C = 0, S > 0 \\ \tan^{-1}\left(S/C\right) + \pi & , \text{ if } C < 0 \\ \tan^{-1}\left(S/C\right) + 2\pi, & \text{if } C \geq 0, S < 0 \\ \text{undefined} & , \text{ if } C = 0, S = 0 \end{cases} \quad (1)$$

where, $C=\sum_{i=1}^{n}\cos\theta_i$ and $\mathrm{S}=\sum_{i=1}^{n}\sin\theta_i$ accomplish the polar-to-rectangular transformation. On the other hand, the measure for the dispersion of the directional data is commonly derived from the circular variance, which is given as

$$V=1-\frac{1}{n}\sqrt{C^2+S^2} \qquad (2)$$

A small value of the circular variance indicates the data have a large concentration around the mean direction.

The percentile measure for directional data is same as that for linear data: it is a measure of the value of a variable below a certain percent of observations. For example, the (100*p*)-th percentile is often called the quantile of order *p*. Let $y_1 \le y_2 \le \ldots \le y_n$ be an order statistic for *n* observations; y_r is then the quantile of order $p=\frac{r}{(n+1)}$ as well as the $\frac{100r}{(n+1)}$ percentile. Thus, the *p*-th percentile of the data is also a quantile of order *p* for the data. Using these descriptive measurements, Table 1 shows the descriptive statistics for the wind directional data in Kudat.

Table 1. Descriptive statistics for Kudat wind direction.

Kudat wind station	
Mean direction	218.21°
Circular variance	0.978
25th percentile	80°
50th percentile	220°
75th percentile	240°

Based on the descriptive statistics in Table 1, the circular mean of the wind direction is approximately 218.21°. However, the circular variance is 0.978, which implies that the data were not well concentrated around their mean direction. Thus, we suspect that the data are either approximately uniformly distributed or have a several-directional mean. The values of 25th, 50th, and 75th percentile are 80°, 220°, and 240°, respectively.

4 Wind direction modelling

For the purpose of modelling the wind direction in a particular area, various circular distributions have been used, such as the von Mises distribution, the generalised von Mises distribution, finite mixtures of von Mises distributions, the wrapped Cauchy distribution, the uniform distribution, and the wrapped round-normal distribution. The von Mises model is among the most commonly used and was found to provide good results. For examples, Kamisan et al.[6] evaluated the best fitting model for the wind directional data in southwesterly Malaysian using four different types of circular probability distribution namely circular uniform distribution, von Mises distribution, wrapped-normal distribution and wrapped-Cauchy distribution. Based on the result of performance indicators, they found that the von Mises distribution was the best circular distribution to describe the southwesterly monsoon wind direction. Carta et al. [12] have showed that a the finite mixture of von Mises is a very flexible model for wind direction studies particularly for the wind direction regimes in zones with several models or prevailing wind directions. In addition, Carta et al. [13] have showed that the finite von Mises is also a flexible model correspond to the wind speed density function in explaining the wind regime behaviours that takes into account the correlation between wind speeds and its directions. In fact, the same result have been showed by Azmani et al. [14] in modelling the sensor data for the cases of a recursive change point estimate of the wind speed and direction. Apart from that, a lot of interesting studies have been done regarding the application of von Mises in modelling the directional data. For example, the Shieh et al. [15] has proposed a bivariate model with von Mises marginal distributions for independence in paired wins direction data. Dobigeon & Tourneret [16] proposed a method of joint segmentation for the wind speed and direction based on the information from the von Mises distribution. Williams et al. [17] has used the von Mises model to represented the wind direction data as a function of a regression model to described about the embedding dispersion of pollution source. Heckenbergerova et al. [18] have developed a method of Particle Swarm Optimization using the information from the finite mixture of von Mises distribution in order to provide an accurate information regarding wind direction distribution. There are still many more research studies that have used the von Mises distribution in modelling the wind directional data. However, most of the studies involving the von Mises distribution does not describe well about the dominant direction of the wind blows in term of the mean direction and their concentration parameter of the von Mises model. Thus, since the von Mises distribution is a flexible model for addressing wind directional data with

several modes. This study attempts to address the issued regarding the strength of a dominant direction of a wind blows based on a von Mises-Fisher distributions (von Mises-Fisher is a von Mises model in *p* dimensions). Apart from that, the suitability of von Mises model will be compare with the circular model based on nonnegative trigonometric sums in order to determine the best statistical model in describing the wind directional data in Kudat.

The von Mises-Fisher distribution can be categorised as a single model or a mixture of several von Mises-Fisher distributions. The suitability of the von Mises model will be compared to the circular distribution based on nonnegative trigonometric sum to determine the best model for describing the wind directional data in Kudat. However, we first present a review of some interesting work by Banerjee et al. [19] regarding the von Mises-Fisher model (single/mixture) and parameter estimates

4.1 Single von Mises-Fisher distributions (vMF)

The single von Mises-Fisher distribution is a probability distribution function whose the total probability is concentrated on the circumference of a unit circle. It was introduced by von Mises in 1918, and Gumbel et al. have emphasised its importance and its similarities to the normal distribution [20]. From a statistical inference viewpoint, the von Mises distribution is the most commonly used for modelling circular data. Let θ be a random variable representing the wind direction in radians, and let $\mathbf{x}' = [\cos\theta_i, \sin\theta_i]'$ be a circular data point in rectangular coordinates. Thus, a *d*-dimensional unit random vector $\mathbf{x}$ is said to have a *d*-variate von Mises-Fisher distribution, which can be written as

$$f(\mathbf{x}\boldsymbol{\mu}, \kappa) = c_d(\kappa) e^{(\kappa \mathbf{x}\boldsymbol{\mu})} \quad (3)$$

Where $\|\boldsymbol{\mu}\| = 1$ is the mean direction parameter, $\kappa > 0$ is the parameter concentration with a larger values of κ imply stronger concentration about the mean direction, and $c_d(\kappa)$ is a normalising constant given by

$$c_d(\kappa) = \frac{\kappa^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa)}, \quad (4)$$

where $I_r(.)$ represents the modified Bessel function of the first kind and order *r*. Next, to derive the maximum likelihood estimator (MLE) for the parameters of single vMF, assume $\mathbf{x}_i$ to be independent and identically distributed, and let $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a finite set of sample units following $f(\mathbf{x}\boldsymbol{\mu}, \kappa)$ The likelihood function can then be written as

$$\begin{aligned} P(X \mid \boldsymbol{\mu}, \kappa) &= P(\mathbf{x}_1, \ldots \mathbf{x}_n \mid \boldsymbol{\mu}, \kappa) \\ &= \prod_{i=1}^{n} f(\mathbf{x}\boldsymbol{\mu}, \kappa) \\ &= \prod_{i=1}^{n} c_d(\kappa) e^{(\kappa \mathbf{x}\boldsymbol{\mu}')} \end{aligned} \quad (5)$$

By taking the logarithm of both side of Equation (5), we obtain

$$\ln P(X \mid \boldsymbol{\mu}, \kappa) = n \ln c_d(\kappa) + \kappa \boldsymbol{\mu}' \mathbf{r} \quad (6)$$

where $\mathbf{r} = \sum_{i=1}^{n} \mathbf{x}_i$. To derive the maximum likelihood estimator for parameters $\boldsymbol{\mu}$ and κ, the log-likelihood needs to be maximised subject to the constraints $\|\boldsymbol{\mu}\| = 1$ and $\kappa > 0$. This maximisation can be performed by applying the Lagrange multiplier λ to the function, which is given as

$$\begin{aligned} L(\boldsymbol{\mu}, \kappa, \lambda; X) &= n \ln c_d(\kappa) \\ &\quad + \kappa \boldsymbol{\mu}' \mathbf{r} + \lambda(1 - \boldsymbol{\mu}' \boldsymbol{\mu}) \end{aligned} \quad (7)$$

Next, by differentiating Equation (7) with respect to $\boldsymbol{\mu}$, κ, and λ and setting each derivative to zero, the equations that the parameter estimates $\boldsymbol{\mu}$, κ, and λ must satisfy are given by

$$\hat{\boldsymbol{\mu}} = \frac{\hat{\kappa}}{2\hat{\lambda}} \mathbf{r} \quad (8)$$

$$\hat{\boldsymbol{\mu}}' \hat{\boldsymbol{\mu}} = 1 \quad (9)$$

$$\frac{n c_d(\hat{\kappa})}{c_d(\hat{\kappa})} = -\hat{\boldsymbol{\mu}}' \mathbf{r} \quad (10)$$

By substituting Equation (8) into Equation (9), the MLEs for $\boldsymbol{\mu}$ and λ are given by

$$\hat{\lambda} = \frac{\hat{\kappa}}{2}\|\mathbf{r}\| \tag{11}$$

$$\hat{\boldsymbol{\mu}} = \frac{\mathbf{r}}{\|\mathbf{r}\|} = \frac{\sum_{i=1}^{n}\mathbf{x}_i}{\left\|\sum_{i=1}^{n}\mathbf{x}_i\right\|} \tag{12}$$

By substituting $\hat{\boldsymbol{\mu}}$ from Equation (12) into Equation (10), the MLE for κ is given by

$$\frac{c'_d(\hat{\kappa})}{c_d(\hat{\kappa})} = -\frac{\|\mathbf{r}\|}{n} = -\bar{r} \tag{13}$$

On the other hand, Banerjee at al. [19] have shown that, $\frac{-c'_d(\hat{\kappa})}{c_d(\hat{\kappa})} = \frac{I_{\frac{d}{2}}(\hat{\kappa})}{I_{\frac{d}{2}-1}(\hat{\kappa})}$; thus, the MLE for κ can be simplified as

$$\frac{I_{\frac{d}{2}}(\hat{\kappa})}{I_{\frac{d}{2}-1}(\hat{\kappa})} = \bar{r} \tag{14}$$

The numerical approach needs to be applied to Equation (14) to obtain the final estimate for κ.

4.2 Mixture of von Mises-Fisher distributions (mvMF)

The single von Mises-Fisher is very useful for modelling unimodal wind directional data. However, in some applications, the observed wind direction data cannot be represented by a unimodal distribution. To overcome this problem, a finite von Mises-Fisher mixture distribution (mvMF), which is comprised of a sum of *H* von Mises probability distributions, has been proposed. The mixture of von Mises-Fisher distributions is given by

$$f(\mathbf{x}|\boldsymbol{\mu}_h,\kappa_h) = \sum_{h=1}^{H}\omega_h f(\mathbf{x}|\boldsymbol{\mu}_h,\kappa_h) = \sum_{h=1}^{H}\omega_h c_d(\kappa_h) e^{(\kappa_h \mathbf{x}'\boldsymbol{\mu}_h)} \tag{15}$$

where $\boldsymbol{\mu}_h, \kappa_h$ are the parameter mean direction and concentration parameter, respectively, for *h=1, 2,…* *H* components of the von Mises distribution, while ω_h is a mixing parameter of nonnegative quantities that sum to one, given by

$$0 \le \omega_h \le 1 \text{ and } \sum_{h=1}^{H}\omega_h = 1,\ h = 1,2\ldots,H \tag{16}$$

$c_d(\kappa_h)$ is a normalising constant given by $c_d(\kappa_h) = \frac{\kappa_h^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(\kappa_h)}$, where $I_r(.)$ represents the modified Bessel function of the first kind and order *r*. The MLEs for the mvMF are very difficult to derive in a standard way. However, Banerjee et al. [19] provided a solution of the parameter estimates for the mvMF distribution based on the expectation maximisation (EM) approach. Let $\boldsymbol{\alpha}_h = (\boldsymbol{\mu}_h, \kappa_h)$ denote the parameters of the von Mises-Fisher distribution, $f_h(\mathbf{x}|\boldsymbol{\alpha}_h)$, for $1 < h < H$. Then, the mvMF distribution can be written as

$$f(\mathbf{x};\Theta) = \sum_{h=1}^{H}\omega_h f_h(\mathbf{x}|\boldsymbol{\alpha}_h) \tag{17}$$

where $\Theta = (\omega_1,\omega_1,\ldots,\omega_H,\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\ldots,\boldsymbol{\alpha}_H)$. According to Banerjee et al. [19], to generate a random sample from this mixture distribution, the *h*-th von Mises distribution is randomly chosen with probability ω_h. Let, $X = \{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}'$ be a data set of *n* independent sample points following Equation (17), and let $Z = \{z_1,z_2,\ldots,z_n\}'$ be the corresponding set of hidden random variables that indicate a particular von Mises distribution from which a sample is generated. In particular, $z_i = h$ if $\mathbf{x}_i$ is generated from $f_h(\mathbf{x}|\boldsymbol{\alpha}_h)$. Thus, the log-likelihood can be written as

$$\ln f(X,Z|\Theta) = \sum_{i=1}^{H}\ln\left(\omega_{z_i} f_{z_i}(\mathbf{x}|\boldsymbol{\alpha}_{z_i})\right) \tag{18}$$

Assume that the posterior distribution, $p(h|\mathbf{x}_i,\Theta)$, of the hidden variables $Z|(X,\Theta)$ are known. Then, the expectation of the log-likelihood over the given posterior distribution *p* is given by

$$E\left[\ln P(X,Z\mid\Theta)\right]=\sum_{i=1}^{n}E_p\left[\ln\left(\omega_{z_i}f_{z_i}\left(\mathbf{x}_i\mid\boldsymbol{\alpha}_{z_i}\right)\right)\right] \quad (19)$$

$$=\sum_{i=1}^{n}\sum_{h=1}^{H}\ln\left(\omega_h f_h\left(\mathbf{x}_i\mid\boldsymbol{\alpha}_h\right)\right)p\left(h\mid\mathbf{x}_i,\Theta\right)$$

$$=\sum_{h=1}^{H}\sum_{i=1}^{n}\left(\ln\omega_h\right)p\left(h\mid\mathbf{x}_i,\Theta\right)$$

$$+\sum_{h=1}^{H}\sum_{i=1}^{n}\ln\left(f_h\left(\mathbf{x}_i\mid\boldsymbol{\alpha}_h\right)\right)p\left(h\mid\mathbf{x}_i,\Theta\right)$$

Next, the parameter Θ is re-estimated to maximise the expectation function. To maximise the expectation function with respect to ω_h, the Lagrangian multiplier λ corresponding to the constraint $\sum_{h=1}^{H}\omega_h=1$ is used, and by taking the partial derivatives with respect to each ω_h from the Lagrangian, the following is obtained

$$\sum_{i=1}^{n}p\left(h\mid\mathbf{x}_i,\Theta\right)=-\lambda\omega_h \quad (20)$$

Next, by summing both sides of Equation (20) over all *h*, Banerjee at al. [19] found that $\lambda=-n$; thus, the parameter estimate for ω_h is given by

$$\hat{\omega}_h=\frac{1}{n}\sum_{i=1}^{n}p\left(h\mid\mathbf{x}_i,\Theta\right) \quad (21)$$

The parameter estimates for $\boldsymbol{\alpha}_h=\left(\boldsymbol{\mu}_h,\kappa_h\right)$ can be derived under the constraints $\left\|\boldsymbol{\mu}\right\|=1$ and $\omega_h\geq 0$ for *h=1, 2, … H*. Let λ_h be the Lagrange multiplier corresponding to the constraint; if $\kappa=0$, then $f\left(\mathbf{x}\mid\boldsymbol{\alpha}\right)$ is the uniform distribution on the sphere, and if $\kappa>0$, then the multiplier for the inequality constraint has to be zero. Thus, the Lagrangian is given as

$$L\left(\left\{\boldsymbol{\mu}_h,\kappa_h,\lambda_h\right\}_{h=1}^{H}\right)=\sum_{h=1}^{H}\sum_{i=1}^{n}\ln\left(f_h\left(\mathbf{x}_i\mid\boldsymbol{\alpha}_h\right)\right)p\left(h\mid\mathbf{x}_i,\Theta\right)$$

$$+\sum_{h=1}^{H}\lambda_h\left(1-\boldsymbol{\mu}_h'\boldsymbol{\mu}_h\right) \quad (22)$$

$$=\sum_{h=1}^{H}\left[\sum_{i=1}^{n}\left(\ln c_d\left(\kappa_h\right)\right)p\left(h\mid\mathbf{x}_i,\Theta\right)+\sum_{i=1}^{n}\kappa_h\boldsymbol{\mu}_h'\mathbf{x}_i p\left(h\mid\mathbf{x}_i,\Theta\right)\boldsymbol{\mu}_h+\lambda_h\left(1-\boldsymbol{\mu}_h'\boldsymbol{\mu}_h\right)\right]$$

By taking the partial derivative with respect to $\left\{\boldsymbol{\mu}_h,\kappa_h,\lambda_h\right\}_{h=1}^{H}$ from Equation (22), and setting it equal to zero, for each *h*, Banerjee et al. [19] obtained

$$\boldsymbol{\mu}_h=\frac{\kappa_h}{2\lambda_h}\sum_{i=1}^{n}\mathbf{x}_i p\left(h\mid\mathbf{x}_i,\Theta\right) \quad (23)$$

$$\boldsymbol{\mu}_h'\boldsymbol{\mu}_h=1 \quad (24)$$

$$\frac{c_d'\left(\kappa_h\right)}{c_d\left(\kappa_h\right)}\sum_{i=1}^{n}p\left(h\mid\mathbf{x}_i,\Theta\right)=\boldsymbol{\mu}_h'\sum_{i=1}^{n}\mathbf{x}_i p\left(h\mid\mathbf{x}_i,\Theta\right) \quad (25)$$

Using Equations (23) and (24), it is found that

$$\lambda_h=\frac{\kappa_h}{2}\left\|\mathbf{x}_i p\left(h\mid\mathbf{x}_i,\Theta\right)\right\| \quad (26)$$

$$\boldsymbol{\mu}_h=\frac{\sum_{i=1}^{n}\mathbf{x}_i p\left(h\mid\mathbf{x}_i,\Theta\right)}{\left\|\sum_{i=1}^{n}\mathbf{x}_i p\left(h\mid\mathbf{x}_i,\Theta\right)\right\|} \quad (27)$$

Next, substituting Equation (27) into Equation (25) provides the parameter estimates for κ_h as given by

$$\frac{c_d'\left(\kappa_h\right)}{c_d\left(\kappa_h\right)}=\frac{\left\|\sum_{i=1}^{n}\mathbf{x}_i p\left(h\mid\mathbf{x}_i,\Theta\right)\right\|}{\sum_{i=1}^{n}p\left(h\mid\mathbf{x}_i,\Theta\right)} \quad (28)$$

where $p\left(h\mid\mathbf{x}_i,\Theta\right)=\dfrac{\omega_h f_h\left(\mathbf{x}_i\mid\Theta\right)}{\sum_{l=1}^{k}\omega_l f_l\left(\mathbf{x}\mid\Theta\right)}$. Readers desiring a detailed discussion of the parameter estimates for the von Mises-Fisher distribution (single/mixture) should consult [19, 21, 22].

5 Results and Discussion

As mentioned above, the objective of this study was to identify the most appropriate distribution for wind direction at the Kudat station to better understand the wind regime in this area. Tables 2 shows the parameter estimates for the von Mises mixture distribution (*H*=1, 2, 3, and 4) at the Kudat station. Figure 4 presents the fitted mvMF (*H*=1, 2, 3, and 4) for the wind direction at the Kudat station. From the figure, it is clear that the single mvMF distribution (*H*=1) failed to model the wind direction data at the Kudat station accurately.

However, as the number of components of the vMF increase, the mvMF models fit the data in a more precise way. As a result, the fitted mvMF model with H=2, 3, and 4 components model the data with similar accuracies. It is quite difficult to determine the suitability precision of the mvMF model based on graphical representations only. Thus, the R^2 coefficient was used to evaluate each mvMF model.

Table 2. The parameter estimates for the mvMF (H=1, 2, 3, and 4) based on the EM algorithm.

mvMF	Parameter estimates			
	μ'		κ	ω
H=1	-0.785452	-0.6189218	0.42187	1
H=2	-0.5153732	-0.8569659	7.242308	0.4787608
	0.2909280	0.9567449	1.628116	0.5212392
H=3	0.02426039	0.9997057	8.096920	0.4945873
	0.56316151	0.8263469	1.002905	0.3759604
	-0.50926732	-0.8606084	12.27827	0.1294524
H=4	0.56772170	0.8232205	11.658204	0.3474064
	-0.31381192	-0.9494852	9.2272398	0.1439757
	-0.63765746	-0.7703200	10.752359	0.1412261
	0.02245815	0.9997478	0.970675	0.3673919

Table 3 and Figure 5 show the R^2 coefficient for each fitted mvMF model. The R^2 values are found to increase significantly for H=2, 3, and 4. From these results, it is clear that the mvMF more precisely model the actual modality of the wind direction histogram as the value of R^2 increases. The R^2 values indicate how much of the observed data the mvMF model is able to describe. Thus, the mvMF distribution with the highest value of R^2 will be able to model the data most accurately. For example, the value of R^2 for the mvMF with $H=1$ component is approximately 0.8954; thus, most of the data can still be modelled by this mvMF. However, the modality of the data cannot be modelled accurately. As the number of components for the mvMF model increase, the R^2 value also increases. This implies that the highest values of R^2 will correspond to the best model for the wind direction in Kudat. However, as seen in Table 3 and Figure 5, the R^2 value does not increase significantly for the mvMF model with H=2, 3, and 4. This result is found to be in agreement with the density plot shown in Figure 4. Thus, we can conclude that the mvMF model with $H \geq 2$ components is able to provide a good fit of the wind directional data in Kudat. By fitting the mvMF model with $H \geq 2$, most of the data, including the modality in the histogram can be modelled accurately. The most parsimonious model for wind direction in Kudat is mvMF (H=2); however, the most accurate model is mvMF (H=4).

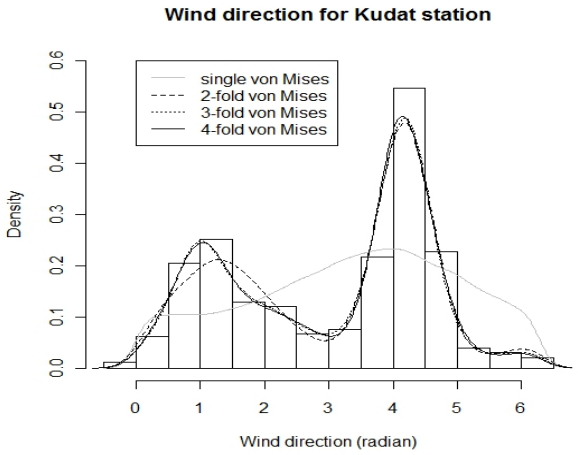

Fig 4. The mvMF (H=1, 2, 3, and 4) for wind direction in Kudat

Table 3. R^2 coefficient for each mvMF model

H	R^2 value
1	0.89540
2	0.99760
3	0.99861
4	0.99862

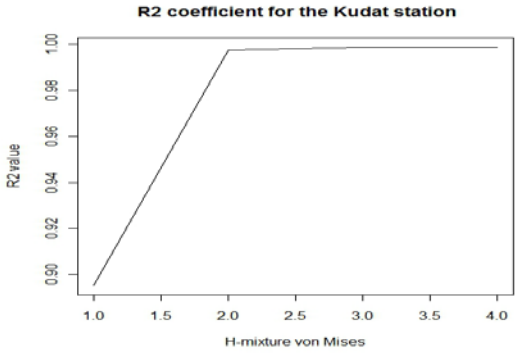

Fig 5. R^2 values for each fitted mvMF model

Based on the mvMF fitted models shown in Figure 3, Table 3, and Figure 5, it is clear that higher R^2 values indicate a better fitting mixture model. However, the weakness of R^2 is that its values will always increase as the number of parameters in the model increases. Thus, it is better to use another method, such as Akaike's information criteria (AIC) or Bayesian information criteria (BIC), for comparison with the R^2 values for each fitted model. In addition, because the mvMF distribution has been shown to adequately fit the data, this study compares the fitted mvMF model

with the circular distribution based on nonnegative trigonometric sums (NNTS) in order to determine the most prominent model for wind direction in Kudat.

5.1 Comparison with circular distribution based on nonnegative trigonometric sums (NNTS)

The nonnegative trigonometric sum series for a circular variable θ has been expressed by Fejer [23] as the squared modulus of a sum of complex numbers, which can be written as

$$\left\| \sum_{k=0}^{M} c_k e^{ik\theta} \right\|^2, \quad \text{for } k = 0, 1, 2, \ldots, M \qquad (29)$$

where $\theta \in 2\pi$, $i = \sqrt{-1}$, and c_k is a complex parameter. Using this series, Fernandez-Duran [24] proposed a new family of distributions for circular random variables, given as

$$f\left(x_j; M, \underline{\theta}\right) = \left\| \sum_{k=0}^{M} \theta_k e^{ikx_j} \right\|^2 = \sum_{k=0}^{M} \sum_{l=0}^{M} \theta_k \overline{\theta}_l e^{i(k-l)x_j} \qquad (30)$$

where the parameters (a_k, b_k) are expressed in terms of the complex parameter $a_k - ib_k = 2\sum_{v=0}^{n-k} c_{v+k}\overline{c}_v$. An additional constraint, $\sum_{k=0}^{n} \|c_k\|^2 = \frac{1}{2\pi} = a_0$, is imposed to make the trigonometric sum integrate to 1. Thus, there are 2**M* free parameters, where the parameter c_0 must be real and positive.

The parameter estimates for this model were conducted using the maximum likelihood estimation method (see [25]). This type of circular distribution has been found to be flexible enough to model directional data sets exhibiting multimodality or skewness. Thus, the fitted mvMF will be compared to this model to determine the best model for wind direction in Kudat. Figure 6 shows the fitted NNTS (*M*=1, 2, 3, and 4). Based on this figure, the fitted NNTS model is able to provide a good result for wind directional data. In particular, for the NNTS model with *M*=4 components provides a similar result to the mvMF model (*H*=4). However, instead of graphical evaluation, Table 4 provides a more meaningful comparison using AIC and BIC values.

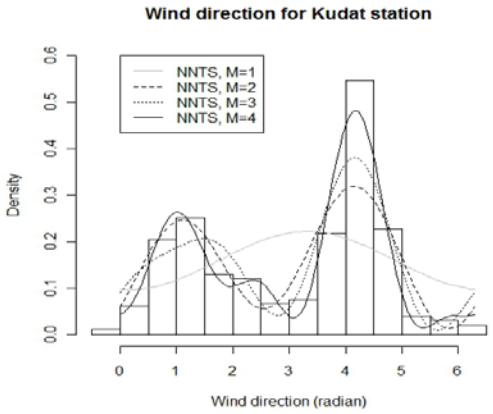

Fig 6. The NNTS model (*M*=1, 2, 3 and 4) for wind direction in Kudat

Table 4. Comparison between the mvMF model and circular distribution based on the NNTS model (*M*=1, 2, 3, and 4).

Model	AIC	BIC	Model	AIC	BIC
mvMF (*H*=1)	31451.2	31472.4	**NNTS (*M*=1)**	31562.1	31576.2
mvMF (*H*=2)	26447.7	26468.9	**NNTS (*M*=2)**	27783.1	27811.4
mvMF (*H*=3)	26210.6	26231.8	**NNTS (*M*=3)**	26269.1	27311.6
mvMF (*H*=4)	26120.3	26141.5	**NNTS (*M*=4)**	26311.6	26368.2
mvMF (*H*=5)	26102.3	26123.5	**NNTS (*M*=5)**	26285.3	26356.1
mvMF (*H*=6)	26091.9	26113.1	**NNTS (*M*=6)**	26154.4	26239.3

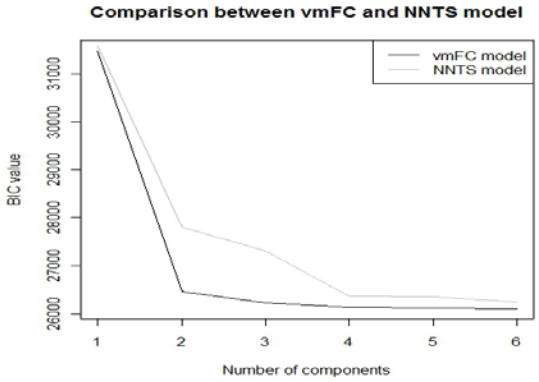

Fig **7.** BIC values for each mvMF and NNTS model

From Table 4 and Figure 7, by comparing each mvMF model, it is clear that the single mvMF has the highest AIC and BIC values, implying that the single mvMF is not a good model for wind direction in Kudat. In fact, this result is similar for the NNTS model with *M*=1 component. However, as the number of components increase for both mvMF and NNTS, the AIC and BIC values decrease, which implies that the use of more components in the mvMF and NNTS models provides a model that more adequately fits the data. These results agree

with those obtained using R^2. In addition, by comparing the values of AIC and BIC for both models, it is found that the AIC and BIC values for the mvMF models are lower than those for the NNTS models for all components. For example, the value of AIC and BIC for the mvMF model with H=4 components are lower than those for the NNTS model with M=1, 2, 3 4, 5, and 6 components. Therefore, the mvMF models were able to provide better results in fitting the wind directional data in Kudat compared to the NNTS models. Thus, the mvMF model is preferred to the NNTS model for fitting the wind directional data in Kudat..In fact, a suitable mathematical equation for the Kudat wind directional data that can written as a mvMF (H=4) model is given by

$$\begin{aligned} f(\mathbf{x}\boldsymbol{\mu},\kappa) = {} & (0.3474064)\frac{11.658^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(11.658)} e^{(11.658\boldsymbol{\mu}_1^T\mathbf{x})} \\ & + (0.1439757)\frac{9.227^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(9.227)} e^{(9.227\boldsymbol{\mu}_2^T\mathbf{x})} \\ & + (0.1412261)\frac{10.752^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(10.752)} e^{(10.752\boldsymbol{\mu}_3^T\mathbf{x})} \\ & + (0.3673919)\frac{0.971^{\frac{d}{2}-1}}{(2\pi)^{\frac{d}{2}} I_{\frac{d}{2}-1}(0.971)} e^{(0.971\boldsymbol{\mu}_4^T\mathbf{x})} \end{aligned} \quad (31)$$

with the parameter of mean directions:

$$\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \\ \boldsymbol{\mu}_4 \end{bmatrix}^T = \begin{bmatrix} 0.56772170 & 0.8232205 \\ -0.31381192 & -0.9494852 \\ -0.63765746 & -0.7703200 \\ 0.02245815 & 0.9997478 \end{bmatrix}^T \quad (32)$$

Because the mvMF has been determined to be a good model for the data, it can be used to describe some characteristics of the wind direction in Kudat. In this study, the parameter $\boldsymbol{\mu}$ for mvMF has been defined in terms of rectangular coordinates. The interpretation of the dominant direction of the wind is not suitable to be described in this way. Thus, by transforming the results into units of degrees, $0 \le \mu < 360$ may be more appropriate. Based on Equation (31), the measured parameters for the mean directions in terms of degrees are 233.37°, 256.06°, 55.52° and 82.65°. In addition, Figure 8 shows a circular density plot for the mvMF with H=4 components. This figure clearly shows that most of the wind was blowing from the north-northeast and the west-southwest and some from the east-southeast. The circular density plot reveals that the wind direction has two different dominant directions: from 190°-270° with mean directions of 233.37° to 256.06° with respect to the parameter concentration $\kappa = 11.658$ and $\kappa = 10.752$, while the other dominant direction are found to be in the range of 30°-90° with mean directions of 55.52° and 82.65° and also the concentration parameter $\kappa = 9.227$ and $\kappa = 0.971$. These implies that a stronger concentration about the mean direction of the wind blow comes from the South-West direction and follow by the minor dominant direction of the wind blows from the North-East direction.

Apart from that, the others direction are found to be quite uniformly distributed. Determining the dominant wind direction will contribute valuable information to planning or forecasting activities in such sectors as wind energy generation, air pollution assessment, climate change, construction, and maritime activities. For example, in wind energy evaluation, based on this information, the wind turbine can be positioned such that the production of energy is maximised.

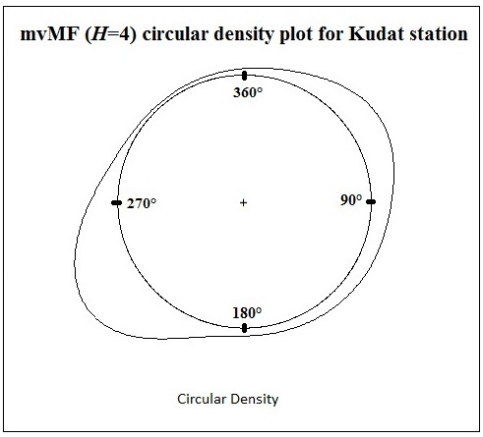

Fig 8. Circular density plot for the mvMF (H=4) model at the Kudat station

6 Conclusions

Our study focused on determining the best statistical model for wind direction in the Kudat region. The single von Mises-Fisher distribution and mixtures thereof were fit to the data. The results obtained showed that von Mises-Fisher distributions with $H \ge 2$ components adequately modelled the wind direction distribution in Kudat. Additionally, the mixture of von Mises-Fisher distributions was compared with the circular distribution based on nonnegative trigonometric sums. The results obtained based on AIC and BIC values indicated that the mixture of von Mises-Fisher distributions are preferable to the circular distribution based on nonnegative trigonometric sums for modelling the wind directional data in Kudat. Circular plots of the

model clearly show that several wind directions are more dominant in Kudat, while the other directions show an approximately uniform dispersion

Acknowledgements

The authors are indebted to the staff of the Malaysian Meteorology Department for providing wind direction data. This research would not have been possible without sponsorship from the Ministry of Higher Education, Malaysia. grant number FRGS/1/2014/SG04/UKM/03/1 and GGPM-2014-056).

References:

[1] S. R. Jammalamadaka, A. SenGupta, 2001. *Topics in circular statistics*, World Scientific Publishing, Singapore, 2001.

[2] A. Zaharim, S. K. Najid, A. M. Razali, K. Sopian, The suitability of statistical distribution in fitting wind speed data, *WSEAS Transactions on Mathematics*, Vol 7, 2008, pp. 718-727.

[3] S. Soraha, S. K. Aggarwal, A review and evaluation of current wind power prediction technologies, *WSEAS Transactions on Power Systems*, Vol. 10, 2015, pp. 1-12.

[4] A. Gjukaj, R. Bualoti, M. Celo, M. Kullolli, Wind power plant data monitoring and evaluating, *WSEAS Transactions on Power Systems*,Vol 8, 2013, pp. 24-34.

[5] N. Masseran, Markov chain model for the stochastic behaviors of wind-direction data, *Energy Conversion and Management*, Vol. 92, 2015, pp. 266-274.

[6] N. A. B. Kamisan, A. G. Hussin, Y. Z. Zubairi,Finding the best circular distribution for southwesterly monsoon wind direction in Malaysia, *Sains Malaysiana* 39, 2010, pp. 387-393.

[7] K. V. Mardia, P. E. Jupp, *Directional Statistics*, John Wiley, Chichester. 1999.

[8] N. I. Fisher, *Statistical analysis of circular data*, Cambridge University Press, Cambridge, 1993.

[9] N. Masseran, A. M. Razali, K. Ibrahim, W. Z. W. Zin, A. Zaharim, On spatial analysis of wind energy potential in Malaysia, *WSEAS Transactions on Mathematics*, Vol. 11, 2012, pp. 467-477.

[10] N. Masseran, A. M. Razali, K. Ibrahim, A. Zaharim, K. Sopian, The probability distribution model of wind speed over east Malaysia, *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 6, 2013, pp. 1774-1779.

[11] N. Masseran, A. M. Razali, K. Ibrahim, A. Zaharim, K. Sopian, Application of the single imputation method to estimate missing wind speed data in Malaysia, *Research Journal of Applied Sciences, Engineering and Technology*, Vol. 6, 2013, pp. 1780-1784.

[12] J. A. Carta, C. Bueno, P. Ramirez, Statistical modelling of directional wind speeds using mixture of von Mises distributions: Case study, *Energy Conversion and Management*, Vol. 49, 2008, pp. 897-907.

[13] J. A. Carta, P. Ramirez, C. Bueno, A joint density function of wind speed and direction for wind energy analysis, *Energy Conversion and Management* Vol. 49, 2008, pp. 1309-1320.

[14] M. Azmani, S. Reboul, J. B. Choquel, M. Benjelloun, A recursive, change point estimate of the wind speed and direction, *IEEE 7th International Conference on Computational Cybernetics*, Palma de Mallorca, Spain, 2009.

[15] G. S. Shieh, R. A. Johnson, Inferences based on a bivariate distribution with von Mises marginals, *Annals of the Institute of Statistical Mathematics*, Vol. 57, 2005, pp. 789-802.

[16] N. Dobigeon, J. Y. Tourneret, Joint segmentation of wind speed and direction using a hierarchical model, *Computational Statistics & Data Analysis*, Vol. 51, 2007, pp. 5603-5621.

[17] B. Williams, W. F. Christensen, C. S. Reese, Pollution sources direction identification: embedding dispersion models to solve an inverse problem, *Environmetrics*, Vol. 22, 2011, pp. 962-974.

[18] J. Heckenbergerova, P. Musilek, P. Kromer, Optimization of wind direction distribution parameters using particle swarm optimization. *Advances in Intelligent System and Computing*, Vol. 334, 2015, pp. 15-26.

[19] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit Hypershere using von Mises-Fisher Distributions, *Journal of Machine Learning Research*, Vol. 6, 2005, pp. 1345-1382.

[20] E. J. Gumbel, J. A. Greenwood, D. Durand, 1953. The circular normal distribution: theory and tables, Journal of the American Statistical Association Vol. 48, 1953, pp. 131-152.

[21] J. A. Mooney, P. J. Helms, I. T. Jolliffe, Fitting mixture of von Mises distributions: a case study involving sudden infant death syndrome,

Computional Statistics & Data Analysis, Vol. 41, 2003, pp. 505-513.

[22] K. Hornik, B. Grun, movMF: An R package for fitting mixture of von Mises-Fisher distributions, *Journal of Statistical Software*, Vol. 58, 2014, pp. 1-31.

[23] L. Fejer, 1915. Uber trigonometrische polynomek, *Journal fur die Riene und Angewandte Mathematik*, Vol. 146, 1915, pp. 53-82.

[24] J. J. Fernandez-Duran, Circular distributions based on nonnegative trigonometric sums, *Biometrics*, Vol. 60, 2004, pp. 499-503.

[25] J. J. Fernandez-Duran, M. M. Gregorio-Dom´ınguez, Maximum likelihood estimation of nonnegative trigonometric sum models using a Newton-like algorithm on manifolds, *Electronic Journal of Statistics*, Vol 4, 2010, pp. 1402-1410