Sains Malaysiana 52(7)(2023): 2139-2148
http://doi.org/10.17576/jsm-2023-5207-19

A Comparison between Two Discordancy Tests to Identify Outlier in Wrapped Normal (WN) Samples

(Perbandingan antara Dua Ujian Percanggahan untuk Mengenal Pasti Data Terpencil dalam Sampel Normal Balutan (WN))

Nurisha Mohd Zulkefli[1], Adzhar Rambli[1,*], Mohamad Ismeth Khan Azhar Suhaimi[1], Ibrahim Mohamed[2] & Raiha Shazween Redzuan[3]

[1]School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia
[2]Institute of Mathematical Sciences, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[3]Centre for Foundation Studies in Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

Received: 17 March 2023/Accepted: 6 July 2023

ABSTRACT

This study focuses on comparing the performance of the Robust Circular Distance (RCDU*) (simplified version) and A statistics in detecting a single outlier in the Wrapped Normal (WN) samples. Firstly, this study proposes a simplified version of RCDU statistic. Then, the paper generates the cut-off points for both statistics taken from WN samples via a simulation study. This study also evaluates the performance of both statistics using the proportion of a correct outlier detection. As a result, for a small sample size, the performance of RCDU* and A statistics do not have a huge difference. However, for a large sample size of n=250, A statistic performs slightly better than RCDU* statistic. As an illustration of a practical example, both statistics successfully detected one outlier present in the wind direction data at Kota Bharu station.

Keywords: Circular data; discordancy tests; outliers; wrapped normal distribution

ABSTRAK

Kajian ini memfokuskan kepada perbandingan prestasi Jarak Berkeliling Teguh (RCDU*) (versi ringkas) dan statistik A dalam mengesan satu data terpencil dalam sampel Normal Balutan (WN). Pertama, kajian ini mencadangkan versi ringkas statistik RCDU. Kemudian, kertas itu menjana titik potong untuk kedua-dua statistik yang diambil daripada sampel WN melalui kajian simulasi. Kajian ini juga menilai prestasi kedua-dua statistik menggunakan perkadaran pengesanan data terpencil yang betul. Akibatnya, untuk saiz sampel yang kecil, prestasi RCDU* dan statistik A tidak mempunyai perbezaan yang besar. Walau bagaimanapun, untuk saiz sampel yang besar n=250, statistik A menunjukkan prestasi yang lebih baik sedikit daripada statistik RCDU*. Sebagai ilustrasi contoh praktikal, kedua-dua statistik berjaya mengesan satu data terpencil hadir dalam data arah angin di stesen Kota Bharu.

Kata kunci: Data pekeliling; data terpencil; taburan normal balutan; ujian percanggahan

Introduction

Directional data can be in the form of circular and spherical data. As such, circular data is measured in the form of two-dimensional or angles orientations in degrees or radians. It represents a unit circle as points on the circumstances. Circular data is commonly used in a wide range of science backgrounds such as Biology, Geography, Geology, and Oceanography. In addition, the uniqueness of circular data lies in the fact that conventional techniques for the analysis of linear data (real line data) cannot be applied due to their bounded range property (Fisher 1993).

When compared to the majority of the data, a small collection of data may generate unusual and inconsistent results. This is an example of a discordancy or outlier problem. It does not have to be excessively large or small, however, it is possible to be in the middle of the data (Jammalamadaka & SenGupta 2001). However, the main concern in circular data is the existence of outliers in data as it may affect the accuracy of the analysis (Hussin et al. 2013). In the worst instance, outliers can lead to incorrect interpretations and inappropriate modelling. The identification of outliers allows for closer examination and the discovery of unexpected knowledge, proving the cruciality of the study on outliers. The decision on ways to deal with an outlier is upon the understanding of the occurrence of the outlier itself. Some may remove it if it is due to human or machine error, and some may use a robust analysis without removing it as if it is due to a natural phenomenon. The study of outliers is important because once outliers have been identified; they can do a further investigation that can lead to some new input.

Numerous outlier detection statistics have been proposed in the literature such as L-statistic, C-statistic, D-statistic, M-statistic, G-Statistic, and RCDU statistic (Abuzaid, Mohamed & Hussin 2009; Collet 1980; Mahmood et al. 2017; Mohamed et al. 2016). The latest study by Mahmood et al. (2017) used RCDU statistic to compare with other statistics (M, A and Cord statistics), however, the result is tested on Von Misses (VM) samples. The RCDU statistic is proposed to identify outliers and it depends on two points which are (1) the outliers in circular data may not be extreme values and (2) an important property of the distribution is to be symmetrical with the mean direction. The wrapped normal distribution is obtained by wrapping a normal distribution. Rambli et al. (2012) compared a few statistics (excluding RCDU statistic) in detecting outliers for the wrapped normal data. Hence, this paper focuses on the comparison of the performance between A and simplified RCDU* statistics to detect outliers in Wrapped Normal (*WN*) distribution via Monte Carlo simulation.

In the next section, the paper presents the *WN* distribution and describes some important properties of the distribution. It further reviews the A statistic and RCDU statistic. The simplified version of the RCDU* statistic is proposed. The cut-off points for both methods are obtained. The performance of RCDU* and A statistics are carried out via simulation study. One example is discussed in the last section where the tests are applied to a dataset concerning the wind direction at Kota Bharu in April 2014. The main purpose is to identify the possible outliers in a dataset.

THE WRAPPED NORMAL (WN) DISTRIBUTION

The Wrapped Normal (*WN*) distribution emerges when the density of the one-dimensional normal distribution is wrapped around the circle infinitely many times. It is also referred to as the circular normal distribution Rambli et al. (2012). The distribution is denoted by where is the mean direction and is the measure of concentration parameter. The probability density function for distribution is given by

$$g(\theta;\mu,\rho)=\frac{1}{\sigma\sqrt{2\pi}}\sum_{k=-\infty}^{\infty}e^{\left[-\frac{(\theta-\mu-2k\pi)^2}{2\sigma^2}\right]} \tag{1}$$

where σ^2 is the circular variance. The *WN* distribution also has alternate density functions as introduced by Jammalamadaka and SenGupta (2001) is;

$$g(\theta;\mu,\rho)=\frac{1}{\sqrt{2\pi}}\left(1+2\sum_{k=1}^{\infty}\rho^{k^2}\cos k\,(\theta-\mu)\right), \quad 0\le\theta<2\pi,\quad 0\le\rho\le 1\,. \tag{2}$$

The *WN* distribution is also unimodal and symmetrical about $\theta = \mu$.

A STATISTIC

The circular distance between each observation in *WN* samples is executed. The circular distance is defined as

$$d_{ij}=1-\cos(\theta_i-\theta_k) \tag{3}$$

by Abuzaid, Mohamed and Hussin (2009) where d_{ij} is a monotone increasing function (θ_i - θ_k) and $d_{ij}\in[0,2]$. Next, to evaluate the summation of all circular distances of the point of interest θ_k to all other points as follows

$$D_k=\sum_{i=1}^{n}(1-\cos(\theta_i-\theta_k)), \quad i=1,2,\ldots,n. \tag{4}$$

If the observation θ_k is an outlier, the value of D_k will increase. Thirdly, the average circular distance, $D_k/(n-1)$ is calculated and can be used to identify the possible outlier. Finally, the value of A statistic is given by

$$A=max_k\left\{\frac{D_k}{2(n-1)}\right\},\quad k=1,2,\ldots,n \tag{5}$$

where $A \in [0,1]$ is a linear measure. The values of statistic A are standardised by dividing the average circular distance by 2. This proposed statistic is based on the relative decrease in the summation of circular distance by omitting the point of interest (possible outlier),θ_k.

Abuzaid, Mohamed and Hussin (2009) propose an alternative statistic to calculate the value of A (called A*) using the definition of circular distance by Jammalamadaka and SenGupta (2001);
A reasonable statistic can be simplified as;

$$D_j* = \sum_{i=1}^{n} \left(\pi - \left|\pi - |\theta_i - \theta_j|\right|\right), \quad j = 1,2,\ldots,n.$$
$$A* = max_j \left\{\frac{D_j*}{(n-1)}\right\}, \quad j = 1,2,\ldots,n \tag{6}$$

where $A \in [0, \pi]$. This formulation is implemented in this research.

SIMPLIFIED ROBUST CIRCULAR DISTANCE (RCDU*) STATISTIC

The RCDU method is proposed by Mahmood et al. (2017) to detect single and multiple outliers using a robust method. The main concept of this statistic is the use of a circular median. First, the circular median, *med*, for each *WN* samples is calculated. Second, the distance between *med* and each of the observations is calculated. There are several possible ways to calculate the circular distance, $dist_{(i)}$ between the θ_i and the circular median, *med* as the result of the circular geometry of the data. Mahmood et al. (2017) consider the cases as follows:

If $0 \le med \le \pi$; $dist_{(i)}$

$$= \begin{cases} |\theta_i - med| & \text{if } |\theta - med| \le \pi \\ 2\pi - \theta_i + med & \text{if } |\theta - med| > \pi \end{cases}$$

If $\pi \le med \le 2\pi$; $dist_{(i)}$

$$= \begin{cases} |\theta_i - med| & \text{if } |\theta - med| \le \pi \\ 2\pi - med + \theta_i & \text{if } |\theta - med| > \pi \end{cases}$$

In this study, the formula of distance from Mahmood et al. (2017) is used to calculate the distance between each observation and the *med.* However, the distance can be simplified to the following formula $dist_{(i)}$*, where

$$dist_{(i)}* = \sum_{i=1}^{n} \left(\pi - \left|\pi - |\theta_i - med|\right|\right), \quad i = 1,2,\ldots,n.$$

The original formulation is simplified to provide an easier calculation. If θ_k is an outlier, then $dist_{(i)}$* is expected to be relatively large. The value of statistic is given by

$$RCDu* = max\ (dist*).$$

THE CUT-OFF POINTS OF RCDU* AND A STATISTICS

The cut-off points of the RCDU* and A statistics have to be obtained before the performance of both statistics are tested using *WN* samples. The cut-off points for both methods can be calculated by using R statistical package ('circular' and 'CircStat') to find the percentage points of the null distribution of no outliers in the circular dataset. The cut-off-points of RCDU* and A statistics will be obtained by considering several values of concentration parameter, ρ in the range 0.5 to 0.975 and sample sizes, n from 10 to 250. Then, for each combination of ρ and n, a sample from $WN(\mu = 0, \rho)$ is generated and the test statistics in each generated random sample are calculated using the statistics. For each combination of n and ρ, the process is repeated 2000 times. The percentage points of the discordance tests are estimated at the 10%, 5% and 1% upper percentiles.

PERFORMANCE OF RCDU* AND A STATISTIC

For each combination of n and ρ, $WN(\mu = 0, \rho)$ samples are generated for n-1 number of observations. One observation is taken from $WN(\alpha + \lambda\pi, \rho)$ where λ is the degree of contamination and $0 \le \lambda \le 1$. This observation is an outlier to the data. The performances of A and RCDU* statistics are tested based on the proportion of correct detection of the outlier. If the statistic has the proportion of correct detection equal to 1, this indicates that the statistic is expected to perform well in detecting the outlier in any real dataset. The performances of A and RCDU* statistics are compared simultaneously to find which statistic performs the best. The proportion of correct detection for both statistics is plotted with fixed ρ and different n and λ. The best statistic is the one that has higher proportion of correct detection in any value of λ.

Results and Analysis

The cut-off points are obtained by using Monte Carlo estimation in R programming and it is done for 2000 replicates. The number of WN samples generated are 10, 30, 50, 100, 150, and 250. The values of ρ used ranges from 0.5 to 0.975. The value of cut-off points for both statistics are shown to be parallel results. It can be affirmed that the behaviour of cut-off points are aligned.

Based on Table 1, it can be seen that for RCDU* and A statistics, the cut-off points increase if the sample size increases. This result applies to all values of concentration parameter, ρ and upper percentiles (10% and 1%). This result is expected due to the hypotheses of the bigger variation of the data, the more distance appears between the observations and the median of the data. When the distance is larger, the cut-off points is also large accordingly. However, this does not apply for ρ=0.5; where the data is poorly concentrated. Hence, the value of cut-off-points is bigger than the concentrated samples.

TABLE 1. Cut-off-points of RCDU* and A statistics

n	Upper percentiles	RCDU* Statistic				A Statistic			
		ρ				ρ			
		0.5	0.9	0.95	0.975	0.5	0.9	0.95	0.975
10	10%	2.838	1.225	0.831	0.576	2.373	1.232	0.863	0.602
	5%	2.947	1.325	0.909	0.638	2.459	1.366	0.947	0.655
	1%	3.052	1.595	1.085	0.752	2.577	1.610	1.130	0.766
30	10%	3.011	1.362	0.938	0.665	2.337	1.379	0.948	0.670
	5%	3.058	1.459	1.029	0.712	2.376	1.473	1.025	0.721
	1%	3.112	1.686	1.168	0.800	2.457	1.686	1.197	0.804
50	10%	3.067	1.423	0.990	0.699	2.318	1.413	0.900	0.7001
	5%	3.094	1.526	1.055	0.750	2.351	1.513	1.073	0.756
	1%	3.123	1.721	1.210	0.829	2.408	1.708	1.180	0.872
100	10%	3.101	1.504	1.058	0.745	2.296	1.508	1.053	0.741
	5%	3.116	1.597	1.134	0.800	2.320	1.612	1.118	0.787
	1%	3.132	1.779	1.280	0.884	2.363	1.786	1.255	0.896
150	10%	3.117	1.543	1.095	0.762	2.282	1.561	1.079	0.777
	5%	3.125	1.625	1.157	0.810	2.304	1.662	1.137	0.821
	1%	3.135	1.813	1.282	0.894	2.340	1.833	1.283	0.929
250	10%	3.126	1.613	1.147	0.795	2.265	1.614	1.135	0.792
	5%	3.132	1.694	1.207	0.839	2.283	1.715	1.183	0.834
	1%	3.137	1.886	1.333	0.934	2.307	1.911	1.305	0.917

PERFORMANCE OF RCDU* AND A STATISTICS

In order to compare the performance of RCDU* and A statistics, the value of the concentration parameter, ρ used is 0.95 with upper percentiles 5%. The outcome of the simulation is in favour of A statistic since this method outperforms RCDU* in all cases, with a very small margin of percentages. The highest size of the sample, the difference more evidently.

For a small sample size, n=10 and n=30, the performances of RCDU* and A statistics do not have a huge difference, since the RCDU* statistic is masked by A statistic (Figure 1). Whilst, with the large sample size of n=250, A statistics performs slightly better than RCDU* statistic as shown in Table 2. Overall, the performance of RCDU* statistic has almost been masked by A statistic. In addition, from contamination values of 0 to 0.3, the performance for both methods are low since the outlier is detected as unclear. Nonetheless, the outlier is successfully detected when the contamination values reach 0.4 to 1. It can be assured that, as the contamination values increase from 0.5 onwards, the performance for both methods is capable in detecting the outlier.

PRACTICAL EXAMPLE OF REAL DATA

The data consists of wind direction at Kota Bharu in April 2014. The data is obtained from the Malaysian Meteorology Department. The main objective was to examine if there is a possible wind direction that deviates away from the usual direction. The data is tabulated in Table 3. The observations are plotted as shown in Figure 2. Figure 2(a) shows that there is a candidate for an outlier (the 28th observation with with $\theta = 0.5262$) presence in the data set. Furthermore, Figure 2(b) indicates that the observations follow a WN distribution.

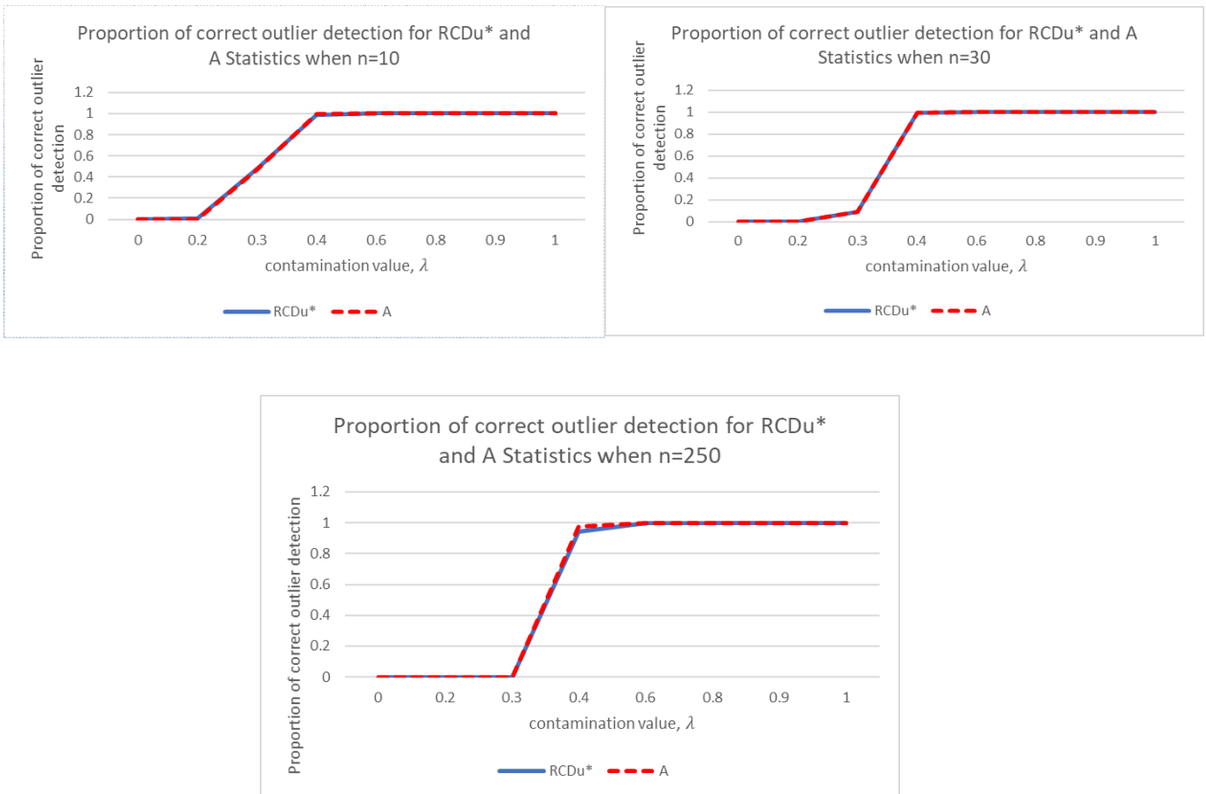


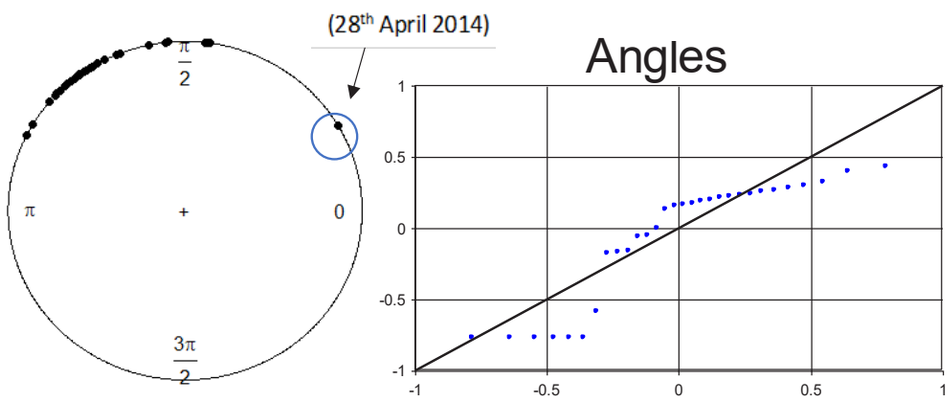
FIGURE 1. Proportion of correct outlier detection for RCDU* and A statistics for all values of λ, and n

TABLE 2. Performance of RCDU* statistic and A statistic

n	λ	$\rho = 0.95$	
		RCDU*	A
10	0.1	0.000	0.000
	0.2	0.007	0.002
	0.3	0.480	0.472
	0.4	0.983	0.992
	0.6	1.000	1.000
	0.8	1.000	1.000
	0.9	1.000	1.000
	1	1.000	1.000
30	0	0.000	0.000
	0.2	0.000	0.000
	0.3	0.086	0.087
	0.4	0.994	0.996
	0.6	1.000	1.000
	0.8	1.000	1.000
	0.9	1.000	1.000
	1	1.000	1.000
250	0	0.000	0.000
	0.2	0.000	0.000
	0.3	0.000	0.000
	0.4	0.945	0.976
	0.6	1.000	1.000
	0.8	1.000	1.000
	0.9	1.000	1.000
	1	1.000	1.000

TABLE 3. Wind direction for Kota Bharu in April 2014

Date	Wind direction (radian)	Date	Wind direction (radian)
1st April	2.391684	16th April	2.267877
2nd April	2.121094	17th April	2.174819
3rd April	2.206873	18th April	2.681867
4th April	2.379025	19th April	2.162481
5th April	1.971244	20th April	2.35695
6th April	1.944247	21st April	2.610076
7th April	1.680474	22nd April	1.775548
8th April	2.317006	23rd April	1.448051
9th April	2.133993	24th April	2.228268
10th April	1.458142	25th April	2.445246
11th April	1.428168	26th April	2.244698
12th April	2.087719	27th April	2.121297
13th April	2.040151	28th April	0.526182
14th April	2.389398	29th April	2.29731
15th April	2.209842	30th April	1.667283



(a) (b)

FIGURE 2. Circular plot (a) & PP-plot (b) of Kota Bharu wind direction

Based on Table 4, the value of the mean direction, $\mu = 2.0793$ and the concentration parameter, $\rho = 0.9176$. The cut-off points of the data are obtained using the same procedures as applied for both RCDU* and A statistics. The cut-off points ($n = 30$ and $\rho = 0.9176$) considered are for 1%, 5% and 10% upper percentile (Table 5).

The test statistics for the candidate of an outlier with direction (the 28th observation with $\theta = 0.5262$) for Kota Bharu wind data in April 2014 are shown in Table 6. Since the test statistics for both methods are greater than the cut-off points for all values of upper percentiles, this can be concluded that the 28th observation is an outlier. After the outlier is omitted, the descriptive statistics (Table 7) for observations of wind direction in Kota Bharu in April 2014 are obtained. In addition, by removing the discordancy, the concentration parameter increases, proving that the data is more concentrated.

TABLE 4. Descriptive statistics for Kota Bharu in April 2014 with one possible outlier

Parameter	Value
Number of observations, n	30
Mean direction, μ	2.0793
Median	2.1687
Concentration parameter, ρ	0.9176
Standard deviation	0.4146

TABLE 5. Cut-off-points for RCDU* and A statistics for wind direction data

Upper percentile	Cut-off points for RCDU* Statistics	Cut-off points for A Statistics
10%	1.2042	1.2494
5%	1.2876	1.3378
1%	1.4901	1.5183

TABLE 6. Test statistics for the wind direction data

Observation	Test statistics	
	RCDU* statistics	A statistics
28th April	1.6425	1.5856

TABLE 7. Descriptive statistics for Kota Bharu in April 2014 without the outlier

Parameter	Value
Number of observations, n	29
Mean direction, μ	2.1156
Median	2.1748
Concentration parameter, ρ	0.9493
Standard deviation	0.3226

The wind direction on 28th April 2014 can be looked at more closely. Many questions have arisen from this result. Was the recording device for the wind direction faulty on that day? Was there any human error during the measurement? Were there any natural phenomena that happened on that day resulting in the wind direction to deviate away from other normal days? These problem statements are important to be addressed once the outlier is identified. Since this observation is a significant outlier, Meteorology Department must investigate this wind direction reading on the 28th of April. The decision on ways to handle the outlier; either to use a robust approach or to eliminate the outlier from the analysis is suggested to be done.

Conclusion

This study found that A statistics has a slightly better performance in detecting single outlier in the WN samples for all sample sizes and concentration parameter values as compared to the RCDU* statistic. However, the performance of the A statistic is only slightly better than RCDU* statistic and this indicates that both methods are good in detecting single outlier in the WN samples. The better performance of A statistic is more visible in a large sample size (n=250). Besides, these methods are also applied to detect a single outlier from a Kota Bharu wind direction data in April 2014. Both methods managed to detect the outlier present in the data. This implied that both statistics can be applied to real data, making them a useful tool to help the Meteorology Department to detect an observation on the inconsistency of the wind direction with the majority of the data. In choosing the best method of detecting a single outlier in the WN samples, the most suitable is the A Statistics for any sample size, n and concentration parameter, . For more details about this study, it is recommended for the potential researchers to apply the RCDU* and A statistics on any different circular distribution. Hence, this study contributes to the practitioner's exploration of the outlier detection especially in circular data that follows WN distribution.

Acknowledgements

We would like to extend our gratitude to the Ministry of Higher Education for the FRGS grant FRGS/1/2019/STG06/UITM/02/1 and Pembiayaan Yuran Penerbitan Artikel (PYPA) scheme for providing financial support for this study.

References

Abuzaid, A.H., Mohamed, I.B. & Hussin, A.G. 2009. A new test of discordancy in circular data. *Communications in Statistics: Simulation and Computation* 38(4): 793-809. https://doi.org/10.1080/03610910802627048

Collett, D. 1980. Data in circular outliers. *Journal of the Royal Statistical Society* 29(1): 50-57.

Fisher, N.I. 1993. *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511564345

Hussin, A.G., Abuzaid, A.H., Ibrahim, A.N.I. & Rambli, A. 2013. Detection of outliers in the complex linear regression model. *Sains Malaysiana* 42(6): 869-874.

Jammalamadaka, S.R. & SenGupta, A. 2001. *Topics in Circular Statistics*. World Scientific.

Mahmood, E.A., Rana, S., Midi, H. & Hussin, A.G. 2017. Detection of outliers in univariate circular data using robust circular distance. *Journal of Modern Applied Statistical Methods* 16(2): 418-438.

Mohamed, I.B., Rambli, A., Khaliddin, N. & Hussin, A.G. 2016. A new discordancy test in circular data using spacings theory. *Communication in Statistics - Simulation and Computation* 45(5): 2904-2916. https://doi.org/10.1080/03610918.2014.932799

Rambli, A., Ibrahim, S., Abdullah, M.I., Hussin, A.G. & Mohamed, I. 2012. On discordance test for the wrapped normal data. *Sains Malaysiana* 41(6): 769-778.

*Corresponding author; email: adzhar@uitm.edu.my