Pak.j.stat.oper.res. Vol.18 No. 3 2022 pp 561-573 DOI: http://dx.doi.org/10.18187/pjsor.v18i3.3982

Pakistan Journal of Statistics and Operation Research

The Effect of Different Similarity Distance Measures in Detecting Outliers Using Single-Linkage Clustering Algorithm for Univariate Circular Biological Data

Nur Syahirah Zulkipli[1], Siti Zanariah Satari[2*], Wan Nur Syahidah Wan Yusoff[3]

* Corresponding Author

1. Centre for Mathematical Sciences, Universiti Malaysia Pahang, Malaysia, syahirahzulkipliwork@gmail.com
2. Centre for Mathematical Sciences, Universiti Malaysia Pahang, Malaysia, zanariah@ump.edu.my
3. Centre for Mathematical Sciences, Universiti Malaysia Pahang, Malaysia, wnsyahidah@ump.edu.my

Abstract

Clustering algorithms can be used to create an outlier detection procedure in univariate circular data. The circular distance between each point of angular observation in circular data is used to calculate the similarity measure to appropriately group observations. In this paper, we present a clustering-based procedure for detecting outliers in univariate circular biological data using various similarity distance measures. Three circular similarity distance measures; Satari distance, Di distance and Chang-chien distance were used to detect outliers using a single-linkage clustering algorithm. Satari distance and Di distance are two similarity measures that have similar formulas for univariate circular data. This study aims to develop and demonstrate the effectiveness of the proposed clustering-based procedure with various similarity distance measures in detecting outliers. The circular similarity distance of SL-Satari/Di and other similarity measures, including SL-Chang, were compared at various dendrogram cutting points. It is found that a clustering-based procedure using a single-linkage algorithm with various similarity distances is a practical and promising approach to detect outliers in univariate circular data, particularly for biological data. According to the results, the SL-Satari/Di distance outperformed the SL-Chang distance for certain data conditions.

Key Words: Similarity measure; Circular distance; Circular data; Outliers; Clustering algorithm

1. Introduction

Data recorded in angular measurements, or circular data, has been used to extend outlier detection procedures from linear cases to circular cases over the last few decades. Outliers are objects or observations in a dataset that are inconsistent with the rest of the data (Collett, 1980). In other words, outliers are observations that deviate significantly or are dissimilar to the others. Many researchers have proposed outlier detection procedures in univariate circular data using various methods such as the row deletion method (Abuzaid et al., 2009; Abuzaid et al., 2012; Collett, 1980; Rambli, 2015), the robust method (Mahmood et al., 2017) and the clustering-based method (Abuzaid, 2012; Ahmed et al., 2019).

A clustering-based procedure can be used to detect outliers by classifying the outliers and inliers. Clustering is used to classify data points into distinct clusters or groups, whereas outlier detection is used to identify data points that do not appear to fit naturally into these distinct clusters. Clustering and outlier detection have a symbiotic relationship. The clustering-based outlier detection method has an advantage in that it can find clusters while also identifying outliers. Combining clustering and outlier detection has the following benefits: (i) the clusters become more compact and semantically coherent, (ii) the clusters become more robust against data perturbations and (iii) the outliers are contextualised by the clusters and more interpretable (Ott et al., 2014).

Clustering-based outlier detection in circular data has been previously proposed using a single-linkage clustering algorithm for bivariate circular data (Di et al., 2017; Satari, 2015; Satari et al., 2021). There is no single-linkage clustering-based outlier detection method for univariate circular data, hence we developed a procedure for detecting outliers in univariate circular data using an agglomerative hierarchical clustering method based on a single-linkage algorithm with various similarity distance measures. The agglomerative hierarchical clustering method produces a dendrogram, which is circular sed to represent the clusters. The single-linkage algorithm is the simplest agglomerative hierarchical clustering method that merges observations by taking the shortest distance.

Circular distance in a circular dataset is defined as the distance between two points on the circumference, i.e., between any two angles (Jammalamadaka and Sengupta, 2001). Meanwhile, circular similarity distance can be constructed from the circular distance in order to determine the similarity measure. Chang-chien et al. (2012) and Hung et al. (2012) developed the mean shift-based clustering algorithm for circular data using the formulation of circular distance introduced by Jammalamadaka and Sengupta (2001). Satari (2015) used the single-linkage clustering algorithm and introduced the circular city-block distance, also known as the Satari distance, to calculate the similarity measure in detecting outliers for bivariate circular data. Di et al. (2017) incorporated the circular Euclidean distance into a single-linkage bivariate circular outlier detection procedure. Outliers are identified when a cluster exceeds a certain height of the dendrogram, which is then cut using the cutting rule proposed by Satari (2015).

In this study, we present a clustering-based procedure for detecting outliers in univariate circular data using a single-linkage clustering algorithm and various similarity distance measures. Univariate circular data is common in biological studies such as ecological and biomedical research. The performance of the proposed procedure was compared to that of its counterparts using five real univariate circular biological datasets: Turtle data (the directions of turtles after laying eggs) (Chang-chien et al., 2012; Hung et al., 2012), Eye data (eye angles obtained from glaucoma patients) (Abuzaid, 2020; Alkasadi et al., 2018; Rambli, 2015), Sea star data (the direction of sea stars after removal from the natural habitat) (Mahmood et al, 2017), Pigeon data (vanishing direction of homing pigeons) (Jammalamadaka & Sengupta, 2001), and Frog data (the homing ability of the frog) (Abuzaid et al., 2009; Abuzaid et al., 2012; Collett, 1980; Zulkipli et al., 2020). These biological data have been used in various studies of outlier detection procedures in univariate circular data. Therefore, the datasets were used in this study to demonstrate the applicability of the proposed clustering-based procedure with various similarity distance measures to detect outliers.

2. Similarity Distance Measures for Circular Data

Jammalamadaka and Sengupta (2001) define the circular distance between any two points as the lesser of the two arc lengths between the points along the circumference, i.e., for any two angles θ_i and θ_j . The circular distance d_0 can be defined as follows (Jammalamadaka and Sengupta, 2001):

$$d_0(\theta_i,\theta_j) = \pi - \left|\pi - \left|\theta_i - \theta_j\right|\right|, \quad i, j = 1, 2, ..., n. \tag{1}$$

Another closely related definition of the circular distance between angles θ_i and θ_j is (Jammalamadaka and Sengupta, 2001):

$$d_0(\theta_i,\theta_j) = \left(1 - \cos\left(\theta_i - \theta_j\right)\right). \tag{2}$$

In a clustering-based procedure, circular distance has been used to identify outliers in circular data (Di et al., 2017; Satari, 2015; Satari et al., 2021). The similarity measure is used in clustering to appropriately group observations. The similarity measure can be calculated using the circular distance between each point observation. In this study, three circular similarity measures were used to develop the proposed clustering-based procedure for detecting outliers in univariate circular biological data. A circular city-block distance (Satari distance) and two circular Euclidean distances (Di distance and Chang-chien distance) were used to calculate the similarity distance in the single-linkage algorithm. The formulas for the Satari, Di and Chang-chien similarity distances are as follows:

Satari distance (Satari, 2015):

$$d_{ij(Satari)} = \sum_{k=1}^{p} \left(\pi - \left| \pi - \left| \theta_{ik} - \theta_{jk} \right| \right| \right) \tag{3}$$

Di distance (Di et al., 2017):

$$d_{ij(Di)} = \sqrt{\sum_{k=1}^{p} \left(\pi - \left| \pi - \left| \theta_{ik} - \theta_{jk} \right| \right| \right)^2} \tag{4}$$

Chang-chien distance (Chang-chien et al., 2012):

$$d_{ij(Chang-Chien)} = \sqrt{\sum_{k=1}^{p} \left(1 - \cos(\theta_{ik} - \theta_{jk}) \right)^2} \tag{5}$$

In equations (3)–(5), d_{ij} is the distance between observations *i* and *j*, *p* is the number of variables, θ_{ik} is the value of the *k*th variable for the *i*th observation and θ_{jk} is the value of the *k*th variable for the *j*th observation, where $i = 1, 2, ..., n$ and $j = 1, 2, ..., n.$ Since there is only one variable in univariate circular data, $p = 1$ in all formulas. Hence, equations (3) and (4) can be simplified further by taking $p = 1$, yielding the same formula as shown in equation 6.

$$\begin{aligned} d_{ij(Di)} &= \sqrt{\left(\pi - \left| \pi - \left| \theta_{ik} - \theta_{jk} \right| \right| \right)^2} \\ &= \left(\pi - \left| \pi - \left| \theta_{ik} - \theta_{jk} \right| \right| \right) \\ &= d_{ij(Satari)} \end{aligned} \tag{6}$$

Only when $p > 1$ in the case of bivariate or multivariate data do the Satari and Di distances yield different similarity distances. Since the Satari and Di distances are the same for univariate circular data, these distances in this study are referred as the Satari/Di distance.

3. Clustering-based Procedure for Outlier Detection in Univariate Circular Data

The single-linkage algorithm, also known as the minimum method, is one of the most basic agglomerative clustering methods that merges two clusters with the smallest minimum pairwise distance at each iteration. This study focuses on the single-linkage algorithm which is sensitive to the presence of outliers (Klutchnikoff et al., 2021). The single-linkage algorithm is as follows (Johnson and Wichern, 2014):

Step 1: Begin with *n* clusters, each containing a single observation $\theta_1, \theta_2, \ldots, \theta_n$.

Step 2: Calculate the distance matrix between all possible pairs of clusters using the circular similarity distances in equations (3)–(5).

Step 3: Find the minimum $d_{\theta_i \theta_j}$ (distance between θ_i and θ_j) in the distance matrix to determine the nearest members and merge the corresponding observations. For example, if θ_1 and θ_2 are the nearest members, then they formed a cluster. The distance between cluster (θ_1, θ_2) and the other clusters or single observations $\theta_3, \theta_4, \ldots, \theta_n$ are then calculated. Let a single observation (θ_3) be a cluster and the distance between clusters (θ_1, θ_2) and (θ_3) is computed using equation 7. The distance of $d_{\theta_1 \theta_3}$ and $d_{\theta_2 \theta_3}$ are referred from the distance matrix in Step 2.

$$d_{(\theta_1, \theta_2)\theta_3} = \min \left\{ d_{\theta_1 \theta_3}, d_{\theta_2 \theta_3} \right\}. \tag{7}$$

Step 4: The rows and columns in the distance matrix responding to the merged cluster(s) are deleted. A new distance matrix is updated by recalculating the distance matrix after forming a new cluster in Step 3.

Step 5: If more than one cluster remains, repeat Steps 3 and 4 until all clusters have been merged into a single cluster.

The single-linkage algorithm then generates a dendrogram or cluster tree as an output. The clusters are represented as branches of a tree. A dendrogram is used to find the optimal number of clusters, and it must be "cut" at a certain height to separate outliers from inliers. In this study, the dendrogram was cut using the cutting rule proposed by Satari (2015), as shown in equation 8. Another cutting rule is shown in equation 9.

$$\bar{h} + 2.06_{s_h}, \text{ at 95\% confidence interval,} \tag{8}$$

$$\bar{h} + 1.69_{s_h}, \text{ at 90\% confidence interval,} \tag{9}$$

where $\bar{h}$ is the average height of the cluster trees of all $N-1$ clusters and s_h is the circular standard deviation of the heights given by:

$$s_h = \sqrt{-2\log \bar{R}_h} \quad , \tag{10}$$

where $\bar{R}_h$ is the mean resultant length of the height for $N-1$ clusters. The cluster that exceeds a certain point of cutting height is classified as a candidate for outliers at 95% and 90% confidence intervals.

4. Illustrative Examples and Results

Five univariate circular biological datasets with outliers were used to demonstrate the applicability of the proposed outlier identification procedure. The datasets are Turtle data, Eye data, Sea star data, Pigeon data and Frog data taken from Collett (1980), Fisher (1993), Jammalamadaka and Sengupta (2001), Mahmood et al. (2017) and Rambli (2015), respectively. These datasets have been widely used in studies of univariate circular data outlier identification. Table 1 details the datasets used in this study.

Table 1: Summary of five historical univariate circular biological datasets

No.	Dataset	Sample size, *n*	Number of outliers	Percentage of outliers	Outlier observation
1	Turtle data	76	7	9%	Observations 57- 60, 74 - 76
2	Eye data	23	2	9%	Observations 10 and 17
3	Sea star data	22	1	5%	Observation 13
4	Pigeon data	15	2	14%	Observations 1 and 15
5	Frog data	14	1	7%	Observation 14

4.1 Turtle data

The Turtle dataset contains 76 observations of turtle directions after laying eggs (Fisher, 1993). Chang-chien et al. (2012) and Hung et al. (2012) applied clustering-based procedures to the Turtle dataset using a mean shift-based clustering algorithm. Figure 1 shows that there are seven suspected outliers in the circular plot of the dataset, namely observations 57, 58, 59, 60, 74, 75 and 76, respectively. Figure 2 shows the results of applying the proposed clustering-based procedure to the Turtle dataset to confirm the suspected outliers.

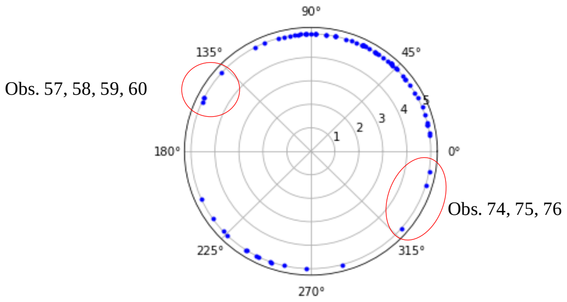
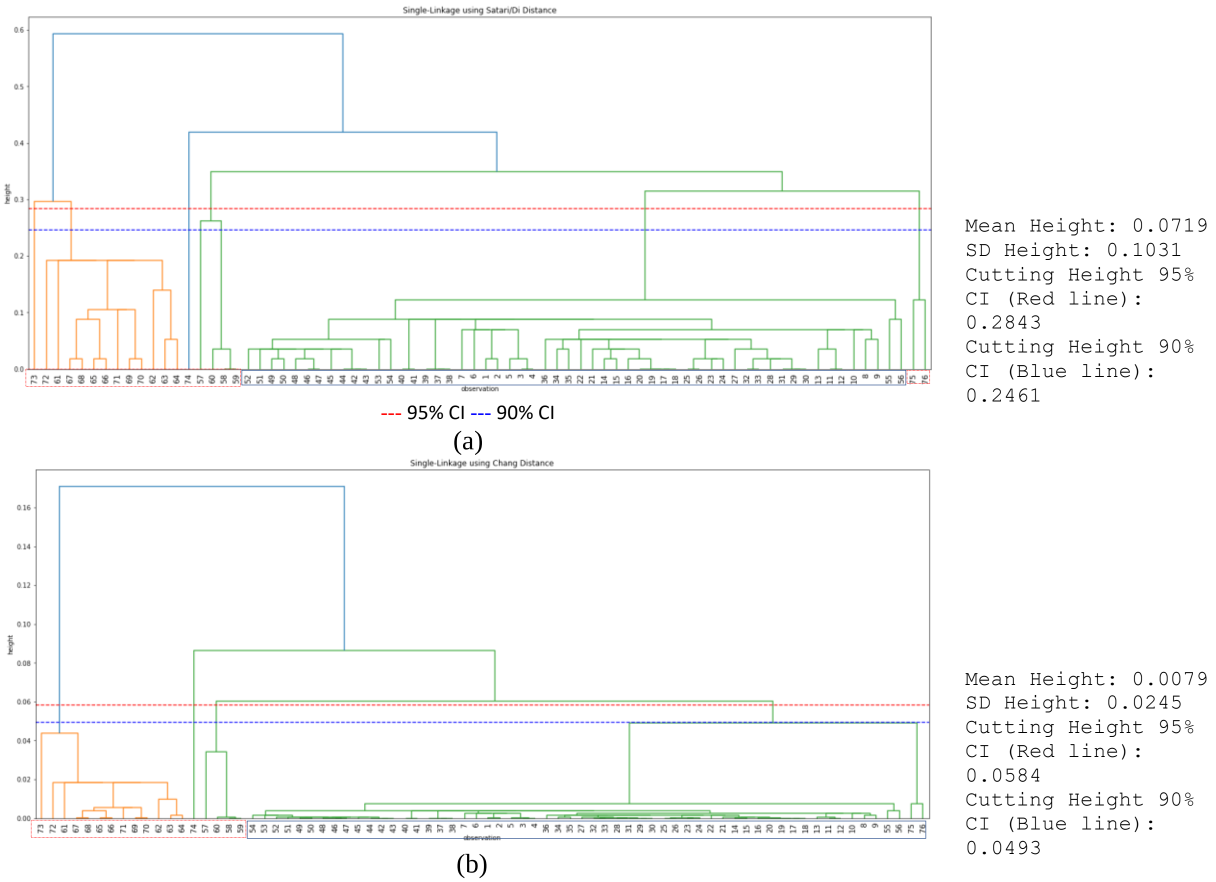


Figure 1: Circular plot of the Turtle data



```
Mean Height: 0.0719
SD Height: 0.1031
Cutting Height 95%
CI (Red line):
0.2843
Cutting Height 90%
CI (Blue line):
0.2461
```

(a)

```
Mean Height: 0.0079
SD Height: 0.0245
Cutting Height 95%
CI (Red line):
0.0584
Cutting Height 90%
CI (Blue line):
0.0493
```

(b)

Figure 2: The dendrogram with corresponding cutting height for the Turtle dataset using a) SL-Satari/Di and b) SL-Chang similarity distances.

As shown in Figure 2(a), the dendrogram of the SL-Satari/Di distance forms three clusters after its height was cut at 95% and 90% confidence intervals. Cluster 1 consists of observations 57 to 74, while Cluster 3 consists of observations 75 and 76, both of which exceed the 95% and 90% confidence intervals of cutting height. Therefore, all observations in Cluster 1 and Cluster 3 were identified as outliers. Cluster 2 consists of the remaining majority of observations and does not exceed the cutting height, indicating that the cluster consists of inliers.

In Figure 2(b), the dendrogram of the SL-Chang distance forms two clusters after its height was cut at 95% and 90% confidence intervals. Cluster 1 which consists of observations 57 to 74 exceeds the cutting height at 95% and 90% confidence intervals, indicating that these observations are outliers. Cluster 2 consists of the remaining majority of observations and does not exceed the cutting height, indicating that the cluster consists of inliers.

4.2 Eye data

The Eye dataset contains 23 observations of the eye angle of posterior corneal curvature of glaucoma patients (Abuzaid, 2020; Alkasadi et al., 2018; Rambli, 2015). Figure 3 shows that there are two suspected outliers in the circular plot of the dataset, namely observations 10 and 17. Rambli (2015) successfully identified these two observations as outliers using the row deletion method. Figure 4 shows the results of applying the proposed clustering-based procedure to the Eye dataset to confirm the suspected outliers.

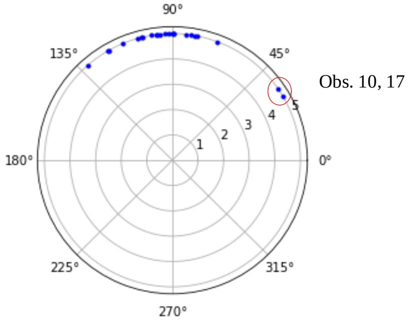
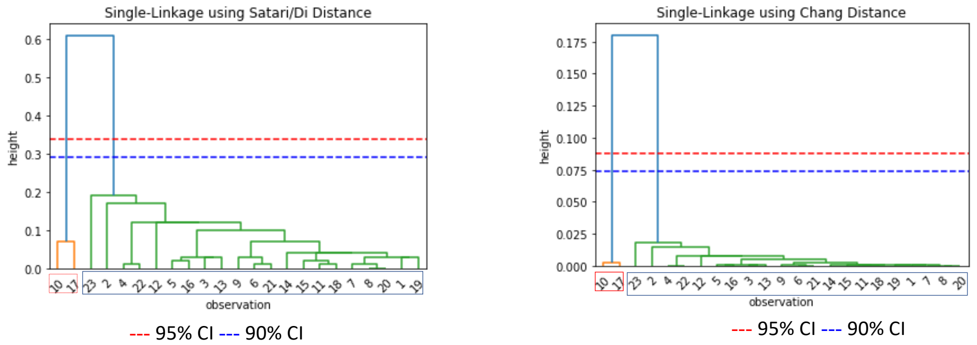


Figure 3: Circular plot of the Eye data



```
Mean Height: 0.0789, SD Height: 0.1259
Cutting Height 95% CI (Red line): 0.3383
Cutting Height 90% CI (Blue line): 0.2917
```

(a)

```
Mean Height: 0.0109, SD Height: 0.0374
Cutting Height 95% CI (Red line): 0.0879
Cutting Height 90% CI (Blue line): 0.0741
```

(b)

Figure 4: The dendrogram with corresponding cutting height for the Eye dataset using a) SL-Satari/Di and b) SL-Chang similarity distances.

Figure 4 shows two clusters were formed after the dendrograms were cut at 95% and 90% confidence intervals for both the SL-Satari/Di and SL-Chang distances. Cluster 1 consists of observations 10 and 17, while Cluster 2 consists of the remaining observations. Observations 10 and 17 were identified as outliers since Cluster 1 exceeds the cutting height at 95% and 90% confidence intervals for both the SL-Satari/Di and SL-Chang distances. Cluster 2 consists of the remaining majority of the observations and does not exceed the cutting height, indicating that the cluster consists of inliers. Both SL-Satari/Di and SL-Chang distances successfully identified observations 10 and 17 as outliers at 95% and 90% confidence intervals.

4.3 Sea star data

The Sea star dataset contains 22 observations of the direction of sea stars after removal from their natural habitat (Fisher, 1993). Mahmood et al. (2017) used the robust method to identify observation 13 as an outlier in this dataset. Figure 5 shows that there is one suspected outlier in the circular plot of the dataset, namely observation 13. Figure 6 shows the results of applying the proposed clustering-based procedure to the Sea star dataset to confirm the suspected outlier.

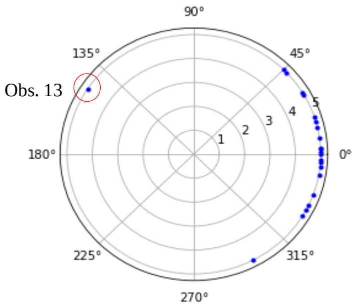
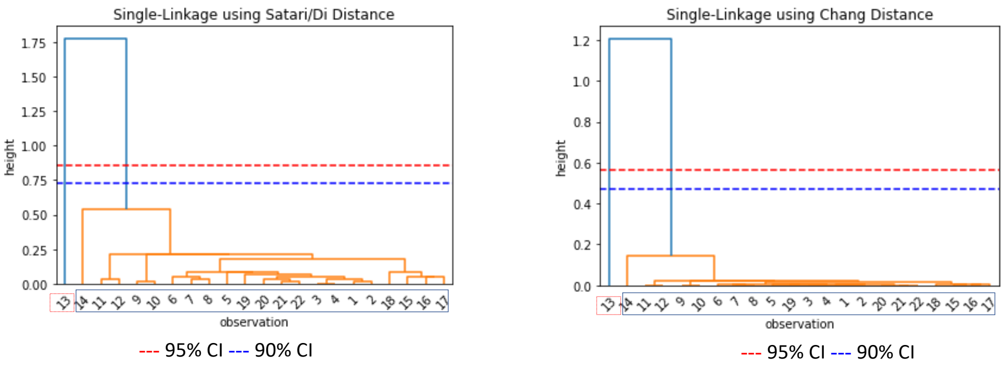


Figure 5: Circular plot of the Sea star dataset



```
Mean Height: 0.1429 SD Height: 0.3475
Cutting Height 95% CI (Red line): 0.8588
Cutting Height 90% CI (Blue line): 0.7302
```

(a)

```
Mean Height: 0.0569 SD Height: 0.2456
Cutting Height 95% CI (Red line): 0.5628
Cutting Height 90% CI (Blue line): 0.4720
```

(b)

Figure 6: The dendrogram with corresponding cutting height for the Sea star dataset using a) SL-Satari/Di and b) SL-Chang similarity distances.

Figure 6 shows two clusters were formed after the dendrograms were cut at 95% and 90% confidence intervals for both the SL-Satari/Di and SL-Chang distances. Cluster 1 consists of observation 13, while Cluster 2 consists of the remaining observations. Observations 13 was identified as an outlier since Cluster 1 exceeds the cutting height at 95% and 90% confidence intervals for both SL-Satari/Di and SL-Chang distances. Cluster 2 consists of the remaining majority of the observations and does not exceed the cutting height, indicating that the cluster consists of inliers. Both SL-Satari/Di and SL-Chang distances successfully identified observation 13 as an outlier at 95% and 90% confidence intervals.

4.4 Pigeon data

The Pigeon dataset contains 15 observations of the vanishing direction of homing pigeons (Jammalamadaka and Sengupta, 2001). Figure 7 shows that there are two suspected outliers in the circular plot of the dataset, namely observations 1 and 15. Figure 8 shows the results of applying the proposed clustering-based procedure to the Pigeon dataset to confirm the suspected outliers.

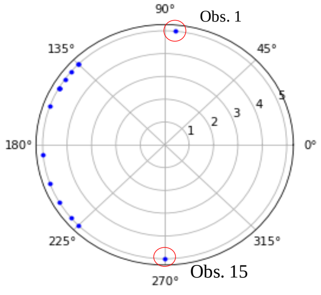
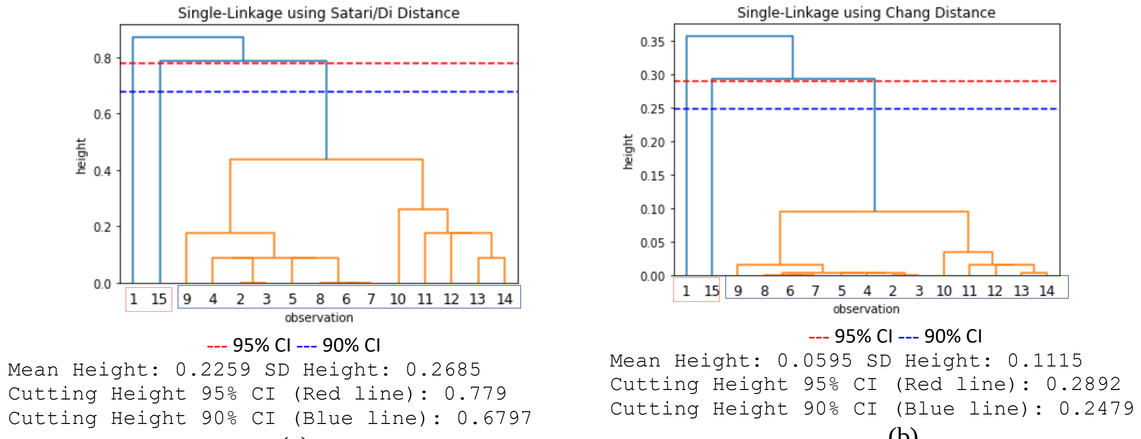


Figure 7: Circular plot of the Pigeon dataset



```
Mean Height: 0.2259 SD Height: 0.2685
Cutting Height 95% CI (Red line): 0.779
Cutting Height 90% CI (Blue line): 0.6797
```

(a)

```
Mean Height: 0.0595 SD Height: 0.1115
Cutting Height 95% CI (Red line): 0.2892
Cutting Height 90% CI (Blue line): 0.2479
```

(b)

Figure 8: The dendrogram with corresponding cutting height for the Pigeon dataset using a) SL-Satari/Di and b) SL-Chang similarity distances.

Figure 8 shows two clusters were formed after the dendrograms were cut at 95% and 90% confidence intervals for both the SL-Satari/Di and SL-Chang distances. Cluster 1 consists of observations 1 and 15, while Cluster 2 consists of the remaining observations. Observations 1 and 15 were identified as outliers since Cluster 1 exceeds the cutting height at 95% and 90% confidence intervals for both the SL-Satari/Di and SL-Chang distances. Cluster 2 consists of the remaining majority of the observations and does not exceed the cutting height, indicating that the cluster consists of inliers. Both SL-Satari/Di and SL-Chang distances successfully identified observations 1 and 15 as outliers at 95% and 90% confidence intervals.

4.5 Frog data

The Frog dataset contains 14 observations of frog homing ability (Collett, 1980). Various methods have been used to detect outliers in the dataset (Collett, 1980; Abuzaid et al., 2009; Abuzaid et al., 2012; Zulkipli et al., 2020). Figure 9 shows that there is one suspected outlier in the circular plot of the dataset, namely observation 14. Figure 10 shows the results of applying the proposed clustering-based procedure to the Frog dataset to confirm the suspected outlier.

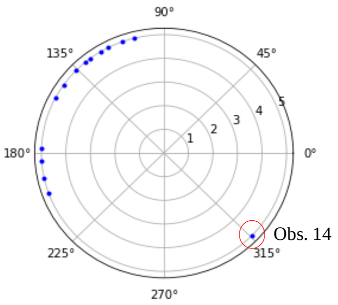
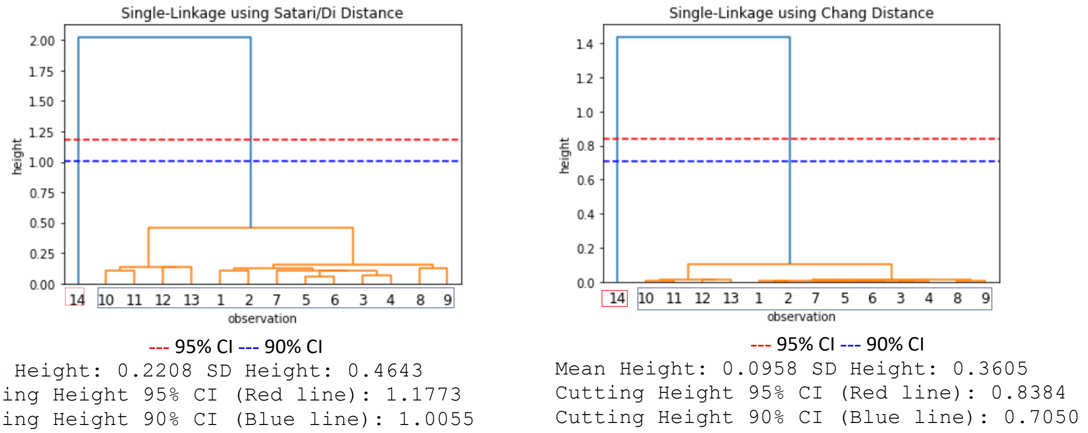


Figure 9: Circular plot of the Frog data



```
Mean Height: 0.2208 SD Height: 0.4643
Cutting Height 95% CI (Red line): 1.1773
Cutting Height 90% CI (Blue line): 1.0055
```

```
Mean Height: 0.0958 SD Height: 0.3605
Cutting Height 95% CI (Red line): 0.8384
Cutting Height 90% CI (Blue line): 0.7050
```

(a) (b)

Figure 10: The dendrogram with corresponding cutting height for the Frog dataset using a) SL-Satari/Di and b) SL-Chang similarity distances.

The Frog dataset contains 14 observations of frog homing ability (Collett, 1980). Various methods have been used to detect outliers in the dataset (Collett, 1980; Abuzaid et al., 2009; Abuzaid et al., 2012; Zulkipli et al., 2020). Figure 9 shows that there is one suspected outlier in the circular plot of the dataset, namely observation 14. Figure 10 shows the results of applying the proposed clustering-based procedure to the Frog dataset to confirm the suspected outlier.

5. Performance Measure and Discussion

This section compares and discusses the performance of the SL-Satari/Di and SL-Chang similarity distances in the single-linkage clustering algorithm for each dataset. The performance of the proposed clustering-based algorithm was measured using three measurements introduced by Sebert et al. (1998) that have been applied in many studies (Collett, 1980; Abuzaid et al., 2009; Abuzaid et al., 2012; Abuzaid, 2013; Rambli, 2015; Mahmood et al., 2017; Satari, 2015; Di et al., 2017; Satari et al., 2021).

i. Probability of all outliers are successfully detected, *pout*.

$$pout = 'success' / out$$

where *'success'* is the number of observations that clustering-based procedure successfully identified all the outliers and *out* is the number of outliers. The closer the *pout* value to 1, the better the proposed outlier detection procedure detects all the outliers.

ii. Probability of outliers are falsely detected as inliers (masking effect), *pmask*.

$$pmask = ' failure ' / out$$

where *'failure'* is the number of outliers in dataset that detected as inliers and *out* is the total number of outliers. The *pmask* value ranges from 0 to 1, with a value close to 0 indicating that no outliers were detected as inliers.

iii. Probability of inliers detected as outliers (swamping effect), *pswamp*.

$$pswamp = ' false ' / (n - out)$$

where *'false'* is the number of inliers in all data set that detected as outliers, *n* is total number of observations and *out* is total number of outliers. The *pswamp* value ranges from 0 to 1, with a value close to 0 indicating that no inliers were detected as outliers.

Table 2 shows performance measures for the proposed clustering-based procedures on the five biological datasets.

Table 2: Result of performance measures for five real datasets

No	Dataset	Circular similarity distance	Cutting rule	Performance Measures		
				pout	*pmask*	*pswamp*
1.	Turtle data	SL-Satari/Di	95%	7 (1.0000)	0 (0.0000)	13 (0.1884)
			90%	7 (1.0000)	0 (0.0000)	13 (0.1884)
		SL-Chang	95%	5 (0.7143)	2 (0.2857)	13 (0.1884)
			90%	5 (0.7143)	2 (0.2857)	13 (0.1884)
2.	Eye data	SL-Satari/Di	95%	2 (1.0000)	0 (0.0000)	0 (0.0000)
			90%	2 (1.0000)	0 (0.0000)	0 (0.0000)
		SL-Chang	95%	2 (1.0000)	0 (0.0000)	0 (0.0000)
			90%	2 (1.0000)	0 (0.0000)	0 (0.0000)
3.	Sea star data	SL-Satari/Di	95%	1 (1.0000)	0 (0.0000)	0 (0.0000)
			90%	1 (1.0000)	0 (0.0000)	0 (0.0000)
		SL-Chang	95%	1 (1.0000)	0 (0.0000)	0 (0.0000)
			90%	1 (1.0000)	0 (0.0000)	0 (0.0000)
4.	Pigeon data	SL-Satari/Di	95%	2 (1.0000)	0 (0.0000)	0 (0.0000)
			90%	2 (1.0000)	0 (0.0000)	0 (0.0000)
		SL-Chang	95%	2 (1.0000)	0 (0.0000)	0 (0.0000)
			90%	2 (1.0000)	0 (0.0000)	0 (0.0000)
5.	Frog data	SL-Satari/Di	95%	1 (1.0000)	0 (0.0000)	0 (0.0000)
			90%	1 (1.0000)	0 (0.0000)	0 (0.0000)
		SL-Chang	95%	1 (1.0000)	0 (0.0000)	0 (0.0000)
			90%	1 (1.0000)	0 (0.0000)	0 (0.0000)

The SL-Satari/Di distance successfully identified all outliers ($pout = 1$) in the Turtle dataset at 95% and 90% confidence intervals and does not misclassify outliers as inliers ($pmask = 0$). The SL-Chang distance successfully identified five outliers ($pout = 0.7143$) but misclassified two outliers (observations 75 and 76) as inliers, indicating a masking effect ($pmask = 0.2857$). It was also found that both the SL-Satari/Di and the SL-Chang distances misclassified 13 inliers (observations 61 to 73) as outliers, indicating a swamping effect ($pswamp = 0.1884$).

The SL-Satari/Di and SL-Chang distances successfully identified observations 10 and 17 as outliers ($pout = 1$) in the Eye dataset at 95% and 90% confidence intervals. There is no masking effect ($pmask = 0$), hence both algorithms do not misclassify any outlier as an inlier. There is also no swamping effect ($pswamp = 0$), implying that the procedures do not misclassify any inlier as an outlier.

On the other hand, the SL-Satari/Di and SL-Chang distances successfully identified observation 13 as an outlier ($pout = 1$) in the Sea Star dataset at 95% and 90% confidence intervals. There is no masking effect ($pmask = 0$), hence both algorithms do not misclassify any outlier as an inlier. There is also no swamping effect ($pswamp = 0$), implying that the procedures do not misclassify any inlier as an outlier.

Meanwhile, the SL-Satari/Di and SL-Chang distances successfully identified observations 1 and 15 as outliers ($pout = 1$) in the Pigeon dataset at 95% and 90% confidence intervals. There is no masking effect ($pmask = 0$), hence both algorithms do not misclassify any outlier as an inlier. There is also no swamping effect ($pswamp = 0$), implying that the procedures do not misclassify any inlier as an outlier.

Lastly, the result of performance for Frog dataset based on three measurements in Table 3 shows that, the SL-Satari/Di and SL-Chang distances successfully identified observation 14 as an outlier ($pout = 1$) in the Frog dataset at 95% and 90% confidence intervals. There is no masking effect ($pmask = 0$), hence both algorithms do not misclassify any outlier as an inlier. There is also no swamping effect ($pswamp = 0$), implying that the procedures do not misclassify any inlier as an outlier.

Table 4 shows a summary of the best single-linkage algorithms for detecting outliers for each univariate circular biological dataset.

Table 4: Summary of the best single-linkage algorithm for five datasets

Datasets	All outliers successfully detected (*pout*)		No outliers falsely detected as inliers (*pmask*)		No inliers falsely detected as outliers (*pswamp*)	
	95%	90%	95%	90%	95%	90%
Frog data	All	All	All	All	All	All
Turtle data	SL-Satari/Di	SL-Satari/Di	SL-Satari/Di	SL-Satari/Di	None	None
Eye data	All	All	All	All	All	All
Sea star data	All	All	All	All	All	All
Pigeon data	All	All	All	All	All	All

Except for the Turtle dataset, the SL-Satari/Di and SL-Chang distances successfully detected all outliers in the remaining four datasets at 95% and 90% confidence intervals. Only SL-Satari/Di on both confidence intervals of cutting height successfully detected all outliers in the Turtle dataset with a sample size of 76 and 9% outliers. Similarly, neither single-linkage algorithm has a masking effect, except for the Turtle dataset, where SL-Chang misclassified two outliers as inliers. For the Turtle dataset, both single-linkage algorithms misclassified inliers as outliers, resulting in a swamping effect. The other four datasets, on the other hand, show no swamping effect for both SL-Satari/Di and SL-Chang at 95% and 90% confidence intervals.

6. Conclusion

This study found that the similarity distance measure for univariate circular data, namely Di distance which utilised Euclidean distance, has a similar effect to the Satari distance, which has been used as city-block distance. In particular, the proposed clustering-based procedure with single-linkage and Satari/Di and Chang distances are both practical to detect outliers in univariate circular biological datasets at various points of cutting height. The performance of each clustering-based outlier detection procedure was evaluated by calculating the probability of successfully detecting all outliers, the probability of misclassifying outliers as inliers (masking effect), and the probability of misclassifying inliers as outliers (swamping effect). The ideal clustering-based outlier detection procedure would detect all outliers with the least amount of masking and swamping effects.

This study also found that all clustering-based procedures using SL-Satari/Di and SL-Chang distances successfully detected all outliers in the Eye, Sea star, Pigeon and Frog datasets with no masking and swamping effects at 95% and 90% of confidence intervals. For the Turtle dataset, the SL-Satari/Di distance detected all outliers with no masking effect but has a swamping effect at the 95% and 90% of confidence intervals. On the other hand, SL-Chang did not successfully detect all outliers at both 95% and 90% of confidence intervals with both masking and swamping effects occurring when detecting the outliers in the Turtle dataset. Under certain data conditions, the SL-Satari/Di distance appears to outperform the SL-Chang distance.

In conclusion, the proposed clustering-based procedure based on single-linkage algorithms with Satari/Di and Chang distances is a practical and promising approach for detecting outliers in univariate circular data, particularly biological data. We propose that this procedure be extended using other agglomerative clustering algorithms with different data conditions. A simulation study can aid in the confirmation of findings in various scenarios and conditions.

Acknowledgement

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: RDU190363, RDU1901168 and PGRS210328).

References

1. Abuzaid, A. H. (2012). Analysis of Mother’s Day celebration via circular statistics. The Philippine Statistician, 61(2), 39–52.
2. Abuzaid, A. H. (2013). On the Influential Points in the Functional Circular Relationship Models. Pakistan Journal of Statistics and Operation Research, 9(3), 333–342.
3. Abuzaid, A. H. (2020). Identifying density-based local outliers in medical multivariate circular data. Statistics in Medicine, 1–6.
4. Abuzaid, A. H., Hussin, A. G., Rambli, A., & Mohamed, I. (2012). Statistics for a New Test of Discordance in Circular Data. Communications in Statistics—Simulation and Computation, 41, 1882–1890.
5. Abuzaid, A. H., Mohamed, I. B., & Hussin, A. G. (2009). A New Test of Discordancy in Circular Data. Communications in Statistics - Simulation and Computation, 38(4), 682–691.
6. Ahmed, H. I. E. S., Abuzaid, A. H., & Awar, I. I. Al. (2019). Detection of Outliers in Circular Data using Kernel Density Function. Life Sciences: An International Journal (LSIJ), 1(1), 1–11.
7. Alkasadi, N. A., Abuzaid, A. H. M., Ibrahim, S., & Yusoff, M. I. (2018). Outliers Detection in Multiple Circular Regression Model via DFBETAc Statistic. International Journal of Applied Engineering Research, 13(11), 9083–9090.
8. Chang-chien, S., Hung, W., & Yang, M.-S. (2012). On mean shift-based clustering for circular data. Soft Comput, 16, 1043–1060.
9. Collett, D. (1980). Outliers in Circular Data. Journal of the Royal Statistical Society, 29(1), 50–57.
10. 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. AIP Conference Proceedings, 1842.
11. Fisher, N. I. (1993). Statistical Analysis in Circular Data. Cambridge University Press.
12. Hung, W. L., Chang-Chien, S. J., & Yang, M. S. (2012). Self-updating clustering algorithm for estimating the parameters in mixtures of von Mises distributions. Journal of Applied Statistics, 39(10), 2259–2274.
13. Jammalamadaka, S. R., & Sengupta, A. (2001). Topics in Circular Statistics. World Scientific Publishing Co. Pte. Ltd. P.
14. Johnson, R., & Wichern, D. (2014). Applied Multivariate Statistical Analysis (Sixth). Pearson.
15. Klutchnikoff, N., Poterie, A., & Rouviere, L. (2021). Statistical analysis of a hierarchical clustering algorithm with outliers. HAL Open Science.
16. Mahmood, E. A., Rana, S., Hussin, A. G., & Midi, H. (2017). Adjusting Outliers in Univariate Circular Data. Pertanika J. Sci. & Technol. 25, 25(4), 1147–1158.
17. Ott, L., Pang, L., Ramos, F. T. & Chawla, S. (2014). On integrated clustering and outlier detection. Advances in Neural Information Processing Systems, 1359-1367.
18. Rambli, A. (2015). A half-circular distribution and outlier detection procedures in directional data. PhD Thesis. University of Malaya.

19. Satari, S. Z. (2015). Parameter Estimation and Outlier Detection for Some Types of Circular Model. PhD Thesis. University of Malaya.
20. Satari, S. Z., Muhammad Di, N. F., Zubairi, Y. Z., & Hussin, A. G. (2021). Comparative Study of Clustering-Based Outliers Detection Methods in Circular- Circular Regression Model. Sains Malaysiana, 50(6), 1787–1798.
21. 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.
22. Zulkipli, N. S., Satari, S. Z., & Yusoff, W. N. S. W. (2020). Descriptive analysis of circular data with outliers using Python programming language. Data Analytics and Applied Mathematics (DAAM), 01(01), 31–36.