The Identification of Outliers in Wrapped Normal Data by using Ga Statistics

Mohammad Illyas Sidik, Adzhar Rambli, Zamalia Mahmud, Raiha Shazween Redzuan, Nur Huda Nabihan Md Shahri

Abstract: ***This paper focuses on identifying outliers in the wrapped normal distribution. It is commonly found and when it is dealing with circular data, the existing of outliers will increase several problems.We will be using the existing statistics, the G_a statistics to identify a single and patch of outliers in the wrapped normal data. A Monte Carlo simulation will be carried out to generate the cut-off point's value. The power performance of the discordancy test in circular data has been investigated. The increment of the contamination level, λ, large value of concentration parameter, ρ and large sample size, n will increase the performance of the outlier detection procedures. In addition, the result shows that the statistics performs well in detecting a patch of outliers in the data. As an illustration a practical example is presented by using the wind direction in Kota Bharu station. As conclusion, the G_a statistics successfully detect outlier presence in this data set.***

Keywords: Circular data, outliers, G_a statistics, wrapped normal distribution, Monte Carlo simulation, wind direction.

I. INTRODUCTION

Linear data can be represented on a real line such as weight (kg), height (cm), and volume of water (litres). However, there exists another type of data which is circular data. It is also known as directional data which represent a two dimensions data. Circular data manages data points disseminated in a circle. It is a continuous point where the starting and ending points are joined together. Besides that, the bounded property and the directions in a circular data are near the contrary end points and are close neighbour. However, in linear data, it is maximally distant compared to circular data (Abuzaid [1]). The range of circular data is of 0 to 2π radian or 0° to 360°.

Revised Manuscript Received on February 05, 2019.

Mohammad Illyas Sidik, Centre of Statistical and Decision Sciences Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia,

Adzhar Rambli, Centre of Statistical and Decision Sciences Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia,

Zamalia Mahmud, Centre of Statistical and Decision Sciences Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

Raiha Shazween Redzuan, Centre for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia

Nur Huda Nabihan Md Shahri, Centre of Statistical and Decision Sciences Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia

There are several circular measure instruments such as the clock, compass, and protractor. The single circular observation represents θ as a point on unit vector or radius on a circle. The directional statistics applications are quite popular and broadly found in various scientific areas including the studies in meteorology, biology, physics, psychology, medicine, geology, natural science, geography, and political science. A lot of contexts are explored in natural science regarding circular data such as Rivest [15] which forecasted the path of ground movement when an earthquake occurs, while Downs and Mardia [4] employed their projected circular regression models on the circular data. Next, Mulder and Klugkist [11] developed the Bayes factor and applied in psychology to evaluate whether deafness improves haptic perception. Besides that, according to Downs *et al*. [5], in medical sciences, the correlations amongst circadian biological rhythms were in a 24 hour clock [2,3,10] and the angles of knee flexion as a measurement of recovery of orthopaedic patients (Jammalamadaka *et al*., [8]) are considered as a circle, while Roy *et al*. [16] proposed joint circular linear mixture model to be used in psoriatic plaque segmentation of a skin images by using the hue and the chroma observation. Rambli [13] was successfully verified in detecting a patch of two outliers on the local eye data set by applying G_a statistics from von Mises distribution (VM). Moreover, in extra solar planets Yanga *et al*. [18]applied the proposed learning-based EM algorithm to cluster exoplanet data which is one of application in spherical data.

Since the response data is circular, the classical longitudinal or functional data statistical tools might not be suitable and that could lead the outputs to misleading decisions. For example, in the meteorological department, the wind direction can be measured by using direction sensor. The direction of wind refers to the course from which the wind blows. It is communicated in degrees measured clockwise from the topographical north.

One of problems is the occurrence of outlying points in a circular data. Outlier is defined as point lies distant from the other observations [6,9]. The presence of outlying points in data set is affecting the estimation of parameter and could lead the outputs in the wrong decision. Moreover, the G_a statistics is based on the spacing theory such as gap or distance between the points in circular data. The statistics considered both sides which are the right and left sides. If the gap is definitely large and far from the other observation, the kind of point will be presented as an outlier.

Besides that, the simulation method application of the distribution of linear data such as Uniform, Gamma, Exponential, Cauchy, Normal, Poisson, Chi-square, and others are accessible. Normal distribution is commonly found in the study due to its significant points. A distribution of wrapped normal sample (WN) is gained by wrapping a normal sample on circumferences of a unit circle (Rambli *et al*. [14]). In circular data, the wrapped normal data is one of the distributions that could detect outliers. The von Mises distribution is also very popular in circular data due to focus in the literature on parameter estimation that describes the prediction, confidence interval and sampling distribution on the circle. Moreover, the applications and analyses of circular data are becoming important and highly demanded for further exploration in future due to distinct scientific areas. Recently, the wrapped normal distribution is a unique case in circular data and its one of the wrapped stable families. For instance, the cardioid distribution, wrapped Cauchy distribution, uniform distribution, and wrapped normal distribution are some of the circular data distributions. In this paper, the wrapped normal distribution will be considered as one of the circular data which can be used to identify the unexpected samples.

In order to study and explore more about the characteristics of the wrapped normal distribution, the bench mark points are obtained and the simulation methods would be performed to study the performance of the statistics in order to detect influential values in circular data. Next, the application of real data set that was attained from the Malaysian Meteorological Department on appropriate statistics approach. In circular data, the issue of outlying observations has not been fully completed. The emphasis in this paper is to investigate the performance of G_a statistics to identify outliers in wrapped normal data.

II. METHODOLOGY

The Test of Outliers Identification in Wrapped Normal Data

A distribution of wrapped normal sample is gained by wrapping a normal sample on circumferences of a unit circle. The normal distribution is represented by $N(\mu_L, \sigma_L^2)$ where μ_L refers to the mean and σ_L^2 refers to the variance while the WN distribution is represented by $WN(\mu, \rho)$, where μ and ρ refer to the mean direction and the concentration parameter, respectively. Its probability distribution function is given by

$$f(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} e^{\frac{-(\theta-\mu-2k\pi)^2}{2\sigma^2}} \qquad (1)$$

where, σ^2 is the variance for circular data.

Whittaker and Watson [17] postulated, a substitute and a more suitable depiction of this density which is

$$f(\theta) = \frac{1}{2\pi}\left[1 + 2\sum_{k=-\infty}^{\infty} \rho^{k^2} \cos k(\theta - \mu)\right], 0 \le \theta < 2\pi, 0 \le \rho \le 1 \qquad (2)$$

The distribution is uni-modal and symmetric about the value $\theta = \mu$. Unlike the von Mises sample, the WN sample holds the additional property, such that, the joints density of two WN variables is also WN.

To be exact, if $\theta_1 \sim WN(\mu_1, \rho_1)$, and $\theta_2 \sim WN(\mu_2, \rho_2)$ are independent, then $\theta_1 + \theta_2 \sim WN(\mu_1 + \mu_2, \rho_1 + \rho_2)$(see Jammalamadaka and SenGupta [7]).

Unlike the wrapped Cauchy distribution, the behaviour of the wrapped normal sample is closer to that of the von Mises sample. As ρ increase from 0 to 1, the points are more concentrated in the direction $\mu = 0$. As the value of ρ gets smaller, the generated data set tends to be uniformly distributed.

The Spacing Theory

The good spacing theories have been reviewed and it can be referred to Pyke [12]. Rambli [13] has developed a different test of outlier identification, represented by G_a statistics for identifying a one, multiple and a patch of outliers and applied it to the VM distribution. Therefore, this paper used the algorithm that been proposed by Rambli [13] to further explore about extreme values in wrapped normal data. The algorithm is given as follows. Firstly, $\theta_1, \theta_2, \dots, \theta_n$ are independently and identically distributed $(i.i.d)$ placed on around a unit circle, while $\theta_{(1)}, \theta_{(2)}, \dots, \theta_{(n)}$ are the equivalent sequenced circular observation. Then, according to Rambli [13], he defined the one-step spacing for i^{th} sequenced sample as

$$G_{1i} = \theta_{(i+1)} - \theta_{(i)}, i = 1, 2, \dots, n, \text{ and } G_{1n} = 2\pi - \theta_{(n)} + \theta_{(1)} \qquad (3)$$

Note that $\{G_{1i},\ i = 1, 2, \dots, n\}$ provides an order of the gaps between sequential samples on the circle. Further essentially, the order is not influenced by the selection of the zero direction. The statistics (3) can be generalised to identify outliers in circular samples. So that, the study defines G_{ai} as the $a-$step spacing for the i^{th} sequenced sample, $a = 1, 2, 3, \dots$ and $i = 1, 2, \dots, n$ such that

$$G_{ai} = \theta_{(i+a)} - \theta_{(i)},\ i = 1, 2, \dots, n-a \text{ and } G_{ai} = 2\pi - \theta_{(i)} + \theta_{(i+a)-n} \qquad (4)$$

where,

$$i = (n+1) - a, (n+2) - a, \dots, n. \qquad (5)$$

This paper has employed statistics (3) and (4) in the implementation of a new outlier identification test, represented by G_a in the identification of a single, multiple and patch of outliers in a wrapped normal distribution.

The G_a Statistics

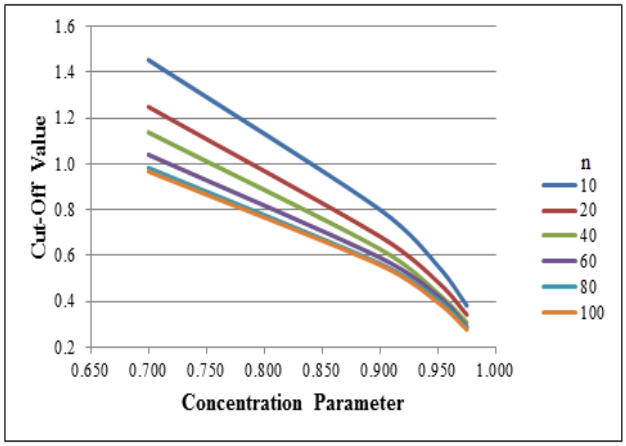
Assume $\theta_1, \theta_2, \dots, \theta_n$ are $(i.i.d)$ taken from a WN sample. Rambli [13] stated that the stages to attain the G_a statistics are elaborated as follows. Firstly, sort the samples as $\theta_1, \theta_2, \dots, \theta_n$ in ascending order. Secondly, select the number of $a-$step spacing, then compute $G_a, i = 1, 2, \dots, n$ as shown in equation (4).

Thirdly, define $G_i = \min(G_{ai}, G_{a,i-a}), i = 1, 2, \dots, n$, which the smaller is the $a-$step spacing on both side of θ_i. Fourthly, define $G_a = \max_{i=1,2,\dots,n}(G_i)$. Since the G_a value surpasses the cut-off point, so the i^{th} sample equivalent to $\max_{i=1,2,\dots,n}(G_{ai})$ is indicated as an outlier.

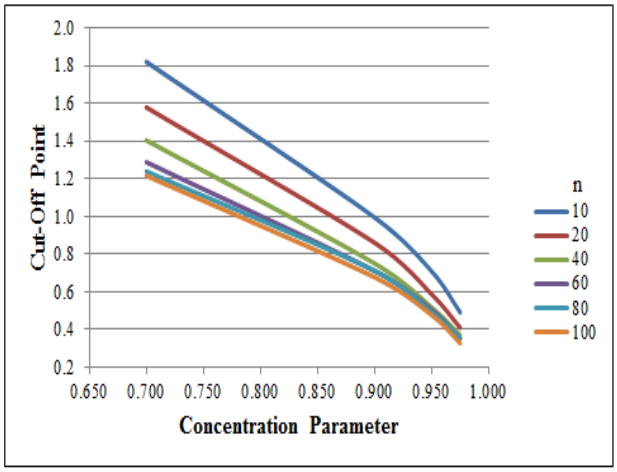
The cut-off points for the G_a statistics

The focus of this research is examining the procedure of the G_a statistics to identify outliers in data sets produced from aWNsample. First of all, the study acquired the cut-off points C_g for the test. The study designed a Monte Carlo study using the SPlus statistical package to determine the proportion points under the null hypothesis of no outliers in the circular observations. The study considered the four values of the concentration parameter, ρ in the range of 0.7 to 0.975 and distinct sample sizes, n between 10 and 100. For every combination of n and ρ, the study generate a sample of $WN(\mu = 0, \rho)$ and calculated the G_a statistics. The study repeated the process 3000 times and the study aimed to evaluate the proportion points of the G_a statistics at the 10%, 5% and 1% upper percentiles when there is no presence of outlier in the data. Commonly, the researcher expects that only 5% significance level in estimation of parameter, forecasting and modelling. Thus, at 95% confidence level we choose the cut-off point to indicate the outlier if the gap exceeds the cut-off point.

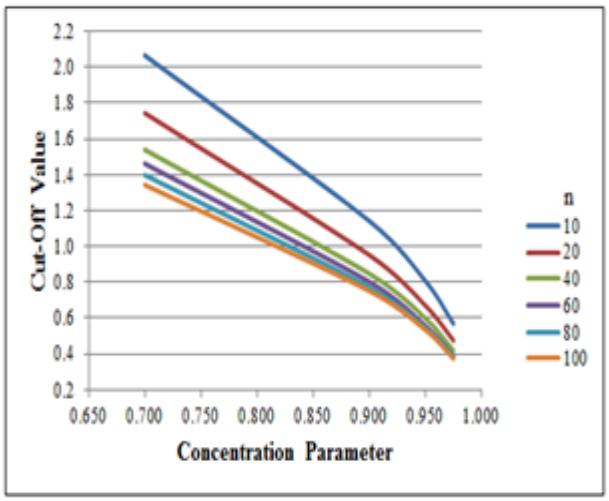
The results of Monte Carlo simulation of cut-off points for the $G_a$statistics where, $a = 1$, 2, and 3 were showed in Figure 1. All the parameter values, considering the value of cut-off point, decreases as the concentration parameter, ρincreases. Besides that, the result shows that as the sample size increase, the cut-off point for all values of the concentration parameter, ρ and the levels of upper percentile decreased.



(a) G_1



(b) G_2



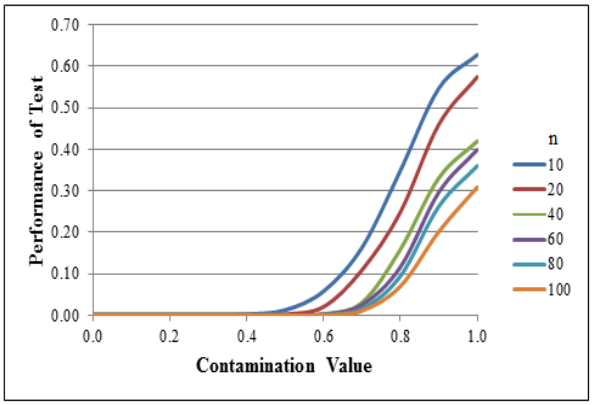
(c) G_3

Fig. 1 The 95 Upper Percentile Level Cut-Off Points of G_a Statistics for Different Sample Size, n

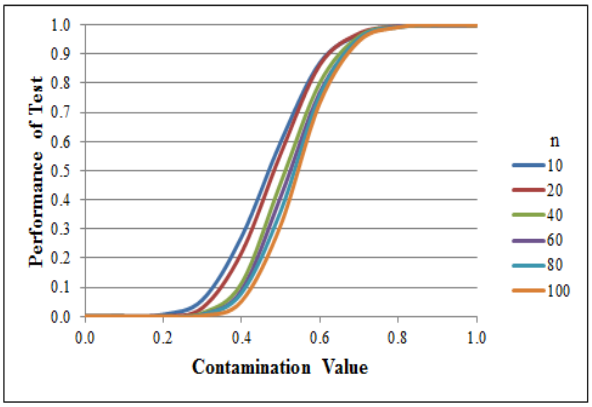
The G_a statistics performance

The suggested G_a statistics is advantageous compared to other outlier detection methods as this statistical method that can be generalised to identify outliers in a circular observation (Rambli [13]). Thus, in this paper, the G_a statistics was used in order to identify outliers in wrapped normal distribution. To examine the performance of the G_a statistics for $a = 1$, 2, and 3, the research produced observations based on distinct sizes $10 \leq n \leq 100$ and concentration values $\rho = 0.7$, 0.9, 0.95, and 0.975. The data are produced in such a way that $n - a$ of the samples comes from $WN(\alpha, \rho)$ and the remainder from $WN(\alpha + \lambda\pi, \rho = 0.975), 0 \leq \lambda \leq 1$. The study established from ρ =0.975, so that the outlying samples are gathered in a single patch. The G_a statistics in each random observation is then computed. If G_a is surpasses cut-off point, it can be concluded that the study has accurately identified the patch of a outliers. The simulation is repeated 3000 times and the percentage of correct identification of the patch of outliers was attained and incorporated into the samples.

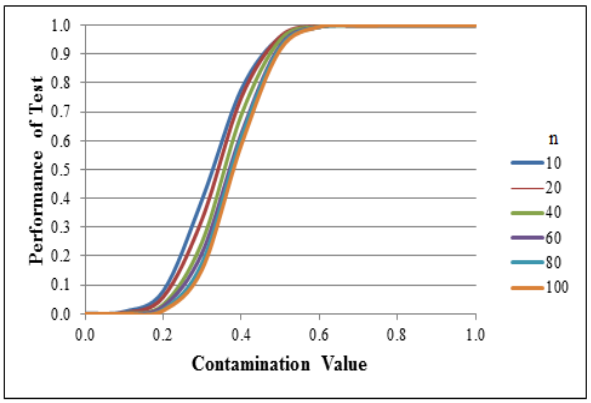
The performance of the G_a statistics where, $a = 1$, 2, and 3 for sample size, n between 10 and 100 while the concentration parameter, ρ between 0.7 and 0.975 are illustrated in Figure 2−4. Moreover, as the value of contamination degree, λ increases, the proportion of correct outlier detection will be increased for all concentration parameter, ρ. The value of outlier correct detection for G_a statistics is higher when the concentration parameter, ρ increases. Other than that, as the sample size, n increases, the proportion of outlier correct detection for G_a statistics is gradually decreased. Thus, the larger concentration parameter, ρ demonstrates better performance in outlier correct detection for G_a statistics.



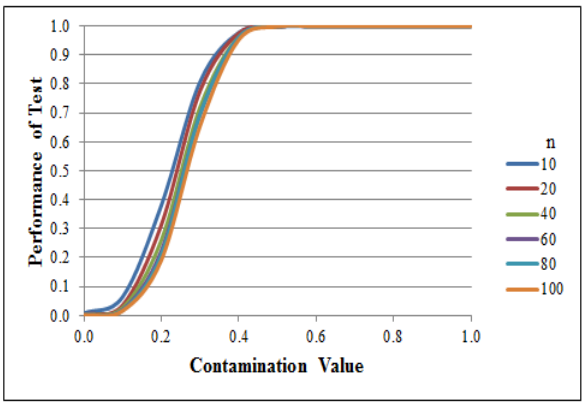
(a) $\rho = 0.7$



(b) $\rho = 0.9$

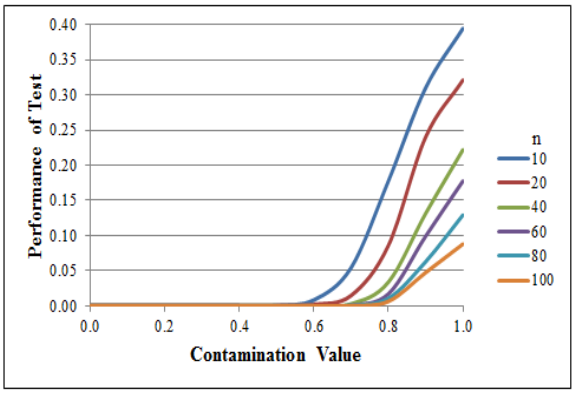


(c) $\rho = 0.95$

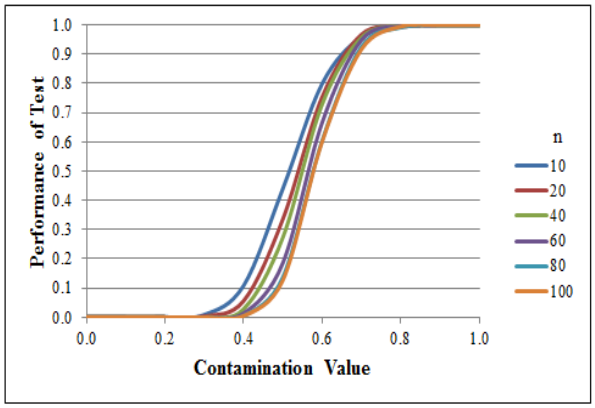


(d) $\rho = 0.975$

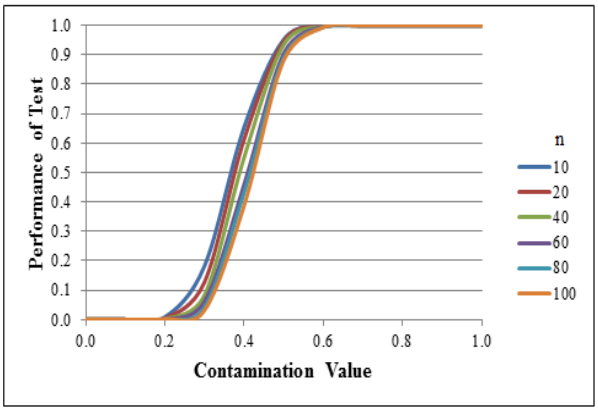
Fig. 2 The Performance of G_1 Statistics for ρ



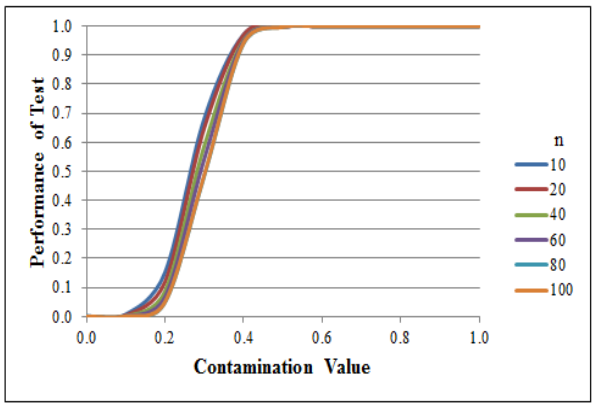
(a) $\rho = 0.7$



(b) $\rho = 0.9$

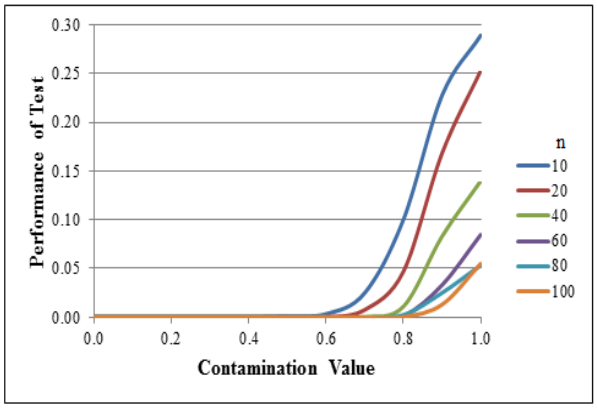


(c) $\rho = 0.95$

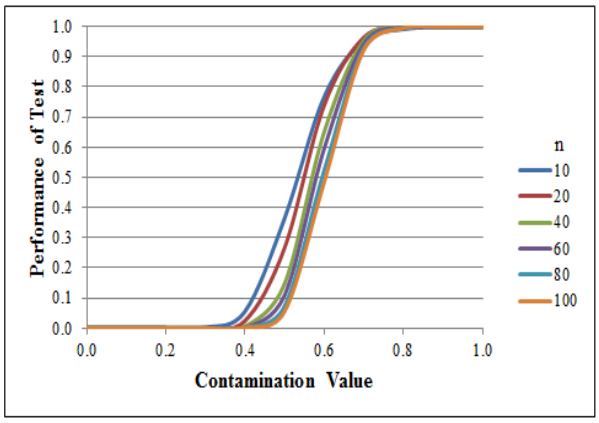


(d) $\rho = 0.975$

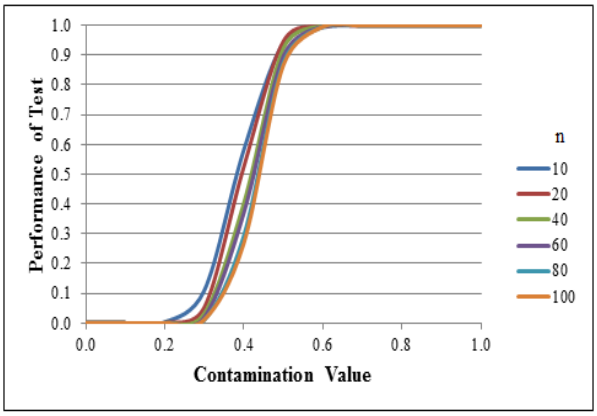
Fig. 3 The Performance of G_2 Statistics for ρ



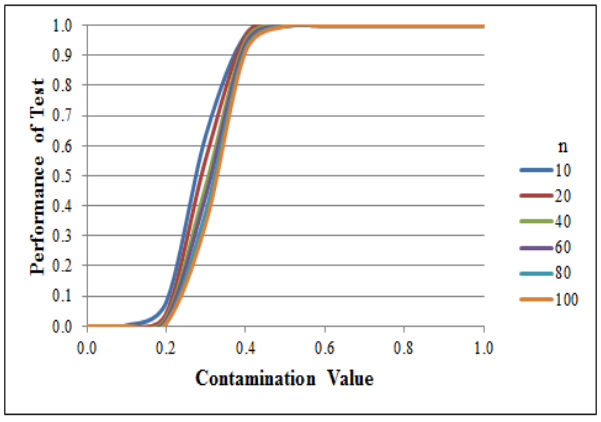
(a) $\rho = 0.7$



(b) $\rho = 0.9$



(c) $\rho = 0.95$

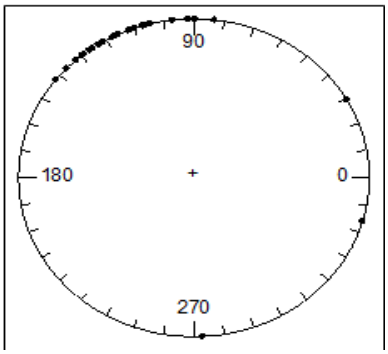


(d) $\rho = 0.975$

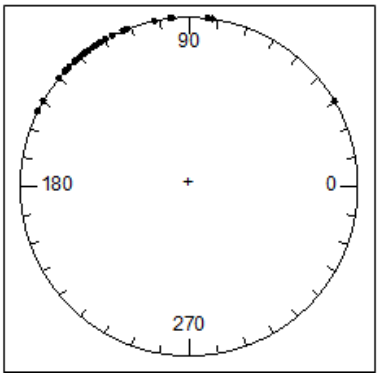
Fig. 4 The Performance of G_3 Statistics for ρ

III. RESULTS

This paper examined the daily wind direction data set form 2014 which was obtained from the Malaysian Meteorological Department. Several plots have been plotted to illustrate the distribution of the data set. Thus, the circular plot in Figure 5 shows that the months of February and April have possible outliers. Figure 6 shows the circular histograms of wind direction in February and April, 2014, respectively. The figures indicate that the wind direction has the highest frequency in the second quadrant which signifies that the wind flowed more towards the South East direction. The mean of the wind direction is 111.278° in February and 119.135° in April. The parameter estimation of wrapped normal concentration parameter, ρ is 0.8132 in February and 0.9176 in April.

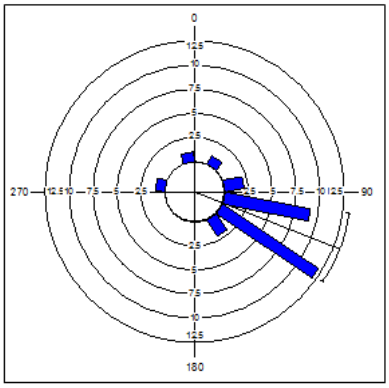


(a) February

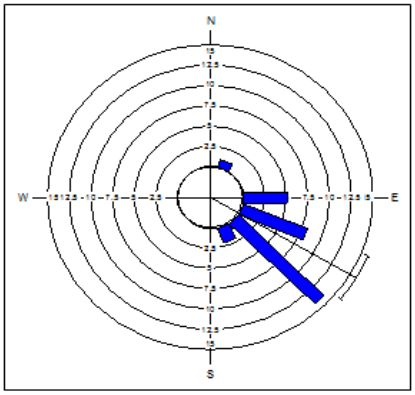


(b) April

Fig. 5 The Wind Direction



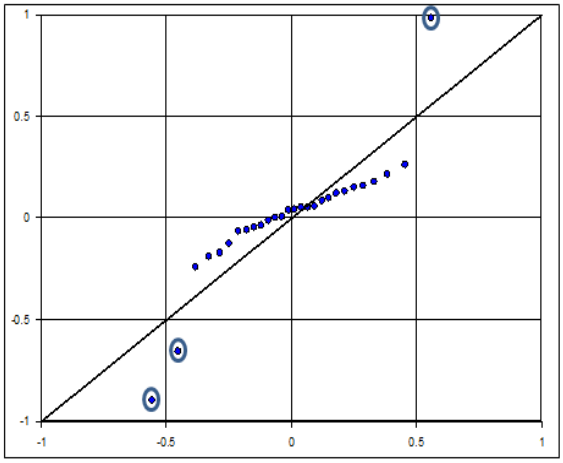
(a) February



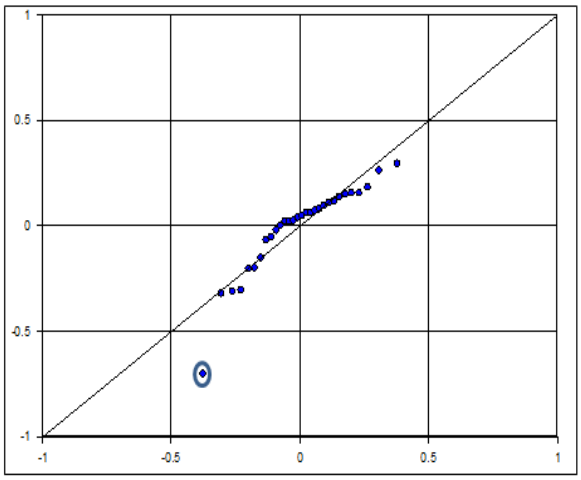
(b) April

Fig. 6 The Circular Histogram of Wind Direction

Figure 7 shows that the quantile probability plots for wind direction in February and April 2014. From Figure 7(a), most of the samples are not lie in the straight line. Moreover, there are three samples located far from the other observations. In Figure 7(b) shows most of observations are lie approximately along the straight line. However, there is one observation located far from the other observations.



(a) February

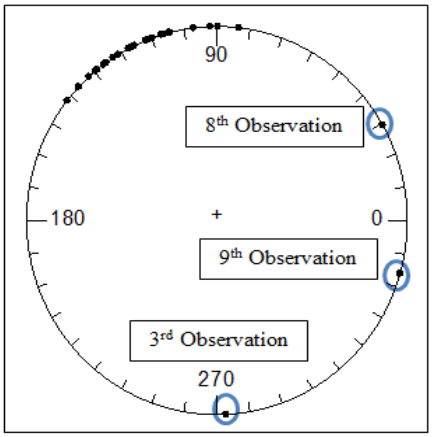


(b) April

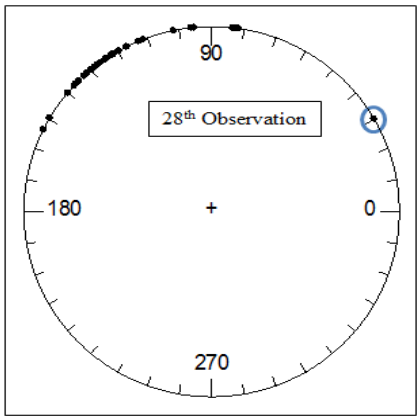
Fig. 7 The Quantile Probability-Probability Plot

Figure 8(a) shows the circular plot of the daily wind direction of February 2014 and it is observed that there are three possible points of outliers which are pointed far from the rest of observations. The researcher used three-step G_a statistics, G_{3i} statistics which is the value for each observation at 3rd, 8th, and 9th observations are $G_{3,3} = 2.4527$ radian, $G_{3,8} = 1.0892$ radian, and $G_{3,9} = 1.8418$ radian respectively. Hence, the cut-off point at 95% confidence level is 1.2664 radian. Thus, it indicates that there are two points at 95% of confidence level which are outliers which are 2.4527 radian and 1.8418 radian since the 3rd and 9th observations exceed the cut-off point. Moreover, there are two outliers which were identified by applying the extension of the three-step G_a statistics, G_{3i} statistics.

Figure 8(b) shows the circular plot of the daily wind direction in April 2014. The figure highlights that there is a possibility of a point of being an outlier which is pointed far from the rest of observations. This paper employed the one-step G_a statistics, G_{1i} statistics where the value at the 28th observation is $G_{1,28} = 0.902$ radian. Hence, the cut-off point at 95% confidence level is 0.5976 radian. Thus, it indicates that one point at 95% of confidence level is an outlier which is 0.902 radian since the 28th observation exceeds the cut-off point. Moreover, there is an outlier which is identified by applying the extension of the one-step G_a statistics, G_{1i} statistics.



(a) February

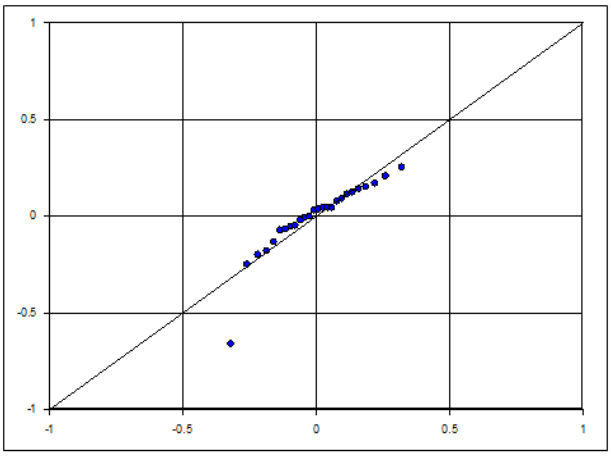


(b) April

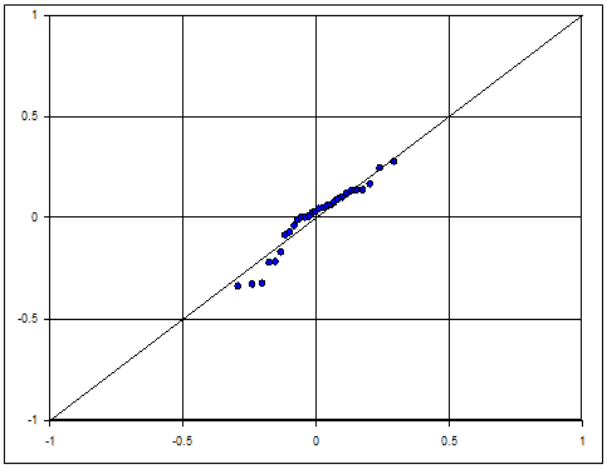
Fig. 8 The Wind Direction

Further, by removing the 3rd and 9th observations, it is indicated the mean of wind direction in February have increased to 112.467° and the values of concentration parameter, ρ significantly increased to 0.9356. Besides that, it is indicate that the removal of the 28th observation shows that the mean of wind direction increased to 121.217° while, it will significantly increase the concentration parameter, ρ to 0.9493.

Figure 9 shows that the quantile probability plots without outliers for wind direction in February and April 2014. Most of observations are lie approximately along the straight line which indicates that the wind direction data set is satisfied towards the wrapped normal distribution.



(a) February



(b) April

Fig. 9 The Quantile Probability-Probability Plot

IV. CONCLUSIONS

In this paper, the discordancy test of the G_a statistics where $a = 1$, 2, and 3 has successfully identified the existence of outliers in circular data via Monte Carlo simulation method. The application of G_a statistics was successful demonstrated the presence of outlier in the wind direction data set. Besides that, by using the simulation approach, the study found that cut-off points and the performance of G_a statistics for wrapped normal distribution is depending on the sample size, n and concentration parameter, ρ. Moreover, in the sample of wrapped normal data, the concentration parameter, ρ is very important. The performance of the test would be much better if the value of the concentration of parameter, ρ is larger. According to Rambli *et al*. [13], the result is expected since the circular data are more concentrated with larger concentration parameter, ρ which resulted in a smaller difference between the two largest values of statistics. The values of cut-off point will decrease if the sample size, n increases. The data should be valid as the sample size, $n$increases and the gap between the circular observations in circular plot become smaller.

Secondly, G_a statistics based on the circular gap were applied. The cut-off values and the performance test were obtained. Monte Carlo discovered that the G_a statistics performed better compared with the other discordancy tests and this test is also easier to be applied and explained by the practitioners. The G_a statistics can be exposed by other circular distributions that have been mentioned in previous chapters. The most important characteristic about G_a statistics is its ability in detecting the outlier not only for a single outlier but more than one and a patch of outliers.

Lastly, there are factors that lead to the outlier occurrences such as human error, systematic error, random error, and environmental factors. In this paper, since the researcher is only focused on the detection of the outlier, this can provide significant information of the outlier to practitioners and meteorological department staff in order to further explore about the occurrences of outliers. For further investigation of the outlier due to the environmental phenomenon, it can be explored by meteorological department staff as the staffs have the significant information about the occurrence of outliers.

ACKNOWLEDGEMENT

The authors would like to gratefully acknowledge to the staff of the Malaysian Meteorological Department, Siti Khadijah Ramli for providing the facilities, knowledge and assistance. We would also like to extend our gratitude to the Universiti Teknologi MARA Research Grants (600-IRMI/PERDANA 5/3 BESTARI (P)(042/2018)) for providing the financial support in this study.

REFERENCES

1. Abuzaid, A. H. (2009). Some problems of outliers in circular data. PhD thesis: Institute of Mathematical Science, University of Malaya.
2. Binkley, S. (1990). The clockwork sparrow: time, clock and calendars in Biological Organisms. Upper Saddle River, NJ: Prentice Hall.
3. Downs, T. (1974). Rotational Angular Correlation. In Biorhythms and Human Reproduction. Ed. M., Ferin, F. Halberg, & L. van der Wiele. New York: Wiley. 97-104.
4. Downs, T. D. & Mardia, K. V. (2002). Circular regression. Biometrika.89(3): 683-697.
5. Downs, T., Leibman, J. & Mackay, W. (1970). Statistical methods for vectorcardiogram orientations. In Proc. XI Int. Vector Cardiography Symposium, Ed. I. Hoffman. Amsterdam: North Holland. 216-222.
6. Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. Technometrics. 11 (1): 1–21.
7. Jammalamadaka, S. R. & SenGupta, A. (2001). Topics in Circular Statistics. World Scientific Press, Singapore.
8. Jammalamadaka, S. R., Bhadra, N., Chaturvedi, D., Kutty, T. K., Majumdar, P. P., & Poduval, G. (1986). Functional assessment of knee and ankle during level walking. In Matusita, K., editor. Data Analysis in Life Science. Indian Statistical Institute, Calcutta, India. 21-54.

9. Maddala, G. S. (1992). Outliers. Introduction to Econometrics (2nd ed.). New York: MacMillan. pp. 88–96.
10. Moore-Ede, M. C., Sulzman, F. M., & Fuller, C. A. (1982). The Clocks that Time Us: Physiology of the Circadian Timing System. Cambridge, MA: Harvard University Press.
11. Mulder, K. &Klugkist, I. (2017). Bayesian estimation and hypothesis tests for a circular Generalized Linear Model. Journal of Mathematical Psychology. 80: 4-14.
12. Pyke, R. (1965). Spacings. Journal of the Royal Statistical Society Series B. 27(3): 395-449.
13. Rambli, A. (2015). A Half-Circular Distribution and Outlier Detection Procedures in Directional Data. PhD. thesis: Faculty of Science, University of Malaya, Kuala Lumpur.
14. 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.
15. Rivest, L. -P. (1997). A decentred predictor for circular–circular regression. Biometrika. 84(3): 717-726.
16. Roy, A., Pal, A., & Garain, U. (2017). JCLMM: A finite mixture model for clustering of circular-linear data and its application to psoriatic plaque segmentation. Pattern Recognition. 66: 160–173.
17. Whittaker, E. T. & Watson, G. N. (1994). A course in Modern Analysis. Cambridge University Press, Cambridge.
18. Yanga, M. S., Chiena, S. J. C., & Hung, W. L. (2017). Learning-based EM clustering for data on the unit hypersphere with application to exoplanet data. Applied Soft Computing. 60: 101–114.