http://journalppw.com
ISSN 2587-0130

Using Some Metric Distance in Local Density Based on Outlier Detection Methods

Shahad Adel Abdulghafoor[1,2], Prof.Lekaa Ali Mohamed[1]

[1]*Baghdad University, College of Management and Economics, Department Of Statistics*
[2]*Ministry of planning, Central Statistical Organization*
shahed.adel1001a@coadec.uobaghdad.edu.iq
lekaa.a@coadec.uobaghdad.edu.iq

Abstract

This study used different metric distances to estimate density functions in outlier detection. We employed multidimensional scaling for dimension reduction using two metric distances (the standardized Euclidean and Minkowski distances). A local density-based method was applied to three methods for outlier detection. We use the criterion for evaluating the performance of outlier approaches in this paper is Precision. The Gaussian local density estimation method uses three nearest neighbors types (KNN, RNN, and SNN). While in SGR, and Volcano use one kind of nearest neighbor (KNN). Extensive experiments on a synthetic dataset have shown that the result of the two distances was approximately equal. The RDOS and the VOL methods are more efficient when we increase the number of nearest neighbours. The average numbers of outliers increase in the SGR method, when we grow NN, the average number of outliers appears weak in the technique.

Keywords: k-nearest neighbor, Reverse nearest neighbor, Shared nearest neighbor, standardized Euclidean distance, Minkowski distance.

1. Introduction

Researchers can discover outliers in data and gain required information that helps in making better decisions. If we don't identify all outliers, it can lead to false assumptions, biased parameter estimation, and inaccurate results. As a result, identifying outliers is critical before modelling and analysis. Sometimes the researcher is interested in Detecting outliers, such as credit card [Carcillo et al, 2021], fraud detection[Yadav et al, 2021], cybersecurity intrusion detection[Kilincer et al, 2021], and medical diagnosis[Sun et al, 2018]. In these cases, the outliers are core data or the researcher interested in cleaning the data from the outliers. Several researchers defined the outliers in many ways. In general, we can determine the outlier as a data point that is considerably different from other data points in a way that arrow suspicious The observation was generated using a different mechanism[Hawkins, 1980].

There are numerous methods and approaches for detecting outliers. The researcher classified them into four groups: Methods of univariate vs multivariate. The earlier works in outlier detection were in univariate methods. The current body of work is concerned with multivariate methods. In the second categorization, there are three scenarios supervised, semi-supervised, and unsupervised learning methods. Supervised learning means learning by example; this kind of learning examines training data and makes new functions based on function applications from training data. Unsupervised learning seeks to discover hidden patterns in unlabeled data. It cannot be used directly to a classification issue as the output values are unknown. Semi-supervised learning lies between the labeled data and unlabeled data. Semi-supervised learning aims to figure out how combining unlabeled and some labeled input affects learning behaviour [Yang et al, 2021]. The third categorization is parametric and nonparametric. The parametric approach or the statistical method presupposes that the underlying distribution for the observation is known, often unsuitable for large datasets with several dimensions. Numerous nonparametric methods exist in the field of data mining, such as the distance-based method and the density-based method. The final classification is global versus local outlier methods. The global approach looks for outliers that are in relation to the rest of the data points in the set. In comparison, the local detection of outliers means searching for outliers around a local spot and how the data point is isolated with respect to its neighborhood.

An overview of approaches for detecting outliers is making assumptions about outliers to the rest of the dataset. Following these assumptions, some researchers discover outliers and categorize them into four types: Model-based methods. They believe that data points created by a statistical model are normal, while data points that do not follow the model are outliers.

Proximity-based methods If the object's nearest neighbors are far distant in feature space, consider the data point to be an outlier. There are two types of proximity-based outlier detection methods: density-based and distance-based. in detecting outliers based on distance, an item is considered an outlier if there aren't enough other points in the neighborhood. Outlier detection based on density An object is identified as an outlier if its density is significantly lower than its neighbors.

Clustering-based methods, large and dense clusters contain normal data points while outliers are members of small or sparse clusters or are not members of any clusters[Zhao et al, 2021].

High dimensional methods identify outliers in subspace or in extended conventional outlier detection, modeling outliers with a high degree of dimension. As a result, the most effective outlier detection methodology is to use the proximity or density method when dealing with high-dimensional data.

Breunig et al. first introduce the notion of a local outlier factor. Give a degree of an outlier to each object. This degree is referred to as an object's local outlier factor value (LOF). It is local in the sense that the degree is determined by how isolated the object is from its surroundings data point. The most method of density outliers detection depends on the framework of the LOF[Breunig at el, 2000].

Shekhar, S. et al. provided a method for detecting spatial outliers in multidimensional traffic data. This approach's statistical model was defined and analyzed also gave a solid spatial outlier detection technique and cost model[Shekhar et al, 2002].

Fan, H. et al. presented nonparametric outlier identification with a new solution and data mining. The output of the outlier algorithm considers the dataset's local and global objects. The algorithm is tested using synthetic and real-life datasets from large building contractors. Moreover, compared to a previous mining algorithm, this method was more effective and superior[Fan et al, 2006].

Gao, J. et al. propose a non-parameter detection of outliers with regression learning Using a Multi-scale Local Kernel Regression approach (MLKR) that computes outlier factors by merging information from several scale neighborhoods[Gao et al, 2010].

Later, the same researchers used variable kernel density estimates to address the shortcomings of the LOF method's accuracy when the data set is enormous. Additionally, they used the weighted density neighborhood estimate for better robustness to parameter variations. They propose the Volcano kernel as a new way of detecting outliers[Gao et al, 2011].

Fink, O. et al. used a Multivariate kernel density estimation technique to find outliers. Another approach employs the "growing neural gas," an unsupervised algorithm based on artificial neural gas (GNG). In the field of railway turnout systems, these two methods are applied. Both techniques are effective in recognizing novel patterns. Furthermore, the GNG was the best choice for dimensionality in input data and online learning[Fink et al, 2015].

Sharma, S. et al. detecting outliers by kernel density estimation and assigning an outlier score to each data point. kernel function gave a smoother for the density estimation. By comparing the local density estimate of each object to the neighbors, an outlier score is assigned to each one. The kernel function and the outlier score are applied to discover the unusual pattern in the data[Sharma et al, 2015].

Tang, B. and He, H. suggest a density-based approach for detecting local outliers. The Outlier Score is calculated based on Relative Density (RDOS). This measure of object local outlierness uses a local KDE technique based on the object's extended nearest neighbors to estimate its density distribution. The researchers used three neighbors (KNN, RNN, and SNN)[Tang et al, 2017].

This research will apply the methods for two metric distances (the Minkowski distance and the standardized Euclidean distance) for volcano kernel, kernel fun. SGR with second-order and RDOS by using multidimensional skilling.

1. Methods

Estimation of local kernel density

The density at a given location for an object is estimated via kernel density estimation based on a set of data points[Tang et al, 2017]. Let D=$\{c_1, c_2, \dots, c_n\}$, where c_1,c_2 and c_3 are data points in the set. Where $c_i \epsilon R^d \ for \ i = 1,2,3,4, \dots, n$, the KDE is estimated as follows :

$$p(c) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{h^d}K(\frac{c-c_i}{h}) \qquad (1)$$

Where $K(\frac{c-c_i}{h})$ denoted by the kernel function with the kernel width h, smoothing kernel is satisfied the following conditions: $\int K(u)du = 1$, $\int u\,K(u)du = 0$, $\int u^2K(u)du > 0$.

The following is a commonly used multivariate Gaussian kernel function in detecting outliers [Tang et al, 2017]:

$$K(\frac{c_i - c_j}{h_j})_{Gaussian} = \frac{1}{(2\pi)^d}exp\left(-\frac{\left\|c_i - c_j\right\|^2}{2 * h_j^2}\right) \qquad (2)$$

Where $\left\|c_i - c_j\right\|$ The Euclidean distance between the points c_i and c_j. d: denotes dimensions, and h: represents the kernel bandwidth.

Density estimation

To estimate the density at a particular location for the point c_i is computed by considering its surrounding neighbors rather than considering all the points in the data set as kernels. As a result, Density estimation for the entire data set may result in the loss of local density differences and an inability to detect local outliers. Furthermore, using the whole data set to determine the outlier degree for each data point in the dataset results in substantial computing costs, particularly in O(n^2), where n is the total number of samples in the data set.

the estimation of the density distribution will improve in the neighborhood for the data point to this method by presenting three types of neighbors (k nearest neighbors, reverse nearest neighbors, shared nearest neighbors)

If $\mathrm{NN_j(c_k)}$ is the $\mathrm{j^{th}}$ denote jth nearest neighbors for the point $\mathrm{c_k}$, let $S_{KNN}(c_k)$ be a set of K nearest neighbors of c_k :

$$\mathrm{KNN(c_i)} = \{\mathrm{NN_1(c_i), NN_2(c_i), \dots, NN_k(c_i)}\} \qquad (3)$$

For the RNNs for the point c_k are those points that take into account c_k as one of their KNNs, we mean that c is one of the RNNs of the c_k as the $\mathrm{NN_j(c) = c_i}$ for each j ≤ k. RNNs have zero, one or more points of data. In recent studies, the RNN has been successfully employed in clustering[Zhu et al, 2016] and classification[Tang et al, 2015] and has been present for best local distribution data information and applied to detect outliers[Jin et al, 2006].

The shared nearest neighbors for the point c_k are those points that share one or more of c_k's nearest neighbors. c would be one of the SNNs for c_k where $NN_j(c) = NN_s(c_k)$ for all $j, s \leq k$.

Figure (1) display the three types of the nearest neighbors

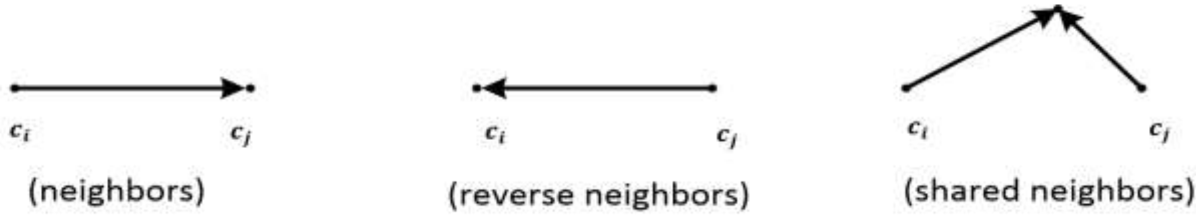


Figure (1) illustration of the three types of nearest neighbors.

We indicate that the set of reverse nearest neighbors as $S_{RNNs}(c_k)$ and the set of the shared nearest neighbors as $S_{SNNs}(c_k)$ for the point c_k.

In the set $S_{KNNs}(c_k)$, there would always be k nearest neighbors. whereas the two sets of $S_{RNNs}(c_k)$ and $S_{SNNs}(c_k)$ is either empty or contains one or more items.

The three sets of data $S_{SNNs}(c_k)$, $S_{RNNs}(c_k)$ and $S_{SNNs}(c_k)$ for the point c_k By merging them, we build an expanded local neighborhood. Defined by:

$$S(c_k) = S_{KNNs}(c_k) \cup S_{RNNs}(c_k) \cup \mathrm{S}_{\mathrm{SNNs}}(c_k) \qquad (4)$$

The location's estimated density for c_k will equal to:

$$p(c_k) = \frac{1}{|S(c_k)|+1} \sum_{c\, \in S(c_k)\cup\{c_k\}} \frac{1}{h^d} K(\frac{c-c_k}{h}) \qquad (5)$$

$|S|$ represents the total number of elements in S.

RDOS calculation

After using KDE to calculate the density at all of the data points in the set. We will use an outlier factor based on relative density (RDOS) to calculate the degree of outlierness. When the data points c_k deviate from their surroundings neighbors, and It can be explained as follows:

$$RDOS_K(c_k) = \frac{\sum_{c_i \in S(c_k)} p(c_i)}{|S(c_k)|p(c_k)} \qquad (6)$$

RDOS: is the average density neighborhood divided by the density of certain data point c_k. When $RDOS_K(c_k)$ is significantly greater than 1, and the data point will be located outside the dense cluster. implying that c_k is an outlier, and when $RDOS_K(c_k)$ is equal to or less than 1, then the data point c_k would be encircled by the same density cloud of neighbors. Implying that c_k is a normal data point.

Using KNN graph to determind the sets of KNNs, RNNs and SNNs by the approximation computantional method for the cost of O(N). for each data point we put a local set of nearest neighbors S with the collection of sets S_{KNNs}, S_{RNNs} and $\mathrm{S}_{\mathrm{SNNs}}$. The density will be calculated locally for the data point c_k according to set S, we determine RDOS for all data points according to the densities of local neighborhoods in S.

The top-n of outliers is defined by sorting the values of RDOS in descending way. When we want to decide whether the data point is an outlier or not, we compare the value of RDOS with the threshold value (τ) (this value is pre-determined according to the researcher experience). If $RDOS_K(c_k)$ is satisfied, then the point is an outlier.

$$\mathrm{RDOS_k}\ (c_k)\ > \tau \qquad (7)$$

Volcano kernel method

This function is presented to avoid the drawbacks in the Gaussian kernel for anomaly estimation. We mean that in some methods that use a Gaussian kernel, we cannot guarantee that the normal data point is approximately equal to one for the outlier score, so we must use a threshold value (τ).

The volcano kernel [Gao et al, 2011],[Hu et al, 2018] is determined as follows:

$$K(c) = \begin{cases} \beta & \text{if } \|c\| \le 1 \\ \beta g(\|c\|) & \text{otherwise} \end{cases} \qquad (8)$$

Where β ensures that kernel function is a probability density function, the condition of K(c) integration is equal to 1. g(c) is a function that decreases monotonically, with the close interval [0,1], and at infinity, equal to zero. The $g(c) = e^{-|c|+1}$ as a standard function in this method.

in a univariate feature space, figure () illustrates the curve of the Volcano kernel function. The kernel value equals a constant β when ||c|| ≤ 1. This guarantees that the outlier scores of samples within a cluster are close to 1. When ||c|| > 1, Kernel value is less than one and decreases monotonically as ||c|| grows.

As a result, the outlier score for anomalies is substantially greater than one.

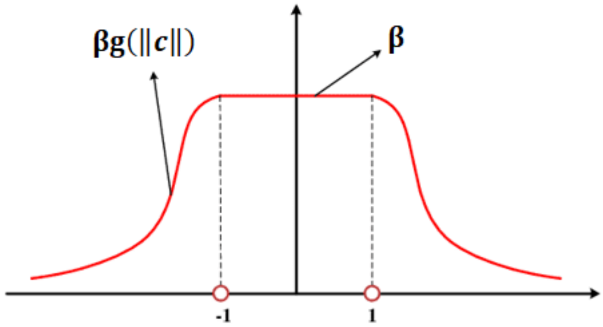


Figure (2) the curve of the Volcano kernel function in univariate space[Gao et al, 2011],[Hu et al, 2018].

The Volcano kernel was created to identify anomalies. Its goal is to develop outlier scores for a normal sample close to 1, and for the anomaly, the outlier scores are more than 1. The neighborhood number k required by our Volcano kernel is smaller than that needed for the Gaussian kernel. The normal samples make up the vast majority of the dataset. The random variable ||c|| has values between -1 and 1. figure (2) shows the densities for our Volcano kernel when ||c|| is between [-1,1]. Using the Volcano kernel to estimate the density of a sample requires fewer neighboring samples than using the Gaussian kernel. As a result, the Volcano kernel requires a smaller k neighbors value.

SGR kernel method

the kernel function SGR of order 2 [Sharma et al, 2015], this function is a PDF the integral of the SGR function is equal to 1, its an asymmetric function, finite even moments, zero odd moments.

$$K_{SGR,h} = \frac{1}{2.374\pi}(5 - 4c^2) \quad for \ |c| < \sqrt{5/4} \qquad (9)$$

Where h: is the optimal bandwidth that counted by minimizing MISE of equation (1) :

$$MISE\big(\hat{f}(c)\big) \approx \frac{1}{4} h^4 M_2^2(K)\, R(f'') + \frac{1}{nh} R(K) \qquad (10)$$

using the partial derivative of MISE to h and as well as setting it to zero:

$$h_{opt} = \left[\frac{R(K)}{M_2^2(K) R(f'')\, n}\right]^{1/5} \qquad (11)$$

Where $R(K) = \int K^2(u)du$, the $M_2(K) = \int u^2 K(u)du$ is the second moment of kernel function of equation (1) and $R(f'') = \int f''(c)^2 dc.$

In this approach, we will use the plug-in bandwidth estimator for h_{opt}. By putting $R(f'') = \Psi_4$ in equation (11) because the formula is unapplicable, so the formula will be:

$$h_{plug-in} = \left[\frac{R(K)}{M_2^2(K)\ \Psi_4\ n}\right]^{\frac{1}{5}} \qquad (12)$$

Where $\Psi_4 = E\left(f^4(c)\right)$ and $\Psi_4(g) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}^4(c_i)$

$$\Psi_r = \frac{-1^{r/2}\, r!}{(2\sigma)^{r+1}\,(r/2)!\,\sqrt{\pi}} \qquad (13)$$

We get the MISE optimal by substituting h_{opt} of equation (11) in the value of MISE in equation (10):

$$AMISE_{opt} = \frac{5}{4}\left(R^2(K)M_2^2(K)R(f'')\right)^{\frac{1}{5}} n^{-4/5} \qquad (14)$$

In this research, we will use Multidimensional skilling. There are two types: metric and non-metric in the metric method, calculate the distance by the Euclidean distance. The non-metric way uses other methods for calculation distance. Multidimensional scaling is highly similar to principal component analysis (PCA). It transforms distances between variables rather than correlation or covariance into a two-dimensional graphic.

The graph will be precisely the same as the PCA graph if we calculate MDS between variables using the Euclidian distance. In other words, clustering based on distance minimization is equivalent to maximizing linear correlations between the points. So we use another way to calculate distances like Minkowski distance, standardized Euclidean distance, Hamming distance, Great circle distance, Manhattan distance, log fold change distance etc.

In this research, we use the two metric distances (Minkowski distance[Wachowicz et al, 2013], standardized Euclidean distance[Rao, 2012]) in the MDS[Manly et al, 2016].

the Minkowski distance can be calculated from the formula:

$$\left(\sum_{i=1}^{n}|x_i - y_i|^n\right)^{1/n} \qquad (15)$$

This formula can be equal to manhattan distance when n=1, Euclidean distance when n=2, and Chebyshev when n= infinity.

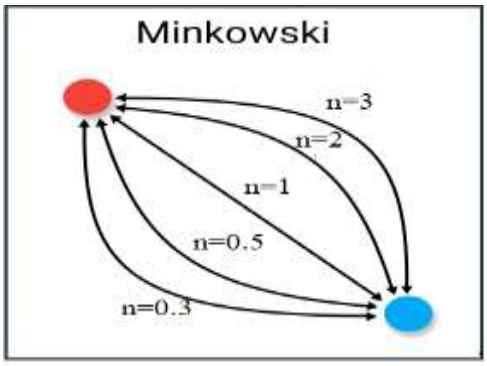


Figure 3: illustrate the Minkowski distance for different data values[Wachowicz et at, 2013].

The standardized Eucleadian distance between two vectors of dimension j is given by:

$$d_{x,y} = \sqrt{\sum_{j=1}^{n}\left(\frac{x_j}{s_j} - \frac{y_j}{s_j}\right)^2}$$

$$= \sqrt{\sum_{j=1}^{n}\frac{1}{s_j^2}\left(x_j - y_j\right)^2}$$

$$= \sqrt{\sum_{j=1}^{n} w_j\left(x_j - y_j\right)^2} \qquad (16)$$

Where

$$w_j = 1/S_j^2$$

the criterion for evaluating the performance of outlier approaches in this paper is Precision(P)[Xu et al, 2018], it is defined as the ratio that divided the number of correct outliers by the total number of points that filtered to be outliers:

$$Precision\frac{m}{t} \qquad (17)$$

m=number of correct outliers that found in the set, t=total number of points that filtered to be outliers.

3. Experiment Analysis and Result

The dataset is about three groups of random numbers naturally generated according to the normal distribution of mean= 0 and variance 0.5. These groups are represented in three different sizes (N=50, N=100, N= 150). Three explanatory variables were generated (P= 3, P=5, P=7). The nearest neighbours are in the range (2 to 10). And the number of iterations for each dataset and explanatory variables is (itr =100) according to Tabel (1). Several experiments were conducted.

Table 1: The order of initial variables is determined by the variable's size and the sample size.

Explanatory variables	Sample size			nearest neighborhood
3	50	100	150	2,3,4,5,6,7,8,9,10
5	50	100	150	2,3,4,5,6,7,8,9,10
7	50	100	150	2,3,4,5,6,7,8,9,10

RDOS, SGR, and Volcano kernel are The methods were applied to the simulated data set of size (N = 50) with three explanatory variables (3, 5, and 7), with Standardized Euclidean distance and Minkowski distance as shown in Table(2) and Tabel (3) as shown below. The average number of outliers for the SGR method increases when K nearest neighbors increase to10. While the RDOS and VOL methods, the average number of outliers decreases when we increase the numbers of K nearest neighbors. Figures 4, 6, 8, 5, 7 and 9 explain the difference between these two distances in the three methods.

The same methods applied for sample size (N=100) and the three variables (3, 5, 7) tables 4, 5 show that when the number of variables is equal to three, in the SGR method, the average number of outliers by using Minkowski distance was higher than Standardized Euclidean distance. While in the RDOS and The VOL, little fluctuated between increasing and decreasing in the two metric distances.

When the number of variables is (5,7), the average number of outliers in the SGR and RDOS methods is smaller in Minkowski distance than the standardized Euclidean distance.

And when we have five variables in the VOL method, the average number of outliers using Minkowski distance is lower at the beginning, then becomes equal to the value of the standardized Euclidean distance as the number of k nearest neighbors increases to 10. And when the number of variables is seven in the vol method, the average number of outliers fluctuated between decreasing and increasing by using the two metric distances as the number of nearest neighbors rose to 10. The figures (10, 12, 14, 11,13 and 15) illustrate the increase and decrease in the methods (SGR, RDOS and EPA).

For sample size (150) for the three different variables (3, 5 and 7) in the SGR method, the average number of outliers in Minkowski distance is larger than the standardized Euclidean distance. Still, for the RDOS and VOL methods, the average is fluctuated between increasing and decreasing using the two distances as the number of k nearest neighbor increases. The figures (16, 18, 20, 17, 19 and 21).

Table 2: The results of 50 observations for the methods with sample 100 replicate for Standardized Euclidean distance

	Average number of outlier								
K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	35	22	27	36	21	30	35	15	32
3	36	18	24	37	15	22	36	12	27
4	36	15	29	37	14	26	36	11	19
5	36	13	21	38	13	16	37	11	12
6	37	12	18	38	13	9	37	11	9
7	37	12	13	39	13	1	38	11	4
8	38	13	6	39	13	3	38	11	9
9	38	13	2	39	13	2	38	11	5
10	38	13	3	39	13	1	39	11	1

Table 3: The results of 50 observations for the methods with sample 100 replicate for Minkowski distance

	Average number of outlier								
K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	35	23	24	35	22	30	34	22	29
3	35	16	24	35	16	27	34	16	26
4	35	14	29	36	14	28	34	15	23
5	36	13	15	36	14	13	35	15	10
6	36	13	7	37	13	10	35	15	5
7	36	13	3	37	13	3	36	15	3
8	36	13	3	37	13	3	36	15	3
9	37	12	1	37	13	6	36	15	3
10	37	13	3	37	13	3	36	15	3

Table 4: The results of 100 observation for the methods with 100 replicate for Standardized Euclidean distance

	Average number of outlier								
K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	66	56	61	69	56	62	72	53	52
3	65	48	49	70	41	63	72	35	50
4	66	39	30	70	30	44	73	28	43
5	66	31	36	71	24	42	74	25	34
6	67	27	37	71	22	17	75	23	42
7	68	25	14	71	21	21	75	23	14
8	68	23	23	72	21	16	76	22	8
9	69	22	12	73	21	7	77	22	6
10	69	22	10	74	21	17	77	22	2

Table 5: The results of 100 observations for the methods with 100 replicate for Minkowski distance

	Average number of outlier								
K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	68	55	61	68	52	56	70	50	58
3	67	47	52	69	39	51	71	34	57
4	68	37	43	70	30	42	71	26	42
5	68	30	43	70	24	47	73	24	38
6	69	25	27	71	22	15	73	23	30
7	70	23	22	71	21	8	74	22	13
8	70	22	7	72	20	13	75	22	9
9	71	22	13	72	20	2	75	21	10
10	71	22	11	72	20	1	76	22	1

Table 6: The results of 150 observation for the methods with 100 replicate for Standardized Euclidean distance

	Average number of outlier								
K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	98	81	73	100	84	94	105	78	87
3	99	73	82	102	74	79	105	64	94
4	99	66	45	102	59	60	106	49	54
5	99	55	60	103	46	67	107	38	60
6	100	47	59	104	38	57	108	34	48
7	101	41	19	105	35	50	109	32	36
8	101	37	29	105	33	45	110	30	46
9	102	35	27	106	32	27	110	30	28
10	102	32	29	107	31	24	111	30	25

Table 7: The results of 150 observations for the methods with 100 replicate for Minkowski distance

K	Average number of outlier								
	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	106	82	83	102	83	69	103	81	79
3	107	66	77	103	72	72	104	61	71
4	107	48	68	105	58	91	105	44	70
5	108	38	58	105	46	56	106	36	63
6	110	34	47	106	38	40	107	32	62
7	111	31	41	107	33	16	107	30	27
8	111	30	41	107	30	29	108	29	33
9	112	29	25	108	29	22	110	28	13
10	114	28	17	108	29	19	110	28	17

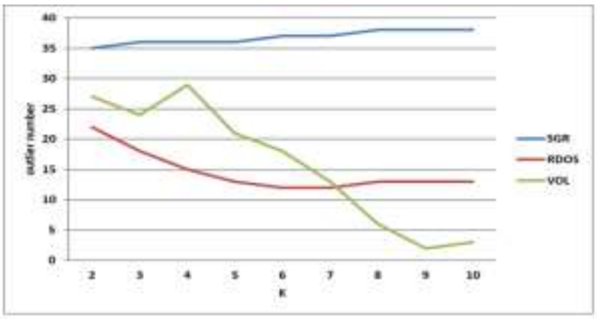


Figure (4) sample size 50 with three variables with Standardized Euclidean distance for SGR, RDOS and VOL

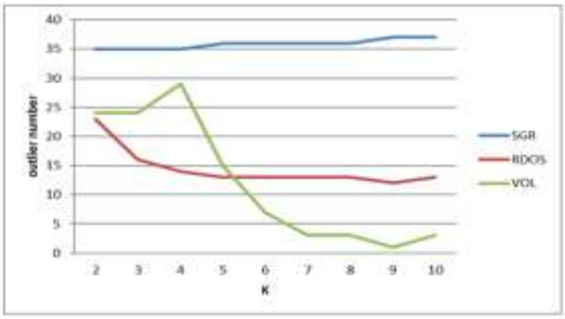


Figure (5) sample size 50 with three variables with Minkowski distance for SGR, RDOS and VOL

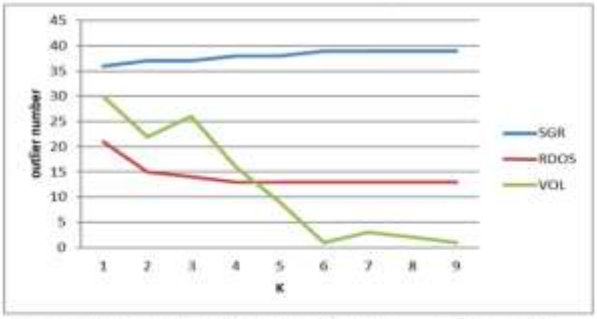


Figure (6) sample size 50 with five variables with Standardized Euclidean distance for SGR, RDOS and VOL

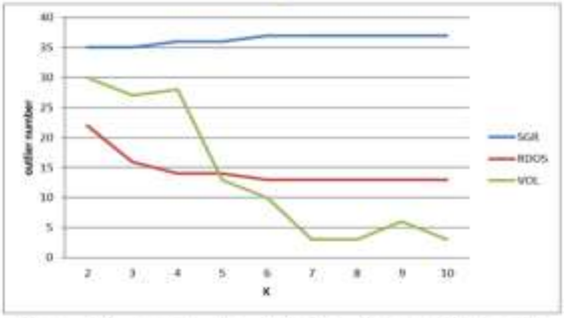


Figure (7) sample size 50 with five variables with Minkowski distance for SGR, RDOS and VOL

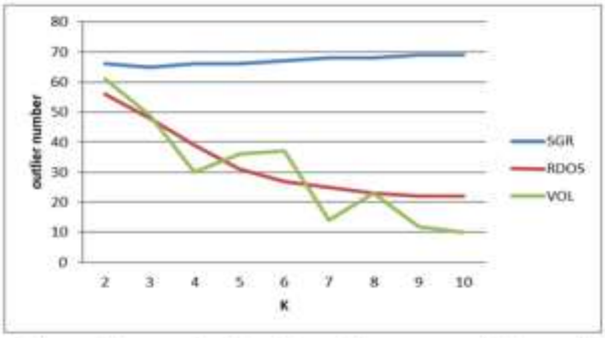


Figure (8) sample size 50 with seven variables with Standardized Euclidean distance for SGR, RDOS and VOL

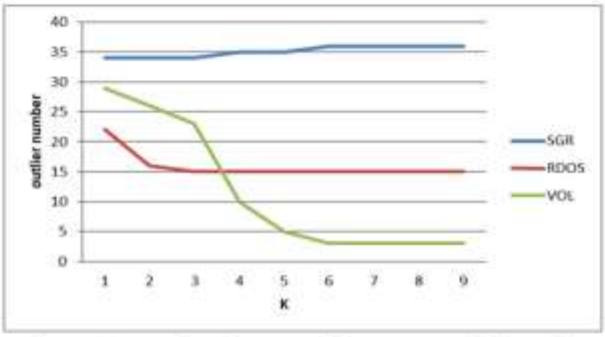


Figure (9) sample size 50 with seven variables with Minkowski distance for SGR, RDOS and VOL

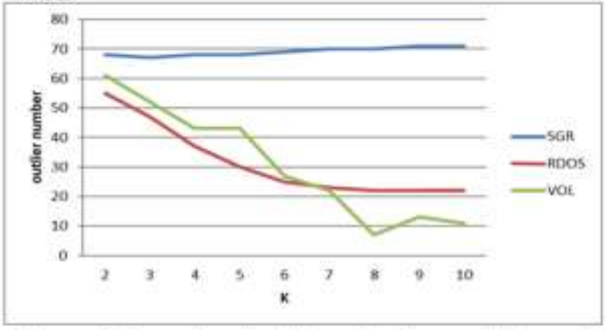


Figure (10) sample size 100 with three variables with Standardized Euclidean distance for SGR, RDOS and VOL

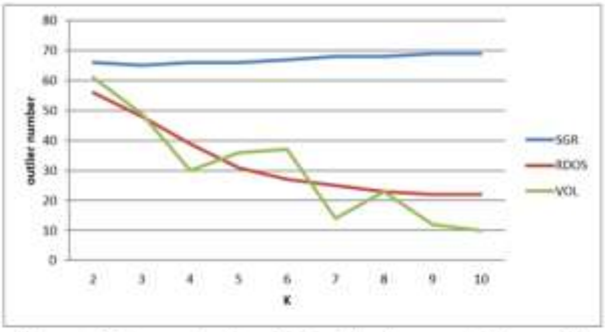


Figure (11) sample size 100 with three variables with Minkowski distance for SGR, RDOS and VOL

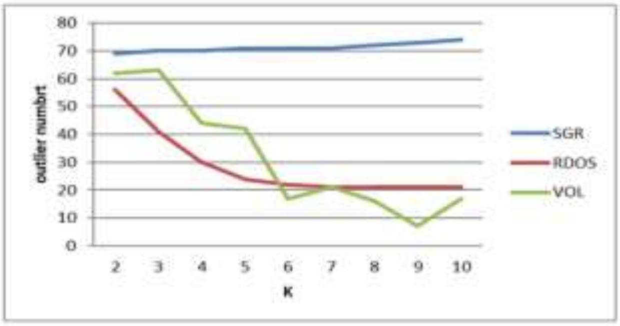


Figure (12) sample size 100 with five variables with Standardized Euclidean distance for SGR, RDOS and VOL

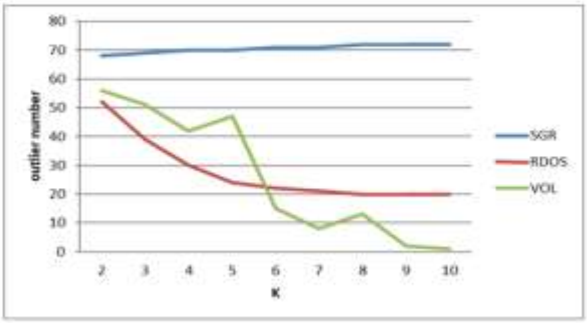


Figure (13) sample size 100 with five variables with Minkowski distance for SGR, RDOS and VOL

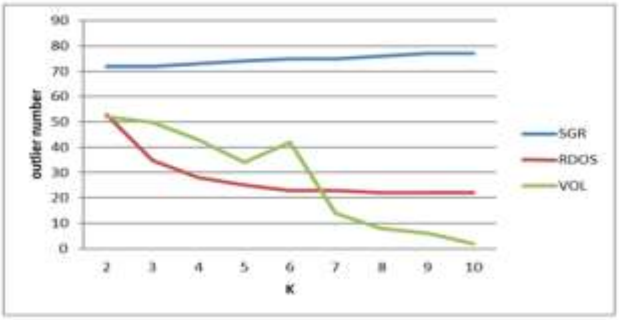


Figure (14) sample size 100 with seven variables with Standardized Euclidean distance for SGR, RDOS and VOL

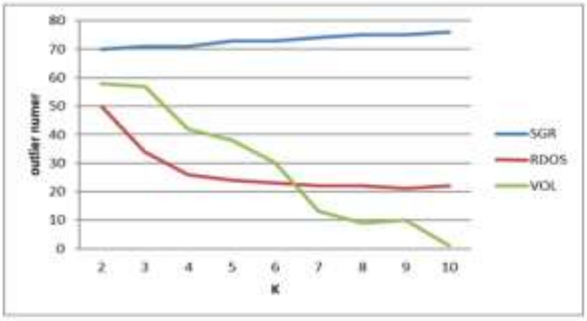


Figure (15) sample size 100 with seven variables with Minkowski distance for SGR, RDOS and VOL

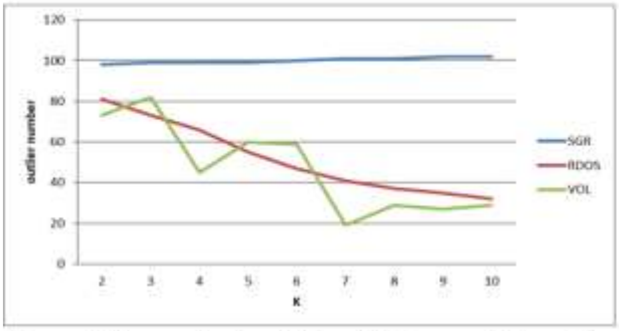


Figure (16) sample size 150 with three variables with Standardized Euclidean distance for SGR, RDOS and VOL

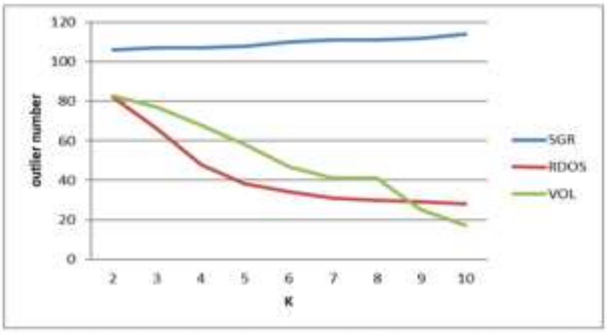


Figure (17) sample size 150 with three variables with Minkowski distance for SGR, RDOS and VOL

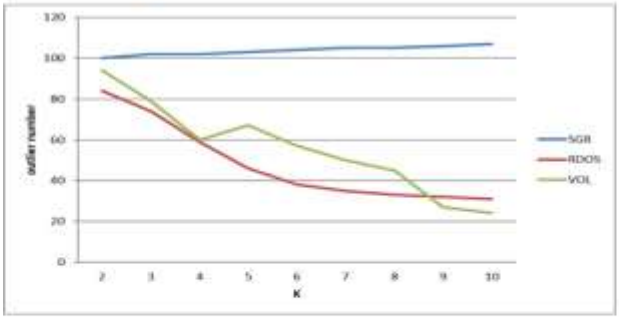


Figure (18) sample size 150 with five variables with Standardized Euclidean distance for SGR, RDOS and VOL

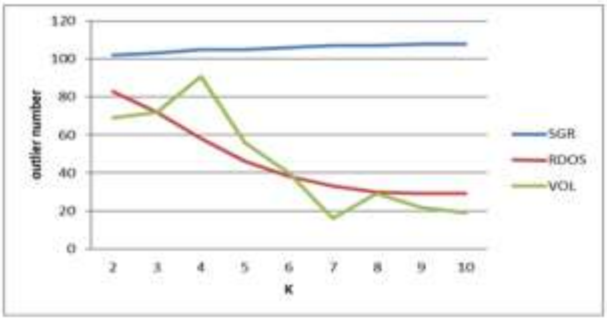


Figure (19) sample size 150 with five variables with Minkowski distance for SGR, RDOS and VOL

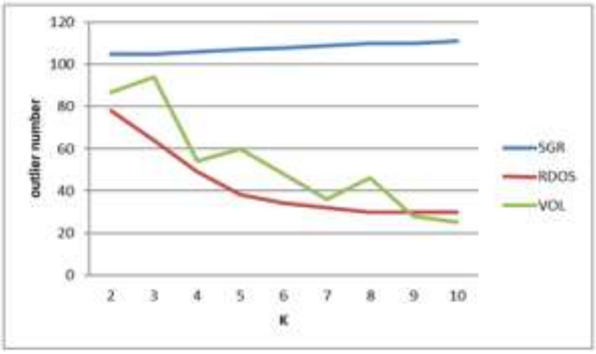


Figure (20) sample size 150 with seven variables with Standardized Euclidean distance for SGR, RDOS and VOL

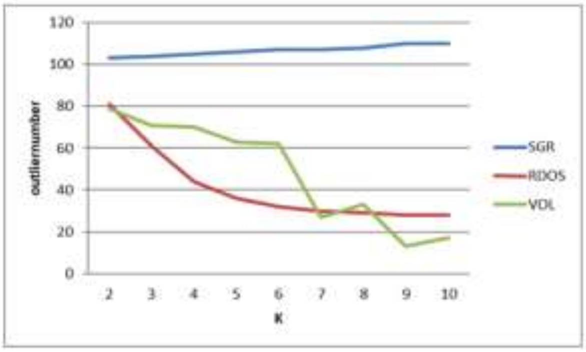
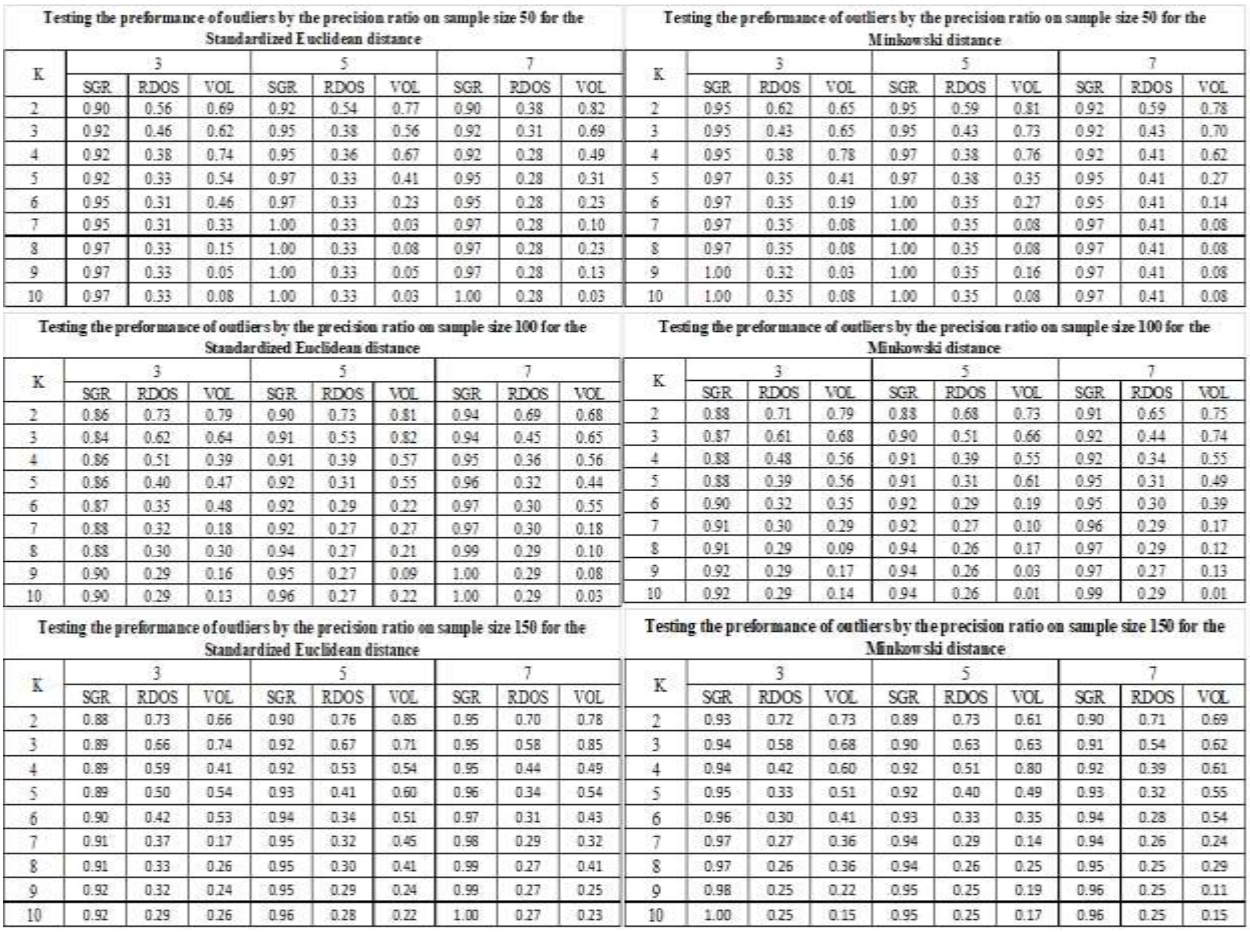


Figure (21) sample size 150 with seven variables with Minkowski distance for SGR, RDOS and VOL

Table 8: the precision ratio of three sample sizes (50, 100 and 150) and with k from (2 to 10).



Testing the preformance of outliers by the precision ratio on sample size 50 for the Standardized Euclidean distance

K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	0.90	0.56	0.69	0.92	0.54	0.77	0.90	0.38	0.82
3	0.92	0.46	0.62	0.95	0.38	0.56	0.92	0.31	0.69
4	0.92	0.38	0.74	0.95	0.36	0.67	0.92	0.28	0.49
5	0.92	0.33	0.54	0.97	0.33	0.41	0.95	0.28	0.31
6	0.95	0.31	0.46	0.97	0.33	0.23	0.95	0.28	0.23
7	0.95	0.31	0.33	1.00	0.33	0.03	0.97	0.28	0.10
8	0.97	0.33	0.15	1.00	0.33	0.08	0.97	0.28	0.23
9	0.97	0.33	0.05	1.00	0.33	0.05	0.97	0.28	0.13
10	0.97	0.33	0.08	1.00	0.33	0.03	1.00	0.28	0.03

Testing the preformance of outliers by the precision ratio on sample size 50 for the Minkowski distance

K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	0.95	0.62	0.65	0.95	0.59	0.81	0.92	0.59	0.78
3	0.95	0.43	0.65	0.95	0.43	0.73	0.92	0.43	0.70
4	0.95	0.38	0.78	0.97	0.38	0.76	0.92	0.41	0.62
5	0.97	0.35	0.41	0.97	0.38	0.35	0.95	0.41	0.27
6	0.97	0.35	0.19	1.00	0.35	0.27	0.95	0.41	0.14
7	0.97	0.35	0.08	1.00	0.35	0.08	0.97	0.41	0.08
8	0.97	0.35	0.08	1.00	0.35	0.08	0.97	0.41	0.08
9	1.00	0.32	0.03	1.00	0.35	0.16	0.97	0.41	0.08
10	1.00	0.35	0.08	1.00	0.35	0.08	0.97	0.41	0.08

Testing the preformance of outliers by the precision ratio on sample size 100 for the Standardized Euclidean distance

K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	0.86	0.73	0.79	0.90	0.73	0.81	0.94	0.69	0.68
3	0.84	0.62	0.64	0.91	0.53	0.82	0.94	0.45	0.65
4	0.86	0.51	0.39	0.91	0.39	0.57	0.95	0.36	0.56
5	0.86	0.40	0.47	0.92	0.31	0.55	0.96	0.32	0.44
6	0.87	0.35	0.48	0.92	0.29	0.22	0.97	0.30	0.55
7	0.88	0.32	0.18	0.92	0.27	0.27	0.97	0.30	0.18
8	0.88	0.30	0.30	0.94	0.27	0.21	0.99	0.29	0.10
9	0.90	0.29	0.16	0.95	0.27	0.09	1.00	0.29	0.08
10	0.90	0.29	0.13	0.96	0.27	0.22	1.00	0.29	0.03

Testing the preformance of outliers by the precision ratio on sample size 100 for the Minkowski distance

K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	0.88	0.71	0.79	0.88	0.68	0.73	0.91	0.65	0.75
3	0.87	0.61	0.68	0.90	0.51	0.66	0.92	0.44	0.74
4	0.88	0.48	0.56	0.91	0.39	0.55	0.92	0.34	0.55
5	0.88	0.39	0.56	0.91	0.31	0.61	0.95	0.31	0.49
6	0.90	0.32	0.35	0.92	0.29	0.19	0.95	0.30	0.39
7	0.91	0.30	0.29	0.92	0.27	0.10	0.96	0.29	0.17
8	0.91	0.29	0.09	0.94	0.26	0.17	0.97	0.29	0.12
9	0.92	0.29	0.17	0.94	0.26	0.03	0.97	0.27	0.13
10	0.92	0.29	0.14	0.94	0.26	0.01	0.99	0.29	0.01

Testing the preformance of outliers by the precision ratio on sample size 150 for the Standardized Euclidean distance

K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	0.88	0.73	0.66	0.90	0.76	0.85	0.95	0.70	0.78
3	0.89	0.66	0.74	0.92	0.67	0.71	0.95	0.58	0.85
4	0.89	0.59	0.41	0.92	0.53	0.54	0.95	0.44	0.49
5	0.89	0.50	0.54	0.93	0.41	0.60	0.96	0.34	0.54
6	0.90	0.42	0.53	0.94	0.34	0.51	0.97	0.31	0.43
7	0.91	0.37	0.17	0.95	0.32	0.45	0.98	0.29	0.32
8	0.91	0.33	0.26	0.95	0.30	0.41	0.99	0.27	0.41
9	0.92	0.32	0.24	0.95	0.29	0.24	0.99	0.27	0.25
10	0.92	0.29	0.26	0.96	0.28	0.22	1.00	0.27	0.23

Testing the preformance of outliers by the precision ratio on sample size 150 for the Minkowski distance

K	3			5			7		
	SGR	RDOS	VOL	SGR	RDOS	VOL	SGR	RDOS	VOL
2	0.93	0.72	0.73	0.89	0.73	0.61	0.90	0.71	0.69
3	0.94	0.58	0.68	0.90	0.63	0.63	0.91	0.54	0.62
4	0.94	0.42	0.60	0.92	0.51	0.80	0.92	0.39	0.61
5	0.95	0.33	0.51	0.92	0.40	0.49	0.93	0.32	0.55
6	0.96	0.30	0.41	0.93	0.33	0.35	0.94	0.28	0.54
7	0.97	0.27	0.36	0.94	0.29	0.14	0.94	0.26	0.24
8	0.97	0.26	0.36	0.94	0.26	0.25	0.95	0.25	0.29
9	0.98	0.25	0.22	0.95	0.25	0.19	0.96	0.25	0.11
10	1.00	0.25	0.15	0.95	0.25	0.17	0.96	0.25	0.15

Table 8 explains the precision ratio of the RDOS, and the VOL decreases as the number of NN approaches 10. In contrast, the precision ratio for the SGR increases as the NN increases.

And the VOL method has the most decreasing ratio

4. Discussion

We may see from the above result that the average number of outliers will significantly increase when we increase the number of neighbors in the GSR method. While the other two methods (RDOS and VOL), when the number of nearest neighbors increases, the average number of outliers decreases.

When we compare the standardized Euclidean distance with Minkowski distance for the sample size equals (50, 100), there is a slight difference in the average number of outliers. The average approximately became stable when nearest neighbors between (6 to 10).

And for sample size (150), the average number of outliers was a little fluctuated as the number of neigbors increased to 10.

References

[1] Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104).

[2] Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information sciences, 557, 317-331.

[3] Fan, H., Zaïane, O. R., Foss, A., & Wu, J. (2006, April). A nonparametric outlier detection for effectively discovering top-n outliers from engineering data. In Pacific-Asia conference on knowledge discovery

and data mining (pp. 557-566). Springer, Berlin, Heidelberg.

[4] Fink, O., Zio, E., & Weidmann, U. (2015). Novelty detection by multivariate kernel density estimation and growing neural gas algorithm. Mechanical Systems and Signal Processing, 50, 427-436.

[5] Gao, J., Hu, W., Li, W., Zhang, Z., & Wu, O. (2010, August). Local outlier detection based on kernel regression. In 2010 20th International Conference on Pattern Recognition (pp. 585-588). IEEE.

[6] Gao, J., Hu, W., Zhang, Z. M., Zhang, X., and O. Wu, "RKOF: robust kernel-based local outlier detection," In Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg, pp.270-283, 2011.

[7] Hawkins D M, "Identification of Outliers", Chapman and Hall., London, Vol 11, 1980.

[8] Hu, W., Gao, J., Li, B., Wu, O., Du, J., & Maybank, S. (2018). Anomaly detection using local kernel density estimation and context-based regression. IEEE Transactions on Knowledge and Data Engineering, 32(2), 218-233.

[9] Jin, W., Tung, A. K., Han, J., & Wang, W. (2006, April). Ranking outliers using symmetric neighborhood relationship. In Pacific-Asia conference on knowledge discovery and data mining (pp. 577-593). Springer, Berlin, Heidelberg.

[10] Kilincer, I. F., Ertam, F., & Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study. Computer Networks, 188, 107840.

[11] Manly, B. F., & Alberto, J. A. N. (2016). Multivariate statistical methods: a primer. Chapman and Hall/CRC.

[12] Rao, R. (2012). Weighted Euclidean distance based approach as a multiple attribute decision making method for plant or facility layout design selection. International Journal of Industrial Engineering Computations, 3(3), 365-382.

[13] Sharma, S., & Jain, R. (2015). SGR: A New Efficient Kernel for Outlier Detection in Sensor Data Minimizing Mise. Sensors & Transducers, 189(6), 97.

[14] Shekhar, S., Lu, C. T., & Zhang, P. (2002). Detecting graph-based spatial outliers. Intelligent Data Analysis, 6(5), 451-468.

[15] Sun, C., Yan, Z., Li, Q., Zheng, Y., Lu, X., & Cui, L. (2018). Abnormal group-based joint medical fraud detection. IEEE Access, 7, 13589-13596.

[16] Tang, B., & He, H. (2017). A local density-based approach for outlier detection. Neurocomputing, 241, 171-180.

[17] Tang, B., & He, H. (2015). ENN: Extended nearest neighbor method for pattern recognition [research frontier]. IEEE Computational intelligence magazine, 10(3), 52-60.

[18] Wahid, A., & Rao, A. C. S. (2020). Rkdos: A relative kernel density-based outlier score. IETE Technical Review, 37(5), 441-452.

[19] Wachowicz, T., & Błaszczyk, P. (2013). TOPSIS based approach to scoring negotiating offers in negotiation support systems. Group Decision and Negotiation, 22(6), 1021-1050.

[20] Xu, X., Liu, H., Li, L., & Yao, M. (2018). A comparison of outlier detection techniques for high-dimensional data. International Journal of Computational Intelligence Systems, 11(1), 652-662.

[21] Yang, X., Song, Z., King, I., & Xu, Z. (2021). A Survey on Deep Semi-supervised Learning. arXiv preprint arXiv:2103.00550.

[22] Yadav, A. K. S., & Sora, M. (2021). Fraud detection in financial statements using text mining methods: A review. In IOP Conference Series: Materials Science and Engineering (Vol. 1020, No. 1, p. 012012). IOP Publishing.

[23] Zhang, L., Lin, J., & Karim, R. (2018). Adaptive kernel density-based anomaly detection for nonlinear systems. Knowledge-Based Systems, 139, 50-63.

[24] Zhao, W., Li, L., Alam, S., & Wang, Y. (2021). An incremental clustering method for anomaly detection in flight data. Transportation Research Part C: Emerging Technologies, 132, 103406.

[25] Zhu, Q., Feng, J., & Huang, J. (2016). Natural neighbor: A self-adaptive neighborhood method without parameter K. *Pattern Recognition Letters*, *80,* 30-36.