Sains Malaysiana 50(6)(2021): 1787-1798
http://doi.org/10.17576/jsm-2021-5006-24

Comparative Study of Clustering-Based Outliers Detection Methods in Circular-Circular Regression Model

(Kajian Perbandingan Kaedah Penetapan Titik Terpencil Berasaskan Kelompok dalam Model Pendaftaran Lingkaran)

SITI ZANARIAH SATARI*, NUR FARAIDAH MUHAMMAD DI*, YONG ZULINA ZUBAIRI & ABDUL GHAPOR HUSSIN

ABSTRACT

This paper is a comparative study of several algorithms for detecting multiple outliers in circular-circular regression model based on the clustering algorithms. Three measures of similarity based on the circular distance were used to obtain a cluster tree using the agglomerative hierarchical methods. A stopping rule for the cluster tree based on the mean direction and circular standard deviation of the tree height was used as the cutoff point and classifier to the cluster group that exceeded the stopping rule as potential outliers. The performances of the algorithms have been demonstrated using the simulation studies that consider several outlier scenarios with a certain degree of contamination. Application to real data using wind data and a simulated data set are given for illustrative purposes. Thus, it has been found that Satari's algorithm (S-SL algorithm) performs well for any values of sample size n and error concentration parameter. The algorithms are good in identifying outliers which are not limited to one or few outliers only, but the presence of multiple outliers at one time.

Keywords: Circular distance; circular-circular regression model; clustering; outliers; stopping rule

ABSTRAK

Kertas ini membincangkan kajian perbandingan beberapa algoritma yang mengesan titik terpencil berganda dalam model regresi bulatan berdasarkan algoritma berkelompok. Tiga ukuran persamaan berasaskan jarak bulatan telah digunakan bagi mendapatkan pokok kelompok menggunakan algoritma aglomeratif hierarki. Satu nilai potongan untuk pokok kelompok berdasarkan min terarah dan sisihan piawai bulatan bagi ketinggian pokok tersebut telah digunakan bagi mengkelaskan kumpulan kelompok yang melebihi titik potongan ini sebagai titik terpencil. Prestasi algoritma ini telah diuji dalam kajian simulasi yang mengambil kira beberapa senario titik terpencil dengan tahap berbeza. Untuk tujuan illustrasi, satu aplikasi data sebenar menggunakan data angin dan satu set data simulasi telah diberikan. Kami mendapati algoritma Satari (Algoritma S-SL) adalah baik untuk sebarang nilai saiz sampel dan parameter menumpu. Algoritma tersebut adalah baik dalam mengenal pasti titik terpencil atau berganda pada satu masa.

Kata kunci: Algoritma berkelompok; jarak bulatan; model regresi bulatan; nilai potongan; titik terpencil

INTRODUCTION

In most model buildings, the presence of influential observation, outliers and missing values cannot be ignored. Outliers frequently occur in real data and may be unnoticeable. If outliers in circular regression model remain undetected, they can lead to erroneous parameter estimations and inferences. Outlier detection in circular-circular regression model may be performed using graphical and numerical techniques. A number of graphical approaches and discordance tests have been proposed over the year.

The first published work on the identification of outlier in circular-circular regression model can only be found in Abuzaid et al. (2009). Several outlier detection methods such as circular plot, P-P plot, Q-Q plot, D-Statistic,

C-Statistic and M-Statistic were performed for Hussin et al. (2004) circular regression model. The D-Statistic, C-Statistic and M-Statistic are some discordance tests for von Mises data as proposed by Collet (1980). As for the complex regression model, Hussin et al. (2013) introduced similar identification techniques to detect outliers in Hussin (2010)'s model. Also in 2012, Abuzaid et al. (2012a) proposed a boxplot version for a circular data set and identified outliers for Hussin et al. (2004) circular-circular regression model. Later, Abuzaid et al. (2012b) introduced a new discordance test called P_j -Statistic that performed well in Down and Mardia (2002) circular-circular-circular regression model.

The effect of the row deletion approach is also oftenly considered in detecting outliers for circular regression models. In 2010, Abuzaid introduced a statistic called the Difference Mean Circular Error Statistic (DMCEs). Rambli et al. (2010) considered this statistic to identify influential observations in Down and Mardia (2002) circular-circular-circular regression model. Likewise, in 2013, Abuzaid et al. modified the DMCEs statistic and proposed another DMCE. The cutoff points and the performance of DMCE procedure for Hussin et al. (2004) were obtained *via* simulation studies. Ibrahim (2013) also gave the cutoff points and the performance of DMCE procedure for Sarma and Jammaladaka (1993) circular regression model. In 2016, Rambli et al. used a trigonometric function by transforming the circular residuals into linear measures and employed DMCEs statistic to identify outlier for linear data.

Outlier detection procedures based on COVRATIO statistic are also widely used for different circular regression models. Influence observation detection procedure based on $\left|COVRATIO_{(-i)}-1\right|$ statistic was introduced by Hussin et al. (2010) in Caires and Wyatt (2003) functional relationship model. Later, Rambli (2011) conducted a similar procedure for Down and Mardia (2002) model. Abuzaid et al. (2011) explored the accessibility of this procedure in Hussin et al. (2004). Hussin and Abuzaid (2012) also studied the performance of this procedure in a new complex linear functional relationship model.

Later on, Ibrahim et al. (2013) extended the usage of COVRATIO statistic to Sarma and Jammaladaka (1993) circular regression model. Each researcher obtained the cut-off points and the performance of the procedure *via* simulation studies. It was shown that COVRATIO statistic has higher power of performance and performs well in detecting influential observations (Ibrahim et al. 2013). In 2015, Rambli et al. extended the COVRATIO statistic to identify outliers by examining the effect of the outliers on the covariance matrix. The COVRATIO statistic shows better performance for large sample size with high concentration parameter. Then, Alkasadi et al. (2016) extended COVRATIO statistic in detecting the outliers for multiple circular regression model.

Based on the referred literature, it can be inferred that the row deletion technique is the well known procedure for outlier detection in circular regression models. This method, however, focuses on identifying a single outlier at a time. Hence, alternative approach such as clustering algorithm has been introduced to detect outliers for circular data which can be used to identify outliers not limited to one or few outliers only but the presence of multiple outliers at one time.

In 2012, Chang-Chien et al. proposed a Mean Shift-based Clustering (MSBC) method to detect outliers in circular data that are based on the single-linkage method. This method utilised circular Euclidean distance (Chang-chien's distance) to cluster the data and detect multiple outliers simultaneously. Later, Satari (2015) proposed a new clustering algorithm called Satari's algorithm (S-SL algorithm) using circular City-block distance (namely Satari's distance) and single-linkage method to detect multiple outliers in Down and Mardia (2002) circular-circular-circular regression model. Satari's algorithm performed very well in detecting outliers in *v*-space data. Noted that, *u* and *v* are fixed independent angle and the dependent random angle, respectively. Then, Di and Satari (2017) modified S-SL algorithm by proposing new circular distance called Di's distance and a new single-linkage algorithm known as D-SL algorithm. D-SL algorithm managed to detect multiple outliers in most conditions investigated for sample sizes of $n = 30$ and $n = 100$ with a concentration parameter of 5 and 10.

To investigate the effect of different outlier scenarios in circular-circular regression model, Di et al. (2017) compared D-SL algorithm and S-SL algorithm with two outlier scenarios which were outliers in *v*-space and *u*-space outlier. Based on the findings, S-SL algorithm performs better in detecting outliers in *v*-space and *u*-space outlier's scenarios. Later, Satari et al. (2017) modified S-SL algorithm by using average linkage with Satari's distance (S-AL algorithm). Simulation study was performed with a sample size of 100 and concentration parameter of 20. However, the result has showed that this method is more sensitive as it can falsely detect clean observation as outliers.

In this study, a comparative study of clustering-based outliers detection methods in circular-circular-circular

regression model is conducted. Several algorithms, including S-SL algorithm and D-SL algorithms were compared to display the applicability of clustering-based algorithm in detecting outlier for Down and Mardia circular-circular-circular regression model. All the algorithms used in this study were cluster-based algorithms using single-linkage, average linkage, and complete linkage method with circular Euclidean distance and circular City-block distance as a similarity measure. Then, three outlier scenarios were used in this study; *u*-space, *v*-space and *uv* spaces outliers. The methods used in this study is summarised in Figure 1.

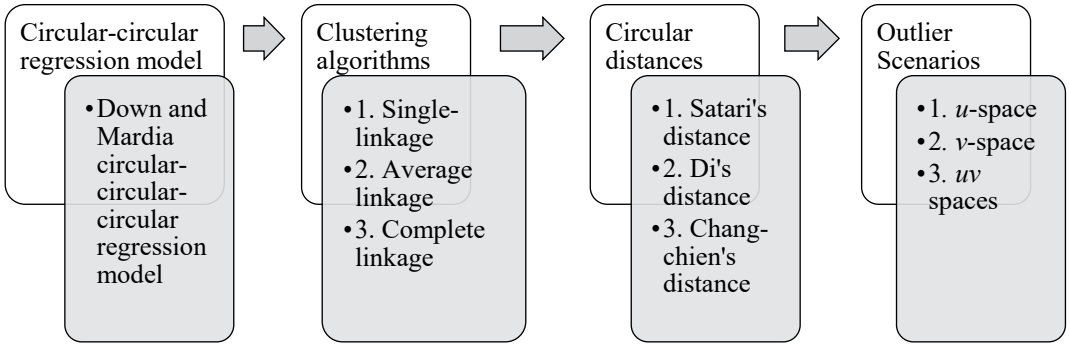


FIGURE 1. The overview of the research method

The performance of the proposed clustering algorithm was assessed *via* simulation studies. Wind data application and a simulated data set have been considered as illustrative examples. Simulation study and illustrative example were used to demonstrate the applicability of the clustering-based methods in various conditions with different outliers scenarios, concentration parameters, sample sizes and degree of contaminations. In short, nine clustering algorithms were used for this comparative study as displayed in Table 1.

TABLE 1. Names of clustering-based algorithms

Circular distance	Clustering algorithms	Names
Di's distance	Single-linkage	D-SL
	Average Linkage	D-AL
	Complete Linkage	D-CL
Satari's distance	Single-linkage	S-SL
	Average Linkage	S-AL
	Complete Linkage	S-CL
Chang-chien's distance	Single-linkage	C-SL
	Average Linkage	C-AL
	Complete Linkage	C-CL

DOWN AND MARDIA CIRCULAR-CIRCULAR-CIRCULAR REGRESSION MODEL

In 2002, Down and Mardia introduced a one-to-one correspondence between independent angle and the mean of dependent angle that employed the Möbius circle mapping for complex variables. The model is given by (1)

$$\tan\frac{1}{2}(v-\beta)=\omega\tan\frac{1}{2}(u-\alpha), \tag{1}$$

where ν is the dependent random angle; u is the fixed independent angle; ω is a slope parameter in the closed interval $[-1,1]$; and α and β are angular location parameters, with the following unique solution (2)

$$v=\beta+2\tan^{-1}\left\{\omega\tan\frac{1}{2}(u-\alpha)\right\}. \tag{2}$$

Parameters (α, β, ω) in the model were estimated using the maximum likelihood estimator method.

CLUSTERING ALGORITHM FOR OUTLIERS DETECTION

Data clustering is a method of creating groups of objects or clusters, in such a way that objects in one cluster are very similar and objects in different clusters are quite distinct (Gan et al. 2007). Clustering methods are sensitive to outliers (Chang-Chien et al. 2012; Sebert et al. 1998). There are three primary decisions the analyst has to make before clustering the multivariate data (Sebert et al. 1998). Firstly, one must decide on what point or variables to use, secondly, the measure of similarity to use, and lastly, the clustering algorithm or method to use. The most common similarity measure for linear variables is the measure of distance, such as Euclidean, Manhattan, Minkowski and Mahalanobis distances (Gan et al. 2007).

As for the clustering method, the most popular one in practice is the hierarchical agglomerative method (Sebert et al. 1998). This method operates on the similarity matrix to construct a tree, depicting specified relationships among entities and produce non-overlapping clusters. Generally, there are four major clustering algorithms for hierarchical agglomerative method, namely linkage (single, complete, average), centroid, median and Wards (Gan et al. 2007).

SIMILARITY MEASURE FOR CIRCULAR VARIABLES

In order to group the variables into their natural groupings, it is necessary to have a measure of 'closeness' or 'similarity' or a measure of dissimilarity between the variables. Measure of distance is commonly used to compute the similarity measure between observation i and j. In this study, we compared three measures of similarity for circular data namely Di's distance (Di & Satari 2017), Satari's distance (Satari 2014) and Chang-chien's distance (Chang-chien et al. 2012). These distances are derived from two circular distances proposed by Jammalamadaka and SenGupta (2001) as in Equations (3) and (4), respectively.

$$d_{ij}=\pi-|\pi-|\theta_{ik}-\theta_{jk}||, \tag{3}$$

where d_{ij} is the distance between observation i and j, θ_{ik} is the value of the kth variable for the ith observation and θ_{jk} is the value of the kth variable for the jth observation where i = 1, 2,…, d and j = 1, 2,…, d.

$$d_0=(\alpha,\beta)=(1-\cos(\alpha-\beta)) \tag{4}$$

where α and β represent the angles corresponding to the distance of two points.

Satari's distance utilised the city-block distance based on circular distance in (3), given as (5),

$$d_{ij(Satari)}=\sum_{k=1}^{p}\left(\pi-\left|\pi-\left|\theta_{ik}-\theta_{jk}\right|\right|\right). \tag{5}$$

whereas, Di's algorithm utilised circular distance that is based on Euclidean distance as follows (6),

$$d_{ij(Di)}=\sqrt{\sum_{k=1}^{p}\left(\pi-\left|\pi-\left|\theta_{ik}-\theta_{jk}\right|\right|\right)^2}. \tag{6}$$

Meanwhile, the circular distance proposed by Chang-chien et al. (2012) used circular distance in (4) with Euclidean distance given as (7):

$$d_{ij(Chang-chien)}=\sqrt{\sum_{k=1}^{p}(1-\cos(\theta_{ik}-\theta_{jk}))^2}. \tag{7}$$

For (5) - (7), d_{ij} is the distance between i and j; p is the number of variables; and θ_{ik} is the value of kth variable for the ith observation and θ_{jk} is the value of the kth variable for the jth observation where i = 1, 2,…, d and j = 1, 2,…, d.

STOPPING RULE FOR OUTLIERS DETECTION

After a clustering algorithm is used on a data set, the user, usually, decides the number of groups (if any) in the data set. Specifically, the cluster tree must be portioned or 'cut' at a certain height. The number of groups depends upon where the tree is cut (Sebert et al. 1998). A comprehensive discussion on the different types of stopping rules can be found in Milligan and Cooper (1985). Most of these

stopping rules have difficulty in a two-cluster scenario that is; it seems that a two-cluster case is the most difficult structure for the stopping rules to detect.

The Mojena's stopping rule is widely used for linear variables and has been the subject of some limited validation research (Blashfield & Morey 1980). Formally, Mojena's stopping rule or 'cut height' is, $\overline{h} + \alpha s_h$, where $\overline{h}$ is the average heights for all $N-1$ clusters, s_h is the unbiased standard deviation of the heights, and α is a specified constant. Mojena (1977) suggested that α should be specified in the range of 2.75-3.50. In this study, we adopted the stopping rule by Mojena by redefining the mean and standard deviation of circular variables that follow the von Mises distribution. This technique was proposed by Satari (2015). A von Mises $VM(\mu,\kappa)$ random variable can be centred at mean direction μ but not rescaled to have unit spread (Fisher 1995). In fact, Fisher (1995) stressed out that there was no circular distribution available with an associated measure of spread, which can be rescaled to have one. Hence, it is an error to measure the spread or range of circular data set from its mean in the formulation of stopping rule.

It is also worthwhile to note that Fisher (1985) suggested a useful way to define or interpret the circular standard deviation of a von Mises distribution from the fact that a fixed probability interval can be calculated around the mean direction μ by assuming that we want an interval around μ containing a specified percentage P% of the von Mises distribution of the form $\mu \pm \theta_P$. Also to relate θ_P with a circular standard deviation σ and let λ_P be a number such that (8)

$$\text{Prob}\left(\mu - \lambda_P\sigma \le \Theta \le \mu + \lambda_P\sigma\right) \simeq P \tag{8}$$

where $\sigma = \sqrt{-2\log\rho}$ is the circular standard deviation for von Mises distribution and ρ is the mean resultant length of the distribution.

The multipliers λ_P are displayed for a selection of possibly P-values. Some appropriate values of λ_P and their ranges of validity in terms of κ, are (9) and (10):

$$P = 0.90,\ \lambda_P = 1.69,\ \kappa \ge 0.65\ (\rho \ge 0.31), \tag{9}$$

$$P = 0.95,\ \lambda_P = 2.06,\ \kappa \ge 0.80\ (\rho \ge 0.37). \tag{10}$$

Practically, we may say that at certain significance level, the circular mean direction μ is situated within a specified range of the circular standard deviation σ. As an example, at significance level 0.05, the circular mean direction μ is situated within $\pm 2.06\sigma$. This finding is reasonable since the von Mises distribution is always symmetric about the mean direction μ.

For the purpose, Satari (2015) chose a specific multiplier $\lambda_P = 2.06$ as the main focus. Consequently, we may say at 95% of confidence level that the cluster group that exceeds the stopping rule (11)

$$\overline{h} + 2.06 s_h \tag{11}$$

is classified as potential outliers where $\overline{h}$ is the average of the heights for all $N-1$ clusters, $s_h = \sqrt{-2\log \overline{R}_h}$ is the circular standard deviation of the heights, and $\overline{R}_h$ is the mean resultant length of the heights.

CLUSTERING ALGORITHMS FOR MULTIPLE OUTLIERS DETECTION IN CIRCULAR-CIRCULAR-CIRCULAR REGRESSION MODEL

In this study, nine clustering methods based on the agglomerative hierarchical method to cluster the points in the predicted values versus residual values plot of bivariate circular variables are considered. This methods are motivated based on the clustering algorithm proposed by Sebert et al. (1998), an approach for identifying a reasonable subset of potential outliers using a single linkage clustering algorithm with the Euclidean distances for the standardised predicted and standardised residual values from a least squares fit for a linear regression model. It is well known that residual plotted against the corresponding predicted values is a useful tool to judge the adequacy of a regression model and also to identify the presence of outliers. Data sets with no outliers will have a linear relationship that can be seen in the plot of residual and predicted values.

Our particular interest is to see how the clustering algorithms work in the circular environment, specifically, for identifying multiple outliers in the Down and Mardia (2002) circular-circular-circular regression model. In summary, the flow chart of comparative study is outlined as follows:

Step 1: Obtain the predicted (*i*) and residual (*j*) values from Down and Mardia circular-circular-circular regression model. Step 2: Obtain the similarity distance between pair of *i* and *j* values from Step 1 by using Di's distance, Satari's distance and Chang-chien's distance. Step 3: Cluster the observation using agglomerative hierarchical clustering algorithms and obtain the cluster tree. Step 4: Compare the analysis of different outlier scenarios. Step 5: Measure the performance of each clustering method in terms of outlier detection using 'success' probability (*pout*), masking effect (*pmask*) and swamping effect (*pswamp*).

Step 6: Results and conclusion.

With a 95% confidence, the cluster group that exceeds the stopping rule may be classified as potential outliers. Usually, the cluster groups with largest observations are considered as inliers since the groups are always situated below the stopping rule while all other observations in cluster groups with minority observations are considered as outliers. We also investigated the power of performance of the proposed algorithm *via* simulation study.

PERFORMANCE INDICATORS

Two fundamental problems with multiple outliers' detection techniques are masking and swamping error. Masking appears if an outlier is not detected or missed throughout the detection procedure. On the other hand, if swamping occurs, a 'clean' or inlying observation is identified as outlier even if they are not influential. Masking is a more serious problem than swamping. When an outlier is missed because of masking, the outlier, if influential, can degrade the performance of the regression model (Hartigan 1975).

For that reason, the performance of clustering algorithms has been investigated by using the power of performance that calculated the 'success' probability together with the masking and swamping errors which occurred *via* simulation study. Similar process was also used by Adnan and Mohamad (2003) and Sebert et al. (1998) for linear cases. 'Success' means that the method had successfully identified all of the outlying observations (no masking occurred). If the method is successful but also includes inlying observations in the candidate set of outliers (swamping occurs), this will be noted as 'false alarm'.

The probability of planted outliers, which was correctly detected, is (12)

$$pout = \frac{"success"}{s} \tag{12}$$

where *"success"* is the number of data set that the method had successfully identified out of all the planted outlying observations. The probability of planted outliers is falsely detected as inliers is (13)

$$pmask = \frac{"failure"}{(out)(s)}, \tag{13}$$

where *"failure"* is the number of outliers in all data set detected as inliers. Also, the probability of clean observations detected as outliers is (14):

$$pswamp = \frac{"false"}{(n-out)s}, \tag{14}$$

where *"false"* is the number of inliers in all data set detected as outliers.

SIMULATION STUDY

Simulation study was done using SPlus statistical package. The sample sizes used were n = 30, 50, 100, and 120, respectively, for the independent circular variable u and circular error e from von Mises distribution. The values of u were chosen from $VM\left(\pi/2, 2\right)$ and assumed to be fixed. For circular error e, the values were chosen from $VM\left(0, \kappa\right)$ with a set of concentration parameters given by $\kappa = 5$, 10, 15, and 20, respectively. Then, from the generated random samples u and e, we calculated the values of the response variable v using the Down and Mardia (2002) circular-circular-circular regression model as given in equation (1) with fixed values of $\alpha = 1.5$, $\beta = 1.5$, and $\omega = 0.5$.

All algorithms were investigated for three outlier scenarios: v-space outlier, u-space outlier and uv spaces outliers. For outlier in v-space, at point [d] of the response variable v, the observation v [d] is contaminated as (15)

$$v^*[d] = v[d] + \lambda\pi \tag{15}$$

where v^* [d] is the contaminated observation at position [d] and λ is the degree of contamination in the range of $0 \le \lambda \le 1$. In other word, this process allows the outliers (contaminated observation) to be placed at a specific distance away from the inliers in the response variable v-space. Hence, for the outlier in u-space, the observation u [d] is contaminated as (16),

$$u^*[d] = u[d] + \lambda\pi \tag{16}$$

For uv spaces outliers, Equations (15) and (16) were run simultaneously to produce outliers at u and v spaces. Subsequently, upon fitting the DM regression model on the simulated data, we managed to obtain the predicted (fitted) values $\hat{v}$ and fitted error $\hat{e} = \hat{v} - v$. In this study, three outliers were randomly planted, $[d_1, d_2, d_3]$ for each data set. Then, these outliers were set to the clustering algorithms in a simulation study that was repeated 1000 times. In the simulation study, the values of *pout*, *pmask*, and *pswamp* as given in (12) to (14) were obtained.

RESULTS AND DISCUSSION

Table 2 and Figure 2 represent some parts of the simulation results and the plots for the power of performance of the new clustering method on Down and Mardia (2002)

circular-circular regression model. Table 2 shows the power of performance for S-SL, D-SL and C-SL at *v*-space using single-linkage method. From Table 2, similar pattern can be seen for all methods where the "success" probability (*pout*) increases significantly with an increase in the level of contamination λ . As sample size *n* increases, the *pout* values gradually increase for most fixed values of the error concentration for parameter κ. Subsequently, as the error concentration parameter κ increases, the *pout* values gradually increase at all fixed values of sample size *n*. In general, as the level of contamination λ, sample size *n*, and error concentration parameter κ increase, the *pout* values approach one. The results show a similar pattern for *u* space and *uv* spaces outliers.

Alternatively, part of the results are represented in Figure 2 in which the plot of "success" probability (*pout*), masking error (*pmask*) and swamping error (*pswamp*) for $n = 120$ and $n = 30$ with concentration parameter $\kappa = 20$ and $\kappa = 5$ for Single-linkage algorithm at *v*-space. From Figure 2, it can be seen that the curve pattern of the "success" probability (*pout*) approaches to one as the level of contamination λ increases. For any fixed *n*, the larger the κ is, the faster the curve point will be approaching one. These results clearly indicate that, if the value of the error concentration parameter κ is high, then the proposed method may detect an outlier at lower contamination value or situated closer to the inlying observations. Nevertheless, if the value of error concentration parameter κ is low, then the methods are only able to detect an outlier at high contamination value or situated far from the inlying observations.

From Figure 2, it can be seen that the curve pattern of the "success" probability (*pout*) approaches one for all values of *n* for any fixed value of κ. The larger *n* value is, the faster the curve approaches to one. Also, it can be seen that the masking error (*pmask*) decreases significantly with an increase in the level of contamination λ . Particularly as sample size *n* increases, the *pmask* values gradually decrease for most of the values of the error concentration parameter κ. Similarly, as the error concentration parameter κ increases, the *pmask* values gradually decrease for most of the values of the sample size *n*.

Additionally, as the level of contamination λ, sample size *n*, and error concentration parameter κ increase, the values of masking error (*pmask*) get closer to zero. Besides, it can be seen that the curve pattern of the masking error (*pmask*) is a decreasing function as the level of contamination λ increases. For any fixed *n*, the bigger the error concentration parameter κ is, the faster the curve decreases and approaches zero. For large values of error concentration parameter with low contamination value, the *pmask* value is low and closer to zero.

However, for small value of the error concentration parameter with low contamination value, the *pmask* value is still high and is not close to zero. This result is expected as at the high level of contamination value, the outlying observation is situated far from the inlying observations. At the same time, this situation may increase the possibility of the proposed method to detect the outliers and decrease the masking error effects. From the figure, it can be seen that the curve pattern of the masking error (*pmask*) for different values of sample sizes *n* are very close to each other and decrease gradually to zero for any fixed value of the error concentration parameter κ. For any fixed κ, the larger the *n* value, the faster the curve descends to zero.

Figure 2 also shows the power of performance of the clustering methods using the swamping error (*pswamp*). It can be seen that the swamping error (*pswamp*) decreases gradually with the increase in the level of contamination λ . In general, the values of swamping error (*pswamp*) are relatively small and less than 0.12 regardless of how large the sample size *n* or error concentration parameter κ are given. Furthermore, as sample size *n* increases, the *pswamp* values gradually decrease for any fixed value of the error concentration parameter κ. From the figure, it can be seen that the curve pattern of the swamping error (*pswamp*) for all different values of error concentration parameters κ are very close to each other and they gradually decrease as the level of contamination λ increases.

In addition, Table 3 summarise the best method based on the sample sizes and concentration parameter. For single-linkage based methods, S-SL algorithm performs better compare to other methods at higher levels of concentration parameter ($\kappa = 15$ and $\kappa = 20$) for all outlier scenarios. Meanwhile, D-SL algorithm is at best when concentration parameter is at 10. Next, for average linkage based methods, S-AL is the best method at lower concentration parameter when $\kappa = 5$ and $\kappa = 10$. Besides, CC-AL is the best method for all outlier scenarios when contamination level is higher. Similar pattern can be seen for complete linkage based methods.

In conclusion, based on the values of "success" probability (*pout*), masking error (*pmask*), and swamping error (*pswamp*), the single-linkage methods perform very well on the simulated random data set. At high levels of contamination λ , S-SL, D-SL and C-SL methods give a high value of *pout*, low value of *pmask*, and low value of *pswamp.* However, the results for average linkage and complete linkage show average performance, we may say that the single-linkage method performed at its best and very effective if the outlying observations were situated far from the remaining inlying observations.

TABLE 2. The power of performance for the clustering method using "success" probability (*pout*) for single-linkage clustering at *v*-space outlier

		pout				*pmask*				*pswamp*			
		$\kappa = 5$		$\kappa = 20$		$\kappa = 5$		$\kappa = 20$		$\kappa = 5$		$\kappa = 20$	
Algorithm	λ	$n = 30$	$n = 120$	$n = 30$	$n = 120$	$n = 30$	$n = 120$	$n = 30$	$n = 120$	$n = 30$	$n = 120$	$n = 30$	$n = 120$
S-SL	0.0	0.026	0.012	0.052	0.014	0.903	0.938	0.861	0.940	0.094	0.064	0.123	0.068
	0.2	0.037	0.017	0.088	0.073	0.852	0.868	0.768	0.711	0.089	0.064	0.113	0.063
	0.4	0.103	0.100	0.550	0.861	0.714	0.612	0.362	0.063	0.089	0.063	0.108	0.052
	0.6	0.395	0.534	0.961	1.000	0.418	0.235	0.025	0.000	0.087	0.057	0.104	0.051
	0.8	0.793	0.930	0.984	1.000	0.130	0.026	0.007	0.000	0.083	0.054	0.104	0.049
	1.0	0.880	0.993	0.980	1.000	0.060	0.002	0.007	0.000	0.077	0.047	0.101	0.049
D-SL	0.0	0.024	0.011	0.048	0.016	0.910	0.950	0.869	0.947	0.093	0.060	0.117	0.058
	0.2	0.042	0.019	0.096	0.062	0.856	0.889	0.781	0.790	0.087	0.059	0.113	0.055
	0.4	0.109	0.092	0.442	0.756	0.740	0.666	0.488	0.145	0.086	0.057	0.109	0.049
	0.6	0.333	0.446	0.896	1.000	0.521	0.317	0.084	0.000	0.085	0.056	0.107	0.049
	0.8	0.682	0.883	0.978	1.000	0.227	0.048	0.011	0.000	0.077	0.049	0.103	0.048
	1.0	0.771	0.987	0.969	1.000	0.127	0.004	0.011	0.000	0.072	0.047	0.091	0.048
C-SL	0.0	0.035	0.020	0.070	0.043	0.934	0.944	0.850	0.927	0.105	0.071	0.141	0.078
	0.2	0.051	0.028	0.114	0.071	0.842	0.894	0.753	0.806	0.102	0.064	0.140	0.075
	0.4	0.123	0.081	0.504	0.679	0.717	0.699	0.418	0.221	0.102	0.057	0.134	0.069
	0.6	0.369	0.392	0.901	0.999	0.463	0.375	0.081	0.000	0.101	0.054	0.133	0.063
	0.8	0.745	0.849	0.979	1.000	0.168	0.069	0.011	0.000	0.096	0.052	0.130	0.062
	1.0	0.857	0.979	0.975	1.000	0.085	0.009	0.011	0.000	0.092	0.051	0.126	0.060

TABLE 3. The best clustering method

		Outlier in *v*-space				Outlier in *u*-space				Outlier in *uv* spaces			
Method		$n = 30$	$n = 50$	$n = 100$	$n = 120$	$n = 30$	$n = 50$	$n = 100$	$n = 120$	$n = 30$	$n = 50$	$n = 100$	$n = 120$
Single-linkage	$\kappa = 5$	D-SL	D-SL	D-SL	D-SL	CC-SL	CC-SL	CC-SL	CC-SL	CC-SL	CC-SL	CC-SL	CC-SL
	$\kappa = 10$	D-SL	D-SL	D-SL	D-SL	D-SL	D-SL	D-SL	D-SL	D-SL	D-SL	D-SL	D-SL
	$\kappa = 15$	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL
	$\kappa = 20$	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL	S-SL
Average linkage	$\kappa = 5$	S-AL	S-AL	S-AL	S-AL	S-AL	S-AL	S-AL	S-AL	S-AL	S-AL	S-AL	S-AL
	$\kappa = 10$	S-AL	S-AL	S-AL	S-AL	D-AL	D-AL	S-AL	S-AL	D-AL	D-AL	D-AL	CC-AL
	$\kappa = 15$	D-AL	D-AL	D-AL	D-AL	D-AL	D-AL	D-AL	D-AL	S-AL	S-AL	S-AL	CC-AL
	$\kappa = 20$	CC-AL	CC-AL	CC-AL	CC-AL	CC-AL	CC-AL	CC-AL	CC-AL	CC-AL	CC-AL	CC-AL	CC-AL
Complete linkage	$\kappa = 5$	S-CL	S-CL	D-CL	CC-CL	S-CL	S-CL	S-CL	S-CL	CC-CL	CC-CL	CC-CL	CC-CL
	$\kappa = 10$	S-CL	S-CL	D-CL	CC-CL	S-CL	D-CL	D-CL	D-CL	D-CL	D-CL	D-CL	D-CL
	$\kappa = 15$	S-CL	S-CL	D-CL	CC-CL	S-CL	D-CL	D-CL	D-CL	D-CL	D-CL	D-CL	D-CL
	$\kappa = 20$	S-CL	S-CL	D-CL	CC-CL	S-CL	CC-CL	CC-CL	CC-CL	D-CL	D-CL	D-CL	D-CL

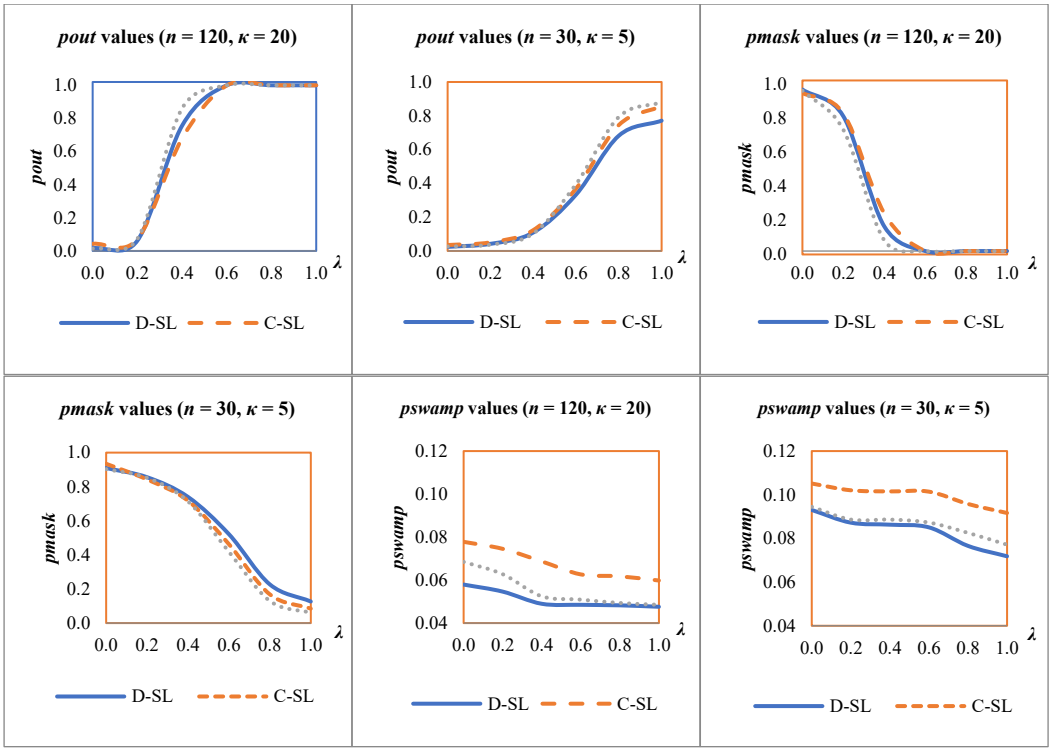


FIGURE 2. Plot of "success" probability (*pout*), masking error (*pmask*) and swamping error (*pswamp*) versus the level of contamination (λ) for $n = 120$ and $n = 30$ with concentration parameter $\kappa = 20$ and $\kappa = 5$ for single-linkage algorithm at *v*-space

PRACTICAL EXAMPLES

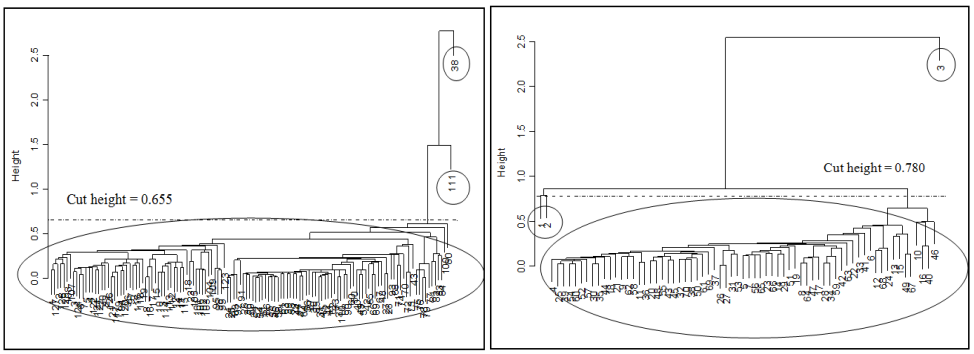
For illustrations, we consider two data sets to demonstrate the applicability of the clustering algorithm. The data sets are the Humberside wind direction data, and a simulated data set generated from a set of true values of parameters.

HUMBERSIDE WIND DIRECTION DATA

Humberside wind direction data is often used by previous researchers to demonstrate the presence of outliers in bivariate circular data sets (Abuzaid et al. 2012b, 2011, 2009; Hussin et al. 2013, 2010; Ibrahim et al. 2013; Rambli 2011; Satari 2015). Thus, it has been established that observations 38 and 111 are outliers. However, none of the referred literature shows that both outlying observations can be detected at one time except in Satari (2015). Here, we cluster the observations using agglomerative hierarchical methods with three circular distances (Satari's distance, Di's distance and Chang-chien's distance) between pairs of predicted values and residuals as the similarity measure and obtained the cluster tree. Based on the stopping rule proposed by Satari (2015), the tree is cut and grouped at the height of $\bar{h} + 2.06 s_h$. For the data set, the average of the cluster tree heights is $\bar{h} = 0.167$ and the circular standard deviation of the heights is $s_h = 0.237$. Therefore, the cutting height of the cluster tree is 0.655.

The cluster tree and corresponding cut height that are shown in Figure 3(a) is an example of using single-linkage method and Satari's distance (S-SL algorithm). It can be seen that there are three groups with Group 1 consisting of observation 38, Group 2 consisting of observation 111, and Group 3 consisting the remaining observations. Group 3 which contains the majority of the observations in the data set are considered as inlying observations. With 95% confidence level, we may say that observations 38

and 111 are identified as outliers. Hence, S-SL algorithm can successfully identify any outlying observations at any conditions with different degree of contamination, suggesting a high applicability of the new clustering technique for detecting multiple outliers in circular-circular regression models.



(a) (b)

FIGURE 3. (a) The cluster tree and corresponding cut height for Humberside wind direction data. (b) The cluster tree and corresponding cut height for simulated data set

SIMULATED DATA

We generated random samples of size $n = 70$ for the independent circular variable *u* and circular error *e* from von Mises distribution. The values of *u* and *e* are chosen from $u \sim VM(0, 2)$ and $e \sim VM(0, 20)$, respectively. Based from the generated random samples *u* and *e*, we calculated the values of the response variable *v* using the Down and Mardia (2002) circular-circular-circular regression model with fixed true parameter values of $\alpha = 1.5$, $\beta = 1.5$, and $\omega = 0.5$. Additionally, we specifically planted three outliers at a certain point $[d_1, d_2, d_3]$ using (8) with different degrees of contamination, $\lambda_1 = 0.4$, $\lambda_2 = 0.6$ and $\lambda_3 = 1.0$, respectively.

We clustered the observations using S-SL algorithm between pairs of predicted values and residuals as the similarity measure and obtained the cluster tree (Figure 3(b)). Based on the proposed stopping rule, the tree is cut and groups formed at the height of 0.780. The cluster tree and corresponding height cut are shown in Figure 3(b). It can be seen that after the cut, there are three groups formed. Group 1 consists of observations 1 and 2. Group 2 consists of observation 3 and Group 3 consists of the remaining observations other than observations 1, 2 and 3. Hence, group 3 contains the majority of the observations in the data set and will be considered as inlying observations. Subsequently, with 95% confidence level, we may say that observations 1, 2, and 3 are identified as outliers. Hence, S-SL algorithm can successfully identify any outlying observations at any conditions with different degrees of contamination, thus suggesting the high applicability of the new clustering technique for detecting multiple outliers in circular-circular regression models.

CONCLUSION

The aim of this study was to compare nine clustering-based outliers detection methods for identifying multiple outliers in circular-circular regression model. The clustering-based procedure is developed for predicted and residual values obtained from the Down and Mardia (2002) model fit of a circular-circular data set. For all methods, three circular distances were used and a cluster tree was obtained using the agglomerative hierarchical methods. At 95% confidence level, the cluster group that exceeds the stopping rule $\bar{h} + 2.06 s_h$ is classified as potential outliers.

It is shown from simulation study and application to real data set that the single-linkage method performed well for any value of sample size n and error concentration parameter, κ especially in S-SL algorithm. Also, the algorithms are good in identifying outliers not limited to one or few outliers only but the presence of multiple outliers at one time.

Furthermore, one may argue that the predicted value versus residual plot is probably sufficient enough to identify multiple outliers in circular regression. Traditionally, predicted values plot against residual is a simple tool and useful to detect any abnormal observations in the data set since we can easily spot a typical group of observation. However, in some cases, the outlying observations are situated too close with inlying observations. If such situation exists, the predicted value versus residual plot is no longer beneficial. Therefore, an alternative method of grouping the observations and classifying them as inlying or outlying, as proposed in this study is needed. Moreover, it can be seen that from the practical example that there was no masking. From the simulation study, it can be observed that the masking error is practically small and almost zero, especially when the outlier is situated far from the inlying observations. Similar finding can be found for the swamping errors in which, masking is a more serious problem than swamping.

ACKNOWLEDGEMENTS

The authors would like to thank the Ministry of Higher Education for providing financial support under Fundamental research grant No. FRGS/1/2016/STG06/UMP/02/5 (University reference RDU160117).

REFERENCES

Abuzaid, A.H. 2010. Some problems of outliers in circular data. University of Malaya. Ph.D. Thesis (Unpublished).

Abuzaid, A.H., Hussin, A.G. & Mohamed, I.B. 2013. Detection of outliers in simple circular regression models using the mean circular error statistic. *Journal of Statistical Computation and Simulation* 83(2): 269-277.

Abuzaid, A.H., Mohamed, I.B. & Hussin, A.G. 2012a. Boxplot for circular variables. *Computational Statistics* 27(3): 381-392.

Abuzaid, A.H., Hussin, A.G., Rambli, A. & Mohamed, I.B. 2012b. Statistics for a new test of discordance in circular data. *Communications in Statistics-Simulation and Computation* 41(10): 1882-1890.

Abuzaid, A.H., Hussin, A.G., Rambli, A. & Mohamed, I.B. 2011. COVRATIO statistic for simple circular-circular regression model. *Chiang Mai Journal of Science* 38(3): 321-330.

Abuzaid, A.H., Hussin, A.G. & Mohamed, I.B. 2009. Identifying single outlier in linear circular-circular regression model based on circular distance. *Journal of Applied Probability & Statistics* 3(1): 107-117.

Adnan, R. & Mohamad, M.N. 2003. Multiple outliers detection procedures in linear regression. *Matematika* 19(1): 29-45.

Alkasadi, N.A., Ibrahim, S., Ramli, M.F. & Yusoff, M.I. 2016. A comparative study of outlier detection procedures in multiple circular regression. In *AIP Conference Proceedings* 1775(1): 1-7.

Blashfield, R.K. & Morey, L.C. 1980. A comparison of four clustering methods using MMPI Monte Carlo data. *Applied Psychological Measurement* 4(1): 57-64.

Di, N.F.M. & Satari, S.Z. 2017. The effect of different distance measures in detecting outliers using clustering-based algorithm for circular regression model. In *AIP Conference Proceedings* 1842(1): 1-13.

Di, N.F.M., Satari, S.Z. & Zakaria, R. 2017. Detection of different outlier scenarios in circular regression model using single-linkage method. *Journal of Physics: Conference Series* 890(1): 1-5.

Caires, S. & Wyatt, L.R. 2003. A linear functional relationship model for circular data with an application to the assessment of ocean wave measurements. *Journal of Agricultural, Biological, and Environmental Statistics* 8(2): 153-169.

Chang-Chien, S.J., Hung, W.L. & Yang, M.S. 2012. On mean shift-based clustering for circular data. *Soft Computing* 16(6): 1043-1060.

Downs, T.D. & Mardia, K.V. 2002. Circular regression. *Biometrika* 89(3): 683-698.

Fisher, N.I. 1995. *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.

Gan, G., Ma, C. & Wu, J. 2007. *Data Clustering: Theory, Algorithms, and Applications*. United States of America: SIAM.

Hartigan, J.A. 1975. *Clustering Algorithm*. New York: John Wiley & Sons Inc.

Hussin, A.G. & Abuzaid, A.H. 2012. Detection of outliers in functional relationship model for circular variables via complex form. *Pakistan Journal of Statistics* 28(2): 205-216.

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

Hussin, A.G., Abuzaid, A., Zulkifli, F. & Mohamed, I. 2010. Asymptotic covariance and detection of influential observations in a linear relationship model for circular data with application to the measurements of wind directions. *ScienceAsia* 36(2010): 249-253.

Hussin, A.G., Fieller, N.R. & Stillman, E.C. 2004. Linear regression model for circular variables with application to directional data. *Journal of Applied Science and Technology* 9(1): 1-6.

Ibrahim, S. 2013. Some outlier problems in a circular-circular regression model. University of Malaya. Ph.D. Thesis (Unpublished).

Ibrahim, S., Rambli, A., Hussin, A.G. & Mohamed, I. 2013. Outlier detection in a circular-circular regression model using COVRATIO statistic. *Communications in Statistics-Simulation and Computation* 42(10): 2270-2280.

Jammalamadaka, S.R. & Sengupta, A. 2001. *Topics In Circular Statistics*. Singapore: World Scientific.

Jammalamadaka, S.R. & Sarma, Y.R. 1993. Circular regression. In *Statistical Sciences and Data Analysis*, edited by Matusita, K. Puri, M.L. & Hayakawa, T. Utrecht: VSP. pp. 109-128.

Milligan, G.W. & Cooper, M.C. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50(2): 159-179.

Mojena, R. 1977. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal* 20(4): 359-363.

Rambli, A. 2011. Outlier detection in circular data and circular-circular-circular regression model. University of Malaya. M.Sc. Thesis (Unpublished).

Rambli, A., Abuzaid, A.H., Mohamed, I.B. & Hussin, A.G. 2016. Procedure for detecting outliers in a circular-circular regression model. *PloS ONE* 11(4): e0153074.

Rambli, A., Yunus, R.M., Mohamed, I. & Hussin, A.G. 2015. Outlier detection in a circular-circular regression model. *Sains Malaysiana* 44(7): 1027-1032.

Rambli, A., Mohamed, I., Abuzaid, A.H. & Hussin, A.G. 2010. Identification of influential observations in circular-circular regression model. In *Proceedings of the Regional Conference on Statistical Sciences (RCSS'10)*. pp. 195-203.

Satari, S.Z., Di, N.F.M. & Zakaria, R. 2017. The multiple outliers detection using agglomerative hierarchical methods in circular regression model. *Journal of Physics: Conference Series* 890(1): 1-5.

Satari, S.Z. 2015. Parameter estimation and outlier detection for some types of circular model. University of Malaya. Ph.D. Thesis (Unpublished).

Sebert, D.M., Montgomery, D.C. & Rollier, D.A. 1998. A clustering algorithm for identifying multiple outliers in linear regression. *Computational Statistics and Data Analysis* 27(4): 461-484.

Siti Zanariah Satari* & Nur Faraidah Muhammad Di
Centre for Mathematical Sciences
College of Computing & Applied Sciences
Universiti Malaysia Pahang
26300 Kuantan, Pahang Darul Makmur
Malaysia

Yong Zulina Zubairi
Centre for Foundation Studies in Sciences
University of Malaya
50603 Kuala Lumpur, Federal Territory
Malaysia

Abdul Ghapor Hussin
Faculty of Defence Sciences and Technology
National Defence University of Malaysia
Sungai Besi Camp
57000 Kuala Lumpur, Federal Territory
Malaysia

*Corresponding author; email: zanariah@ump.edu.my

Received: 7 May 2019
Accepted: 14 October 2020