A synthetic data generation procedure for univariate circular data with various outliers scenarios using Python programming language

A synthetic data generation procedure for univariate circular data with various outliers scenarios using Python programming language

N S Zulkipli[1]**, S Z Satari** [1] **and W N S Wan Yusoff**[1]
[1]Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Pahang, Malaysia

Email: syahirahzulkipliwork@gmail.com

Abstract. Synthetic data is artificial data that is created based on the statistical properties of the original data. The aim of this study is to generate a synthetic or simulated data for univariate circular data that follow von Mises (*VM*) distribution with various outliers scenario using Python programming language. The procedure of formulation a synthetic data generation is proposed in this study. The synthetic data is generated from various combinations of seven sample size, n and five concentration parameters, κ. Moreover, a synthetic data will be generated by formulating a data generation procedure with different condition of outliers scenarios. Three outliers scenarios are proposed in this study to introduce the outliers in synthetic dataset by placing them away from inliers at a specific distance. The number of outliers planted in the dataset are fixed with three outliers. The synthetic data is randomly generated by using Python library and package which are 'numpy', 'random' and 'vonMises'. In conclusion, the synthetic data of univariate circular data from von Mises distribution is generated and the outliers are successfully introduced in the dataset with three outliers scenarios using Python. This study will be valuable for those who are interested to study univariate circular data with outliers and choose Python as an analysis tool.

1. Introduction

Data such as daily expenses, sales profit and driving distance are recorded and these kinds of data are known as linear data. However, there is another data type that has a direction or cyclic time which refers to circular data such as wind direction, animals' direction and time arrival of patients at hospital [1]. Many researchers in applied sciences such as [1 - 4] are interested to dig deeper on circular data. Since early 1970, researchers working hard to spread the knowledge of circular data and develop statistical procedures that specialised for circular data. This type of data commonly found in the area of meteorology and biology where researchers are interested to investigate the direction of wind and animals. Recently, [5] has reviewed circular biological data and the authors found that there is huge opportunity to be explored in a biological field especially on the biomedical data that involve circular or angular values. Identifying outlier for univariate circular data is important especially in biomedical research and health informatics. Biomedical research is a research that involve biological, physical science and medical. Recently, circular biomedical data is used in [6] and [7] where the authors conducted a study related with outlier identification. Till date, study on outlier identification for

univariate circular biomedical data is limited. Thus, [5] believe that there is a need to explore further on univariate circular data related to biomedical research especially for the abnormality study.

A synthetic data is an artificial data that imitates real data and it is created based on the statistical properties of the original data. Study that requires a simulation study will imitates the real data to create the synthetic data. Moreover, the limitation in real data especially in biological application such as small sample size motivates data analyst to create synthetic data. In academic field, creating synthetic data is very important because it helps researcher to develop a new statistical procedure and test it performance. Furthermore, synthetic data is mostly used in simulation study process. In addition, synthetic data generation closely related with the investigation of outliers. By creating synthetic data with existence of outliers can helps the researcher to investigate the performance of outlier identification method through simulation study. This has been done by a few studies to analyse the performance of the outliers detection method for circular data such as by [8 - 10]. Besides that, synthetic data is needed in the formulation of outliers scenario where the outliers are placed in the synthetic dataset by shifting the specific parameters.

In this study, Python programming language is used to generate the synthetic data. [11] stated that, the usage of Python language has been soaring since the early 2000s in industrial applications and research, while R programming still a popular language for traditional data analytical procedures. R is frequently used by the researchers to create and generate synthetic circular data since the package for circular statistics which also known as 'circular' and 'CircStats' has been released in 2017 and 2018, respectively. Till date, Python has rapidly developed huge libraries for data science such as 'Numpy', 'Pandas', 'Scipy', 'Matplotlib', 'StatsModel' and 'Seaborn'. These Python libraries extremely popular among data analyst and statisticians. Unfortunately, Python packages such as 'pycircstat' and 'spicy.stats' that specialised for circular statistics are not fully covered all functions for circular statistics. Moreover, [12] found that the existing packages for circular data are limited to a few functions. They have been successfully proposed Python coding for the descriptive analysis of the circular data especially with the existence of outliers. In 2018, special package for von Mises distribution named 'vonMises' is successfully developed which consists of few functions such as random data generation from von Mises distribution and its probability density function.

Thus, in this study we aim to propose the formulation of synthetic data generation with three outliers scenarios which based on the parameters in *VM* distribution by using 'vonMises' package from Python. The procedures of synthetic data generation and the formulation of outliers scenario for univariate circular data that follow *VM* distribution will be discussed in the next section.

2. The Formulation of Synthetic Data and Proposed Outliers Scenario

A synthetic data generation procedure for univariate circular data is formulated in this study with three proposed outliers scenarios. Synthetic data is generated from all combinations of different sample size, n and concentration parameter, κ. According to [13], an "outlier scenario" is refers to the setting of outlying observations relative to the inlying or "clean" observations. To create the synthetic data, the "clean" observations were randomly generated from von Mises distribution $VM(\mu,\kappa)$ with mean direction, μ is fixed with 0. Five values of concentration parameter, $\kappa = 5,10,15,20,100$ with seven sample size, $n = 10,15,20,30,50,100,150$ are considered to generate the synthetic univariate circular dataset. Three outliers will be introduced in the dataset.

In this study, the outliers will be placed in the dataset by shifting the specific parameters which are mean direction, μ and concentration parameter, κ. In the first scenario (Outlier Scenario 1), the outliers were placed away from the inlying observations at specified distance by shifting the mean direction with different degree of contamination, λ in the range of $0 \le \lambda \le 1$ which defined by

$$\bar{\theta}_{out} = \mu + \lambda\pi \quad (1)$$

where mean direction, μ is fixed. Six contamination levels of $\lambda = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0$ will be set to the simulated data. Note that, if $\lambda = 0,$ there is no outliers are placed away from its initial location. For the second scenario (Outlier Scenario 2), the outliers were placed away from the inlying observations at specified distance by the shifting concentration parameter, κ by multiplying κ with six contamination levels of $\lambda = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0.$ The outliers defined as

$$\kappa_{out} = \lambda\kappa \tag{2}$$

where concentration parameter, κ is fixed. Lastly, for the third scenario (Outlier Scenario 3), the outliers were placed away from the inlying observations at specified distance by shifting both conditions in Outlier Scenario 1 and Outlier Scenario 2 simultaneously.

The following are the general steps for the procedure of the synthetic data generation for univariate circular data:

i. The inliers are randomly generated from $VM(\mu, \kappa)$ with mean direction, μ is fixed with 0 and concentration parameter is set at $\kappa = 5, 10, 15, 20, 100.$
ii. Three outliers are randomly generated according to the outliers scenarios formulation in equation (1) and equation (2) where $\lambda = 0.0, 0.2, 0.4, 0.6, 0.8, 1.0.$ The mean direction and concentration parameter, κ are fixed corresponding to condition used in (i).
iii. The inliers and outliers are combined in one dataset.

The outlier scenarios and synthetic datasets design that proposed in this study are summarised in table 1 and table 2, respectively. From the combination of seven sample size, n and five concentration parameters, κ, there will be 35 synthetic datasets generated for each outliers scenario.

Table 1. Summary of synthetic datasets design.

Condition	Level
Sample size, n	10, 15, 20, 30, 50, 100 and 150
Concentration parameter, κ	5, 10, 15, 50 and 100
Outliers planted	3 outliers
Outlier scenario	Scenario 1, Scenario 2 and Scenario 3

Table 2. Summary of outliers scenario.

Outliers Scenario	Condition	Formula
1	Shifting mean, μ	$VM(\mu + \lambda\pi, \kappa)$
2	Shifting concentration parameter, κ	$VM(\mu, \lambda\kappa)$
3	Shifting both in scenario 1 and 2	$VM(\mu + \lambda\pi, \lambda\kappa)$

3. Proposed Python Coding

In this section, the coding for generating a synthetic data with various outliers scenarios is proposed using Python programming language. The proposed coding is summarised in table 3 with the corresponding formula for each outliers scenario. The libraries and package that will be used to generate the synthetic data of univariate circular data are 'numpy', 'random' and 'vonMises'. All listed libraries and package must be imported into Python before running the code. The 'vonMises' package can be installed from https://pypi.org/project/vonMises/.

Table 3. Python coding for synthetic data generation of univariate circular data with outliers scenarios.

Outliers Scenario	Formula	Python Coding

1	Equation (1)	`def sce1_function(n,mu,kappa,lambda,out):`
		`  x=vm.rvonmises(n-out,mu,kappa)`
		`  xo=vm.rvonmises(out,mu+(lambda *np.pi),kappa)`
		`  xall=np.concatenate([x,xo])`
		`sce1_function(n,mu,kappa,lambda,out)`

2	Equation (2)	`def sce2_function(n,mu,kappa,lambda,out):`
		`  x=vm.rvonmises(n-out,mu,kappa)`
		`  xo=vm.rvonmises(out,mu,lambda*kappa)`
		`  xall=np.concatenate([x,xo])`
		`sce2_function(n,mu,kappa,lambda,out)`

3	Equation (1) & (2)	`def sce3_function(n,mu,kappa,lambda,out):`
		`  x=vm.rvonmises(n-out,mu,kappa)`
		`  xo=vm.rvonmises(out,mu+lambda *np.pi,lambda*kappa)`
		`  xall=np.concatenate([x,xo])`
		`sce3_function(n,mu,kappa,lambda,out)`

4. Results and Discussion

As an illustration, few datasets are generated from combinations of different value of sample size, n, concentration parameter, κ and contamination level, λ which summarised in table 4.

Table 4. Dataset to illustrate the generation of synthetic data for univariate circular data.

Dataset	Combinations of n, κ, λ		
	n	κ	λ
Dataset 1	10	5	0.2
Dataset 2	100	5	0.2
Dataset 3	10	100	0.8
Dataset 4	100	100	0.8

Suppose four datasets are randomly generated from different combinations of n, κ and λ. Dataset 1 and Dataset 2 are generated from $VM(0,5)$ with $\lambda = 0.2$ where $n = 10$ and $n = 100$, respectively. Dataset 3 and Dataset 4 are randomly generated from $VM(0,100)$ with $\lambda = 0.8$ where $n = 10$ and $n = 100$, respectively. In this study, Dataset 1 and Dataset 2 are considered as dataset that have small κ and small λ, while, Dataset 3 and Dataset 4 are considered as dataset that have large κ and large λ regardless the sample size. The following subsections are the result and discussion of the synthetic datasets in table 4 generated for univariate circular data with three different outliers scenarios using Python.

4.1. Outlier Scenario 1

The formulation of Outlier Scenario 1 is defined in equation (1). To illustrate the Outlier Scenario 1, the outliers are planted in the dataset using value of $\lambda = 0.2$ and 0.8 with concentration parameter, $\kappa = 5$

and $\kappa = 100$ which considered as small and large concentration parameters, respectively. There are four datasets that have been generated to illustrate the generation of synthetic data with Outlier Scenario 1 and visualised in circular plot. Three outliers have been successfully introduced in the datasets by placing away the outliers from the mean direction, $\mu = 0$ according to the chosen level of contamination.

By plotting circular plot in figure 1 - figure 4, the three outliers are placed away from its initial location. Figure 1 and figure 2 are the circular plot for Dataset 1 and Dataset 2 which data are generated with $\kappa = 5$ and $\lambda = 0.2.$ Meanwhile, Dataset 3 and Dataset 4 are generated with $\kappa = 100$ and $\lambda = 0.8$ which are plotted in figure 3 and figure 4, respectively. In this study, $\kappa = 5$ is considered as small and it cause the data less concentrated towards the centre compared to the data with $\kappa = 100$ which consider large concentration parameter. Regardless the sample size, Dataset 3 and Dataset 4 have very concentrated data which the data are concentrated towards the centre, $\mu = 0.$

Moreover, the outliers are planted far away from the inliers when the contamination level, $\lambda = 0.8$ is used compared to the dataset that use $\lambda = 0.2.$ In addition, sample size does affect the location of planted outliers. As shown in figure 1, the three outliers are planted far from each other and located at different quadrant when Dataset 1 is generated from small sample size, $n = 10$ with small κ and λ values. Figure 2 shows that, the three outliers are planted close to each other in quadrant 1 when Dataset 2 is generated from large sample size, $n = 100$ with small κ and λ compared to Dataset 1. However, when κ and λ are large, a small or large sample size does not affect the location of the planted outliers. Figure 3 and figure 4 shows that the three outliers are planted close to each other and successfully located far from the inliers.

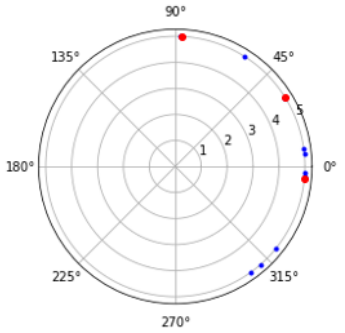


Figure 1. Circular dot plot of Dataset 1 for Outlier Scenario 1.

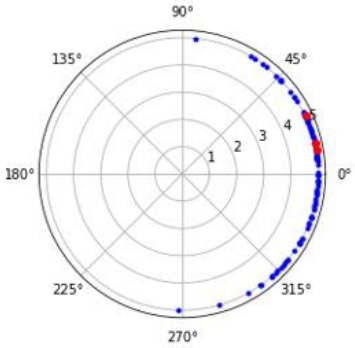
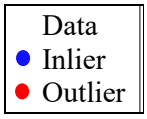


Figure 2. Circular dot plot of Dataset 2 for Outlier Scenario 1.



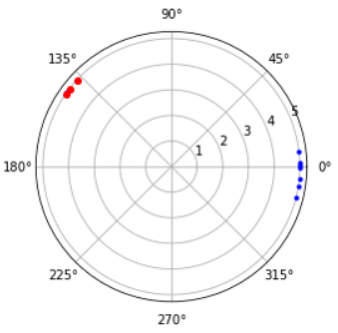


Figure 3. Circular dot plot of Dataset 3 for Outlier Scenario 1.

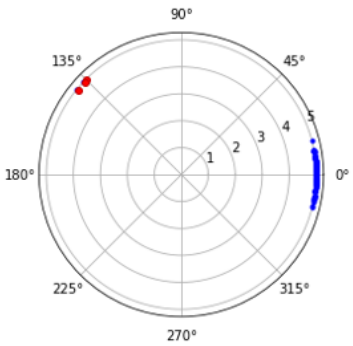


Figure 4. Circular dot plot of Dataset 4 for Outlier Scenario 1.

4.2. Outlier Scenario 2

The formulation of Outlier Scenario 2 formulated in equation (2). The following are the result and explanation of the synthetic dataset generated for univariate circular data with Outlier Scenario 2 using datasets design in table 4. The synthetic datasets with Outlier Scenario 2 are successfully generated and visualised in figure 5 – figure 8. Figure 5 and figure 6 are the circular plot for Dataset 1 and Dataset 2 which data are generated with $\kappa = 5$ and $\lambda = 0.2$ where $n = 10$ and $n = 100,$ respectively. Meanwhile, Dataset 3 and Dataset 4 are generated with $\kappa = 100$ and $\lambda = 0.8$ which are plotted in figure 7 and figure 8, where $n = 10$ and $n = 100,$ respectively. By plotting circular plots in figure 5 – figure 8, the three outliers are placed away from its initial location by shifting their concentration parameters.

Figure 5 and figure 6 show that, the data are well spread or less concentrated since κ is small. Dataset 1 in figure 5 shows the three outliers are spreading far from each other and located at different quadrant when the λ is small for small $n.$ However, Dataset 2 in figure 6 shows the three outliers are located close to each other in the same quadrant for large n when λ is small. Due to small value of $\lambda,$ we found that, the outliers planted in Dataset 1 and Dataset 2 are consistent with the inliers. From the observation in figure 7 and figure 8, Dataset 3 and Dataset 4 have very concentrated data which the data are concentrated towards the centre, $\mu = 0$ since the data are generated with large $\kappa.$ It is found that, when κ and λ are large, a small or large sample size does not affect the location of planted outliers. Figure 7 and figure 8 shows the three outliers are planted close to each other. However, the outliers are planted consistent with the inliers when the outliers are introduced by shifting the value of concentration parameter, κ for both conditions, $\lambda = 0.2$ and $0.8.$ The outcome is contradicted with the outcome in Outlier Scenario 1 because we use different outlier scenario formulation.

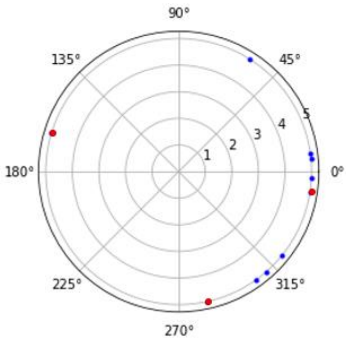


Figure 5. Circular dot plot of Dataset 1 for Outlier Scenario 2.

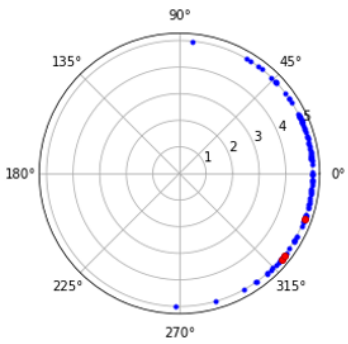
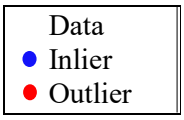


Figure 6. Circular dot plot of Dataset 2 for Outlier Scenario 2.



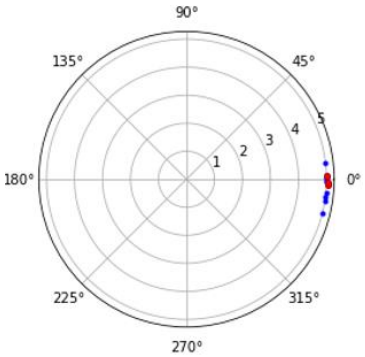


Figure 7. Circular dot plot of Dataset 3 for Outlier Scenario 2.

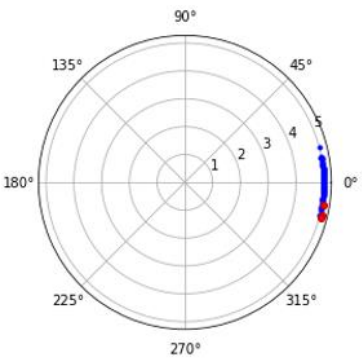


Figure 8. Circular dot plot of Dataset 4 for Outlier Scenario 2.

4.3. Outlier Scenario 3

The third Outlier Scenario is created by using both conditions in Outlier Scenario 1 and Outlier Scenario 2. It means that the formulation of Outlier Scenario 3 is formulated by applying both equation (1) and equation (2) to calculate the specific distance of outliers. The following are the result and explanation of the synthetic dataset generated for univariate circular data with Outlier Scenario 3. Four datasets in table 4 have been generated and visualised in figure 9 – figure 12 to illustrate the generation of synthetic data with Outlier Scenario 3. Figure 9 and figure 10 are the circular plot for Dataset 1 and Dataset 2 which data are generated with $\kappa = 5$ and $\lambda = 0.2$ where $n = 10$ and $n = 100$, respectively. Meanwhile, Dataset 3 and Dataset 4 are generated with $\kappa = 100$ and $\lambda = 0.8$ which are plotted in figure 11 and figure 12, where $n = 10$ and $n = 100$, respectively. By plotting circular plot in figure 9 – figure 12, the three outliers have been successfully introduced in the datasets by shifting both values of μ and κ.

In figure 9 and figure 10, the data are well spread or less concentrated since κ is small. Dataset 1 in Figure 9 shows the three outliers are spreading far from each other and located at different quadrant when the λ is small for small n. In contrary, Dataset 2 in figure 10 shows the three outliers are located close to each other in the same quadrant for large n when λ is small. It is found that, the outliers planted in Dataset 1 and Dataset 2 are consistent with the inliers due to small value of λ. From the observations in figure 11 and figure 12, Dataset 3 and Dataset 4 have a very concentrated data which the data are concentrated towards the centre, $\mu = 0$ since the data are generated with large κ. From figure 11 and figure 12, it can be seen that, when κ and λ are large, a small or large sample size does not affect the location of planted outliers. Figure 11 and figure 12 show that the three outliers are planted close to each other and successfully located far away from the inliers.

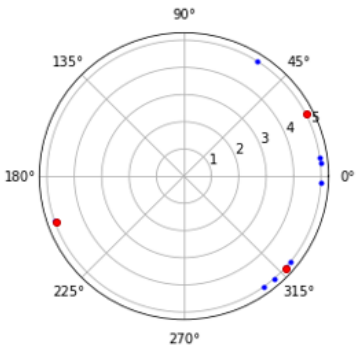


Figure 9. Circular dot plot of Dataset 1 for Outlier Scenario 3.

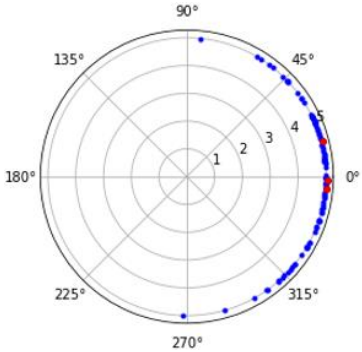
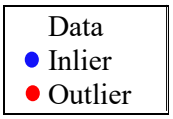


Figure 10. Circular dot plot of Dataset 2 for Outlier Scenario 3.



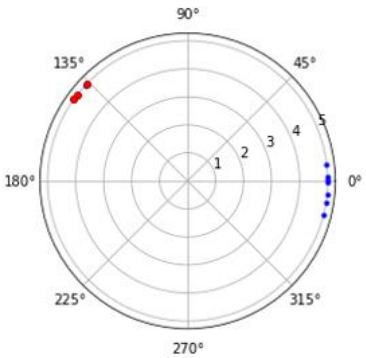


Figure 11. Circular dot plot of Dataset 3 for Outlier Scenario 3.

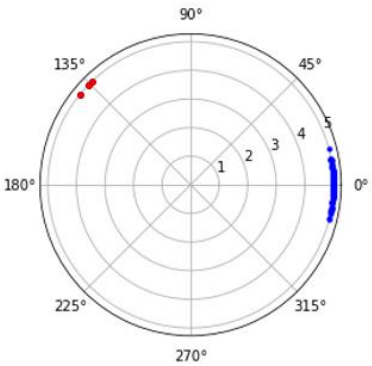


Figure 12. Circular dot plot of Dataset 4 for Outlier Scenario 3.

To simplify, Outlier Scenario 3 is applying both Outlier Scenario 1 and Outlier Scenario 2. Table 5 concludes the findings from the previous section. Generally, the results for Outlier Scenario 3 approximately similar with the results in Outlier Scenario 1 because we use the same outlier scenario formulation by shifting the mean direction together with the concentration parameter.

Table 5. Findings from three outliers scenarios.

Outlier Scenario		Sample size, n	
		Small	Large
1	$\kappa = 5$ $\lambda = 0.5$	Outliers planted far from each other and located consistent with inliers	Outliers planted close to each other and located consistent with inliers
	$\kappa = 100$ $\lambda = 0.8$	Outliers planted close to each other and located far away from inliers	Outliers planted close to each other and located far away from inliers
2	$\kappa = 5$ $\lambda = 0.5$	Outliers planted far from each other and located consistent with inliers	Outliers planted close to each other and located consistent with inliers
	$\kappa = 100$ $\lambda = 0.8$	Outliers planted close to each other and located consistent with inliers	Outliers planted close to each other and located consistent with inliers
3	$\kappa = 5$ $\lambda = 0.5$	Outliers planted far from each other and located consistent with inliers	Outliers planted close to each other and located consistent with inliers
	$\kappa = 100$ $\lambda = 0.8$	Outliers planted close to each other and located far away from inliers	Outliers planted close to each other and located far away from inliers

5. Conclusion

In conclusion, we successfully proposed a synthetic data generation procedure for univariate circular data with three outliers scenarios using Python. A synthetic data is generated from all combinations of different sample size, n and concentration parameter, κ with three different outliers scenarios. Generally, the outcome from Outlier Scenario 1 approximately the same with outcome from Outlier Scenario 3 for any conditions. However, the outcome from Outlier Scenario 2 is contradicted with Outlier Scenario 1 and Outlier Scenario 3 when the sample size is large. A synthetic data with various outliers scenarios is very crucial to generate because it will be useful for the development of a new procedure of outlier detection method for univariate circular data. At the same time, the synthetic circular data can be used to mimics the real data in many applications. Hence, the characteristics of the real data can be investigated further for improvement and observations.

Acknowledgments

Authors would like to thank all the associate editors and referees for their thorough reading and valuable suggestions which led to the improvement of this paper. The Ministry of Higher Education Malaysia and Universiti Malaysia Pahang are acknowledged for the financial support received for this study. (FRGS/1/2019/STG06/UMP/02/6 and UMP Internal Grant: RDU1901168 and RDU190363).

References

[1] Jammalamadaka S R and Sengupta A 2001 *Topics in Circular Statistics* (World Scientific Publishing Co. Pte. Ltd. P.)
[2] Fisher N I 1993 *Statistical Analysis in Circular Data* (New York, USA: Cambridge University Press)
[3] Mardia K V 1975 Statistics of directional data *Journal of the Royal Statistical Society* B **37** pp 349–393
[4] Best D J and Fisher N 1981 The bias of the maximum likelihood estimators of the von Mises-Fisher concentration parameters *Communication in Statistics- Simulation and Computation* **10**(5) pp 493–502
[5] Satari S Z and Ku Khalif K M N 2020 Review on outliers identification methods for univariate circular biological data *Advances in Science, Technology and Engineering Systems* **5**(2) pp 95–103
[6] Mohamed I B, Rambli A, Khaliddin N, and Ibrahim A I N 2016 A New Discordancy Test in Circular Data Using Spacings Theory *Communications in Statistics - Simulation and Computation* **45**(8) pp 2904–16
[7] Alkasadi N A, Abuzaid A H M, Ibrahim S and Yusoff M I 2018 Outliers Detection in Multiple Circular Regression Model via DFBETAc Statistic *International Journal of Applied Engineering Research* **13**(11) pp 9083–90
[8] Mokhtar N A, Zubairi, Y Z and Hussin A G 2017 A clustering approach to detect multiple outliers in linear functional relationship model for circular data *Journal of Applied Statistics* **45**(6) pp 1041–51
[9] Satari S Z, Di N F M and Zakaria R 2019 Single-linkage method to detect multiple outliers with different outlier scenarios in circular regression model *AIP Conference Proceedings* **2059**
[10] Di N F M, Satari S Z, and Zakaria R 2019 Outlier detection in circular regression model using minimum spanning tree method *Journal of Physics: Conference Series* **1366**(1)
[11] Philip Chen C L and Zhang C Y 2014 Data-intensive applications, challenges, techniques and technologies: A survey on Big Data *Information Sciences* **275** pp 314–347
[12] Zulkipli N S, Satari S Z and Wan Yusoff W N S 2020 Descriptive analysis of circular data with outliers using Python programming language *Data Analytics and Applied Mathematics (DAAM)* **1**(1) pp 31–36
[13] Sebert D M, Montgomery D C and Rollier D A 1998 A clustering algorithm for identifying multiple outliers in linear regression *Computational Statistics and Data Analysis* **27**(4) pp 461–484