Sains Malaysiana 52(8)(2023): 2431-2451
http://doi.org/10.17576/jsm-2023-5208-19

A New Single Linkage Robust Clustering Outlier Detection Procedures for Multivariate Data

(Suatu Prosedur Baharu Pengesanan Data Terpencil Berasaskan Pengelompokan Rangkaian Tunggal Teguh bagi Data Multivariat)

SHARIFAH SAKINAH SYED ABD MUTALIB[1,2], SITI ZANARIAH SATARI[1,*] & WAN NUR SYAHIDAH WAN YUSOFF[1]

[1]Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia
[2]Faculty of Computer, Media and Technology Management, University College TATI, Jalan Panchur, Telok Kalong, 24000 Kemaman, Terengganu, Malaysia

Received: 3 January 2023/Accepted: 1 August 2023

ABSTRACT

Outliers are abnormal data, and the detection of outliers in multivariate data has always been of interest. Unlike univariate data, outlier detection for multivariate data is insufficient with a visual inspection. In this study, we developed a new single linkage robust clustering outlier detection procedure for multivariate data. A robust estimator, Test on Covariance (TOC) is used to robustified the similarity distance measure, producing robust single linkage clustering. The performance of the new single linkage robust clustering outlier detection procedure is investigated via a simulation study using three outlier scenarios and historical multivariate datasets as illustrative examples. Three performance measures are used, which are *pout*, *pmask,* and *pswamp*. The performance of the new single linkage robust clustering procedure also compared with single linkage clustering using Euclidean and Mahalanobis distances as similarity distance measures as well as TOC. It is found that the new single linkage robust clustering procedure performs well in Outlier Scenario 3 when the mean and covariance matrix are shifted. The new procedure also performs well by successfully detecting all outliers, does not have masking effects in two out of five datasets and does not have swamping effect in all datasets. In conclusion, the new single linkage robust clustering outlier detection procedure is a practical and promising approach and good for simultaneously identifying multiple outliers in multivariate data.

Keywords: Multivariate data; outliers; single linkage clustering; Test on Covariance; robust clustering

ABSTRAK

Data terpencil ialah data tidak normal dan pengesanan data terpencil untuk data multivariat sentiasa menjana minat. Tidak seperti data univariat, pengesanan data terpencil untuk data multivariat tidak mencukupi dengan pemeriksaan visual. Dalam kajian ini, kami membangunkan satu prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh bagi data multivariat. Penganggar teguh, *Test on Covariance* (TOC) digunakan untuk meneguhkan ukuran jarak persamaan, menghasilkan pengelompokan rangkaian tunggal teguh. Prestasi prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh disiasat melalui kajian simulasi menggunakan tiga senario data terpencil dan set data sedia ada multivariat sebagai contoh ilustrasi. Tiga ukuran prestasi digunakan, iaitu *pout*, *pmask* dan *pswamp*. Prestasi prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh juga dibandingkan dengan pengelompokan rangkaian tunggal menggunakan jarak *Euclidean* dan *Mahalanobis* sebagai ukuran jarak persamaan beserta *TOC*. Didapati bahawa prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh berprestasi baik dalam Senario Data Terpencil 3 apabila min dan matriks kovarians dianjakkan. Prosedur baru juga berfungsi dengan baik apabila berjaya mengesan semua data terpencil dan tidak mempunyai kesan *masking* dalam 2 daripada 5 set data dan tidak mempunyai kesan *swamping* dalam semua set data. Kesimpulannya, prosedur baru pengesanan data terpencil berasaskan pengelompokan rangkaian tunggal teguh ialah pendekatan yang praktikal dan menjanjikan, serta bagus untuk mengesan data terpencil yang berkelompok secara serentak dalam data multivariat.

Kata kunci: Data multivariat; data terpencil; pengelompokan rangkaian tunggal; pengelompokan teguh; *Test on Covariance*

Introduction

Outliers are any abnormal or minority data points that differ from the majority of the data. Human or mechanical error, among other things, are potential causes of outliers (Rousseeuw & Hubert 2011; Wang, Bah & Hammad 2019). Outlier detection is an area of study in a multivariate analysis that has generated interest over the years. In contrast to outlier detection in univariate data, which can be accomplished using visual inspection, visual inspection of multivariate data is insufficient to detect outliers.

Robust distance has been widely used to detect outliers for multivariate data. The classical mean and covariance matrix in Mahalanobis distance is replaced with robust estimators to obtain robust distance. The reason for using robust estimator in detecting outliers instead of the classical estimator is that the classical mean and covariance matrix have masking and swamping effect (Hadi, Rahmatullah Imon & Werner 2009; Rousseeuw & Hubert 2011; Werner 2003). However, robust distance only looks for outliers at a single point, whereas outliers may appear in groups or are clustered (Garcia-Escudero et al. 2010). As a result, many studies have suggested using clustering methods, also called cluster-based outlier detection, to find multiple outliers in multivariate data (Christy, Gandhi & Vaithyasubramanian 2015; Ijaz, Attique & Son 2020; Siti Zanariah et al. 2021).

The clustering method aims to group objects into clusters where the observations are most similar and dissimilar. Outliers in clustering are defined as observations that are far from any clusters or are distantly located from each cluster's centre (Hardin & Rocke 2004; Zhang 2013). According to Aggarwal (2017), a complementary relationship exists between clustering and outlier detection. Studies that used clustering to detect outliers have been done by Almeida et al. (2007), Duan et al. (2009), Jiang, Tseng and Su (2001), and Yoon, Kwon and Bae (2007). The classical clustering methods, however, have a sensitivity to outliers, noise and can have serious issues (García-Escudero et al. 2008; Saxena et al. 2017). Therefore, to solve these issues, robust clustering methods have been proposed. Numerous studies, including Balcan, Liang and Gupta (2014), Dotto et al. (2018), and Olukanmi and Twala (2017), have suggested robust clustering. However, only Balcan, Liang and Gupta (2014) use single linkage as one of the methods in their studies. Nevertheless, the scope of Balcan, Liang and Gupta (2014)'s study is to robustify their algorithm in the presence of noise. In contrast, Dotto et al. (2018) and Olukanmi & Twala (2017) robustify center-based clustering methods to detect outliers. Each study has a different scope and uses different clustering methods, and robust clustering methods are still evolving, such as in Peña (2018) and Sharma and Seal (2021).

Therefore, this study develops a new robust clustering procedure based on single linkage clustering to detect outliers for multivariate data. There are various clustering methods, as can be seen in Gan, Ma and Wu (2007), Saxena et al. (2017), and Xu and Tian (2015) and we choose to use single linkage clustering in detecting outliers for multivariate data. The single linkage method gives the most accurate picture of the structure data compared to the other linkage methods and is the easiest mathematically in clusters (Sebert, Montgomery & Rollier 1998; Siti Zanariah 2015). Studies that used a single linkage method to detect outliers for multivariate data can be seen in Almeida et al. (2007) and Melendez-Melendez et al. (2019). The single linkage clustering will be robustified using robust distance as a similarity measure. A new robust estimator proposed by Abd Mutalib, Satari and Yusoff (2019) is used to calculate the similarity distance measure. This new procedure is named single linkage robust clustering outlier detection procedure. The new procedure will then be tested to detect outliers in simulated and historical datasets. The new single linkage robust clustering procedure will also be compared with the single linkage clustering procedure using existing similarity distance measures, Euclidean and Mahalanobis distances. For abbreviation, we named each single linkage clustering method: New single linkage robust clustering (RDT-SL), single linkage using Euclidean distance (ED-SL), and single linkage using Mahalanobis distance (MD-SL).

Test on Covariance Estimator

Test on Covariance (TOC), a new robust estimator developed by Abd Mutalib, Satari and Yusoff (2019), is less sensitive to the presence of outliers. The performance of TOC has been examined by Abd Mutalib, Satari and Yusoff (2021a, 2021b) and Abd Mutalib, Satari and Yusoff (2019), using simulation studies and historical multivariate datasets. The studies have produced promising results, showing that TOC can successfully identify outliers in multivariate datasets. TOC's performance has been compared to that of other existing robust estimators, and under certain conditions,

TOC can outperform or be comparable to other robust estimators.

The motivation for the development of TOC came from Salleh (2013), who emphasized that further research is needed to determine the conditions under which two covariance matrices are equal. Covariance Matrix Equality (CME) and Index Set Equality (ISE) are the two new robust estimators that have been developed by Salleh (2013) due to the issue with Minimum Vector Variance (MVV). In addition, MVV has been proposed as a solution to the Fast Minimum Covariance Determinant (FMCD) problem.

CME, ISE and MVV basically modified the final step of the FMCD algorithm. The equality between two covariances is tested in the final stage of the FMCD, CME, MVV, and ISE algorithm. Based on these ideas, TOC also modified the final step in the FMCD algorithm. The TOC algorithm is given as follows. Step 1: Select an arbitrarily subset H_{old} containing h different observations, where h is the smallest integer $\geq (n+p+1)/2$, where p is the number of variables and n is sample size. Step 2: Compute the mean vector $\bar{X}_{H_{old}}$ and covariance matrix $S_{H_{old}}$ of all observations belonging to H_{old} . Step 3: Compute $d^2_{H_{old}}(i) = \left(X_i - \bar{X}_{H_{old}}\right)' S^{-1}_{H_{old}} \left(X_i - \bar{X}_{H_{old}}\right)$ for i = 1,2, ..., n. Step 4: Sort $d^2_{H_{old}}(i)$ for $i = 1,2,\ldots,n$ in increasing order $d^2_{H_{old}}(\pi(1)) \leq d^2_{H_{old}}(\pi(2)) \leq \ldots \leq d^2_{H_{old}}(\pi(n))$ where π is a permutation on i = 1,2, ..., n. Step 5: Define $H_{new} = \{X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(h)}\}$ and then calculate $\bar{X}_{H_{new}}$, $S_{H_{new}}$ and $d^2_{H_{new}}(i)$ for $i = 1,2,\ldots,n$. Step 6 $_{\text{TOC}}$: If H_0 is rejected, calculate $\bar{X}_{H_{new}}$ and let $H_{old} := H_{new}$, $\bar{X}_{H_{old}} := \bar{X}_{H_{new}}$ and $S_{H_{old}} := S_{H_{new}}$. Then go to Step 3. Otherwise, the process is stopped.

The robust estimator of TOC is obtained by testing the equality of two covariance structures. Equation (1) is used to test the hypothesis of $H_0 : \Sigma_{old} = \Sigma_{new}$ versus $H_1 : \Sigma_{old} \neq \Sigma_{new}$.

$$u = \upsilon \left[\sum_{i=1}^{p} (\lambda_i - \ln \lambda_i) - p \right] \tag{1}$$

where υ = n - 1, $p = 1,2,\ldots,k$ and $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $\Sigma_{new}\Sigma^{-1}_{old}$. H_0 is rejected if $u > \chi^2 \left[\alpha, \frac{1}{2} p(p+1) \right]$ as stated in (Rencher 2002). The final robust estimator of TOC will be used to calculate robust distance as similarity measures for the new procedure.

SIMILARITY MEASURES FOR MULTIVARIATE DATA

A measure of how close the observations are to each other is needed since clustering aims to group related observations into one group. Distance is a convenient measurement, and this measurement is frequently referred to as a similarity measure. This study uses three distances as similarity measures: Euclidean distance (ED), Mahalanobis distance (MD), and robust distance using TOC (RDT). ED and MD are existing similarity measures that are widely used in clustering to detect outliers in multivariate data such as in studies done by Badaró et al. (2021), Evans, Love and Thurston (2015), Melendez-Melendez et al. (2019), and Yesilbudak (2016). ED is usually used in clustering because ED is easy to compute and interpret, however ED does not take into account the correlation in the data (De Maesschalck, Jouan-Rimbaud & Massart 2000). In multivariate data, the distance of an observation from the centre and the shape of the data must be considered (Cabana Lillo & Laniado 2021). The covariance matrix characterizes the shape of multivariate data, and the MD is a well-known measure that takes it into account (Cabana, Lillo & Laniado 2021; De Maesschalck, Jouan-Rimbaud & Massart 2000).

The ED between observations i and j is defined as

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2} \tag{2}$$

where d_{ij} is the distance between observations i and j; x_{ik} is the value for ith observation of the kth variable; x_{jk} is the value for jth observation of the kth variable; and p is the number of variables. While the MD between observations i and j is defined as

$$d_{ij,MD} = \sqrt{\left(\mathbf{x}_i - \mathbf{x}_j\right)\mathbf{S}^{-1}\left(\mathbf{x}_i - \mathbf{x}_j\right)^T} \tag{3}$$

where $d_{ij,MD}$ is the distance between observations i and j; $\mathbf{x}_i$ and $\mathbf{x}_j$ are two data points in p-dimensional space; and $\mathbf{S}^{-1}$ is the inverse sample covariance matrix for the dataset.

Based on MD, a new robust distance is proposed using TOC estimator (RDT). The sample covariance matrix of TOC is obtained as follows,

$$\mathbf{S}_{TOC} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \tag{4}$$

Let us consider data with two observations and two variables.

$$\mathbf{x} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

In the case of two variables, the sample covariance matrix for TOC is

$$\mathbf{S}_{TOC} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} s_1^2 & r_{12}s_1s_2 \\ r_{21}s_2s_1 & s_2^2 \end{bmatrix}$$

with

$$\mathbf{S}_{TOC}^{-1} = \begin{bmatrix} s_2^2/\det(\mathbf{S}_{TOC}) & -r_{12}s_1s_2/\det(\mathbf{S}_{TOC}) \\ -r_{21}s_2s_1/\det(\mathbf{S}_{TOC}) & s_1^2/\det(\mathbf{S}_{TOC}) \end{bmatrix} \quad (5)$$

where $\det(\mathbf{S}_{TOC}) = s_1^2 s_2^2 (1 - r_{12}^2)$ is the determinant of the covariance matrix. The RDT between observations 1 and 2 is,

$$\left[(x_{11} - x_{21}) \ (x_{12} - x_{22}) \right]$$

$$\begin{bmatrix} s_2^2/\det(\mathbf{S}_{TOC}) & -r_{12}s_1s_2/\det(\mathbf{S}_{TOC}) \\ -r_{21}s_2s_1/\det(\mathbf{S}_{TOC}) & s_1^2/\det(\mathbf{S}_{TOC}) \end{bmatrix} \begin{bmatrix} (x_{11} - x_{21}) \\ (x_{12} - x_{22}) \end{bmatrix}$$

$$= \frac{s_2^2 (x_{11} - x_{21})^2 - 2(x_{11} - x_{21})(x_{12} - x_{22}) r_{12}s_1s_2 + s_1^2 (x_{12} - x_{22})^2}{\det(\mathbf{S}_{TOC})}$$

$$= \frac{\left(s_2^2 (x_{11} - x_{21})^2 (1 - r_{12}^2) + s_1^2 (x_{12} - x_{22})^2 \right.}{\left. - 2(x_{11} - x_{21})(x_{12} - x_{22}) r_{12}s_1s_2 + s_2^2 (x_{11} - x_{21})^2 r_{12}^2 \right)}{s_1^2 s_2^2 (1 - r_{12}^2)}$$

$$= \frac{(x_{11} - x_{21})^2}{s_1^2} + \frac{(x_{12} - x_{22})^2}{s_2^2 (1 - r_{12}^2)} - \frac{2(x_{11} - x_{21})(x_{12} - x_{22}) r_{12}}{s_1s_2 (1 - r_{12}^2)} + \frac{(x_{11} - x_{21})^2 r_{12}^2}{s_1^2 (1 - r_{12}^2)}$$

$$= \frac{(x_{11} - x_{21})^2}{s_1^2} + \left(\frac{(x_{12} - x_{22})}{s_2\sqrt{1 - r_{12}^2}} - \frac{r_{12}(x_{11} - x_{21})}{s_1\sqrt{1 - r_{12}^2}} \right)^2$$

so that,

$$d_{12,TOC} = \sqrt{\left(\frac{x_{11} - x_{21}}{s_1} \right)^2 + \left[\left\{ \left(\frac{x_{12} - x_{22}}{s_2} \right) - r_{12} \left(\frac{x_{11} - x_{21}}{s_1} \right) \right\} \frac{1}{\sqrt{1 - r_{12}^2}} \right]^2} \quad (6)$$

Equation (6) demonstrates the subtraction of the portion of the second variable that is already explained by the first variable which means, RDT corrects for the correlation within the data because RDT is based on MD (De Maesschalck, Jouan-Rimbaud & Massart 2000).

Therefore, based on the MD, a new robust similarity distance measure namely RDT is proposed as follows.

$$d_{ij,TOC} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)\mathbf{S}_{TOC}^{-1}(\mathbf{x}_i - \mathbf{x}_j)^T} \quad (7)$$

where $d_{ij,TOC}$ is the distance between observations i and j; $\mathbf{x}_i$ and $\mathbf{x}_j$ are two data points in p-dimensional space and $\mathbf{S}_{TOC}^{-1}$ is the inverse sample covariance matrix for TOC.

A distance matrix $(n \times n)$ is produced from these three distances and used as a similarity measure. This distance matrix is based on the number of observations, n. The performance of the RDT as a similarity measure is compared with ED and MD.

CUTTING RULE FOR OUTLIERS DETECTION

A dendrogram, also known as a cluster tree, can visually represent the outcomes of single linkage clustering from the similarity measure. In the tree diagram, the branches represent clusters. The nodes where the branches merge along a similarity axis show the level at which fusion occurs. When using hierarchical clustering, we can select N clusters from the dendrogram by cutting across the branches at a specific level of the similarity measure used by one of the axes. After applying a clustering algorithm, the user must decide how many groups there are in a dataset. To be more precise, cutting or dividing the cluster tree at a specific height is necessary and determines how many groups there will be (Sebert, Montgomery & Rollier 1998).

The method to choose the number of groups for hierarchical clustering procedures is provided by Mojena (1977). Mojena's cutting rule is given by $\bar{h} + \alpha * s_h$ where $\bar{h}$ is the average height for all N - 1 clusters; s_h is the unbiased standard deviation of the heights; and α is a specified constant. According to Mojena (1977), the values of α should fall between 2.75 and 3.50. The best overall performance for Mojena's cutting rule, according to a comprehensive study by Milligan and Cooper (1985), occurs when α is 1.25. Therefore, the cutting rule used in this study was equal to $\bar{h} + 1.25 * s_h$.

A NEW SINGLE LINKAGE ROBUST CLUSTERING OUTLIERS DETECTION PROCEDURE FOR MULTIVARIATE DATA

In this section, a new single linkage robust clustering procedure is developed and shown in Figure 1. The procedure starts with obtaining a robust covariance estimator TOC. Next, by using TOC, the similarity measure matrix between observations *i* and *j* using Equation (7) is obtained. In step 3, we used a single linkage clustering algorithm to cluster the observations. The steps for single linkage clustering as follow. Step 1: Starts with *N* clusters, each containing one multivariate observation. Step 2: Compute the similarity matrix, $\mathbf{D}=\{d_{ij}\}$. Another similarity measure was also obtained in Step 2 using ED and MD. The effectiveness of the new single linkage robust clustering is evaluated against single linkage clustering using ED and MD as similarity measures. Step 3: Find the smallest distance from the similarity measure matrix, $\mathbf{D}=\{d_{ij}\}$ and merge the corresponding cluster. Step 4: The entries are updated in the distance matrix by (a) deleting the rows and columns corresponding to the merged cluster, and (b) adding a row and column giving the distances between the merged cluster and the remaining clusters. Step 5: Repeat Steps 3 and 4 for *N*-1 times until all observations are in a single cluster after the algorithm terminates. A dendrogram or cluster tree is obtained from the single linkage clustering in Step 3. To determine the number of clusters, we used the cutting rule $\bar{h}+1.25*s_h$ in Step 4. After the dendrogram is cut at the specific height, we can identify the outliers and inliers. Inliers are determined when the cluster group contains the largest observations, and outliers are determined if the cluster group contains minority observations.

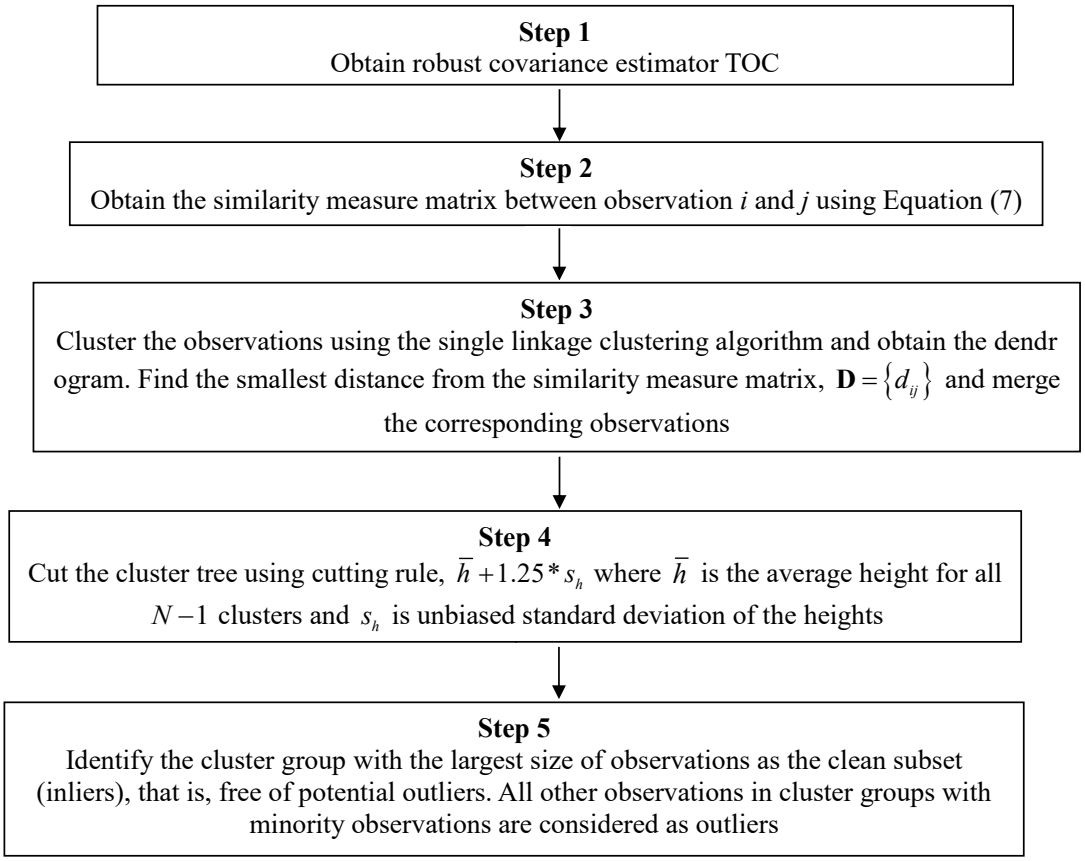


FIGURE 1. New single linkage robust clustering outlier detection procedure for multivariate data

PERFORMANCE MEASURES

Three measurements are used to evaluate the new single linkage robust clustering (RDT-SL) performance compared with the single linkage clustering using MD (MD-SL) and ED (ED-SL). The performance of TOC in detecting outliers is also added for comparisons. The measurements are *pout* (success probability), *pmask* (masking effect) and *pswamp* (swamping effect). Sebert, Montgomery and Rollier (1998) gave an illustration of how the performance ofthe methods was assessed. Siti Zanariah, Nur Faraidah Muhammad Di and Roslinazairimah (2019) and Siti Zanariah et al. (2021) tested their proposed method for detecting outliers using the performance measures from Sebert, Montgomery and Rollier (1998).

Each performance measure's formula is shown in Equation (8) through (10). The probability of finding all outliers successfully (*pout*) is given in Equation (8).

$$pout = \frac{"success"}{s} \tag{8}$$

where "*success*" represents the number of data sets in which the method successfully identified all outliers, and *s* represents the total number of simulations. The probability that outliers are misidentified as inliers (*pmask*) is given in Equation (9).

$$pmask = \frac{"failure"}{(out)(s)} \tag{9}$$

where "*failure*" is the number of outliers detected as inliers in all datasets, and *out* is the number of outliers. The probability of inliers being identified as outliers (*pswamp*) is given in Equation (10).

$$pswamp = \frac{"false"}{(n-out)(s)} \tag{10}$$

where "*false*" represents the number of inliers in all data sets identified as outliers, and *n* represents the sample size.

The range of values for *pout*, *pmask,* and *pswamp* is 0 to 1. The best similarity measurements will have the highest value of *pout* as the value approaches one and the lowest value of *pmask* and *pswamp* when the value approaches zero (Santos-Pereira & Pires 2002; Satari, Muhammad Di & Zakaria 2019). Masking is a more serious problem than swamping and should be avoided. Studies by Abd Mutalib, Satari and Yusoff (2021a, 2021b) investigated the performance of TOC n = 30,50,100,200 and p = 2,3,5 for various outliers percentage and outlier scenarios. Both studies found that TOC performs very well for *pswamp* where TOC became the best estimator and showed similar performance with other robust estimators for more than half of the studied cases. TOC also has the lowest probability of misclassifying inliers as outliers regardless of the distance between outliers and inliers. However, TOC does not perform very well for *pout* and *pmask*. TOC only performs very well and similarly to other robust estimators when the distance between outliers and inliers increases.

SIMULATION STUDY

The simulation study was conducted using the R statistical package. To represent small and moderate sample sizes, the sample sizes used are n = 30,50 and 100. For each sample size, three different number of variables, p = 3,5, and 10 are used. The percentage of outliers will be ε = 0.05,0.1,0.15,0.2, and 0.25. The chosen sample sizes, number of variables and percentage of outliers were taken from Cabana, Lillo and Laniado (2021), Cerioli, Riani and Torti (2011), Fauconnier and Haesbroeck (2009), Filzmoser, Maronna and Werner (2008), Herwindiati, Djauhari and Mashuri (2007), Kosinski (1999), Rocke and Woodruff (1996), and Wada, Kawano and Tsubaki (2020). .

Each combination of *n, p* and ε is set for three outlier scenarios. The outlier scenarios are generated from the following mixture *p*-variate normal distributions (Filzmoser, Maronna & Werner 2008; Herwindiati, Djauhari & Mashuri 2007; Kalina & Tichavský 2021),

$$(1-\varepsilon)N_p(\vec{\mu}_0,\Sigma_0)+\varepsilon N_p(\lambda\vec{\mu}_1,\delta\Sigma_1) \tag{11}$$

Where $\Sigma_0=\Sigma_1=I_p$, $\vec{\mu}_0=(0\ 0\ldots0)'$ and $\vec{\mu}_1=(1\,1\ldots1)'$ is
of dimension p. Outliers are generated from $N_p(\lambda\vec{\mu}_1,\delta\Sigma_1)$ and inliers are generated from $N_p(\vec{\mu}_0,\Sigma_0)$. The coefficients λ and δ in Equation (11) will determine the outlier scenarios.

Outlier Scenario 1 is determined by the values of λ with a fixed value of δ = 1and is given in Equation (12). In Outlier Scenario 1, the values of λ show the separation between outliers and inliers by shifting the mean.

$$(1-\varepsilon)N_p(\vec{\mu}_0,\Sigma_0)+\varepsilon N_p(\lambda\vec{\mu}_1,\Sigma_1) \tag{12}$$

While Outlier Scenario 2 is determined by the values of δ with a fixed value of λ and is given as Equation (13). Covariance matrix for outliers has a different covariance matrix than the rest of the data in this outlier scenario (Filzmoser, Maronna & Werner 2008). The values of δ show the separation between outliers and inliers determined by shifting the covariance.

$$(1-\varepsilon)N_p(\vec{\mu}_0,\Sigma_0)+\varepsilon N_p(0,\delta\Sigma_1) \quad (13)$$

The last outlier scenario is Outlier Scenario 3, which combines Outlier Scenarios 1 and 2. In this outlier scenario, both mean and covariance are shifted. Outlier Scenario 3 is determined by the value of λ and δ and is given in Equation (14).

$$(1-\varepsilon)N_p(\vec{\mu}_0,\Sigma_0)+\varepsilon N_p(\lambda\vec{\mu}_1,\delta\Sigma_1) \quad (14)$$

Results and Discussion

A selection of the simulation results and the plots for performance measures are shown in Table 1 and Figure 2, respectively. Table 1 shows the performance measures for RDT-SL, ED-SL, MD-SL and TOC for Outlier Scenario 1 with $p = 3$. Findings in Table 1 shows that for any fixed values of ε, n, and p, the *pout* values for all single linkage clustering increase when the values of λ increase. It demonstrates that as the distance between outliers and inliers increases by shifting the mean, all single linkage clustering perform better at identifying outliers. Similar results were also obtained for TOC where the *pout* values increased as the values of λ increased for any fixed values of ε, n, and p.

From the simulation study, the *pout* values increase when the values of λ and δ increase for any fixed values of ε, n, and p in Outlier Scenario 2 and 3. These findings show that as the distance between outliers and inliers increases by shifting the covariance and the mean and covariance simultaneously, all single linkage clusterings perform better at identifying outliers. ED-SL and RDT-SL have shown the *pout* values 1.0000 for Outlier Scenarios 1 and 3 for certain conditions. However, no single linkage clustering has *pout* value of 1.0000 in Outlier Scenario 2. These results show that ED-SL and RDT-SL successfully detect all outliers when shifting the mean, and mean and covariance matrix simultaneously for certain conditions.

From Figure 2, the line pattern of the *pout* approaching one as the values of λ increase for all single linkage clusterings in Outlier Scenario 1. The same pattern has also been found in Outlier Scenarios 2 and 3. It can be seen that the *pout* values are low as n increases for $\lambda = 2$. From Figure 2, for any fixed value of ε and p, it is observed that the smaller the n, the faster the point will be approaching one. These results indicate that all single linkage clustering have a high probability of detecting outliers when the sample size is small, as the number of outliers is small for any fixed percentage of outliers and number of variables.

The *pmask* values in Table 1 show that for any fixed values of ε , n, and p, the *pmask* values for all single linkage clustering decrease when the values of λ increase. It demonstrates that all single linkage clustering perform better to avoid misclassifying outliers as inliers as the distance between outliers and inliers increases by shifting the mean. The same findings were also found in Outlier Scenarios 2 and 3. All single linkage clustering are shown good performance by not misclassifying outliers as inliers as the distance between outliers and inliers increases by shifting the covariance and the mean and covariance simultaneously. TOC also shows similar results where the *pmask* values decreased as the values of λ increased for any fixed values of ε, n, and p.

From the simulation results for all Outlier Scenarios, it is also found that the *pmask* values are below 1.0000 for all single linkage clusterings. This indicates that all single linkage clusterings have a low probability of misclassifying outliers as inliers in all Outlier Scenarios. All single linkage clusterings have shown the *pmask* values 0.0000 for Outlier Scenario 1 for certain conditions. Only ED-SL and RDT-SL have shown the *pmask* values 0.0000 for Outlier Scenario 3. However, no single linkage clustering have *pmask* value of 0.0000 in Outlier Scenario 2. These results show that ED-SL and RDT-SL do not have masking errors when shifting the mean, and the mean and covariance matrix for certain conditions. While MD-SL does not have a masking error when shifting the mean for certain conditions.

Figure 2 shows the line pattern of the *pmask* approaching zero as the values of λ increases for all single linkage clusterings in Outlier Scenario 1. The same pattern also has been found in Outlier Scenarios 2 and 3. In particular, it can be seen that the *pmask* values are below 0.4000. From Figure 2, for any fixed value of ε and p, the smaller the n, the faster the point will approach zero. These results indicate that all single linkage clustering have low masking error when sample size is small as the number of outliers is small for any fixed percentage of outliers and number of variables.

Results from Table 1 show that the *pswamp* values for all single linkage clusterings decrease when the values of λ increase for any fixed value of ε, n, and p in Outlier Scenario 1. It demonstrates that, as the distance between outliers and inliers increases, all single linkage clusterings perform better at avoiding the misclassification of inliers as outliers. As the distance between outliers and inliers increases, the performance of single linkage clustering improves. The same results can be found in Outlier Scenarios 2 and 3. TOC also shows similar results where the *pswamp* values decreased as the values of λ increased for any fixed values of ε, n, and p.

From Figure 2, the *pswamp* values gradually decrease as the values of λ increase. In general, the *pswamp* values are relatively small and less than 0.1000 regardless of how large ε, n, and $p$for Outlier Scenario 1. The *pswamp* values for Outlier Scenario 2 are below 0.2000 and for Outlier Scenario 3 are below 0.1000.

TABLE 1. The performance measures of different single linkage clustering and TOC in Outlier Scenario 1 (p= 3)

ε	n	λ	*pout*				*pmask*				*pswamp*			
			ED-SL	MD-SL	RDT-SL	TOC	ED-SL	MD-SL	RDT-SL	TOC	ED-SL	MD-SL	RDT-SL	TOC
0.05	30	2	**0.6496**	0.5673	0.6441	0.9587	**0.2046**	0.2623	0.2063	0.0208	**0.0689**	0.0778	0.0701	**0.1924**
		4	**0.9989**	0.9943	0.9985	**1.0000**	**0.0006**	0.0030	0.0008	**0.0000**	**0.0095**	0.0322	0.0163	**0.1526**
	50	2	**0.5509**	0.4696	0.3848	0.8093	**0.2031**	0.2500	0.3085	0.0684	**0.0746**	0.0802	0.0884	**0.1550**
		4	**0.9993**	0.9966	0.9989	**1.0000**	**0.0003**	0.0012	0.0004	**0.0000**	**0.0191**	0.0390	0.0213	0.1437
	100	2	0.4224	0.3637	**0.4328**	**0.7383**	0.1874	0.2163	**0.1826**	**0.0593**	0.0815	0.0841	**0.0810**	**0.1559**
		4	**0.9996**	0.9987	**0.9996**	**1.0000**	**0.0001**	0.0003	**0.0001**	**0.0000**	**0.0351**	0.0493	0.0371	**0.1246**
0.1	30	2	0.4902	0.3525	**0.4977**	0.1815	0.2561	0.3509	**0.2460**	0.4357	**0.0659**	0.0760	0.0670	0.2264
		4	**0.9982**	0.9971	0.9974	**1.0000**	**0.0006**	0.0100	0.0009	**0.0000**	**0.0102**	0.0370	0.0120	0.1632
	50	2	**0.3639**	0.2372	0.3157	0.3013	**0.2588**	0.3428	0.2893	0.2116	**0.0713**	0.0767	0.0738	**0.1207**
		4	0.9991	0.9848	**0.9995**	**1.0000**	0.0002	0.0046	**0.0001**	**0.0000**	0.0213	0.0448	**0.0176**	**0.1029**
	100	2	**0.1857**	0.1019	0.1075	0.1815	**0.2583**	0.3357	0.3308	0.1580	**0.0773**	0.0799	0.0799	0.1518
		4	**0.9991**	0.9895	0.9985	**1.0000**	**0.0001**	0.0016	0.0001	**0.0000**	**0.0380**	0.0541	0.0430	**0.1072**
0.15	30	2	**0.3333**	0.1351	0.2821	0.1027	**0.3326**	0.5069	0.3757	0.3655	**0.0657**	0.0765	0.0693	0.1551
		4	**0.9978**	0.9050	0.9977	**1.0000**	**0.0006**	0.0455	0.0006	**0.0000**	0.0124	0.0476	**0.0116**	**0.1131**
	50	2	**0.2160**	0.0842	0.1272	**0.0000**	**0.3397**	0.4846	0.4325	**0.1230**	**0.0713**	0.0772	0.0759	0.1542
		4	**0.9977**	0.9436	0.9960	**0.3494**	**0.0004**	0.0198	0.0006	**0.0000**	**0.0242**	0.0513	0.0291	**0.1211**
	100	2	**0.0910**	0.0301	0.0740	0.0281	**0.3478**	0.4688	0.3771	0.2108	**0.0779**	0.0805	0.0784	**0.1185**
		4	**0.9988**	0.9728	0.9985	**0.9997**	**0.0001**	0.0060	0.0001	**0.0000**	**0.0408**	0.0579	0.0423	**0.0793**
0.2	30	2	**0.2752**	0.0836	0.0750	0.0004	**0.3785**	0.5775	0.5894	0.7230	**0.0680**	0.0804	0.0888	0.1489
		4	**0.9977**	0.8581	0.7783	**1.0000**	**0.0006**	0.0717	0.1171	**0.0000**	**0.0134**	0.0528	0.0580	0.1340
	50	2	**0.1772**	0.0423	0.0456	0.0065	**0.3978**	0.5772	0.5669	0.4010	**0.0739**	0.0799	0.0834	0.1579
		4	**0.9985**	0.9125	0.9961	0.5083	**0.0002**	0.0400	0.0005	0.0657	**0.0264**	0.0573	0.0319	**0.0783**
	100	2	**0.0564**	0.0077	0.0130	**0.0000**	**0.4552**	0.5988	0.5797	0.5384	**0.0794**	0.0822	0.0821	0.1034
		4	0.9979	0.9519	**0.9983**	**1.0000**	0.0001	0.0155	**0.0001**	**0.0000**	0.0441	0.0640	**0.0431**	0.0819
0.25	30	2	**0.2154**	0.0354	0.0582	0.1766	**0.4593**	0.6840	0.6496	0.1959	**0.0751**	0.0875	0.0897	**0.2745**
		4	**0.9975**	0.7312	0.9931	0.9999	**0.0006**	0.1680	0.0025	**0.0000**	**0.0175**	0.0650	0.0263	0.1213
	50	2	**0.1260**	0.0183	0.0370	**0.3749**	**0.4926**	0.6776	0.6379	**0.0733**	**0.0797**	0.0862	0.0850	**0.2478**
		4	**0.9979**	0.8508	0.9920	0.9550	**0.0003**	0.0829	0.0020	0.0035	**0.0304**	0.0661	0.0436	0.1866
	100	2	**0.0361**	0.0023	0.0008	0.1058	**0.5559**	0.6913	0.7049	0.0857	**0.0829**	0.0862	0.0879	0.2863
		4	**0.9981**	0.9185	0.9973	**1.0000**	**0.0001**	0.0382	0.0001	**0.0000**	**0.0488**	0.0704	0.0541	0.2148

In addition, Tables 2 - 4 summarise the best single linkage clustering for Outlier Scenarios 1, 2 and 3. All in these tables refer to all single linkage clustering. From Table 2, RDT-SL performs better in detecting outliers (*pout*) when the percentage of outliers is low and the distance between outliers and inliers, $\lambda = 4$. TOC show quite good performance in *pout* for all percentage of outliers for any fixed value of n and p when the percentage of outliers is low and the distance between outliers and inliers, $\lambda = 4$. The same pattern can also be seen in *pmask* for all single linkage clustering and TOC. For *pswamp* values, RDT-SL shows good performance when the percentage of outliers is between 5% and 20%. The performance of TOC in *pswamp* is good for all percentage of outliers for any fixed values of ε, n, λ and p. From Table 2, it is also found that TOC shows good performance compared to RDT-SL in *pout*, *pmask* and *pswamp* when the percentage of outliers is 25% for fixed values of n, λ and p.

The best single linkage clustering for Outlier Scenario 2 are summarised in Table 3. RDT-SL shows good performance in detecting outliers for all ε and δ. However, for $n = 100$, most of the single linkage clustering does not show good performance in detecting outliers. 'None' in Table 3 means that the *pout* value for all single linkage clustering is 0.0000, indicating that no outliers are detected. Overall, TOC shows good performance in *pout* for 5% to 20% of outliers. While for *pmask* value, RDT-SL shows good performance when $\delta = 2$ and $p = 3$ for all ε. This result indicates that RDT-SL has a low masking error when the number of variables is small and the distance between outliers and inliers is small by shifting the covariance matrix. TOC show good results for *pmask* when the percentage of outliers are 5% to 20% for any fixed values of ε, n, δ and p. However, for 25% of outliers, TOC only performs for $n = 30$. Lastly for *pswamp* values, RDT-SL and TOC shows good performance for all ε and δ.

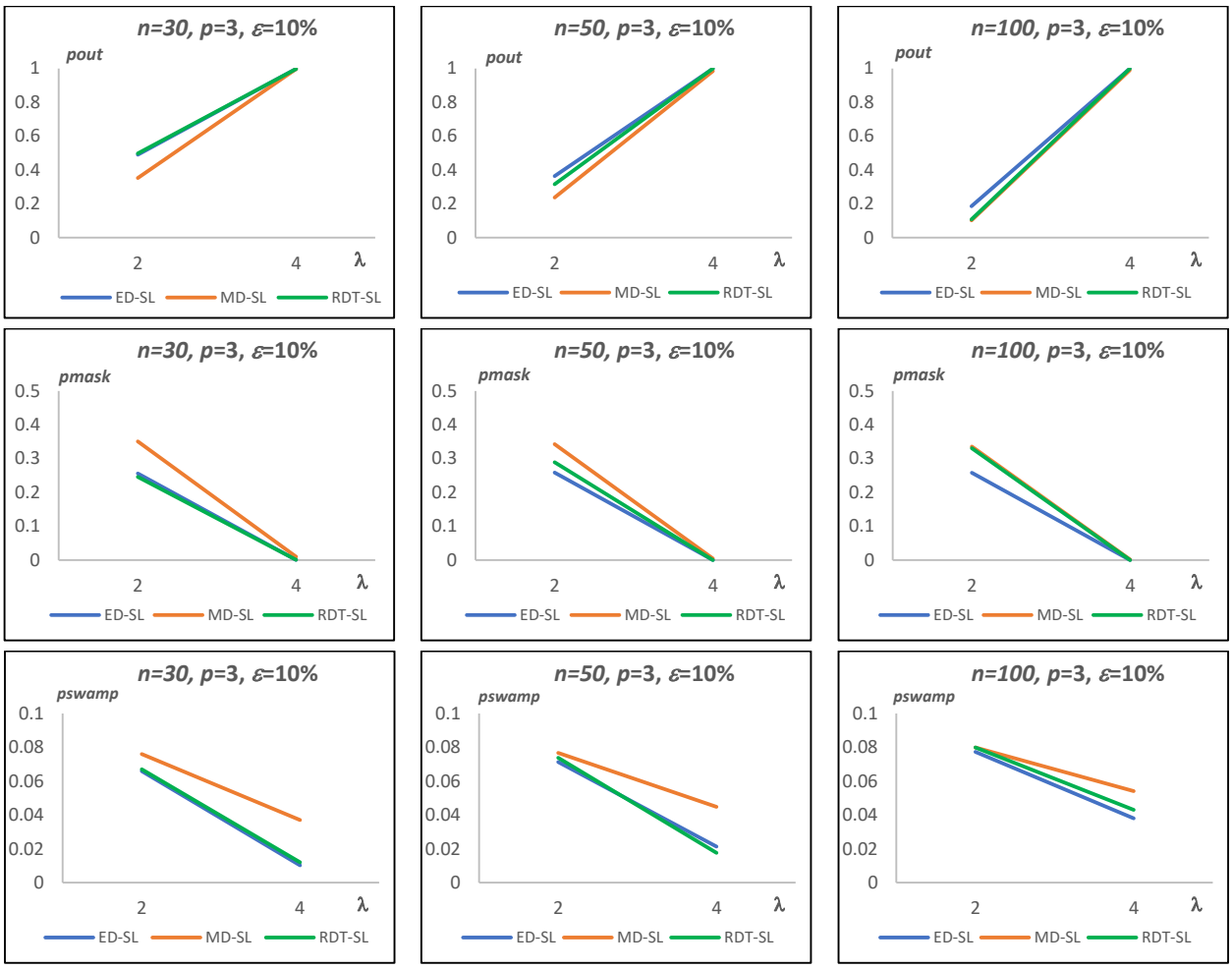


FIGURE 2. Plot of "success" probability (*pout*), masking error (*pmask*) and swamping error (*pswamp*) versus distance of outliers and inliers (λ) for all sample sizes, n with number of variable, $p = 3$ and percentage of outliers, $\varepsilon = 10\%$

TABLE 2. Comparison of the best single linkage clustering and TOC performance for Outlier Scenario 1

ε	n	λ	*pout*			*pmask*			*pswamp*		
			$p = 3$	$p = 5$	$p = 10$	$p = 3$	$p = 5$	$p = 10$	$p = 3$	$p = 5$	$p = 10$
0.05	30	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC
		4	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL,** TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC
	50	2	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL
		4	ED-SL, TOC	ED-SL, **RDT-SL,** TOC	ED-SL, **RDT-SL,** TOC	ED-SL, TOC	**ALL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL, TOC	ED-SL
	100	2	**RDT-SL**, TOC	ED-SL	ED-SL, TOC	**RDT-SL**, TOC	ED-SL	ED-SL, TOC	**RDT-SL**, TOC	**RDT-SL**, TOC	ED-SL
		4	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	**ALL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL
0.1	30	2	**RDT-SL**	ED-SL, TOC	ED-SL	**RDT-SL**	**RDT-SL**, TOC	ED-SL	ED-SL	**RDT-SL**	ED-SL
		4	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC
	50	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC
		4	**RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**	**RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**	**RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC
	100	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL
		4	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL
0.15	30	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL
		4	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	**RDT-SL**, TOC	ED-SL	ED-SL
	50	2	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC
		4	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC
	100	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC
		4	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL
0.2	30	2	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC
		4	ED-SL, TOC	ED-SL	ED-SL, **RDT-SL**	ED-SL, TOC	ED-SL	ED-SL, **RDT-SL**	ED-SL	ED-SL	ED-SL
	50	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL
		4	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL
	100	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL
		4	**RDT-SL**, TOC	ED-SL	ED-SL	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL	**RDT-SL**	ED-SL	ED-SL, TOC
0.25	30	2	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL
		4	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC
	50	2	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL
		4	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL
	100	2	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC
		4	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL	ED-SL	ED-SL	ED-SL, TOC

TABLE 3. Comparison of the best single linkage clustering and TOC performance for Outlier Scenario 2

ε	n	δ	*pout*			*pmask*			*pswamp*		
			$p = 3$	$p = 5$	$p = 10$	$p = 3$	$p = 5$	$p = 10$	$p = 3$	$p = 5$	$p = 10$
0.05	30	2	**RDT-SL**	ED-SL	ED-SL	**RDT-SL**, TOC	ED-SL	ED-SL	MD-SL-SL	ED-SL, TOC	ED-SL
		10	MD-SL-SL	MD-SL-SL	MD-SL	MD-SL	MD-SL	MD-SL	ED-SL	ED-SL, TOC	ED-SL
	50	2	**RDT-SL**	**RDT-SL**, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL	ED-SL, TOC	ED-SL	ED-SL
	100	2	ED-SL, MD-SL	MD-SL, **RDT-SL**	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC
		10	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL
0.1	30	2	**RDT-SL**	**RDT-SL**	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	MD-SL	MD-SL	MD-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	ED-SL	ED-SL, TOC	ED-SL
	50	2	**RDT-SL**	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	MD-SL	MD-SL	MD-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	ED-SL, TOC	ED-SL	ED-SL
	100	2	None, TOC	**ALL**	**RDT-SL**	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL
		10	ED-SL	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**
0.15	30	2	**RDT-SL**, TOC	**RDT-SL**	ED-SL	**RDT-SL**, TOC	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL
		10	MD-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	MD-SL	MD-SL	ED-SL	ED-SL, TOC	ED-SL
	50	2	ED-SL, MD-SL	MD-SL	MD-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC
		10	ED-SL	MD-SL	MD-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**
	100	2	None	None	None	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	MD-SL	**RDT-SL**	MD-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	ED-SL	ED-SL, TOC	**ALL**
0.2	30	2	**RDT-SL**	**RDT-SL**	ED-SL	**RDT-SL**	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL	ED-SL
		10	**RDT-SL**, TOC	MD-SL	MD-SL	MD-SL, TOC	MD-SL	MD-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC
	50	2	None	None	None	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC
		10	**RDT-SL**, TOC	**RDT-SL**, TOC	MD-SL, TOC	MD-SL, TOC	MD-SL, TOC	MD-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**
	100	2	None	None	None	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	None	None	ED-SL	ED-SL	MD-SL	MD-SL	ED-SL, **RDT-SL**	ED-SL, TOC	**ALL**, TOC
0.25	30	2	None	None, TOC	None	**RDT-SL**	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL	ED-SL
		10	**ALL**	ED-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	MD-SL, TOC	ED-SL	ED-SL	ED-SL
	50	2	None	None	None	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL
		10	**RDT-SL**	MD-SL, **RDT-SL**	MD-SL, **RDT-SL**	ED-SL	MD-SL	MD-SL	ED-SL	ED-SL	**RDT-SL**
	100	2	None	None	None	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC
		10	None	None	None	ED-SL	ED-SL	MD-SL	ED-SL, TOC	**ALL**	**ALL**

The summary of results for best single linkage clustering for Outlier Scenario 3 is shown in Table 4. RDT-SL shows good performance in detecting outliers (*pout*), low masking error (*pmask*) and low swamping error (*pswamp*) for all ε, λ and δ when shifting the mean and covariance simultaneously. As can be seen from Table 4, RDT-SL is the best single linkage clustering or have a similar performance with another single linkage clustering in most of the conditions. Compared with TOC, TOC also shows good performance in *pout*, *pmask* and *pswamp* in most conditions for all ε, λ and δ when shifting the mean and covariance simultaneously.

In conclusion, RDT-SL performs the best in Outlier Scenario 3. RDT-SL shows high *pout* values and low values of *pmask* and *pswamp* when the mean and covariance are shifting simultaneously. The new single linkage robust clustering (RDT-SL) outlier detection procedure for multivariate data was also at its best when the outliers were situated far from the inliers.

TABLE 4. Comparison of the best single linkage clustering and TOC performance for Outlier Scenario 3

ε	n	δ	λ	***pout***			***pmask***			***pswamp***		
				p = 3	p =5	p = 10	p = 3	p =5	p = 10	p = 3	p = 5	p = 10
0.05	30	2	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL
			4	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL	ED-SL
		10	2	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC
			4	MD-SL	MD-SL	ED-SL, TOC	MD-SL	MD-SL	ED-SL, TOC	**RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC
	50	2	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC
			4	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC
		10	2	MD-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	MD-SL	MD-SL	ED-SL	ED-SL	ED-SL
			4	MD-SL, TOC	MD-SL	ED-SL, TOC	MD-SL, TOC	MD-SL	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL
	100	2	2	**RDT-SL**	**RDT-SL**	ED-SL	**RDT-SL**	**RDT-SL**	ED-SL	**RDT-SL**, TOC	**RDT-SL**, TOC	ED-SL
			4	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	**RDT-SL**	ED-SL
		10	2	MD-SL	MD-SL	MD-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL
			4	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC

0.1	30	2	2	**RDT-SL**	ED-SL, TOC	ED-SL	**RDT-SL**	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL
			4	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL	ED-SL
		10	2	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL, TOC	ED-SL, TOC	ED-SL	ED-SL
			4	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, **RDT-SL**, TOC
	50	2	2	**RDT-SL**, TOC	**RDT-SL**	ED-SL	ED-SL, TOC	**RDT-SL**	ED-SL	**RDT-SL**, TOC	**RDT-SL**	ED-SL
			4	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**	ED-SL, **RDT-SL**	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	**RDT-SL**	ED-SL	ED-SL, TOC
		10	2	ED-SL	MD-SL	MD-SL	MD-SL	MD-SL	MD-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC
			4	ED-SL	**RDT-SL**	ED-SL, TOC	ED-SL	**RDT-SL**	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**
	100	2	2	**RDT-SL**	**RDT-SL**	ED-SL	**RDT-SL**	**RDT-SL**	ED-SL	**RDT-SL**, TOC	**RDT-SL**, TOC	ED-SL, TOC
			4	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	**RDT-SL**, TOC
		10	2	ED-SL, TOC	ED-SL	MD-SL	ED-SL, TOC	ED-SL	MD-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC
			4	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**
0.15	30	2	2	ED-SL	**RDT-SL**	ED-SL, TOC	ED-SL	**RDT-SL**	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL
			4	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL	ED-SL
		10	2	ED-SL	ED-SL, TOC	MD-SL	MD-SL	MD-SL, TOC	MD-SL	ED-SL, TOC	ED-SL	ED-SL, TOC
			4	**RDT-SL**	**RDT-SL**, TOC	ED-SL	**RDT-SL**	**RDT-SL**, TOC	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**	ED-SL, **RDT-SL**
	50	2	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC
			4	**RDT-SL**	ED-SL, TOC	ED-SL	**RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	**RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC
		10	2	ED-SL	ED-SL	MD-SL, TOC	ED-SL	ED-SL	MD-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, **RDT-SL**
			4	ED-SL, TOC	**RDT-SL**	**RDT-SL**, TOC	ED-SL, TOC	**RDT-SL**	**RDT-SL**, TOC	ED-SL	ED-SL, **RDT-SL**	ED-SL, **RDT-SL**
	100	2	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC
			4	ED-SL, TOC	**RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	**RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	**RDT-SL**, TOC	**RDT-SL**, TOC
		10	2	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	**ALL**
			4	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	**ALL**

0.2	30	2	2	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL
			4	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	2	ED-SL	ED-SL	MD-SL	ED-SL	MD-SL	MD-SL	ED-SL, TOC	ED-SL, **RDT-SL**	ED-SL, **RDT-SL**
			4	**RDT-SL**, TOC	ED-SL	**RDT-SL**, TOC	**RDT-SL**, TOC	ED-SL, TOC	**RDT-SL**, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC
	50	2	2	ED-SL	**RDT-SL**, TOC	ED-SL	ED-SL	**RDT-SL**, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
			4	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	MD-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, **RDT-SL**, TOC	**ALL**
			4	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	**ALL**
	100	2	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
			4	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC
		10	2	ED-SL	**RDT-SL**	ED-SL, TOC	ED-SL	**RDT-SL**	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	**ALL**
			4	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	**RDT-SL**, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**, TOC	**ALL**
0.25	30	2	2	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC
			4	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC
		10	2	ED-SL	ED-SL, TOC	**RDT-SL**, TOC	ED-SL	ED-SL, TOC	**RDT-SL**, TOC	ED-SL, TOC	ED-SL	ED-SL, **RDT-SL**, TOC
			4	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, **RDT-SL**	ED-SL, **RDT-SL**
	50	2	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL
			4	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	2	**RDT-SL**	ED-SL	ED-SL	**RDT-SL**	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	**ALL**, TOC
			4	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, **RDT-SL**	**ALL**
	100	2	2	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL
			4	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL, TOC	ED-SL
		10	2	**RDT-SL**	ED-SL	ED-SL	ED-SL	ED-SL	**RDT-SL**, TOC	ED-SL, TOC	**ALL**, TOC	**ALL**, TOC
			4	ED-SL	ED-SL	ED-SL, TOC	ED-SL	ED-SL	ED-SL, TOC	ED-SL, TOC	**ALL**	**ALL**

ILLUSTRATIVE EXAMPLES

In order to investigate the applicability of the new single linkage robust clustering outlier detection procedure (RDT-SL) in real datasets, five historical multivariate datasets are used as illustrative examples. The datasets are Brain and Weight, Stackloss, Bushfire, Hawkins-Bradu Kass, and Milk. The majority of multivariate data outlier detection studies, including those by Becker and Gather (1999), Hadi (1992), Kosinski (1999), Pan, Fung and

Fang (2000), Rocke & Woodruff, (1996) and Rousseeuw and van Zomeren (1990) have adopted these datasets. These historical datasets already had known outliers. A summary of these datasets can be found in Table 5.

This section displays only dendrograms for the new single linkage robust clustering outlier detection procedure (RDT-SL). Three performance measures are also used for each dataset to assess how well the new single linkage robust clustering outlier detection procedure performs for historical data. The performance of the proposed procedure was also compared with single linkage clustering using ED and MD as similarity distance measures. The performance measures are as below: i. Number of outliers successfully detected. Outliers that were successfully found are counted and displayed in percentage. The proposed procedure is better if the percentage is closer to 100%. ii. Number of outliers falsely detected as inliers (masking effect). Any outliers mistakenly identified as inliers will be counted and displayed as percentages. The proposed procedure is better when the masking effect percentage is lower. iii. Number of inliers falsely detected as outliers (swamping effect). Any inliers mistakenly identified as outliers will be counted and displayed as percentages. The proposed procedure is better when the swamping effect percentage is lower.

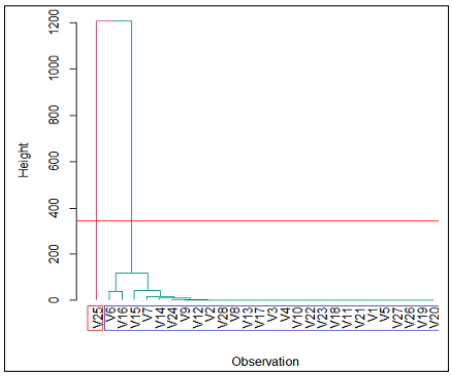
Figure 3 shows the dendrogram for five historical datasets using the new single linkage robust clustering outlier detection procedure. Table 6 shows performance measures and a comparison of the proposed procedure (RDT-SL) with a single linkage using ED (ED-SL) and MD (MD-SL). Previous findings from TOC study by Abd Mutalib, Satari and Yusoff (2021b) is also added in Table 6.

Figure 3(a) shows the dendrogram for RDT-SL to the Brain and weight (BW) dataset. The BW dataset contain two variables: body weight and brain weight for 28 species of animals. This dataset has only three outliers: observations 6th, 25th, and 16th (Atkinson & Mulira 1993; Hadi 1992; Pan, Fung & Fang 2000). From Figure 3(a), the RDT-SL only detects observation 25th as outliers, while observations 6th and 16th are misclassified as inliers. From Table 6, all single linkage clustering can only identify one outlier and misclassify two outliers as inliers. RDT-SL and ED-SL do not have a swamping effect. However, two inliers in the dataset were incorrectly labelled as outliers by MD-SL. Meanwhile, TOC shows excellent performance in detecting outliers for the BW dataset and does not have a masking effect. Despite that, TOC shows poor performance in the swamping effect where five inliers are detected as outliers.

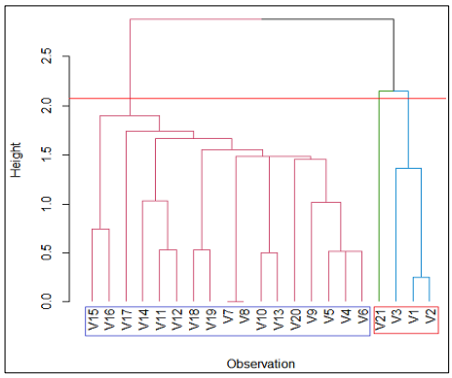
Figure 3(b) shows the dendrogram for the Stackloss dataset. Stackloss data is a dataset obtained from a 21-day experiment measuring the oxidation of ammonia into nitric acid (Becker & Gather 1999). The dataset includes one response variable (Stackloss) and three explanatory variables (rate of incoming ammonia, cooling water temperature, and acid concentration) (Becker & Gather 1999; Hadi 1992). Only three explanatory variables are tested for outlier in this study. Observations 1st - 3rd, and 21st are outliers (Hadi 1992; Pan, Fung & Fang 2000). Figure 3(b) shows that RDT-SL successfully identified all outliers in the Stackloss dataset and did not experience any masking and swamping effect. From Table 6, the Stackloss dataset shows that all single linkage clusterings and TOC successfully identified all outliers and have no masking effect. However, six inliers and one inlier are incorrectly classified as outliers by MD-SL and TOC, respectively.

TABLE 5. Summary of the historical multivariate datasets

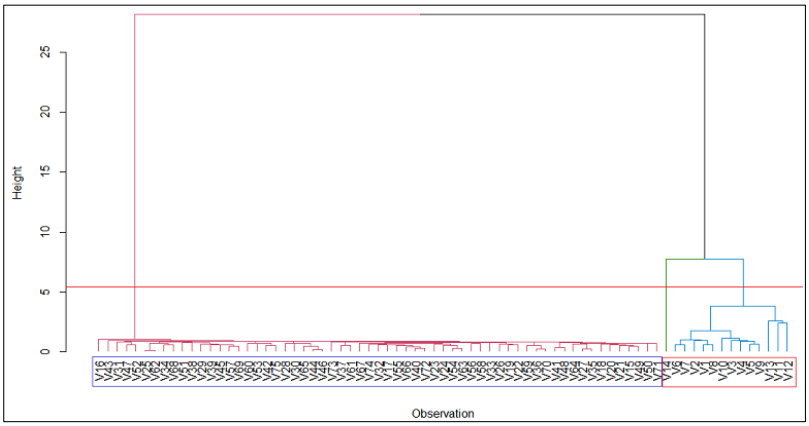
Dataset	n	p	Number of outliers	ε	Outlier observations
Brain and weight data (BW)	28	2	3	11%	Observations 6th, 16th and 25th
Hawkins-Bradu Kass (HBK)	75	3	14	19%	Observations 1st - 14th
Stackloss	21	3	4	19%	Observations 1st - 3rd and 21st
Bushfire	38	5	13	34%	Observations 7th - 11th and 31st - 38th
Milk	86	8	17	20%	Observations 1st - 3rd, 12th, 13th - 17th, 27th, 41st, 44th, 47th, 70th, 74th, 75th and 77th



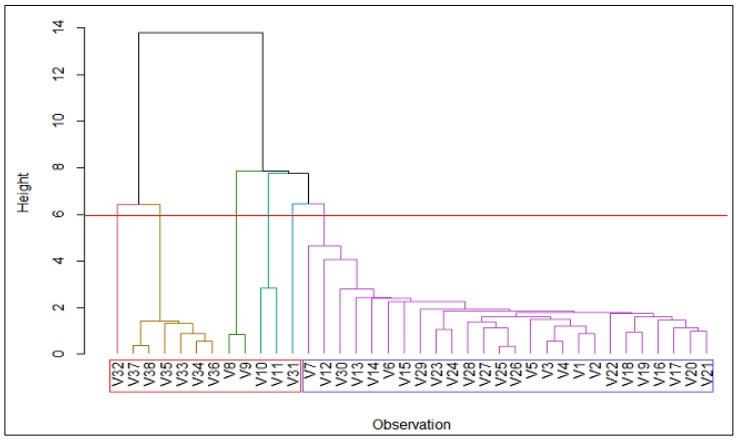
(a)



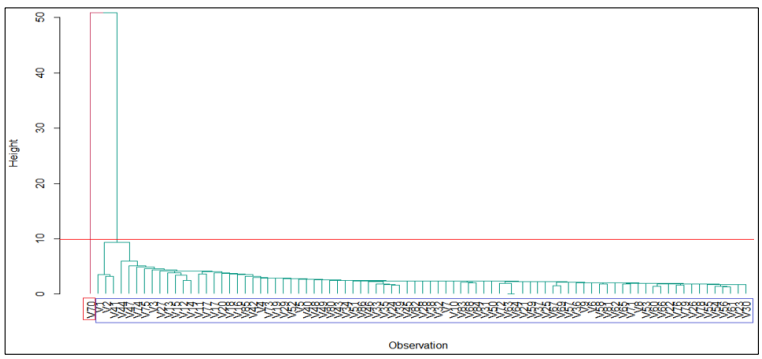
(b)



(c)



(d)



(e)

FIGURE 3. The dendogram using RDT-SL for (a) Brain and Weight dataset (b) Stackloss dataset (c) Hawkins-Bradu Kass dataset (d) Bushfire dataset (e) Milk dataset

TABLE 6. Performance measures for historical data

Dataset	Performance measures	Single linkage clustering procedure			TOC
		ED-SL	MD-SL	RDT-SL	
Brain and weight dataset	Number of outliers successfully detected	1 (33.3%)	1 (33.3%)	1 (33.3%)	3 (100%)
	Number of outliers falsely detected as inliers (masking effect)	2 (66.7%)	2 (66.7%)	2 (66.7%)	0 (0%)
	Number of inliers falsely detected as outliers (swamping effect)	0 (0%)	2 (8%)	0 (0%)	5 (20%)
Stacklos dataset	Number of outliers successfully detected	4 (100%)	4 (100%)	4 (100%)	4 (100%)
	Number of outliers falsely detected as inliers (masking effect)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Number of inliers falsely detected as outliers (swamping effect)	0 (0%)	6 (35.3%)	0 (0%)	1 (5.9%)
Hawkins Bradu Kass dataset	Number of outliers successfully detected	14 (100%)	14 (100%)	14 (100%)	14 (100%)
	Number of outliers falsely detected as inliers (masking effect)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Number of inliers falsely detected as outliers (swamping effect)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Bushfire dataset	Number of outliers successfully detected	12 (92.3%)	12 (92.3%)	12 (92.3%)	13 (100%)
	Number of outliers falsely detected as inliers (masking effect)	1 (7.7%)	1 (7.7%)	1 (7.7%)	0 (0%)
	Number of inliers falsely detected as outliers (swamping effect)	2 (8%)	0 (0%)	0 (0%)	3 (12%)
Milk dataset	Number of outliers successfully detected	7 (41.2%)	4 (23.5%)	1 (5.9%)	17 (100%)
	Number of outliers falsely detected as inliers (masking effect)	10 (58.8%)	13 (76.5%)	16 (94.1%)	0 (0%)
	Number of inliers falsely detected as outliers (swamping effect)	3 (4.3%)	0 (0%)	0 (0%)	5 (7.2%)

The Hawkins-Bradu-Kass (HBK) dataset was created artificially by Hawkins, Bradu and Kass (1984). This dataset has four variables and 75 observations (one response and three explanatory variables). Only three explanatory variables will be tested for outlier in this study. Observations 1 through 14 in this dataset are identified as outliers (Hadi 1992; Pan, Fung & Fang 2000; Rocke & Woodruff 1996; Rousseeuw & van Zomeren 1990). Figure 3(c) demonstrates that RDT-SL identified all outliers in the Hawkins-Bradu Kass dataset without suffering from a masking effect. Additionally, the RDT-SL also does not have a swamping effect. The result of the performance measure for HBK dataset in Table 6 shows that ED-SL, MD-SL and TOC also successfully identified all outliers and did not have masking and swamping effect.

The dendrogram for the RDT-SL of Bushfire dataset is shown in Figure 3(d). The dataset was used to locate bushfire scars and was taken from Maronna and Yohai (1995). The dataset includes satellite measurements for five frequency bands, with 38 pixels for each band, and consists of 13 outliers. Kosinski (1999) and Rocke and Woodruff (1996) classify observations 7th - 11th and 31st - 38th as outliers. From the dendrogram, the RDT-SL misclassifies one outlier (observation 7th) as an inlier. The RDT-SL has a masking effect but no swamping effect. According to Table 6, all single linkage clusterings can identify 12 out of 13 outliers in the Bushfire dataset. One outlier was incorrectly classified as an inlier by all single linkage clustering. ED-SL has a swamping effect when misclassifying two inliers as outliers. Meanwhile, TOC successfully detected all outliers in the Bushfire dataset, having no masking effect but misclassifying three inliers as outliers, which means TOC suffered from the masking effect.

The fifth dataset is the Milk dataset from Daudin, Duby and Trecourt (1988) which consists of 86 milk containers and eight variables. Density, fat, protein, casein, cheese dry substance measured in a factory, cheese dry substance measured in a lab, milk dry substance, and cheese produced are the eight variables. In this dataset, there are 17 outliers, making the percentage of outliers is 20%. Outliers are observations 1st through 3rd, 12th through 17th, 27th, 41st, 44th, 47th, 70th, 74th, 75th, and 77th (Atkinson 1994; Kosinski 1999; Rocke & Woodruff 1996). Figure 3(e) shows that the RDT-SL detects only one outlier, which means that 16 outliers are misclassified as inliers. However, no inliers are misclassified as outliers which means no swamping effect. Performance measures in Table 6 show that ED-SL performs well compared to other single linkage clusterings to detect outliers. All single linkage clusterings have masking effect, and RDT-SL has the highest masking effect. As for the swamping effect, MD-SL and RDT-SL do not have swamping effect. The performance of TOC for Milk dataset is excellent when all outliers are successfully detected and do not have masking effect. However, TOC has a swamping effect where five inliers are detected as outliers. Hence, RDT-SL does not perform well for the dataset with more than five variables with 20% outliers.

Table 6 shows that all single linkage clusterings have successfully detected outliers for two datasets (HBK and Stackloss). However, all single linkage clusterings do not successfully detect all outliers in BW, Bushfire, and Milk datasets. The same results are observed for the masking effect. RDT-SL shows good performance for the swamping effect, indicating no inliers misclassifying as outliers in all datasets. TOC performance in Table 6 shows excellent performance in detecting outliers where TOC successfully detect all outliers and does not have masking effect in five historical datasets. However, TOC only shows no swamping effect in the HBK dataset.

Conclusion

This study aimed to develop a new robust clustering procedure to detect outliers for multivariate data. The clustering method used is single linkage clustering and the new procedure is named the new single linkage robust clustering outlier detection procedure. Single linkage clustering is robustified in this procedure using robust distance as a similarity measure and is named RDT-SL. The new procedure's performance was compared with single linkage clustering using ED (ED-SL) and MD (MD-SL) as similarity measures and robust estimator TOC.

The performance of the RDT-SL is investigated via simulation studies and applied to historical multivariate datasets. Three outlier scenarios were used in the simulation studies, and it was found that the RDT-SL performed well in Outlier Scenario 3 when both the mean and covariance were shifted simultaneously. The RDT-SL also performed well as the distance between outliers and inliers increased. The RDT-SL also performs well in historical multivariate datasets. Out of 5 datasets, the RDT-SL can detect all outliers and does not have a masking effect in 2 datasets where the sample size, $n < 100$ and the number of variables, p = 3 with percentage of outliers less than 20%. The RDT-SL also show good performance for the Bushfire dataset where $n < 100$ and the number of variables, p = 5 with percentage of outliers less than 35%. However, RDT-SL shows poor performance when the number of variables, $p < 3$ or $p < 5$. The RDT-SL also does not have a swamping effect in all datasets. In contrast to TOC's performance, TOC successfully detects all outliers and does not have a masking effect in all datasets. However, TOC only shows no swamping effect in the HBK dataset. In conclusion, the new single linkage robust clustering outlier detection procedure (RDT-SL) is a practical and promising approach to detecting outliers for multivariate data, especially for $n < 100$ and p = 3, 5. This new procedure can be extended using other agglomerative clustering algorithms.

Acknowledgements

The authors would like to thank the Ministry of Higher Education, Malaysia (FRGS/1/2019/STG06/UMP/0216), Universiti Malaysia Pahang (UMP) (RDU190168) and University College TATI (UCTATI).

REFERENCES

Abd Mutalib, S.S.S., Satari, S.Z. & Yusoff, W.N.S.W. 2021a.. Comparison of robust estimators for detecting outliers in multivariate data. *Journal of Statistical Modeling and Analytics* 3(2): 36-64.

Abd Mutalib, S.S.S., Satari, S.Z. & Yusoff, W.N.S.W. 2021b. Comparison of robust estimators for detecting outliers in multivariate datasets. *Journal of Physics: Conference Series* 1988: 1-9.

Abd Mutalib, S.S.S., Satari, S.Z. & Yusoff, W.N.S.W. 2019. 2019. A new robust estimator to detect outliers for multivariate data. *Journal of Physics: Conference Series* 1366(1): 012104. https://doi.org/10.1088/1742-6596/1366/1/012104

Aggarwal, C.C. 2017. *Outlier Analysis*. 2nd ed. Springer. https://doi.org/10.1016/b978-012724955-1/50180-7

Almeida, J.A.S., Barbosa, L.M.S., Pais, A.A.C.C. & Formosinho, S.J. 2007. Improving hierarchical cluster analysis: A New method with outlier detection and automatic clustering. *Chemometrics and Intelligent Laboratory Systems* 87(2): 208-217. https://doi.org/10.1016/j.chemolab.2007.01.005

Atkinson, A.C. 1994. Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association* 89(428): 1329-1339. https://doi.org/10.1080/01621459.1994.10476872

Atkinson, A.C. & Mulira, H.M. 1993. The stalactite plot for the detection of multivariate outliers. *Statistics and Computing* 3(1): 27-35. https://doi.org/10.1007/BF00146951

Badaró, J.P.M., Campos, V.P., Oliveira Campos da Rocha, F. & Lima Santos, C. 2021. Multivariate analysis of the distribution and formation of trihalomethanes in treated water for human consumption. *Food Chemistry* 365: 130469. https://doi.org/10.1016/j.foodchem.2021.130469

Balcan, M-F., Liang, Y. & Gupta, P. 2014. Robust hierarchical clustering. *Journal of Machine Learning Research* 15: 4011-4051. https://doi.org/10.1109/IMSCCS.2006.167

Becker, C. & Gather, U. 1999. The masking breakdown point of multivariate outlier identification rules. *Journal of the American Statistical Association* 94(447): 947-955. https://doi.org/10.1080/01621459.1999.10474199

Cabana, E., Lillo, R.E. & Laniado, H. 2021. Multivariate outlier detection based on a robust Mahalanobis distance with shrinkage estimators. *Statistical Papers* 62: 1583-1609. https://doi.org/10.1007/s00362-019-01148-1

Cerioli, A., Riani, M. & Torti, F. 2011. Accurate and powerful multivariate outlier detection. *Int. Statistical Inst.: Proc. 58th World Statistical Congress*. pp. 5608-5613.

Christy, A., Gandhi, M.G. & Vaithyasubramanian, S. 2015. Cluster based outlier detection algorithm for healthcare data. *Procedia Computer Science* 50: 209-215. https://doi.org/10.1016/j.procs.2015.04.058

Daudin, J.J., Duby, C.D. & Trecourt, P. 1988. Stability of principal component analysis studied by the bootstrap method. *Statistics: A Journal of Theoretical Applied Statistics* 19(2): 241-258. https://doi.org/10.1080/02331888808802095

De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. 2000. Tutorial: The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* 50: 1-18. www.elsevier.comrlocaterchemometrics.

Dotto, F., Farcomeni, A., García-Escudero, L.A. & Mayo-Iscar, A. 2018. A reweighting approach to robust clustering. *Statistics and Computing* 28(2): 477-493. https://doi.org/10.1007/s11222-017-9742-x

Duan, L., Xu, L., Liu, Y. & Lee, J. 2009. Cluster-based outlier detection. *Annals of Operations Research* 168: 151-168. https://doi.org/10.1007/s10479-008-0371-9

Evans, K., Love, T. & Thurston, S.W. 2015. Outlier identification in model-based cluster analysis. *Journal of Classification* 32(1): 63-84. https://doi.org/10.1007/s00357-015-9171-5

Fauconnier, C. & Haesbroeck, G. 2009. Outliers detection with the minimum covariance determinant estimator in practice. *Statistical Methodology* 6(4): 363-379. https://doi.org/10.1016/j.stamet.2008.12.005

Filzmoser, P., Maronna, R. & Werner, M. 2008. Outlier identification in high dimensions. *Computational Statistics and Data Analysis* 52(3): 1694-1711. https://doi.org/10.1016/j.csda.2007.05.018

Gan, G., Ma, C. & Wu, J. 2007. *Data Clustering: Theory, Algorithms, and Applications.* Philadelphia: Society for Industrial and Applied Mathematics.

Garcia-Escudero, L.A., Gordaliza, A., Matran, C. & Mayo-Iscar, A. 2010. A review of robust clustering methods. *Advances in Data Analysis and Classification* 4(2): 89-109. https://doi.org/10.1007/s11634-010-0064-5

García-Escudero, L.A., Gordaliza, A., Matrán, C. & Mayo-Iscar, A. 2008. A general trimming approach to robust cluster analysis. *The Annals of Statistics* 36(3): 1324-1345. https://doi.org/10.1214/07-AOS515

Hadi, A.S. 1992. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society. Series B (Methodological)* 54(3): 761-771.

Hadi, A.S., Rahmatullah Imon, A.H.M. & Werner, M. 2009. Detection of outliers. *Wiley Interdisciplinary Reviews: Computational Statistics* 1(1): 57-70. https://doi.org/10.1002/wics.6

Hardin, J. & Rocke, D.M. 2004. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis* 44(4): 625-638. https://doi.org/10.1016/S0167-9473(02)00280-3

Hawkins, D.M., Bradu, D. & Kass, G.V. 1984. Location of several outliers in multiple-regression data using elemental sets. *Technometrics* 26(3): 197-208. https://doi.org/10.1080/00401706.1984.10487956

Herwindiati, D.E., Djauhari, M.A. & Mashuri, M. 2007. Robust multivariate outlier labeling. *Communications in Statistics-Simulation and Computation* 36(6): 1287-1294. https://doi.org/10.1080/03610910701569044

Ijaz, M.F., Attique, M. & Son, Y. 2020. Data-driven cervical cancer prediction model with outlier detection and over-sampling methods. *Sensors* 20: 1-22.

Jiang, M.F., Tseng, S.S. & Su, C.M. 2001. Two-phase clustering process for outliers detection. *Pattern Recognition Letters* 22(6-7): 691-700. https://doi.org/10.1016/S0167-8655(00)00131-8

Kalina, J. & Tichavský, J. 2021. The minimum weighted covariance determinant estimator for high-dimensional data. *Advances in Data Analysis and Classification*. https://doi.org/10.1007/s11634-021-00471-6

Kosinski, A.S. 1999. A procedure for the detection of multivariate outliers. *Computational Statistics and Data Analysis* 29(2): 145-161. https://doi.org/10.1016/S0167-9473(98)00073-5

Maronna, R.A. & Yohai, V.J. 1995. The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association* 90(429): 330-341. https://doi.org/10.1080/01621459.1995.10476517

Melendez-Melendez, G., Cruz-Paz, D., Carrasco-Ochoa, J.A. & Martínez-Trinidad, J.F. 2019. An improved algorithm for partial clustering. *Expert Systems with Applications* 121: 282-291. https://doi.org/10.1016/j.eswa.2018.12.027

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. https://doi.org/10.1007/BF02294245

Mojena, R. 1977. Hierarchical grouping methods and stopping rules: An evaluation. *The Computer Journal* 20(4): 259-363.

Olukanmi, P.O. & Twala, B. 2017. K-means-sharp: Modified centroid update for outlier-robust k-means clustering. *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference, PRASA-RobMech 2017*. pp. 14-19. https://doi.org/10.1109/RoboMech.2017.8261116

Pan, J-X., Fung, W-K. & Fang, K-T. 2000. Multiple outlier detection in multivariate data using projection pursuit techniques. *Journal of Statistical Planning and Inference* 83(1): 153-167. https://doi.org/10.1016/s0378-3758(99)00091-9

Peña, M. 2018. Robust clustering methodology for multi-frequency acoustic data: A review of standardization, initialization and cluster geometry. *Fisheries Research* 200: 49-60. https://doi.org/10.1016/j.fishres.2017.12.013

Rencher, A.C. 2002. *Methods of Multivariate Analysis*. New York: John Wiley & Sons, Inc. https://doi.org/10.2307/2669873

Rocke, D.M. & Woodruff, D.L. 1996. Identification of outliers in multivariate data. *Journal of the American Statistical Association* 91(435): 1047-1061. https://doi.org/10.1080/01621459.1996.10476975

Rousseeuw, P.J. & Hubert, M. 2011. Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1(1): 73-79. https://doi.org/10.1002/widm.2

Rousseeuw, P.J. & van Zomeren, B.C. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85(411): 633-639. https://doi.org/10.2307/2289999

Salleh, R.M. 2013. A robust estimation method of location and scale with application in monitoring process variability. PhD Thesis. Universiti Teknologi Malaysia (Unpublished).

Santos-Pereira, C.M. & Pires, A.M. 2002. Detection of outlier in multivariate data: A method based on clustering and robust estimators. In *Compstat,* edited by Härdle, W. & Rönz, B. Physica, Heidelberg. pp. 291-296. https://doi.org/10.1007/978-3-642-57489-4_41

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

Satari, S.Z., Muhammad Di, N.F. & Zakaria, R. 2019. Single-linkage method to detect multiple outliers with different outlier scenarios in circular regression model. *AIP Conference Proceedings* 2059: 020003. https://doi.org/10.1063/1.5085946

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. https://doi.org/10.17576/jsm-2021-5006-24

Saxena, A., Prasad, M., Gupta, A., Bharill, N., Prakash Patel, O.P., Tiwari, A., Er, M.J., Ding, W. & Lin, C-T. 2017. A review of clustering techniques and developments. *Neurocomputing* 267: 664-681. https://doi.org/10.1016/j.neucom.2017.06.053

Sebert, D.M., Montgomery, D.C. & Rollier, D.A. 1998. A clustering algorithm for identifying multiple outliers in linear regression. *Computational Statistics & Data Analysis* 27(4): 461-484. https://doi.org/10.1016/S0167-9473(98)00021-8

Sharma, K.K. & Seal, A. 2021. Outlier-robust multi-view clustering for uncertain data. *Knowledge-Based Systems* 211: 106567. https://doi.org/10.1016/j.knosys.2020.106567

Wada, K., Kawano, M. & Tsubaki, H. 2020. Comparison of multivariate outlier detection methods for nearly elliptical distributions. *Austrian Journal of Statistics* 49(2): 1-17. https://doi.org/10.17713/ajs.v49i2.872

Wang, H., Bah, M.J. & Hammad, M. 2019. Progress in outlier detection techniques: A survey. *IEEE Access* 7: 107964-108000. https://doi.org/10.1109/ACCESS.2019.2932769

Werner, M. 2003. Identification of multivariate outliers in large data sets. MSc. University of Colorado (Unpublished).

Xu, D. & Tian, Y. 2015. A comprehensive survey of clustering algorithms. *Annals of Data Science* 2(2): 165-193. https://doi.org/10.1007/s40745-015-0040-1

Yesilbudak, M. 2016. Partitional clustering-based outlier detection for power curve optimization of wind turbines. In *5th International Conference on Renewable Energy Research and Applications (ICRERA)*. pp. 1080-1084.

Yoon, K-A., Kwon, O-S. & Bae, D-H. 2007. An approach to outlier detection of software measurement data using the K-means clustering method. *First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007)*. pp. 443-445. https://doi.org/10.1109/ESEM.2007.49

Zhang, J. 2013. Advancements of outlier detection: A survey. *ICST Transactions on Scalable Information Systems* 13(1): 1-26. https://doi.org/10.4108/trans.sis.2013.01-03.e2

*Corresponding author; email: sharifahsakinah84@gmail.com