Republic of Iraq
Ministry of Higher Education
and Scientific Research
University of Al- Qadisiyah
College of Administration & Economics
Department of Statistics

Multivariate outliers detection by statistical depth functions with an application

A Thesis

Submitted to the council of the college of administration and economics, university of al-qadisiyah in partial fulfillment of the requirements for the degree of master of statistics

By

Hadeel Kamel Habib

Supervised By

Ass.Prof. Dr. Mohammed Al-Guraibawi

2٠٢٢ AD 1443 AH

صدق الله العلي العظيم

﴿سورة البقرة ٣٢﴾

Supervisor certification

I certify that the preparation for this thesis entitled "**Multivariate outliers detection by statistical depth functions with application**" was prepared by **Hadeel Habib Kamel** under my Supervision at the Statistics Department, College of Administration and Economics/ University of Al- Qadisiyah, in partial fulfillment of the requirements for the degree of Master of Statistics.

Supervisor
Signature:
Date: / /2022
Name**:**
Dept. of Statistics
College of Administration and Economics

In view of available recommendations, we forward this thesis for debate by the examining committee.

Head of the Statistics Department
Signature:
Date: / /2022
Name**:**
Dept. of Statistics
College of Administration and Economics

Linguistic Supervisor Certification

This is to certify that I read the thesis entitled "**Multivariate outliers detection by statistical depth functions with application**" and corrected every grammatical and stylistic mistake, therefore, this thesis is qualified for debate.

Signature:

Name:
Date: / / 2022

Examining Committee Certification

We certify that we have read this thesis entitled "**Multivariate outliers detection by statistical depth functions with application**" and as examining the student **Hadeel Habib Kamel** in its contents and that in our opinion it meets the standard of the dissertation for the Degree of Master of Statistics.

Signature:
Name:
Date: / /2022
Address:
University of @
(Chairman)

Signature:
Name:
Date: / /2022
Address:
University of @
(Member)

Signature:
Name:
Date: / /2022
Address:
University of @
(Member)

Signature:
Name:
Date: / /2022
Address: College of Sciences
University of @
(Member & Supervisor)

Approval of the College Board
The Council of the College of @ met ……..in its session: / /2022 and decided to grant it a master's degree in @.

Signature:

Dean of the College of Administration and Economics:

Address: College of Administration and Economics /University of Al-Qadisiyah

Date: / /2022

Dedication

This work is dedicated to…

My father and My mother

I will always be your proud daughter, I ask God to protect you from all harm and I will not disappoint you

My beloved husband and children

Whom I can't force myself to stop loving.

My beloved brothers and sister

My best friend Shatha

Who stands by me when things look very difficult.

Hadeel

Acknowledgments

First and foremost, I would like to thank God for letting me through all the difficulties throughout my research work to complete the research successfully.

I would like to express my deep and sincere gratitude to my research supervisor, Dr. Mouhammed , for giving me the opportunity to do research and providing invaluable guidance throughout this research. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. Working and studying under their guidance was a great privilege and honor. I am extremely grateful for what he has offered me.

I would also like to give special thanks to my husband and my family for their continuous support and understanding when undertaking my research and writing my project.

I would like to thank the staff of the statistics department at the University of Al-Qadisiyah for their kind help and co-operation throughout my study period.

Finally, I want to express my gratitude to my friends and colleagues for their support and advice to me during my study years.

Hadeel ✍

Summary

In this study, the robust data depth function was used to diagnose outliers in a multivariate model. A group of depth functions are used, which depend on the mahalanobis distance and robust variance covariance matrix (MRCD), such outlyingness depth, spatial depth, elliptical depth and triangle depth methods. In addition, for bivariate model we suggested using bagplot for detection outliers in the dataset, that is depend on statistical data depth function and tukey median. Simulation study, artificial data and real data were used to evaluate the proposed methods. The study showed that the suggested methods have good performance for detection outliers compared with some existing methods. We suggested (MRCD) as proposed method.

Abbreviations

DM Depth Median

GESD Generalized Extreme Studentized Deviate

MAD Median Absolute Deviation

MCD Minimum Covariance Determinant

MRCD Minimum Regularized Covariance Determinant

MD Mahalanobis Distance

MAD Median Absolute Deviation

MDE Mutation Detection Enhancement

MDO Medium Density Overlay

MDT minimization of drive test ()

MDE Minimization of drive Error

MDS Multi-dimensional scaling

RMD Robust Mahalanobis Distance

SDF Statistical Depth Function

OGK Orthogonalized Gnanadesikan-Kettenring

ODD Outlier Detection and Description

CONTENTS.

List of Tables

List of Figures

Chapter One
General Introduction

Chapter One
General Introduction

1.1 Introduction

As information technology has advanced, most fields are gradually joined the scope of big-data. Furthermore, with the complexity of studies, data processing problems have emerged. The greater variables that need to be collected, classified, and processed, the higher the probability of errors, and thus the greater the likelihood of outliers appearing in the dataset. In fact, data analysis is influenced by a wide range of interlocking and uncertain factors, one of the most important of which is the appearance of outliers. Therefore, the researcher must detect these outliers before starting the analyses. A group of methods have been proposed to detect outliers in the dataset, there are methods that used to detect outliers in one-dimensional data, such as the three sigma criteria and box plot [74], etc. Unfortunately, these methods do not work with multivariate data. The researchers proposed a set of methods to detect outliers in multivariate data. One of the most widely used methods in this field is the Mahalanobis distance (MD) method. Furthermore, the high leverage also can be used to detect outliers in multivariate data. Despite the wide use of the MD method, it is also sensitive to outliers because it relies on the traditional mean and variance-covariance matric. To address this problem, a number of robust methods have been proposed by researchers such as MVE estimator, MCD estimator, and so on [38]. Thus, in this thesis, we propose using the minimum regularized covariance determinant (*MRCD*) with some identification depth function such as mahalanobis depth distance and pagplot to identify outliers with good proprieties. The proposed methods proved their efficiency in diagnosing outliers by applying them to a set of real data and experimenting with simulations. The disadvantage of these classical methods is that the affine invariance is not achieved which roughly means that the point depth with respect to the distribution depends on the underlying coordinate system.

Disadvantages. Although Mahalanobis distance is included with many popular statistics packages, some authors question the reliability of results, which is what we are trying to avoid in our proposed method. In the Mahalanobis distance (*MD*), for example, the classical mean and covariance matrix suffer from masking (***means that an outlier is undetected because of the pres- ence of another adjacent ones***) and swamping(***is that a good obser- vation is incorrectly identified as an outlier because of the presence of another clean subset***.) effects. When outliers were not identified, masking effects occurred, and swamping effects occurred when inliers were identified as outliers.[71]

1.2 Problem of the study

The main problem of the study is in the process of correct diagnosis of outliers. Ordinary methods suffer from some problems for diagnosis such as (masking and swamping).

1.3 Objectives of the study

In this thesis, we are going to satisfy the following objectives:

1. Using robust variance- covariance matrix to get robust mean and robust variance such as the minimum regularized covariance determinant (*MRCD*).
2. Applying some statistical outlyingness depth functions.
3. Robustify some diagnostic methods by using robust variance- covariance method (MRCD) to avoid masking and swamping problems.
4. Using bagplot based on MRCD as diagnostic method to identify outliers.
5. Design a simulation experiment to verify the efficiency of the proposed methods
6. Applying the proposed methods with real data (PM10 pollution dataset).

1.4 Literature Review

In contemporary statistics, linear regression models represent a large and highly developed field. One of the most widely applied models in this field is the multiple linear regression model. The most common method(MD, MVE and MCD) used in multiple linear regression models is to establish a functional relationship between two or more quantitative variables so that one or more explanatory variables can predict a response variable. Robust statistical depth function approaches for diagnostic outliers in simple and multiple linear regression models are the key issues in this chapter's research reviews.

Regina Y. Liu (1990) presented a significant novel variety of depth functions, the “simplicial depth”, and confirm the general role of a depth functions as providing a center outward ranking of dataset. She introduced a new ideas of data depths. This ideas to emerge naturally out of a fundamental concept underlying affine geometry, namely that of a simplex, and it satisfies the requirements one would expect from a notion of data depths. Thus it leads to an affine invariant, center outward ranking of the data point.[53]

Donoho D. & and Gasko M. (1992) explored the properties of the location depth and of the deepest location for finite data set, where the deepest location is a point with maximal ldepth which it the center of gravity the center of gravity of the innermost ldepth region [19]

Johnson et al. (1998) proposed accounting to only a small fraction of observations while creating the first one depth contour, which results in a lower complexity for small one. A dataset is described by a finite number of depth contours in halfspace depth. [35]

Rousseeuw P. & Hubert Mia. (1999) suggested the algorithm of regression depths. They viewing depths as a property of a fit, rather than a property of an point. In general, they describe the depths of a (candidate) fit θ to a given data Z_n of size n (θ, Z_n) to be the lowest number of points of Z_n that would need to be removed to make θ a nonfit.[62]

Robson G.(2003) mention that the contaminants are outliers caused by human error or the presence of a separate generation mechanism and a different distribution. Outlying observations are typically not contaminants except in the case of heavy-tailed distributions such as Student's t. The contamination of samples drawn from the normal distribution, which is not prone to outliers, was addressed. It is also assumed that nothing about the distribution's parameters is known a priori, which is usually the case.[55]

Miller et al. (2003) compute half-space depth for bivariate dataset, and the half-space depths defends a dataset by a finite number of depths contour with complexity O(n^2), and the depths of a single points may then be computed with complexity ($\log^2 n$).[47]

Bremner et al. (2006) use a primal-dual technique to determine the halfspace depth by incrementally updating the upper and lower boundaries using a heuristic until they coincide. [10]

Bremner et al. (2008) proposed the output sensitive depths-calculating procedure that explains the task as two maximum sub-system issue for $d > 2$. [11]

Zonoid depth was pioneered by Mosler et al. (2009). They took advantage of the notion to divide Rd into direction cones, and later techniques for determining depth and depths region, including the half-space depth, did the same. [48]

A directed relation between multivariate quantile areas and half-space depths trimmed region is shown by Hallin et al. (2010). [27]

When bivariate depth and depth lines continually add points to the data set, updating depth becomes a fascinating problem, which Burr et al. (2011) explores. [12]

Lok W. & Lee S. (2011) suggested a novel depths function depend on inter-point distances, that has the distinct property of respecting multivariate in dataset. With specification of an appropriate inter-point distance, our depth function also applies to infinite-dimensional data. Where the conventional center outward ordering depths function are founded to be inadequate. [42]

In order to demonstrate that their envelope coincides with the appropriate half-space depths trimmed region, Kong and Mizera(2012) use direction quantiles, which are half-spaces that correspond to quantiles on univariate projections. For $d > 2$, the areas of depth are precisely computed.[36]

Ieva and Paganoni (2013) used Multivariate Functional Principal Component Analysis to reduce the dimensionality of their data. It entails multiplying the respective scores by the information's contain in the covariance of the signal and their first derivative. Projecting dataset and derivative onto there relevant Karhunen Loève bases yields scores. [7]

Liu and Zuo (2014) use a breadth first search technique to cap R^d and "QHULL" to identify the direction cones in order to precisely determine the half-space depth. For the precise computation of the half-space depths, he offers two more, seemingly quick procedures. This algorithm is one of them; it is called a refined combinatorial algorithm.[40]

López-Pintado S. et al. (2014) proposed Simple depth-of-range concepts with multivariate model that extended bivariate depths function of range, providing simple and natural criteria for measuring path centrality within a samples of curves. depend on these depth, a sample of multivariate curves could be order from the center out and system statistics can be determined. The suggested depth has characteristics, like a stability and consistency.[43]

Ieva F. et al. (2015) adjusted a method to compare 2 independent samples of multivariate functional dataset that vary in expressions of variance factors. The concept of depth measurement has been generalized to this type of data, taking advantage of the role of contrast factors in weighting the components that determine depth. It was applied to electrocardiogram signal targets at comparing physiological subjects and affected patients with left bundle branch block. Also, the suggested depths scales calculated on the dataset were used to perform a non-parametric comparison test between these two groups. They are also presented in a "generalized regression model" that aims to classify ECG signals.[34]

Katie Evans et al. (2015) devised a method to identify outlying observations in model depend clustering based on normal mixture model that influence cluster structure and number, without identifying clusters amid a wide range of noisy observations. The outliers are those with a minimum membership proportion or for which the cluster specific covariance with and without the point is very different. The method demonstrated its ability to detect true outliers without incorrectly identifying many not outlier and improve performances compared to other methods.[22]

Reyes A. & Cuesta-Albertos J. (2015) proposed an adjustment of the first algorithm in "Hubert et al. (2015)" containing in basing it on the random Turkey's

depths, where the random Turkey's depths are statistical depths that approximates the Turkey's depths. It requires of a very little number of projection to get equivalent outcomes to those of the Turkey's depths. So, the random Turkey's depths are high fast to calculate, make it the depths to go for, not only when the dimension of the space is moderate or high, but also when it is low due to its computationally effectiveness. In addition, the random Turkey's depths be devised from the Turkey's depths the nice properties that made it recognized. Also he proposed a simplest and more usual criteria of variations.[52]

Rainer D. & Pavlo M. (2016) proposed a conceptual framework for calculating the half-space depth, which gives a whole class of procedures. The data to each of these tuple is projected onto the corresponding orthogonal complement, and the half-space depths were calculated as the sum of the depth in these two orthogonal sub-spaces and all suggested procedures are qualified of dealing with dataset that is not in general mode and even with ties.[51]

Data depth, according to Making O. & Adewumi A. (2017), is an alternative to several parametric methodologies in evaluating large amounts of multivariate data. A nonparametric classification strategy depend on several dataset depths function conceptions is addressed, and certain features of these approaches are investigated. The performance of various depth functions in maximum depth classifiers is explored using simulation and real data in the agriculture business. [45]

Dutta S. & Genton M. (2017) Use depth based estimate to put up regression estimate, and examine their performance with respect to existing estimators. To raise the efficiency of the estimators, a reweighted estimators depend on strong MD from the remaining vector has been suggested. The approach is widely stable than current approaches that are generated using sub-samples of dataset from an empirical point of view. The outcomes multivariate regression techniques are

arithmetically feasible, and has been shown to play best than many common robust multivariate regression methods when applied to diverse simulation dataset beside a real reference dataset. When the dimension of the data is too high compared to the sample size, meaningful concepts of data depth can still be used along with corresponding depths value to create a robust estimators in the sparse environment.[20]

Hubert M. et al. (2017) created classifications of multivariate and functional data in order to combine novel stability, robustness, and computational feasibility. On the basis of the halfspace depth, the bag distance (*BD*) has been proposed. It meets the majority of the features of a norm and can also represent asymmetry. Instead of delving into the facts. In addition, a DistSpace transformation based on bd or an outlyingness metric is proposed, followed by k-nearest neighbor (*kNN*) categorization of the changed data points. This combines *kNN*'s wide applicability and endurance with its stability and simplicity. The concept was tested against other approaches using actual and simulated data.[33]

Baghfalaki T. & Ganjali M.(2017) proposed a robust generalized estimating equations (*RGEE*) that depend on depths function and extend the method to robust weighted generalized estimating equations (*RWGEE*), which express centrality of points with respect to the whole sample with a smallest depths (largest depths) for the observation far from the center [1].

Harsh, A.et al. (2018) introduced and perform an adjusted "onion peeling" procedure to identify top-k outliers in the Gaussian two dimension dataset. The notation of "onion peeling", is to build a convex hull around all the observations in the bulk of data and then get the observations that place in the edge of the convex hull. These observations form the first “peel‟ should be remove from the data. By repeat the same procedures give more and more "peels", each of them contain

some of observations. The researchers adjusted this primary notation to identify the k largest outliers in a given two dimensions Gaussian dataset. For selection of k are influenced by the spatial geometry of the dataset and is user defined. The convex hull is the smaller convex set that contain all of the observation in the dataset [29].

Cabana E. et al. (2021) proposed a set of robust Mahalanobis distances based on the concept of shrinking to detect multivariate outliers. Shrinkage is best determined by estimating robust intensities and scaling factors. And some properties were investigated, including equation value and hash. When the normal assumption is not satisfied, the behavior in a simulation and a real data set shows the appropriateness of the method. The advantages of our proposal have been demonstrated by the giving high correct identification rate and low false identification rate in a big number of cases, as well as significantly shorter computation time[13].

González-De La Fuente et al. (2022) studied a statistical depths function with regard to compact convex stochastic datasets, that is proportionate with the multivariate Tukey's depths and the Tukey's depths for fuzzy set. Furthermore, it produes a various perspective to the existing half-space depths with regard to compact stochastic datasets. They produced a group of properties for the "statistical data depth", that constitute the axiomatic idea of multivariate, functional, and fuzzy depth functions and other common properties of depths [24].

Chapter Two

Theoretical Background

Chapter Two

Theoretical Background

2.1 Introduction:

The topic of detecting outliers in the data has taken a wide area in statistical research because of the sensitivity of the normal distribution to outliers, and despite the existence of robust methods, but because of the problems of masking and swamping, most of the clssical methods suffer from these problems in addition, excessive sensitivity of some methods to abnormal values. In this thesis we will applying the following diagnostic methods:-

- Mahalanobis distance (MD)
- Robust mahalanobis distance based on MVE which is depending (RMD)
- The proposed mahalanobis distance outlyingness method based on MRCD (MDO)
- The proposed robust spatial depth method based on MRCD (MDS)
- The proposed robust elliptical depth method based on MRCD (MDE)
- The proposed robust triangle depth method based on MRCD (MDT)

Therefore, in this thesis, it was discussed the above diagnostic methods to make sure the efficiency of the suggest methods and to get a resilient and robust detection methods.

2.2 Brief Overview of Outlier Detection

No observation can be guaranteed to constitute an entirely reliable manifestation of a phenomenon under examination, as Beckman and Cook (1983) pointed out. Observations that differ from the majority of the data, however, need consideration. The phrases outliers, discordant observations, Extreme values, "contaminants, enormous, massive and dirty are just a few that have been used to describe such observations in the literature. According to Beckman and Cook

(1983). Additionally, they used Edgeworth's definition of such observations from 1887, which goes as follows: “Discordant observations may be characterized as those which present the appearance of varying in relation to their law of frequency from other observations with which they are coupled”. Unquestionably, unusual observations have a lengthy record and over the years, the concept of outliers has become a little hazy. It is obvious that by identifying the "faulty" points in a dataset, one can gain a good understanding of the phenomenon being studied. However, it is much easier to spot inconsistent observations, which helps with inference and future predictions. According to Hadi, Imon, and Werner (2009), the "chicken-and-egg" conundrum analogously describes the research or the identification of outliers. Francis Bacon (1620) is credited with the following quote, which was derived from Billor, Hadi, and Velleman (2000): "Whoever understands the methods of Nature will more readily discern her flaws; while, on the other hand, whoever is aware of her deviations will be able to characterize her better”.

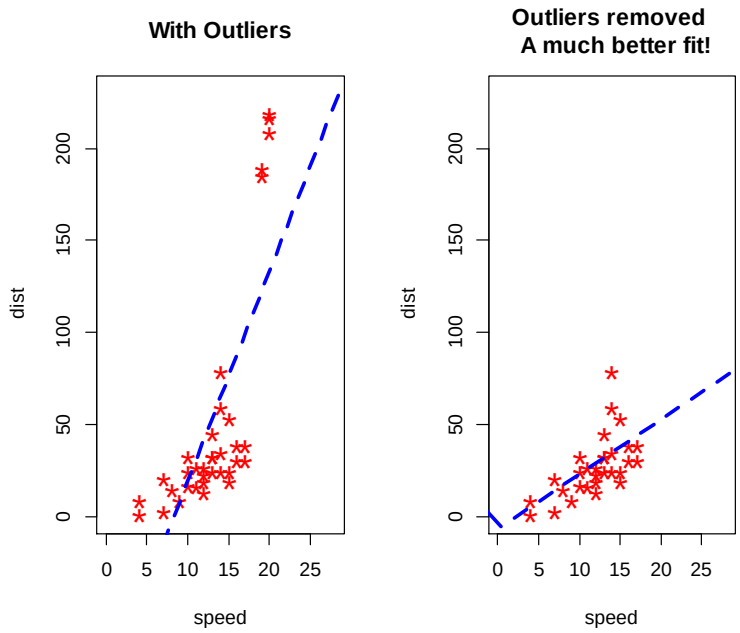


Figure (1): Shows the improvement of the fitting when the outliers are removed.

Figure 1 shows after the outliers have been removed, observe the change in the slope of the best fit line. If we had trained the model using the outliers (left chart), our predictions for higher values of speed would have been inflated (high error) due to the higher slope.

In the context of data analysis, it goes without saying that the study of outliers has attracted a lot of interest, ("see for instance, Tukey (1977), Barnett (1978), Hawkins (1980), Davies and Gather (1993), Barnett and Lewis (1994), Schwertman, Owens, and Adnan (2004), Schwertman and de Silva (2007), Dang and Serfling (2010) and Cerioli (2010)"). This is due to the fact that their non-identification or misidentification can significantly impact data processing, resulting in distortion and perhaps erroneous results. However, in certain cases, the unusual data itself are of importance because they may give a new information's or finding.

2.3 Mahalanobis Distance

The Mahalanobis Distance (MD) is a method for calculate of how far away the observation x_i is from the center of bulk of data Harsh M. et al (2018)[29]. let $\mathbf{X}$ be a $(n \times k)$ design matrix and $x_i = (x_{i1,}, x_{i2}, \dots, x_{ip})'$ be the i^{th} point, where n, is a size of sample and k is the number of predictors and " $'$ " stand for transpose. The arithmetic mean, M_x and covariance matrix, C_x are defined as

$$M_x = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \dots \qquad (2.1)$$

$$C_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - T_i)^t(x_i - T_i) \quad \dots \quad (2.2)$$

The traditional MD for the i^{th} case is defined as follows

$$MD_i = \sqrt{(x_i - T_i)^t C_x^{-1}(x_i - T_i)}\,, i = 1,2, \dots, n \quad \dots \quad (2.3)$$

Where, $T_i = (T_1, T_2, \dots, T_n)$ is a mean.

The MD_i is calculated for each point $i = 1,2, \dots, n$ and compare it with the cut-off point $\left(\sqrt{\chi^2_{k+1,0.95}}\right)$. Observations that exceeds the cut-off point is identify as outliers Harsh M.et al.(2018).

2.4 Bagplot of the Statistical Depth Function

The statistical depth function (*SDF*) is an approach suggested by John Tukey (1975). The *SDF* is determined how close an arbitrary point of the space has existed to an implicitly specified location of a data cloud. In addition, Tukey introduced a "depth median" (*DM*) which is the 'deepest' point in a specific data cloud (Tukey, 1975). The DM is the deepest point which is enclitic by a "bag" including half points with the largest depth. There are a lot of subjects like economics, social sciences that cannot be modeled easily, due to our knowledge of economic rules is not sufficient for effective parametric modeling or the datasets containing outliers or missing data, hence the *SDF* is the effective approach to deal with it [1, 7]. In the *SDF*, the bagplot is a modified shape for the well-known boxplot proposed by Rousseeuw, Ruts, and Tukey [42]. For the bivariate model, the graph of the boxplot replaces to the convex hull, which is the bagplot. As shown in figure 2, in the bag, there are fifty percent of all observations. The fence split up observations within and outside the fence. The loop is stated as the convex polygon that holds observations inside the fence, if observations lie in a straight line we will get the traditional boxplot. Increasing the bag by a proportion of 3 results in the "fence" as shown in figure 2. Observations between the bag and the fence are flagged by a light gray loop, whereas points outside the fence are identified as outliers. [1].

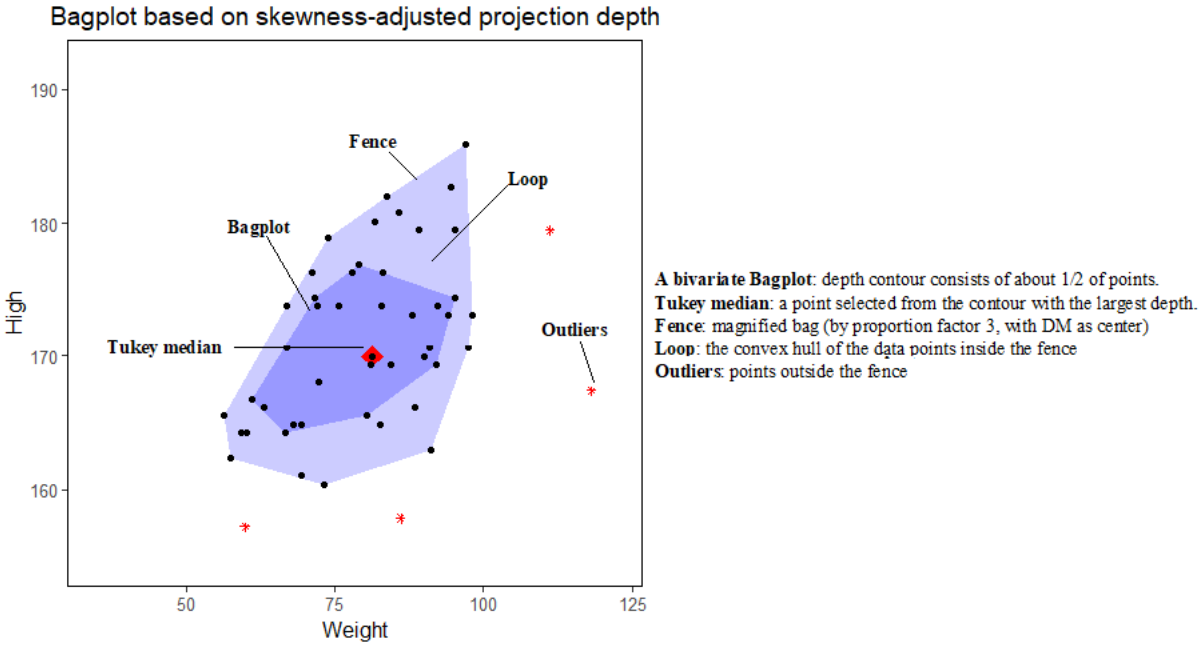


Figure 2: Bagplot based on adjusted projection depth [42]

2.5 Data Depth

A depth function may identify as "is a real-valued function that produces a center-outward ordering of the multivariate data". It is interesting to know, it can be used to detect unusual point in the data set. Recently, many depth function were suggested in the literature such as, Tukey half-space, projection depth, and Mahalanobis depth [1, 33, 43]. The depth data has desirable statistical properties due to it depend on the depth function $D(x; p)$. The depth function $D(x; p)$ has the following properties:

1- Affine invariant(*is one that does the same thing with or without an affine (linear usually) pre-conditioning.*), $D(x; P) = D(Ax + b;\ PAx + b)$ for every nonsingular matrix $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^{d}$.

2- Vanishes at infinity: $D(x; P) \Longrightarrow 0 \quad if\ \|x\| \Longrightarrow 0$, where $\|x\|\ is\ a\ norm$.

3- Upper semicontinuity: $\{x \in \mathbb{R}^{d}: D(x; P) \geq \alpha\}$ is closed.

4- Monotonicity*(is a function between ordered sets that preserves or reverses the given order.*) relative to deepest point.

The classical variance methods are extremely sensitive to outlying observations. A popular approach to deal with this issue is to use robust variance-covariance matrix, converianc matrix is very sensitive to outliring. One of the common practice to obtain robust estimators is to use concept of depth statistics [33]. Applying depth in construct for getting such estimator is easy due to depth has "center outward ordering". The depth has ability to increase at the center of dataset and minimize along all direction for that center [7]. Observations those extreme with respect to the bulk of data will be down weighted by depth.

The value of the depth is proportional "inversely" to the distance from the center for dataset, as the data is closer to the center the greater its depth. On the contrary, the lower the depth value, the further away from that center it is. With the foregoing, it's clear to see that the depth of a point can shift to its outlying-ness and conversely. It will be concluded from the above that the process of converting data depth, will help us to identify points with far outliers on the basis of specific cut-off point [42].

2.6 Univariate Outlier Detection:

The boxplot is an approach that most often utilized univariate to visualize outlier identification tools among researchers. It aids in the visualize of the shape, desperation, and skewness of the observation distributed, as well as unexpected values or outliers. Tukey's (1977) version of the boxplot is known as the typical boxplot.

Let $X_i,\ i = 1,2,\dots,n$ be an observations from a sample size n and $X_{\{i\}}$, $i = 1,2,\dots,n$ represent the order statistics of distribution. The typical boxplot's "lower fences" (*LF*) and "upper fences" (*UF*) are specified by:

$$LF = Q_1 - k(Q_3 - Q_1) \text{ and } UF = Q_3 + k(Q_3 - Q_1)$$

where Q_1 and Q_3 are the quartiles, which were originally defined as:

$$Q_1 = X_f \text{ and } Q_3 = X_{n-f+1}$$

Where $f = \frac{1}{2}\left[\frac{n+3}{2}\right]$, [.] represent the largest integer function, and k is called the fence constant and chosen typically decided to be 1.5 or 3(standard deviations), and so the outlying observations are those that lie under the lower fence or over the higher barrier (or potential outliers). Rosner (1983) introduced the (*GESD*) as a new approach for detecting outliers when the a dataset distributed as normal. In this technique, the practitioner can chose the large number of suspected unusual data. Brant (1990) did a comparison study to compare this method with the classical boxplot to determine that variants chosen appropriately for these principles perform similarly.

2.7 Multivariate Outlier Detection:

Multivariate data analysis are vital in different application contexts, and detecting suspected data in high dimension is difficult for a variety of purposes. These involve difficulties with vision and a lack of natural observational organization. Furthermore, it is not suitable to detect suspected observations for each dimension individually, because may covariates are frequently overlapping, and maybe data that diagnose to be outliers on the univariate may not be outliers on the multivariate model, and vice versa. There are two basic techniques to higher dimensional outlier identification/detection which are:

- Distance-based methods.
- Projection pursuit methods.

In this study, we will focus on distance-based methods, where the mahalanobis distance depth-based outlier identification method, which has gained popularity over the past 20 years, provides a more comprehensive framework.

2.8 Distance Based Outlier Detection

Calculating each multivariate observation's distance from the data's "center" and then sorting these scalar numbers is the primary notion underlying distance-based outlier identification algorithms. Outliers may exist at the farthest points. For

each multivariate observation $x_i \in \mathbb{R}^p$ $(p = 1,2,\dots,n)$, the distance based algorithms calculates traditional mahalanobis distance ("Mahalanobis, 1936"). According to a "center" of some multivariate data X, the observation x_i's classical Mahalanobis distance (CMD) from that point is:

$$CMD_i = \sqrt{(x_i - \mathrm{T})^T \sum^{-1} (x_i - \mathrm{T})}$$

where $\boldsymbol{T}$ and $\sum$ represent the typical sample average and sample variance matrix of dataset $\mathbf{X}$, respectively, where the multivariate data $\mathbf{X}$ represents n observations on k variables:

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

and

$$\mathbf{T} = \bar{\mathrm{X}} = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_n \end{bmatrix} \quad , \quad \mathrm{S} = \begin{bmatrix} S_{11} & \cdots & S_{1k} \\ \vdots & \ddots & \vdots \\ S_{k1} & \cdots & S_{kk} \end{bmatrix}$$

Also $\boldsymbol{S}$ (represent the sample variance-covariance matrix of the dataset $\mathbf{X}$) can be computed as follows:

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^t \, (x_k - \bar{x})'$$

where

$$x_k = \begin{bmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{bmatrix} , \qquad \bar{x}_k = \frac{1}{n} \sum_{j=1}^{n} x_{jk}$$

Observations with high CMD_i^2 values are frequently labeled as probable outliers. To determine whether an observation is an suspected data or not, it is usual to find the values of the squared distance CMD_i^2. When **X's** have multivariate normal. for clean data, the CMD_i^2 has χ_p^2 distribution. As a result, for a significance level α *(is defined as the fixed probability of wrong elimination of null hypothesis when in fact, it is true.),* every observation with CMD^2 value greater than $\chi_{p;\alpha}^2$ [100 (1-α)$^{\text{th}}$

percentile of the chi-square(k) dist.] is considered as outlier. Unfortunately, when the data contains several outliers, this approach of diagnose outliers is mostly affected by "masking" and "swamping" problems (see Becker and Gather, 1999). According to Hadi (1992) the extreme value do not necessarily has huge CMD_i value. A little value of CMD_i will emerge from a limited number of outliers that "attract" $\boldsymbol{T}$ in their direction and "inflate" $\sum$ (leading to masking). Sample mean and sample variance are also "attracted" into the direction of the cluster, "away" from certain data, which might product in some "excellent" observations having large CMD_i and so potentially being deemed outliers ("swamping"). To address this issues, robust version of MD named (RMD) were developed in the literature ("Rousseeuw and Zomeren, 1990"). The RMD is given for each observation $x_i \in \mathbb{R}^p (i = 1,2, \dots, n)$ as:

$$RD_i = \sqrt{(x_i - \mathrm{T})^T \sum^{-1} (x_i - \mathrm{T})}$$

where $T \in \mathbb{R}^p$ is a "robust estimator" of mean and $\sum$ is a "robust estimator" of variance-covariance matrix. Once more, for the sake of identify whether a point is an outlier or it is a clean, it is necessary to understand the distribution of the robust squared distances. This allows for the determination of a critical value (or at least approximately). It was suggested by Maronna and Zamar (2002) to approximate a chi-Square distribution for the robust Mahalanobis squared distances RD_i^2, where

$$D_i^2 = x_{p,0.95}^2 \left(\frac{RD_i^2}{median\ \{RD_i^2\}} \right)$$

According to "Hardin and Rocke (2005)", a scaled F- distribution can more accurately characterize the distributions of the robust distances RD_i^2, but unfortunately, this approach ha time-consuming. It is also preferable to select estimators $\boldsymbol{T}$ and $\sum$ with high breakdown points. The robust variance-covariance matrices *MCD*, and *MVE*, estimators were among the first to have significant breakdown points in location and scale estimation (see "Rousseeuw, 1984; Rousseeuw and Leroy, 1987, Rousseeuw and van Zomoren, 1990"). Although, the

(*MCD*) method estimates the mean and variance using the a subset of given data with the smallest sample variance determinant, it has disadvantage, where, is that it does not work well when the number of dimensions is greater than the sample size. As an alternative method, the "Minimum Regularized Covariance Determinant" (*MRCD*) can be used. Rather than of MCD, the covariance matrix of the *MRCD* is a convex combination of base matrix and, the sample scatter matrix of the subset. The *MRCD* matrix can be found for any dimension, is well conditioned by design, and retains *MCD*'s good robust features. It should be noted that the outlier identification methods discussed above presuppose a "multivariate normal distribution", which is difficult to explain in application, and the previously mentioned distributional outcome has not demonstrated to satisfy with not multivariate normal distribution.

2.9 Detecting Outliers by Using Data Depth:

Each multivariate observation should be given an univariate model that indicates there "location" in relation to the bulk of dataset as part of a more comprehensive framework for multivariate outlier detection. For instance, we may compute the “robust” MD of every observation from a “center” of the dataset and determine the cut-off depend on an ("approximate") distribution of the distances to identify outliers. (For further details (see "Rousseeuw and van Zomeren (1990), Rocke and Woodruff (1996), Filzmozer, Maronna, and Werner (2008), Hardin and Rocke (2005), and Cerioli (2010)"). The robust MD can calculated by use high breakdown, affine equivariant robust estimators of location and scatter, such as the *"MCD"* and *"MVE*" (see "Rousseew 1984; Rousseew and Leroy 1985"). Due to their strong breakdown ability and affine equivariant feature, these estimators can adapt to affine transformations of the data, such as rotations, translations, and changes in scale. They can also withstand up to 50% of outliers. The depth values may then be used to search for outliers in the data. The multivariate data are ordered "center-outward" by a depth function. In the literature, a number of depth

functions have been suggested. For examples the halfspace depth proposed by Tukey (1975), the simplified depth (Liu, 1990), the MD depth ("Liu and Singh, 1993; based on Mahalanobis, 1936"), the spatial depth ("Serfling, 2002; based on Chaudhuri, 1996"), the elliptical depth ("Elmore, 2005"), lately, the triangle's depth (Liu and Modarres, 2010). An observation gets closer to the data cloud's center and is therefore less outlying the deeper it is in relation to the bulk of dataset, or the higher its depth. Conversely, an observation's distance from the center increases with decreasing depth. Outlyingness and depth are therefore "inversely" correlated. The more outlyingness a point is, the lowers its depth, and the greater its outlyingness. It is clear from what came before that an observation's depth might become its outlyingness (and vice versa). Once a data set's depths have been converted to outlyingnesses, one may discover outliers by selecting observations that, depending on a cut-off, have extremely high outlyingness values, and determining the cutoff value, over which an observation's outlyingness must exceed in order for the observation to be labeled as a probable outlier, is the following stage in the procedure of outliers diagnose. When dealing with multivariate normal data, "Dang and Serfling -201") explored a few outlying function and provided cutoff value for the CMD outlying, halfspace or Tukey's outlyingness, and the "Stahel-Donoho outlyingness". They also provided a small example to demonstrate the various outlyingness functions' ability to find outliers:
They also provided a small example to demonstrate the ability of different remote functions to find outliers:

- The CMD Outlyingness,
- The RMD outlyingness based on MCD,
- Spatial depth,
- Elliptical depth,
- Triangle depth.

They used the same assumption that most classical outlier identification systems do. As previously noted, the outlyingness values may be used to a boxplot to detect

the extreme values, which correspond to the (possible) outliers. The majority of the outlyingness values stated above did not respond well to our attempts to use boxplots (the "classic boxplot", the "adjusted boxplot," and the "modified adjusted boxplot"): We note that the higher fences head for greater than 1, as a result, no outlier is found (due to outlyingness value is specified to be between (0,1)). Be aware that this strategy can be applied to multivariate normal data as well as skew-normal data. An analogous strategy was applied in one of Dang and Serfliang's (2010) examples.

2.10 Multivariate Normal Data and Outlier Detection:

The underlying data in most conventional outlier identification algorithms is assumed to have a multivariate normal distribution. As a result, their effectiveness may be called into doubt when the data is distorted. Transforming some or all of the variables and using standard theoretical procedures to the modified data is a popular solution to this problem. However, it is difficult to determine which transformation(s) to perform and which variables to select in order to make the observations as "multivariate normal". Furthermore, as Hubert. and Van. (2008) point out, this technique requires additional pre-processing and produces variables which is frequently un-interpretable. They suggested outliers identification approach influenced by the Stahel-Donoho estimators for multivariate data. Hubert & van der Veeken (2008) begin by adjusting the "Stahel-Donoho outlyingness" (*SDO*) to accommodate for a symmetry, giving rise for the idea of "adjusted outlyingness" (*AO*). Recall that the "Stahel-Donoho outlyingness" for an observation with regard to the dataset is given as follows for univariate data:

$$SDO^{(1)}(x_i - X_n) = \frac{|x_i - med(X_n)|}{MAD(X_n)},$$

Where $X_n = \{x_1, x_2, \dots, x_n\}$ and $med(X_n)$ is the median of the X_n and $MAD(X_n) = med_i(|x_i - med(X_n)|)$ is a "median absolute deviation". $MAD(X_n)$ is sometimes increased by a correction factor for un-biasedness (standard deviation

σ) at normal samples. Additionally provided was the point x_i's multivariate Stahel-Donoho outlyingness with regard to data X_n:

$$SDO_i = SDO^{(p)}(x_i, \mathbf{X}_n) = \sup_{\|u\|=1} SDO^{(1)}(u'x_i, u'\mathbf{X}_n)$$

"Hubert M. & Van der V. (2008)[31]" defined the *AO* of x_i concerning to dataset X_n as follows for univariate data[31]:

$$AO_i = AO^{(1)}(x_i, \mathrm{X}_n) = \begin{cases} \dfrac{x_i - med(\mathrm{X_n})}{w_2 - med(\mathrm{X}_n)} & if\ x_i > med(\mathrm{X}_n) \\ \dfrac{med(\mathrm{X}_n) - x_i}{med(\mathrm{X}_n) - w_1} & if\ x_i < med(\mathrm{X}_n) \end{cases}$$

where w_1 and w_2 are the lower and the upper fences of the "adjusted boxplot", respectively, identified by[31]:

$\left[Q_1 - 1.5e^{(-4MC)}IQR, Q_3 + 1.5e^{(3MC)}IQR\right]$ when $MC \geq 0$ and

$\left[Q_1 - 1.5e^{(-3MC)}IQR, Q_3 + 1.5e^{(4MC)}IQR\right]$ when $MC < 0$.

The scaling factor $(w_2 - med(\mathrm{X}_n))$ of $x_i - med(\mathrm{X_n})$ for observations in the higher tail is bigger than the scaling factor $(med(\mathrm{X}_n) - w_1)$ for points in the lower tail when the data distribution is rightly skewed (*MC*> 0). This avoids misclassifying typical data in the upper tail as outliers. Similar to the *SDO*, the adjusted outlyingness (*AO*) of point x_i with regard to data X_n is defined in the multivariate case[31].

$$AO_i = AO^{(p)}(x_i, \mathbf{X}_n) = \sup_{\|u\|=1} AO^{(1)}(u'x_i, u'\mathbf{X}_n)$$

2.11 Outlier Detection Based on Robust Mahalanobis Depth Functions

The outliers detection in a multi-variant situation was a difficult task. In such case, it's usual to assign a univariate (scalar) number to each data point that describes its 'location' within the data cloud. The location is frequently expressed in respect to a distance function, and extreme points were referred to as outliers. Based on the division of distances, the cut-off (breakpoint) value is chosen, which

is one of the methods of calculating the Mahalanobis distance for each data point. Potential outliers are defined as sites with distances larger than the cut-off. For an illustration of this strategy, Mahalanobis distances robust versions were computed by specific robust estimators as a generalization. The depths are computed through the depth functions, it is real valued function that allows the arrangement of multivariate dataset. Certain depth function has been suggested in the literature (see more information below). The outlying will increased when the lower the depths values of dataset. Contrary to depths and depths functions, the ideas of outlyingness functions were comparable. Unsurprisingly, the more a point's outlyingness, the more remote it is. After finding the outlyingness of the observations, the next stage is to determine which point, if any, has'severe' outlyingness. In this case, a breakpoint value may be beneficial. An outlier was defined as any observation having a greater outlyingness than the cut-off value. Dang and Serfling explored several outlyingness functions and offered breakpoint values for traditional Mahalanobis distance outlyingness, halfspace (or Tukey) outlyingness, and Stahel–Donoho outlyingness. They, like other traditional outlier detection algorithms, presumed that the essential division is multi-variant normal. For such distributions, it appears that very little work has been done. Hubert and Van der Veeken, Use a boxplot to alter skewness to discover outliers from outlyingness data, for example. This thesis emphasizes the following:

(i) Mahalanobis distance robustness Outlyingness (using estimators of minimum Regularized covariance determinant (*MRCD*)),

(ii) Mahalanobis' robust spatial outlyingness (using estimators of minimum Regularized covariance determinant (*MRCD*))

(iii) "Robust triangle depth Outlyingness" that depends on the "Liu and Modarres of triangle" depth.

(iv) "Robust elliptical Outlyingness" that depends on the "Elmore of robust elliptica"l depth.

In the literature, this type of comparison isn't currently available, but it will provide information based on the efficacy and relative strengths of several approaches.

2.11.1 Some outlyingness functions and their corresponding data depth functions

Tukey (1975) first introduces the concept of information profundity. To define the profundity work, as given in Dang and Serfling (2010), it is an assumed likelihood dissemination F on R^p, any work $D(x,F)$ that gives a centre-outward F - based requesting of perceptions $x \in R^p$ could be respected like profundity work. To observe the point that profundity work $D(x,F)$ is used to measure how much 'deep' or 'central' is the point x in respect to dissemination F. The more profound a point, the less it being probable an exception. Taking after Tukey's (1975) halfspace profundity, a few profundity, capacities have been proposed within the writing, counting the Mahalanobis profundity by Tukey the spatial profundity by "Serfling (2002)" depending on "Chaudhuri (1996)", the curved profundity by "Elmore (2005)", and most as of late, the triangle profundity by "Liu and Modarres (2010)". "Zuo and Serfling (2000)" outlined some ideal characteristics in their study on statistical depth functions. Affine invariance, maximality at the center, monotonicity with respect to the deepest point, and vanishing at infinity are all on this list. The great impact of these characteristics is affine invariance, that is the depth of an observation $x \in R^d$ with regard to the distribution F most not be affected by the "underlying coordinate system".

a- Outlyingness and Mahalanobis depth

The Mahalanobis distance is the foundation for the Mahalanobis depth function [39]. The robust *MCD* based *MD* outlyingness functions of data X is

RMDO(*x*, *X*). The robust *MD* depth functions *RMDE*(., *X*) is connected to robust *MCD* by the formula *RMDO*(*x*, *X*) =1-*RMDE*(*x*, *X*). Dang X. and Serfling R. (2010)[17] also made advantage of this outlyingness function.

For $x \in R^d$, we have

$$RMDE(x,\Sigma) = \left[1 + ((x - X_{MRCD})^T \Sigma_{MRCD}^{-1}(x - X_{MRCD}))^{1/2}\right]^{-1}$$

and subsequently

$$RMDO(x,\Sigma) = 1 - \left[1 + ((x - X_{MRCD})^T \Sigma_{MRCD}^{-1}(x - X_{MRCD}))^{1/2}\right]^{-1}$$

An observation is therefore a candidate for being an outlier if:

$$MD > \sqrt{\chi^2_{0.975}(m)}$$

where *m* being the number of variables Dang X. and Serfling R. (2010)[17].

b- Spatial depth and outlyingness:

Serfling R. (2002)[70] explicitly proposed the concept of spatial depth, which is based on Chaudhuri and Koltchinskii's concept of a spatial quantile. The study of our simulation that based on *MCD* is resilient Mahalanobis spatial outlyingness function, referred to *RMSO*, is represented through:

$$RMSO(x,F) = \left\| E_F\left(S\left(C(F_X)^{-1/2}(x - X_{MRCD})\right)\right)\right\|$$

where $C(F_X)$ being the *MCD* estimator of scatter Serfling R. (2002)[70], and the estimated value (E_F) and the vector sign function (S) in R^d represented by

$$S(x) = \begin{cases} \dfrac{d}{\| x \|} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0 \end{cases}$$

c- Outlyingness and Elliptical depth:

For a given random data $X_1, X_2, ..., X_n$ which distributed as F on R^d, assume that $e(X_i, X_j)$ is the elliptical region defined by:

$$e[X_i, X_j] = [t : (X_i - t_{MRCD})^T C_F^{-1} (X_j - t_{MRCD}) \le 0].$$

The function of sample triangle depth defined by Elmore R.(2005)[21] as follows:

$$EDE(x, C_{F_n}) = \frac{1}{\binom{n}{2}} \sum_{i<j}^{n} I[x \in e(X_i, X_j)]$$

Whereas; $I(A)$ represent the event A's standard indicator function.

The robust elliptical depth outlyingness function, referred to by *REDO*, used in the simulation research, is constructed.

$$REDO\ (x, C_{F_n}) = 1 - EDE\ (x, C_{F_n})$$

where $C(F_n)$ being the *MCD* estimator of scatter.

d- "Triangle depth" and "outlyingness":

Triangle depth was recently introduced by Liu Z. & Modarres R.(2010)[41]. The depth function of their sample triangle is described by:

$$TDE(x, F_n) = \frac{1}{\binom{n}{2}} \sum_{i<k}^{n} I\left(x \in \{t_{MRCD} : \| X_i - X_k \| > \max(\| t_{MRCD} - X_i \|, \| t_{MRCD} - X_k \|)\}\right)$$

Where the Euclidian norm is denoted by $\|.\|$. The triangle depth becomes affine-invariant when the Euclidian distance is replaced with the Mahalanobis distance in

$$\|X_i - X_k\| > \max\left(\|x - X_k\|, \|x - X_k\|\right)$$

according to Liu and Modarres (2010)[41]. The robust triangle depth function $RTDE(x, F_n)$ is the resultant depth function. The simulation study's associated robust triangle depth outlyingness function is given by:

$$RTDO\ (x, F_n) = 1 - RTDE\ (x, F_n)$$

2.11.2 Minimum Regularized Covariance Determinant (*MRCD*)

The objective of the "Minimum Covariance Determinant" (*MCD*) method ("Rousseeuw, 1984, 1985") is to locate *h* (*h*<*n*, *n* sample size) points that have the lower determinant of sample variance-covariance matrix. This method is a very reliable estimator of multivariate location and scatter. In order for the covariance matrix of any *h*-subset to be non-singular, the dimension *k* (number of independent variables) must satisfy the condition that $k < h$. In fact, it is frequently advised to use $n > 5k$ for the estimator's accuracy, for example in Rousseeuw et al. (2011). Due to this restriction, high breakdown approach which called "fat data," that have fewer rows (points) than columns, are not readily available (variables).

The modifying the *MCD* to apply to high dimensions in order to close this gap. The main ideas are to replace the subset based variance with a regularized variance estimate, which is specified as a weight average of the sample variance of the h subset and a pre-determined positive definite target matrix. The regularize variance based on the *h* subset, which results in the smallest overall determinant, is then the proposed Minimum Regularized Covariance Determinant (*MRCD*) estimator.

The good breakdown properties of the *MCD* estimator are preserved, and the *MRCD* estimator is a good conditioned by construction, in addition to being available for high dimensions. Due to the covariance matrix is guaranteed to be invertible, it can be used for graphical modeling, linear discriminant analysis, and computing robust distances.[50]

Furthermore, we will generalize Rousseeuw and Van Driessen's (1999) *C*-step theorem by demonstrating that the objective function is reduced when the h subset is concentrated to the h observations with the smallest robust distance computed from the regularized covariance. The suggested fast *MRCD* estimation algorithm is theoretically supported by the *C*-step theorem.

2.12 Mahalanobis Depth Functions Based on *MRCD*

The outliers detection in a multi-variant situation was a difficult task. In such case, it's usual to assign a univariate (scalar) number to each data point that describes its 'location' within the data cloud. The location is frequently expressed in respect to a distance function, and extreme points were referred to as outliers. Based on the division of distances, the cut-off (breakpoint) value is chosen, which is one of the methods of calculating the Mahalanobis distance for each data point. Potential outliers are defined as sites with distances larger than the cut-off. For an illustration of this strategy, Mahalanobis distances robust versions were computed by specific robust estimators as a generalization based on *MRCD*. The depths are computed through the depth functions, as it is a real valued function that allows the arrangement of multivariate data. The outlying will increased when the lower the depth of a data point. Unsurprisingly, the more a point's outlyingness, the more remote it is. After finding the outlyingness of observations, the next stage is to determine which point, if any, has'severe' outlyingness. In this case, a breakpoint value may be beneficial. An outlier is defined as any observation having a greater outlyingness than the cut-off value. In this thesis we emphasizes the following methods:

(i) Mahalanobis distance robustness Outlyingness based on *MRCD*

(ii) Mahalanobis' robust spatial outlyingness based on *MRCD*

(iii) "Robust triangle depth Outlyingness" that depends on the function of "Liu and Modarres" of triangle depth, and based on *MRCD*

(iv) "Robust elliptical Outlyingness" that depends on the function of Elmore of robust elliptical depth based on *MRCD*.

In the literature, this type of comparison isn't currently available, but it will provide information based on the efficacy and relative strengths of several approaches.

Chapter Three
Proposed Model and Experimental Results

Chapter Three
Proposed Model and Experimental Results

3.1 Introduction:

In this thesis, we investigate of the aim of study, which was to diagnose the outliers by proposed a new technique that involve merge the high efficiency "robust variance - covariance matrix" such as "*MRCD*" with some outlyingness depth function methods to obtain more accurate and effective approach to identify outliers in the dataset comparing with some existing methods. For evaluation of methods, different type of data were used such as, simulation data, artificial data and real data.

3.2 Simulation study:

In the simulation experiment, in order to evaluate our suggested methods, the proposed and existing methods were applied in different scenarios. The performance of methods of study, we apply some methods that illustrate in the Section 2, such that, *MD*, *RMD*, *MDO*, *MDS*, *MDT*, and *MDE*. To obtain comprehensive results for the simulation experiment, different sample sizes were used (50, 100, 150, 200), with different percentages contamination (0.05, 0.1, 0.15, 0.20). The multivariate linear model were used with response variable that distributed as normal distribution and 5 explanatory variables that distributed as normal distribution with vector of means (1, 2, 2.5, 3, 3.5) and constant variance equal to one. The vector of coefficients that use to generate the model is (0.5, 0.75, 1, 1.25, 1.5). The error term were distributed as standard normal distribution. The contamination of data were done by replace the clean observations by some others that generate from different distribution for error term such as normal distribution with mean 5 and variance 3 with different ratio (0.05, 0.10, 0.15, 0.20). For consistency of results, we repeat the experiments 500 times. To evaluate the

performance of method for ability to identify the outliers, we compute the ratio of correct identification for outliers and ratio of swamping for methods in all cases of studying. The swamping problem happen when we diagnose a good observation as outlier. The correct identification ratio and swamping ratio are computed as follows;

$$\text{Correct identification ratio} = \frac{\text{sum of outliers that identified correctly}}{\text{sum of outliers}}$$

$$\text{Swamping ratio} = \frac{\text{sum of clean data that identified wrongly as outliers}}{\text{sum of outliers}}$$

3.2.1 The results and discussion of simulation experiment:

Tables 1-5 present the correct identification ratios and swamping ratios for methods of study in various sizes of samples and various contamination ratios. From the results of tables, we found that the traditional mahalanobis method has the worst performance compared to other methods, due to it has low ability to correctly diagnose outliers, but at the same time, it did not diagnose any clean value as outliers, as the swamping ratio in all cases was equal to zero, except when the contamination ratio 0.05, where the swamping percentage was 0.20. as well, we note that the *RMD* method has high diagnostic ratios for correctly diagnose outliers compared with the traditional md method, but at the same time it has high swamping ratios at most contamination rates. On the other hand, we found that all of the proposed methods have good performance due to they have high ratios of correctly detection of outliers without any swamping ratios. Furthermore, it is attractive to see that the robust outlyingness depth method (*MDO*) has exhalant performance with 100% ratio of detection of outliers without any swamping observation. We used the R-programing to analyze the results and conduct simulation experiments.

Table 1: Correct identification ratio and swamping ratio for methods of study with n=50 and different contamination ratios

Contamination		**0.05**	**0.10**	**0.15**	**0.20**
MD	Correct identification ratio	40.0	26	37.1	25
RMD		85.0	52	74.3	50
MDO		100.0	100	100.0	100
MDS		100.0	100	84.3	61
MDT		100	100	75.7	52.0
MDE		100	76	57.1	42.0
MD	Swamping Ratio	0.2	0	0	0
RMD		16.0	10	6.0	0
MDO		0.2	0	0	0
MDS		0	0	0	0
MDT		0	0	0	0
MDE		0	0	0	0

Table 2: Correct identification ratio and swamping ratio for methods of study with n=100 and different contamination ratios.

Contamination		0.05	0.10	0.15	0.2
MD	Correct identification ratio	42	31	26.0	25.5
RMD		84	62	52.0	51.0
MDO		100	100	100.0	100.0
MDS		100	100	77.3	60.0
MDT		100.0	99.0	71.3	54.5
MDE		100	79.0	55.3	44.0
MD	Swamping Ratio	0	0	0.0	0.0
RMD		15	10	5.0	0.0
MDO		0	0	0.0	0.0
MDS		0	0	0	0
MDT		0	0	0	0
MDE		0	0	0	0

Table 3: Correct identification ratio and swamping ratio for methods of study with n=200 and different contamination ratios

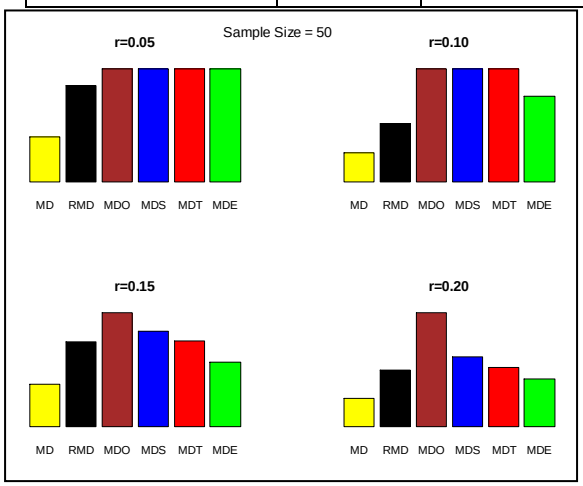
contamination		0.05	0.10	0.15	0.2
MD	Correct identification ratio	46	33	28.0	23.2
RMD		92	66	56.0	46.5
MDO		100	100	100	100
MDS		100	100	77.3	60.0
MDT		100	97.5	71.7	57.5
MDE		100	75	58.3	46.2
MD	Swamping Ratio	0	0	0	0
RMD		15	10	5.0	0.0
MDO		0	0	0.0	0.0
MDS		0	0	0	0
MDT		0	0	0	0
MDE		0	0	0	0

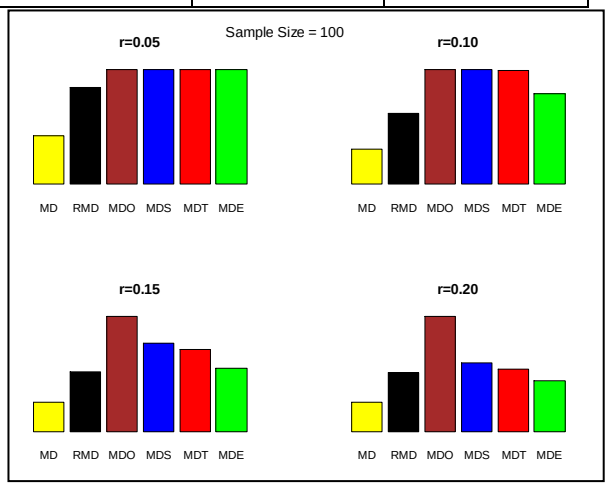
Table 4: Correct identification ratio and swamping ratio for methods of study with n=300 and different contamination ratios

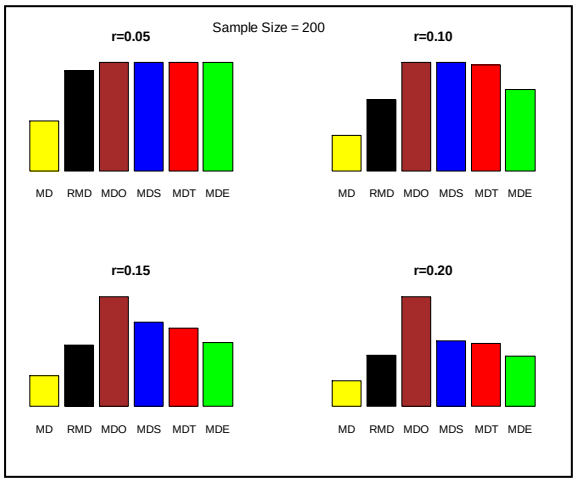
Contamination		0.05	0.10	0.15	0.2
MD	Correct identification ratio	38	35.7	27.3	22.8
RMD		76	71.3	54.7	45.7
MDO		100	100	100	100
MDS		100	100	77.6	59.5
MDT		100	98.3	72.2	53.3
MDE		100	81.7	55.6	42.5
MD	Swamping Ratio	0	0	0	0
RMD		15	10.0	5.0	0.0
MDO		0	0	0.0	0.0
MDS		0	0	0	0
MDT		0	0	0	0
MDE		0	0	0	0

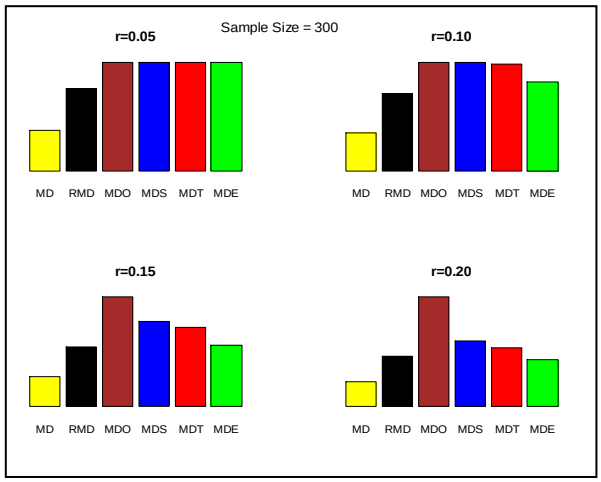
Table 5: Correct identification ratio and swamping ratio for methods of study with n=500 and different contamination ratios

Contamination		**0.05**	**0.10**	**0.15**	**0.2**
MD	Correct identification ratio	45.6	36	29.7	23.4
RMD		91.2	72	59.5	46.8
MDO		100	100	100	100
MDS		100	100	77.2	59.9
MDT		100	100	73.3	52.5
MDE		100	78.8	56.7	42.5
MD	Swamping Ratio	0	0	0	0
RMD		15	10	5.0	0
MDO		0	0	0.0	0.0
MDS		0	0	0	0
MDT		0	0	0	0
MDE		0	0	0	0









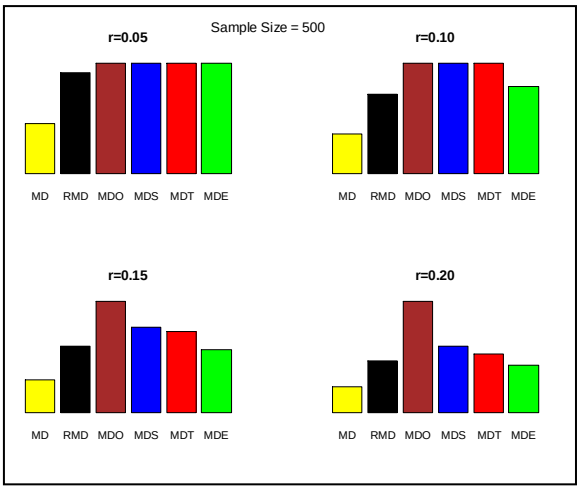


Figure 3: Histogram represents Correct identification ratios and swamping ratios for methods of study with different sample size of sample and different contamination ratios

3.3 Example and discussion

In this section, we consider two examples to investigate the targets of this study and to assess the performance of depth concept comparable with classical Mahalanobis distance to identify outliers.

3.3.1 First example (Artificial data)

The first example is a simulation study that presented in Table 6, where the variables are generated normally with a specified location parameter and fixed scale. The data demonstrate the relationship between independent and response variables with 14 observations. Suppose, "Spending" represents the response variable and "*GDP*" represent the independent variable, then the regression equation is given by:

$$\text{Spending} = \beta_0 + \beta_1\ (GDP) + \ \varepsilon$$

For the purpose of examining the depth approach, we contaminated the first three points in the response variable with an arbitrary big value. Table 7 and Figure 4 (a and b) present the classical *MD* values and depth values for the points of the dataset. It's clear to see that observations numbered (1, 2, and 3) are assigned

correctly as outliers according to depth values, whereas, the *MD* approach fails to identify those points as outliers. On the other hand, we observed from Figure 5 that the boxplot fails to diagnose the outliers (Figure 5-a), whereas the bagplot is identified those outliers correctly as shown in Figure (5-b).

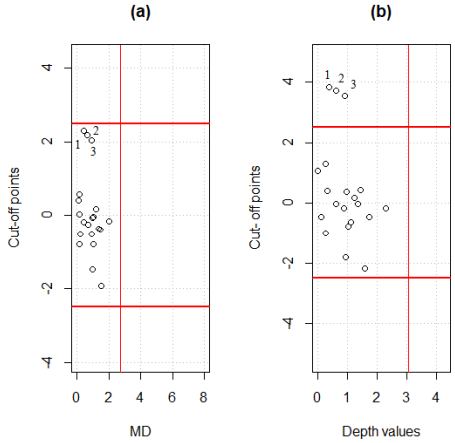
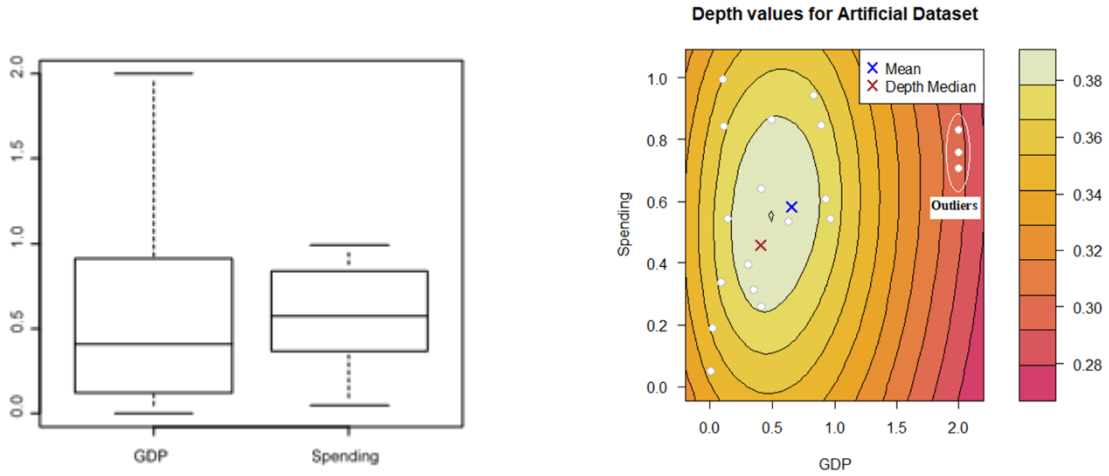


Figure 4: plot of:
(a) *MD* against cut-off point (b) Depth values against cut-off point

Table 6: An artificial data

no.	x	y
1	12985.30	101843.9
2	9325.50	103546.6
3	7643.90	109393.1
4	5897.00	111455.8
5	10061.80	120626.5
6	45230.20	124702.8
7	55398.20	132687.0
8	58103.30	142700.2
9	73137.40	162587.5
10	3269.98	174990.0
11	55823.50	175335.4
12	46369.40	183616.3
13	62810.30	208932.1
14	70435.80	205268.1



(a) **(b)**

Figure 5: (a) boxplots and (b) bagplot

3.3.2 The second example

The second example is related to real data on *GDP* and public spending for the period from 2004 to 2017 in the Republic of Iraq. Data were collected from the Central Statistical Organization of Iraq. By drawing the spread plot of values based on the *MD* and depth values, we find that the *MD* has diagnosed one outlier (point 10), while the depth values have diagnosed three outliers (10, 13, and 14). By applying the regression many times depending as follows:

Table 7: depth values and *MD* values for generated dataset

Observation N0.	MD (2.716)	Depth values (2.35)
1	0.46	**4.58**
2	0.92	**4.35**
3	0.65	**4.49**
4	0.14	0.92
5	1.99	-0.50
6	0.70	-0.12
7	1.01	0.02
8	1.54	-1.03
9	0.46	0.02
10	0.97	-0.80
11	1.35	0.47
12	0.92	-0.47
13	0.10	0.82
14	1.05	-0.10
15	1.21	0.15
16	1.47	-0.52
17	0.22	-0.09
18	0.98	0.63
19	0.14	-0.50
20	0.18	0.35

From Table 8, we note that the standard error of the regression model is (29390), this value decreases when deleting the tenth observation that was diagnosed by relying on the *MD*, while we find that this value decreases significantly when deleting the three observations that were diagnosed based on the depth values, where its value was (19070) and this It indicates that the depth approach has a good potential to detect abnormal values in the data

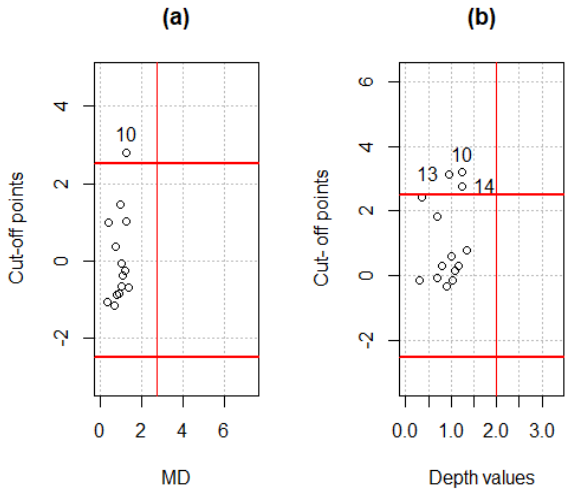


Figure 6

3.4 Real Data

PM10 is thick pendent particulate matter, stiff or fluid, that has less than ten micrometer for thickness. It can be including smolder, dust, smut, acid, and metal. Commonly, these particulate matter are floating dust in the air. These particles can remain suspended in the air for different periods ranging from several days to weeks, which in turn leads to their transmission over long distances. However, these dust particles fall to the ground due to gravity and rain, which is considered one of the natural sources of air purification. There is another type of dust particles that differs from PM10 in size and is called PM2.5. PM2.5 is very fine and more dangerous than PM10 because it can pass from the lungs into the bloodstream through inhalation, but in some cases PM10 can pass through the lungs too. Both PM2.5 and PM10 have significant damage to the environment and human health. Where short-term health symptoms can appear, such as: difficulty breathing, coughing, burning in the eyes and nose, throat irritation, chest tightness, pain, and general respiratory fatigue. Long-term symptoms are more serious health concerns, such as "lung tissue damage, asthma, heart failure, cancer, adverse birth outcomes, chronic obstructive pulmonary disease , and premature death".

The Government of Al- Diwaniyah constructed Al-Diwaniyah Environment Station as building was placed in the city center of the, as well as the Directorate was equipped with a mobile station from the development projects of the regions to know the extent of the increase in these gases to control it and by knowing the types and quantity of pollutants from the gases emitted from the air that are emitted by factories and electrical stations because the used fuel contains sulfur compounds, poor fuel and the emission of large quantities of lead compounds from them. Measure and control the amount of significant pollution in the air. In This study, we collected pollution data (PM10) from Al-Diwaniyah Environment Station for the period from 2019-2021 as shown in Table 7, Appendix A. The data contain one response variable (PM10) and eight independent variables x_1 to x_8. The independent variables are illustrated in Table 8.

Table 8: independent variables of PM10

X_1	**TSP** - The term "total suspended particulate" (TSP) refers to all small solid particles that have been discharged, recorded, or otherwise noticed in the atmosphere. Total suspended particles are thought to be the main cause of smog formation, environmental contamination, and air pollution.
X_2	**O_3** - Three oxygen atoms make up the highly reactive gas known as ozone
X_3	**CO_2** - "Is a chemical substance composed up of molecules with one carbon atom double-bonded covalently to two oxygen atoms in each one of the molecules. At room temperature, it exists as a gas"
X_4	**CO** - Is a colorless and odorless gas that is a chemical molecule made up of one carbon and one oxygen atom
X_5	**SO_2** - The chemical substance with the formula SO_2 is known as sulfur dioxide (IUPAC's suggested spelling) or sulphur dioxide (traditional Commonwealth English)
X_6	**NO_2** - Chemically, nitrogen dioxide has the formula NO_2

$\mathbf{X_7}$	**NO** - Nitrogen oxide[6] or nitrogen monoxide, sometimes known as nitric oxide, is a colorless gas with the chemical formula NO
$\mathbf{X_8}$	**NOx** - "Is usually used to include two gases-nitric oxide (NO), which is a colourless, odourless gas and nitrogen dioxide (NO_2), which is a reddish-brown gas with a pungent odour"

3.4.1 Results and discussion of PM10 dataset

Table 9 illustrate the values of *MD*, *RMD*, *MDO*, *MDS*, *MDT*, and *MDE*. The observation declare as outliers when the value of mahalanobis distance exceed the cut-off point. Cut-off values are indicated in bold in the title of the Table 9 below each method whereas, the outliers that diagnosed for each method are indicated in bold in the same table. The *MD* identified just one outliers (case 27), whereas, the *RMD* identified three (cases 26, 27, and 28). On the other hand, the proposed methods (*MDO*, *MDS*, *MDT*, and *MDE*) diagnosed two outliers (cases 26, and 27) as shown in Figure 7. To investigate which method correctly diagnose the right outliers, we remove the suspected observations for each method, then we compute the standard error, the best model that has lowest standard error. From Table 10, we can clearly see that the best model when we remove cases 26 and 27, conclude that the proposed methods have correctly diagnose the suspected outliers while the *RMD* swamp one more observation (case 28), whereas the *MD* mask one observation (case 27). The results of this example are consistent with the finding of simulation study.

Table 9: The values of mahalanobis distance for method of study with cut-off values

No.	MD **9.926**	RMD **2.474**	MDO **0.788**	MDS **0.919**	MDT **0.960**	MDE **0.907**
1	0.036	0.401	0.103	0.091	0.438	0.481
2	1.194	1.052	0.412	0.452	0.551	0.592
3	3.620	1.946	0.575	0.668	0.698	0.700
4	0.278	0.732	0.266	0.253	0.470	0.487
5	0.819	1.002	0.433	0.546	0.732	0.571
6	3.613	2.002	0.578	0.652	0.679	0.687
7	1.592	2.038	0.735	0.808	0.886	0.835
8	0.645	1.028	0.582	0.525	0.619	0.632
9	0.589	0.995	0.630	0.564	0.675	0.651
10	0.169	0.933	0.383	0.505	0.638	0.586
11	3.620	1.946	0.575	0.668	0.698	0.700
12	0.819	1.002	0.433	0.546	0.732	0.571
13	0.411	1.039	0.361	0.451	0.568	0.562
14	0.274	1.070	0.489	0.703	0.781	0.790
15	2.129	1.896	0.528	0.685	0.752	0.786
16	0.620	1.148	0.660	0.645	0.741	0.702
17	1.478	1.515	0.697	0.758	0.803	0.767
18	0.039	0.420	0.119	0.156	0.446	0.489
19	1.646	1.329	0.613	0.607	0.673	0.670
20	0.819	1.002	0.433	0.546	0.732	0.571
21	0.770	0.973	0.599	0.506	0.625	0.611
22	0.640	0.755	0.360	0.315	0.484	0.508
23	0.354	0.881	0.420	0.549	0.665	0.667
24	0.263	0.699	0.552	0.409	0.551	0.562
25	2.386	2.459	0.785	0.912	0.946	0.895
26	2.709	**2.857**	**0.797**	**0.941**	**1.000**	**0.943**
27	**22.405**	**9.085**	**0.916**	**0.971**	**1.000**	**0.943**
28	3.095	**2.508**	0.738	0.898	0.946	0.889
29	1.383	1.212	0.472	0.676	0.760	0.784
30	0.675	0.846	0.505	0.370	0.505	0.530
31	3.613	2.002	0.578	0.652	0.679	0.687
32	0.169	0.933	0.383	0.505	0.638	0.586
33	1.401	1.954	0.733	0.789	0.851	0.800
34	1.459	1.470	0.446	0.526	0.581	0.613
35	0.819	1.002	0.433	0.546	0.732	0.571
36	3.448	2.360	0.591	0.792	0.863	0.865

Table 10: Standard error values for the regression model

Model	standard error
Full data	52.91
Full data except point(26)	53.42
Full data except point(26, and 27)	**48.57**
Full data except point(26, 27, and 28)	54.70

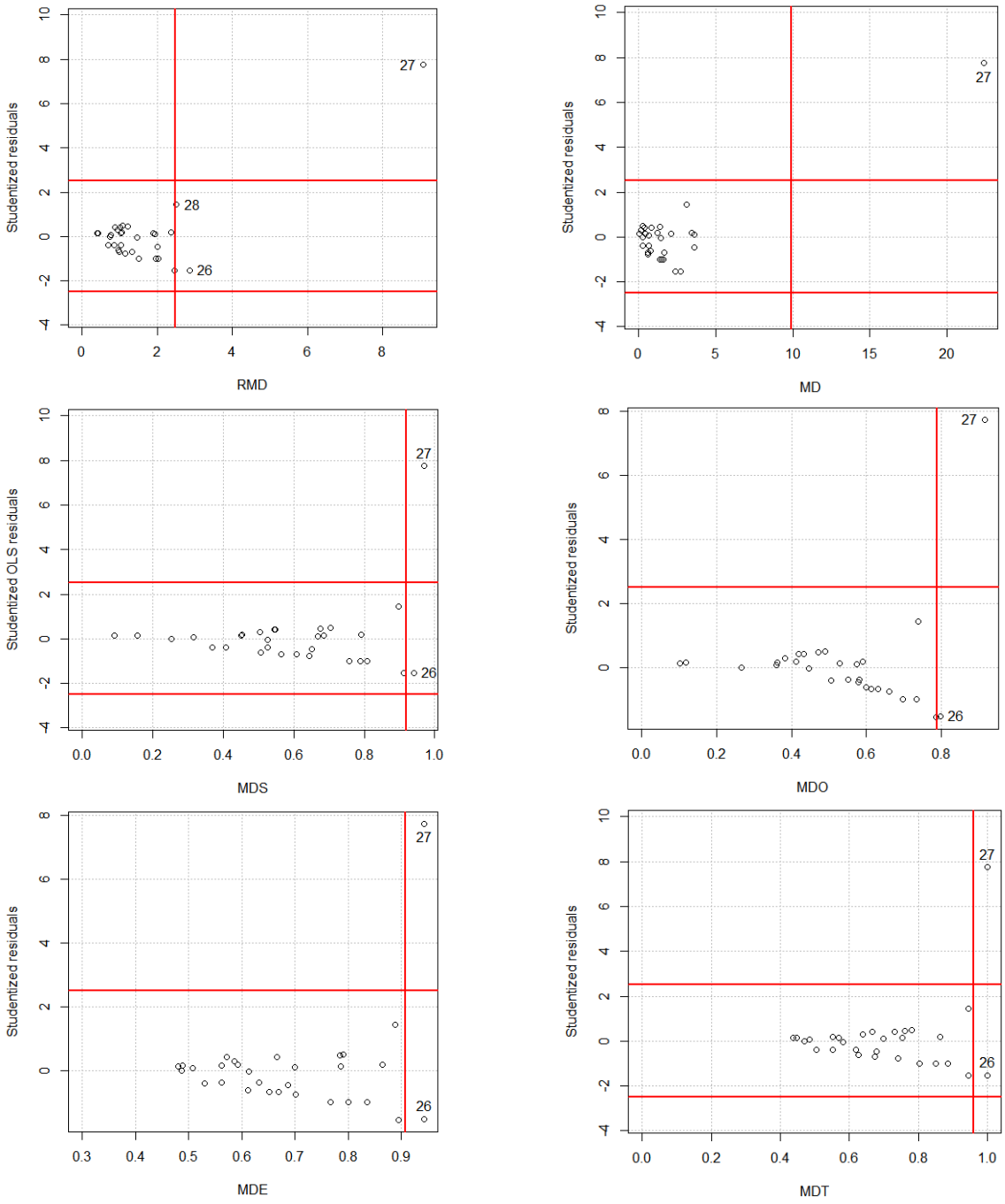

Studentized residuals
RMD
MD
Studentized OLS residuals
MDS
MDO
MDE
MDT
27
28
26

Figure 7

Chapter Four

Conclusion and Future Work

4.1 Conclusion

Through the applying the suggested diagnosis methods for simulation study and real data, and comparing them with some existing methods, we conclude the following:

1- The classical method, such that classical mahalanobis distance, has bad performance for diagnose outliers, but unfortunately, it suffers from masking problem.
2- The robust method, such that robust mahalanobis distance based on MVE, has good performance for diagnose outliers comparing with classical mahalanobis distance, but it suffers from swamping problem.
3- For univariate model, the suggest bagplot has succeed in diagnosing outliers ,whereas, the classical boxplot fails to diagnose them correctly.
4- For multivariate model, the proposed diagnostic methods such that *MDO*, *MDS*, *MDT*, and *MDE* have high ratios of diagnosis outliers with the lowest ratios of masking and swamping issues.
5- Form the simulation experiments, we conclude that the robust mahalanobis depth outlyingness method (MDO) has supermom performance for diagnostic outliers follows by robust mahalanobis depth spatial method (MDS).
6- From real data (PM10 dataset), we conclude that the proposed methods have correctly diagnose the suspected outliers while the RMD has swamp problem and the MD has mask problem.

4.2 Future Work

Several ideas are addressed and proposed as potential directions for future study in outlier detection, with the goal of enhancing the classical methods while focusing on more outlier applications and scenarios.

Several ideas are discussed and suggested for future work that aims to offer new development of the classical methods for outlier detection:

- How to use the *MRCD* method to become more efficient with different sample sizes and different levels of pollution.
- Validation of the MRCD method with a high dimensional data set.
- Using the Cut -off point with different methods to give more specific results.

References

References

[1] Baghfalaki, T. & Ganjali, M. (2017). Robust Weighted Generalized Estimating Equations Based on Statistical Depth. Communications in Statistics-Simulation and Computation. 46. 10.1080/03610918.2016.1277746.

[2] Barnett, V. (1978). "The Study of Outliers: Purpose and Model," Applied Statistics, 27, 3, 242- 250.

[3] Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data. 3rd Edition, John Wiley & Sons, Kluwer Academic Publishers, Boston/Dordrecht/London.

[4] Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. The Annals of Mathematical Statistics 22 (1), 107–111.

[5] Beckman, R. J., & Cook, R. D. (1983). Outliers. Technometrics, 25, 119–163.

[6] Billor, N. et al. (2000). "BACON: blocked adaptive computationally efficient outlier nominators." Computational Statistics & Data Analysis 34: 279-298.

[7] Berrendero, J.R. & Justel, A. & Svarc, M. (2011). Principal components for multivariate functional data. Computational Statistics & Data Analysis. 55. 2619-2634. 10.1016/j.csda.2011.03.011.

[8] Boudt, K. & Cornilly, D. & Holle, F. & Willems, J. (2020). Algorithmic portfolio tilting to harvest higher moment gains. Heliyon. 6. e03516. 10.1016/j.heliyon.2020.e03516.

[9] Brant, C. C. (1990). Native ethics and rules of behaviour. The Canadian Journal of Psychiatry / La Revue canadienne de psychiatrie, 35(6), 534–539.

[10] Bremner, D. & Fukuda, K. & Rosta, V. (2006). Primal-dual algorithms for data depth. 10.1090/dimacs/072/12.

[11] Bremner, D. & Chen, D. & Iacono, J. & Langerman, S. & Morin, P. (2008). Output-sensitive algorithms for Tukey depth and related problems. Statistics and Computing. 18. 259-266. 10.1007/s11222-008-9054-2.

[12] Burr, M. & Rafalin, E. & Souvaine, D. (2011). Dynamic Maintenance of Half-Space Depth for Points and Contours.

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

[14] Cerioli, A. (2010), "Outlier Detection With High-breakdown Estimators," Journal of the American Statistical Association, 105, 147–156.

[15] Chaudhuri, P. (1996), "On a Geometric Notion of Quantiles for Multivariate Data," Journal of the American Statistical Association, 91, 862–872.

[16] Croux, C., & Haesbroeck, G. (1999). Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. Journal of Multivariate Analysis, 71(2), 161-190.

[17] Dang X., and Serfling R. (2010), "Nonparametric Depth-based Multivariate Outlier Identifiers, and Masking Robustness Properties," Journal of Statistical Planning2018 and Inference, 140, 198- 213.

[18] Davies, L., & Gather, U. (1993). The Identification of Multiple Outliers: Rejoinder. Journal of the American Statistical Association, 88(423), 797–801. https://doi.org/10.2307/2290768

[19] Donoho, D. & Gasko, M. (1992). Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. Ann. Stat.. 20. 10.1214/aos/1176348890.

[20] Dutta, S. & Genton, M. (2017). Depth-weighted robust multivariate regression with application to sparse data. Canadian Journal of Statistics. 45. 10.1002/cjs.11315.

[21] Elmore, R.T. (2005), "An Affine-invariant Data Depth Based on Random Hyperellipses," Workshop, Colorado State University

[22] Evans K, Love T, Thurston SW. (2015). "Outlier Identification in Model-Based Cluster Analysis". J Classif. 32(1):63-84. doi: 10.1007/s00357-015-9171-5.

[23] Filzmoser, P., Maronna, R. and Werner, M. (2008) Outlier Identification in High Dimensions. Computational Statistics & Data Analysis, 52, 1694-1711.

[24] González-De La Fuente, L., Nieto-Reyes, A., Terán, P. (2022). Properties of Statistical Depth with Respect to Compact Convex Random Sets: The Tukey Depth. Mathematics, 10, 2758. https://doi.org/10.3390/ math10152758.

[25] Hadi A. S. (1992), "Identifying Multiple Outliers in Multivariate Data," Journal of Royal Statistical Society, Series B, 54 (3), 761-771.

[26] Hadi, A. & Imon, A. & Werner, M. (2009). Detection of outliers. Wiley Interdisciplinary Reviews: Computational Statistics. 1. 57 - 70. 10.1002/wics.6.

[27] Hallin, M. & Paindaveine, D. & Siman, M. (2010). Multivariate Quantiles and Multiple-Output Regression Quantiles: From L1 Optimization to Halfspace Depth. The Annals of Statistics. 38. 635-669. 10.1214/09-AOS723.

[28] Hardin J, and Rocke D.M. (2005), "The Distribution of Robust Distances," Journal of Computational and Graphical Statistics, 14, 928–946.

[29] Harsh, A., Ball JE, Wei P. (2018). "Onion-Peeling Outlier Detection in 2-D data Sets.", International Journal of Computer Applications, 139(3)pp. 26-31.

[30] Hawkins, D. (1980) Identification of Outliers. Chapman and Hall, Kluwer Academic Publishers, Boston/Dordrecht/ London.

[31] Hubert, M., & Van der Veeken, S. (2008). Outlier detection for skewed data. Journal of Chemometrics, 22.

[32] Hubert, M., Vandervieren, E. (2008), "An Adjusted Boxplot for Skewed Distributions," Computational Statistical Data Analysis, 52, 5186-5201.

[33] Hubert, M. & Rousseeuw, P. & Segaert, P. (2017). Multivariate and functional classification using depth and distance. Advances in Data Analysis and Classification. 11. 445–466. 10.1007/s11634-016-0269-3.

[34] Ieva, F. & Tarabelloni, N. & Paganoni, A. & Biasi, R. (2015). Use of Depth Measure for Multivariate Functional Data in Disease Prediction: An Application to Electrocardiograph Signals. The International Journal of Biostatistics. 11. 10.1515/ijb-2014-0041.

[35] Johnson, T., Kwok, I., and Ng, R. (1998). Fast computation of 2- dimensional depth contours. In: Agrawal, R., and Stolorz, P. (eds.), Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, AAAI Press, New York, 224–228.

[36] Kong, L. & Mizera, I. (2008). Quantile tomography: Using quantiles with multivariate data. Statistica Sinica. 22. 10.5705/ss.2010.224.

[37] Li, X. , Deng, S. , Li, L. and Jiang, Y. (2019) Outlier Detection Based on Robust Mahalanobis Distance and Its Application. Open Journal of Statistics, 9, 15-26. doi: 10.4236/ojs.2019.91002.

[38] Liu R.Y. (1990). "On a Notion of Data Depth Based on Random Simplices," The Annals of Statistics, Ann. Statist. 18(1), 405-414.

[39] Liu, R.Y., Singh, K. (1993). A quality index based on data depth and multivariate rank tests. J. Amer. Statist. Assoc. 88, no. 421, 252–260.

[40] Liu, X. & Zuo, Y. (2014). Computing Halfspace Depth and Regression Depth. Communications in Statistics: Simulation and Computation. 43. 10.1080/03610918.2012.720744.

[41] Liu, Z. and Modarres, R. (2010). A Triangle Test for Equality of Distribution Functions in High Dimensions (2010). Journal of Nonparametric Statistic, Vol. 22, no. 6, 1-11.

[42] Lok, W. S. & Lee S. M. (2011) A new statistical depth function with applications to multimodal data, Journal of Nonparametric Statistics, 23:3, 617-631, DOI: 10.1080/10485252.2011.553953.

[43] Lopez-Pintado, S. & Sun, Y. & Lin, J. & Genton, M. (2014). Simplicial band depth for multivariate functional data. Advances in Data Analysis and Classification. 8. 321-338. 10.1007/s11634-014-0166-6.

[44] Lopuha¨a, H. and Rousseeuw, P. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. The Annals of Statistics 19, 229–248.

[45] Makinde, O. S. & Adewumi, A. D. (2017). A comparison of depth functions in maximal depth classification rules. Journal of Modern Applied Statistical Methods, 16(1), 388-405. doi: 10.22237/jmasm/1493598120.

[46] Maronna, R.A., & Zamar, R.H. (2002). Robust Estimates of Location and Dispersion for High-Dimensional Datasets. Technometrics, 44, 307 - 317.

[47] Miller, K. & Ramaswami, S. & Rousseeuw, P. & Sellarès, J. & Souvaine, D. & Streinu, I. & Struyf, A. (2003). Efficient computation of location depth contours by methods of computational geometry. Statistics and Computing. 13. 153-162. 10.1023/A:1023208625954.

[48] Mosler, K., Lange, T., and Bazovkin, P. (2009), "Computing zonoid trimmed regions of dimension d > 2", Computational Statistics and Data Analysis 53, 2500-2510.

[49] Öllerer, V. & Croux, C. (2014). Robust High-Dimensional Precision Matrix Estimation. 10.1007/978-3-319-22404-6_19.

[50] Öllerer, V. and Croux C. (2015). Robust high-dimensional precision ¨ matrix estimation. In Modern Nonparametric, Robust and Multivariate Methods. Springer, pp. 325–350.

[51] Rainer D., Pavlo M.. (2016). Exact computation of the halfspace depth, Computational Statistics & Data Analysis, Volume 98, PP. 19-30, ISSN 0167-9473, https://doi.org/10.1016/j.csda.2015.12.011.

[52] Reyes, A. & Cuesta-Albertos, J. (2015). M. Hubert, P. Rousseeuw and P. Segaert: Multivariate functional outlier detection. Statistical Methods & Applications. 24. 10.1007/s10260-015-0319-6.

[53] Regina Y. Liu (1990). "On a Notion of Data Depth Based on Random Simplices," The Annals of Statistics, Ann. Statist. 18(1), 405-414.

[54] Rocke D. M., and Woodruff D. L. (1996), "Identification of Outliers in Multivariate Data," Journal of the American Statistical Association, 91, 1047–1061.

[55] Robson, G. (2003). "Multiple outlier detection and cluster analysis of multivariate normal data.".

[56] Rosner, B. (1983), "Percentage Points for the Generalized ESD Many-outlier Procedure," Technometrics, 25, 165-172.

[57] Rousseeuw, P. (1984). Least median of squares regression. Journal of the American Statistical Association 79 (388), 871–880.

[58] Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (Eds.), Mathematical Statistics and Applications, Vol. B, pp. 283–297.

[59] Rousseeuw, P. J., and Leroy A. M. (1987), Robust Regression and Outlier Detection, New York: John Wiley.

[60] Rousseeuw, P. J., and van Zomeren, B. C. (1990), "Unmasking multivariate outliers and leverage points," Journal of the American Statistical Association, 85, 633-651.

[61] Rousseeuw, P. & Ruts, I. (1996). Algorithm AS 307: Bivariate Location Depth. Journal of the Royal Statistical Society Series C Applied Statistics. 45. 516-526. 10.2307/2986073.

[62] Rousseeuw, P. & Hubert, M. (1999). Regression Depth. Journal of the American Statistical Association. 94. 388-402. 10.1080/01621459.1999.10474129.

[63] Rousseeuw, P. and Van Driessen, K. (1999) A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics, 41, 212-223.

[64] Rousseeuw, P. & Croux, C. (1993). Alternatives to Median Absolute Deviation. Journal of the American Statistical Association. 88. 1273 - 1283. 10.1080/01621459.1993.10476408.

[65] Rousseeuw, P. & Struyf, A. (2004). Characterizing Angular Symmetry And Regression Symmetry. Journal of Statistical Planning and Inference. 122. 161-173. 10.1016/j.jspi.2003.06.015.

[66] Rousseeuw, P. and Hubert, M. (2011), "Robust Statistics for Outlier Detection," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1, 73–79

[67] Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2012). Robustbase: Basic Robust Statistics. R package version 0.9-7. URL http://CRAN.R-project.org/package=robustbase

[68] Schwertman, N. C., Owens, M. A., and Adnan, R. (2004), "A Simple More General Boxplot Method for Identifying Outliers," Computational Statistics and Data Analysis, 47, 165–174.

[69] Schwertman, N. C., and de Silva, R. (2007), "Identifying Outliers With Sequential Fences," Computational Statistics and Data Analysis, 51, 165-174.

[70] Serfling, R. (2002). A Depth Function and a Scale Curve Based on Spatial Quantiles.

[71] Sharifah S., Siti Z. and Wan N., (2021). Comparison of Robust Estimators for Detecting Outliers in Multivariate Datasets. Journal of Physics: Conference Series, 1988, 012095 doi:10.1088/1742-6596/1988/1/012095

[72] Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. The Annals of Mathematical Statistics 21 (1), 124–127.

[73] Struyf, A. & Rousseeuw, P. (1999). Halfspace Depth and Regression Depth Characterize the Empirical Distribution. Journal of Multivariate Analysis. 69. 135-153. 10.1006/jmva.1998.1804.

[74] Tukey, J.W. (1975). Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, vol. 2, pp. 523–531.

[75] Tukey, J.W. (1977) Exploratory Data Analysis. Addison-Wesley , (Vol. 2)

[76] Woodbury, M. A. (1950). Inverting modified matrices. Memorandum report 42, 106.

[77] Won, J.H., Limو J., Kimو S.J., and Rajaratnamو B. (2013). Condition-number-regularized covariance estimation. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (3), 427-450.

[78] Zuo, Y., and Serfling, R. (2000), "General Notions of Statistical Depth Function," Annals of Statistics, 28, 461–482.

Appendix

Monthly air pollution data PM10 in Diwaniyah for the period from 2019-2021

Y(pm10)	X1(TSP)	X2(O3)	X3(So2)	X4(CO2)	X5(CO)	X6(NOx)	X7(NO2)	X8(NO)
Ug/m3	/Ug/m3	Ppm	ppm	ppm	ppm	ppm	ppm	Ppm
0.2	0.44	0.011	0.11	584.495	0.276	0.021	0.011	0.01
0.1	0.43	0.14	0.014	592.324	0.19	0.015	0.009	0.005
0.0131	0.33	0.031	0.021	589.214	0.123	0.014	0.007	0.012
0.27	0.45	0.028	0.022	567.788	0.07	0.018	0.0018	0.005
0.13	0.293	0.013	0.021	613.767	0.268	0.018	0.009	0.008
0.41	0.71	0.027	0.019	524.729	0.227	0.02	0.014	0.006
0.13	0.411	0.0288	0.011	489.772	0.09	0.019	0.023	0.002
0.14	0.37	0.015	0.013	541.003	0.369	0.042	0.003	0.006
0.25	0.36	0.0236	0.018	509.837	0.298	0.0143	0.008	0.007
0.245	0.47	0.028	0.02	596.247	0.121	0.015	0.01	0.005
0.0131	0.33	0.028	0.022	589.214	0.123	0.014	0.007	0.012
0.13	0.293	0.028	0.022	613.767	0.268	0.018	0.009	0.008
0.28	0.534	0.034	0.008	583.002	0.268	0.018	0.023	0.003
0.2	0.35	0.034	0.022	617.208	0.083	0.015	0.009	0.004
0.368	0.512	0.0189	0.018	576.647	0.28	0.022	0.019	0.004
0.231	0.022	0.03	0.012	503.472	0.01	0.013	0.01	0.006
0.29	0.38	0.0317	0.011	482.195	0.073	0.0119	0.009	0.005
0.2	0.465	0.0312	0.007	585.467	0.1037	0.0207	0.017	0.008
0.33	0.48	0.0283	0.021	507.87	0.1386	0.018	0.0101	0.008
0.13	0.293	0.14	0.014	613.767	0.268	0.0143	0.011	0.018
0.28	0.38	0.0237	0.012	514.622	0.295	0.018	0.01	0.008
0.13	0.315	0.02	0.008	583.002	0.0376	0.016	0.009	0.007
0.17	0.37	0.028	0.016	611.372	0.196	0.019	0.016	0.013
0.18	0.38	0.016	0.01	538.577	0.32	0.0186	0.0092	0.01
0.243	0.55	0.016	0.011	437	0.373	0.033	0.016	0.018
0.141	0.467	0.019	0.014	444	0.783	0.012	0.003	0.003
0.334	0.534	0.023	0.008	971	0.299	0.005	0.003	0.002
0.1	0.42	0.031	0.02	699.47	0.352	0.016	0.011	0.005
0.1	0.031	0.016	0.007	618.32	0.268	0.017	0.01	0.007
0.29	0.54	0.01	0.009	533.218	0.19	0.022	0.011	0.011
0.41	0.71	0.0288	0.011	524.729	0.227	0.019	0.023	0.006
0.245	0.47	0.028	0.022	596.247	0.121	0.014	0.01	0.005
0.147	0.486	0.023	0.02	487.238	0.297	0.0198	0.096	0.006
0.342	0.46	0.0291	0.013	562.738	0.27	0.016	0.008	0.006
0.13	0.293	0.034	0.022	613.767	0.268	0.015	0.009	0.008
0.41	0.48	0.038	0.012	577.349	0.278	0.018	0.018	0.005

Data source: Diwaniyah Environment Directorate.

الخلاصة:

في هذه الدراسة ، تم استخدام دوال عمق البيانات الحصينة لتشخيص القيم الشاذة في نموذج متعدد المتغيرات. تم استخدام مجموعة من دوال العمق ، والتي تعتمد على مسافة mahalanobis ومصفوفة التباين والتباين المشترك الحصينة (MRCD) ، مثل طريقة دالة البعد ودالة العمق المكاني ودالة العمق الإهليلجي ودالة عمق الدوال المثلثية بالإضافة إلى ذلك و بالنسبة للنموذج ثنائي المتغير اقترحنا استخدام bagplot للكشف عن القيم المتطرفة في مجموعة البيانات والتي تعتمد على دالة عمق البيانات الإحصائية ومتوسط tukey. تم استخدام دراسة المحاكاة ووبيانات حقيقية لتقييم الطرق المقترحة. أوضحت الدراسة أن الطرق المقترحة لها أداء جيد في الكشف عن القيم الشاذة مقارنة ببعض الطرق الموجودة.

جمهورية العراق
وزارة التعليم العالي والبحث
العلمي
جامعة القادسية/كلية الإدارة
والاقتصاد
قسم الإحصاء

الكشف عن القيم المتطرفة المتعددة من خلال دوال العمق الإحصائي مع التطبيق

رسالة

مقدمة إلى مجلس كلية الإدارة والاقتصاد/جامعة القادسية كجزء من

متطلبات نيل درجة ماجستيرعلوم في الاحصاء

من قبل
هديل كامل حبيب
بإشراف
الدكتور محمد عبد الحسين محمد الغريباوي