SDFS: A Standardization Technique for Nonparametric Analysis

Avimanyou Kumar Vatsa
Department of Computer Science, Fairleigh Dickinson University, Teaneck, NJ 07666, USA, avatsa@fdu.edu

Abstract: Due to availability of computational tools for data acquisition, it is very easy to collect many dimensions from an object. Nevertheless, data acquisition from an object in an experiment may have a low number of dimensions. The analysis of low dimensional data has break-through role. But raw and sparse nature of dataset imposes new challenges and requirements for data analysis due to their special and unique characteristics. In the process of overall characterization of low-dimensional data, the data pre-processing plays crucial role. One of the first processes is normalization and standardization process. Therefore, in this paper, I would like to propose novel standardization technique called SDFS (**S**tandardization for **D**istribution **F**ree **S**tatistics) for nonparametric data analysis. This technique is robust for small sample size with missing values of data points, which commonly exist in real time experiments lead to sparse low-dimensional data. The comprehensive experimental evaluation shows that SDFS standardization is significantly outperforms on existing standardization methods.

Keywords: Low-dimensional, Standardization, Clustering, Nonparametric, Sparse dataset

Introduction

In the era of data analysis, the evaluation and analysis of real-world data needs systematic application of statistical and/or logical techniques to unravel and illustrate the knowledge. The techniques required for analysis process depends on the nature of real-world raw dataset. The real-world raw dataset can be gathered from variety of sources and from different applications like research digital photography (Kelly 2015a), surveillance videos, field phenotyping and plant phenotyping (Kelly 2015) usually have high or low dimensions.

The analysis of low-dimensional data, with missing data values and small sample size, refer to the possibility of limited amounts of available data and dimensions. The analysis process addresses the many data mining challenges associated with it. These challenges depend on whether data has distribution or not. If It has not any kind of data distribution, called nonparametric data analysis. This analysis makes fewer assumptions and are more flexible, robust and applicable to non-quantitative data (Hopkins 2018). But it must need data preprocessing for easier analysis.

Commonly the data preprocessing method includes data cleaning, normalization, transformation, standardization, feature extraction and selection. One of the first step of data preprocessing is normalization of data because dataset dimensions have different units and scales. So, normalization process makes all dimensions are on same scale and range for fair comparison among them. In this paper, we considered the min-max normalization, a linear transformation, to preserve the behavior of the original data (Han 2012). In general, the min-max normalization uses following formula for range [new_{max_a}, new_{min_a}]:

$$V`_i = (v_i - min_a).(new_{max\_a} - new_{min\_a})/(max_a - min_a) + new_{min\_a}.$$

The next step is data standardization, it is the crucial step of data preprocessing. Because the variances of the normalized data dimensions are different. Standardization brings all attributes into proportion with one another for fair comparison among them (Standardization, Jajuga 2000, and Milligan 88). The dimension with large variances tends to have a larger effect on the resulting analysis result than dimension with smaller variances. But the biggest challenge is that there are too many different kinds of standardization methods are available. However, selecting the best standardization method is very dataset dependent. The problem with existing standardization methods is that they are meant for parametric analysis. Therefore, in this paper we are contributing novel standardization method, called **S**tandardization for **D**istribution **F**ree **S**tatistics (**SDFS**), for nonparametric analysis (Vatsa 2015, Vatsa 2017). SDFS is really useful for the preprocessing of sparse and small sample size dataset. It has been proven to be plausible way to address the problem of knowledge discovery, optimization, low dimensional data characterization like clustering and other data mining problems.

The motivation behind this method is the dataset of Dr. Ann Stapleton lab greenhouse experiment. The experiment was designed for maize plants in the greenhouse. The design principle was for ninety different inbred lines (different genetic backgrounds) of plants and forty plants of each line. These plants were treated with nine different growth conditions of nitrogen fertilizer and water. There were four plants under each growth conditions except for growth condition five, which had eight plants.

The motivation to know how these nine growth conditions determine the plant phenotype, observable and measurable traits of plant, led to proceed further. Therefore, they measured three prime phenotypes, well observable and measurable of each plant. The phenotypes - plant height, canopy spread, and stem diameter - were measured before and after applying the different growth conditions. Many plants died during the experiment, so the number of surviving plants is often much lower. Nevertheless, this dataset has not any kind of distribution. So, it needs nonparametric analysis for computing optimal growth condition of each inbred lines for each growth condition. Moreover, we were also looking for similar phenotypes among all ninety inbred lines (Vatsa 2015). But we were not able to solve these two problems because were not able to preprocess the data in right way. Therefore, I proposed SDFS, a novel standardization technique.

The rest of the paper is organized as follows. Sections 2, contain related works of other existing standardization methods. The proposed work is described in section 3. The results and discussion of the proposed technique and comparison with existing techniques are explained in section 4. The scalable feature of proposed techniques discussed in section 5. Finally, our conclusions are given in section 5 and possible future work are also addressed.

Related Work

Walesiak **\cite{walesiak90, walesiak99}** states that the standardization method is as follows

$$z_{ij} = bx_{ij} + a, (b > 0)$$

where z_{ij} (x_{ij}) denotes the value (standard value) of the j^{th} variable for the i^{th} object. The particular (often used) case of transformation is the one where: $b=1/\sigma$, $a = -\mu/\sigma$

where μ is location parameter and σ is spread or scatter parameter.

This can be written as $z_{ij} = (x_{ij} - \mu)/\sigma$
However, the general existing standardization methods for parametric analysis are given as **(stdize)**

$$f'(x) = \alpha f(x)+ \beta (f \circ g(x))/ \gamma$$

where:

- f \`(x) is the standardized variates;
- f(x) is the measured attributes;
- α and β are constants;
- f o g(x) shifts the attributes;
- γ is a rescaling function.

γ rescaling functions are hard to choose because the well-known standardization methods consider some descriptive statistics, such as sample mean, sample variance, weighted mean, weighted sum, sum, standard deviation, standard deviation about origin median, Median absolute deviation from median (MAD), Inter quartile range (IQR), range, midrange, Euclidean distance, Biweight one-step M-estimate (Owen 2010), Biweight A-estimate (Owen 2010, Goodall 83, and Kafadar 83} (A-estimators and M-estimators of location are independent of the underlying probability distribution function of the data because they minimize an objective function that is dependent on the distances from the observed values to the estimate. It is similar to Maximum Likelihood Estimators (MLEs)), Huber one-step M-estimate (Bickel 75), Huber A-estimate (Bickel 75, Goodall 83, Iglewicz 83) (These identify robust member of group in asymptotic behavior and has independent identically distributed errors with F distribution symmetric about zero), Wave one-step M-estimate (Iglewicz83), Wave A-estimate \cite{iglewicz83}, AGK estimate (a noniterative univariate form of the estimator) (ACECLUS) (Gnanadesikan 82), Mid-minimum spacing, Minimum spacing and L(p) or Minkowski metric (It is flat space metric used in special relativity. It is combination of Euclidean space and time into a four-dimensional manifold where the space time interval between any two events is independent of the inertial frame of reference in which they are recorded) (Walesiak 99, Jajuga99, Wikipedia).

In our experimental dataset the sample size is very small and so I can't determine its distribution, called nonparametric analysis. Assuming a normal or some other distribution will be hazardous for the data analysis. Therefore, there is need to propose novel standardization technique for nonparametric analysis.

Proposed Standardization, SDFS

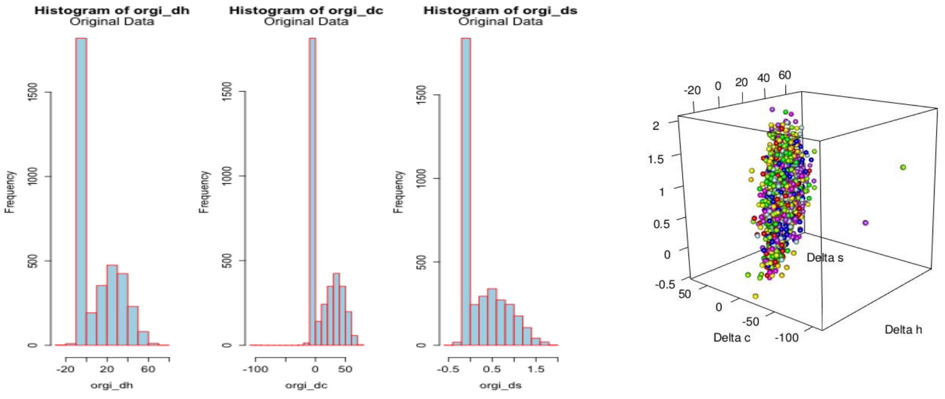
In this case the sample size of each combination of lines and growth conditions is too small to test the distribution of the data, so assuming a distribution would be hazardous for data analysis. Therefore, I have proposed a novel method of standardization for the analysis of sparse and non-parametric data:
$f_s(x) = (f(x) - \min f(x))/f(x)$,
where \$$f_s(x)$ represents the value of standardized variates, and f(x) and min f(x) represents measured variates and minimum value of measured variates across groups, respectively.

Results and Discussion

In order to test the effect of proposed standardization techniques, SDFS, we compared SDFS with the other existing standardization techniques like Mean, Median, Sum, Euclen, USTD, STD, Range, Midrange, Maxabs, IQR, MAD, ABW, AHUBER, AWAVE, AGK, Spacing and L. Moreover, according to the experimental results we evaluated SDFS in terms of the data distribution (by Histogram), spatial data representation using three - Dimensional Plots and non-parametric clustering, called MODECLUS, output by three dimensional plots. The analysis of these results is explained in following sections.

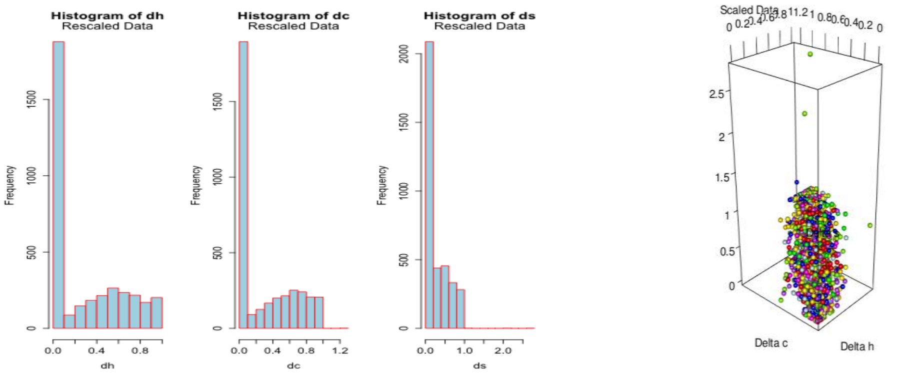
Data Distribution of Original and Rescaled Data

The data distribution and frequency of original and rescaled data on three attributes; Δh, Δc, and Δs; are represented in (Figure 1). The histograms of original (Figure 1a) and rescaled (Figure 1c) data is illustrated that these variates are asymmetric and have mixtures of positive, negative and zero data values. It also illustrates how wide data the range, shape and central location of the data are. Therefore, on the evidence of distribution and on the basis of normal quantile-quantile plots (not shown), we can infer that these three attributes are not normally distributed. Moreover, the three-dimensional plots of the three attributes original (Figure 1b) and rescaled by min-max normalization (Figure 1d) between 0 and 1, depicted in Figure 1, show that the data points are so compact and overlapping each other that they cannot be visually separated. Therefore, there is need to standardize this dataset to unravel the unseen information hidden in the dataset.



(a) Original

(b) Original



(c) Rescaled (d) Rescaled

Figure. 1. Distribution of Original and Rescaled Data

Distribution of Data Standardized by Conventional and Proposed Methods

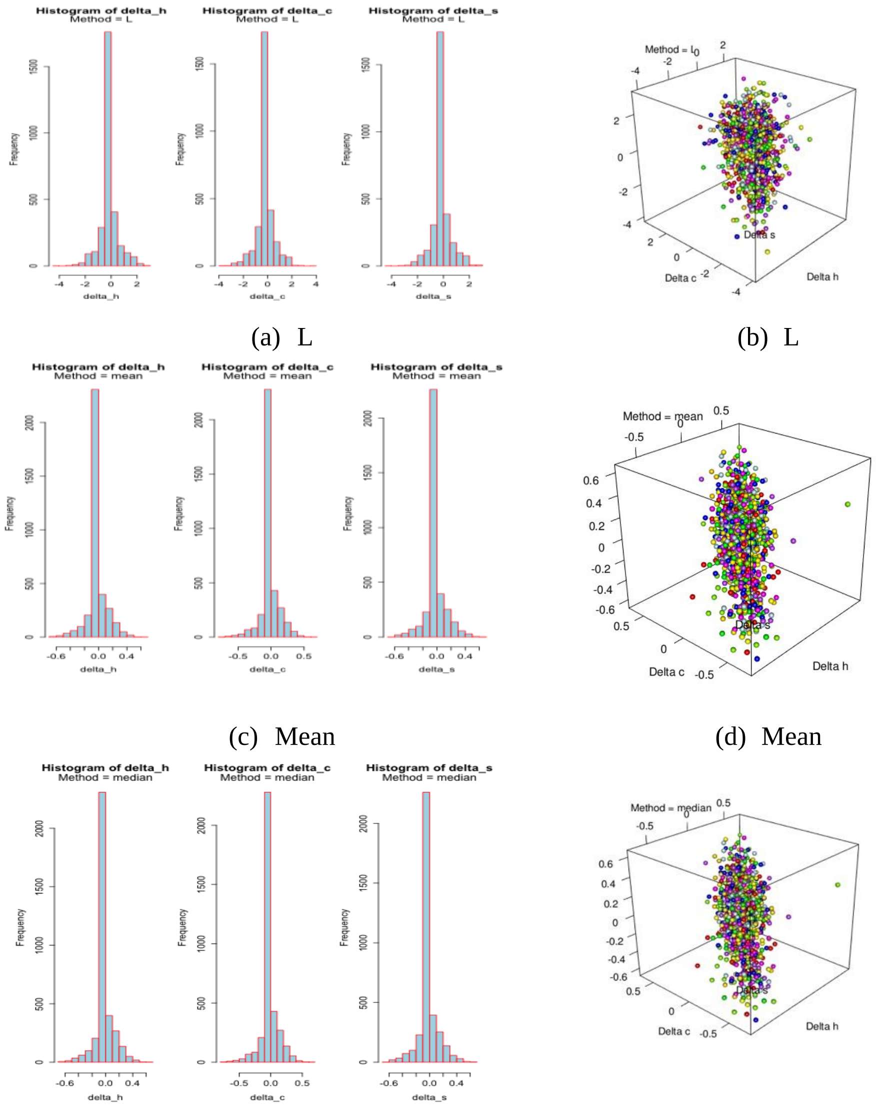
The histograms and three-dimensional plots of three standardized attributes; Δh, Δc, and Δs; are represented in Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, and Figure 7. These histograms show how wide data the range, shape and central location of the data are. The peaks of the heights in the histograms show that the frequency of data values of dataset. But both sides of the peaks do not have an equal number of data frequencies, which would make them look bell shaped. These histograms are categorized in three categories based on their symmetry and data values range.

The first category based on similarity among histograms and three-dimensional plots. This group shows standardization effect illustrated in Figure 2. The effect of standardization methods; L, Mean, and Median; are represented by histograms (Figure 2(a), Figure 2(c), and Figure 2(e)) and it shows the data distribution and frequency of three attributes. These histograms are asymmetric and have mixtures of positive, negative and zero data values. Whereas three dimensional plots (Figure 2(b), Figure 2(d), and Figure 2(f)) of this group shows that the data points are so diagonally compact and overlapped each other that they cannot be visually separated. Therefore, these standardization methods have not any positive effect on rescaled dataset.

The second category is represented in Figure 3, Figure 4, Figure 5 and Figure 6 (a and c). These figures show that the data distribution is asymmetric and have mixture of positive, negative and zero values for standardization methods, STD and AGK. However, Euclen, AHUBER, AWAVE, IQR, MAD, Maxabs, USTD, Range and SDFS methods have data values between zero and one. The three-dimensional plots of STD, Euclen, AGK, ABW, AHUBER, AWAVE, IQR, MAD, Maxabs, SUM and USTD are very dense center at top corner and overlapping to each other such that these standardizations have not any effect on rescaled data.

On other hand the third category is represented by methods Midrange (Figure 7 (c) and Figure 7 (d)) and Spacing (Figure 7 (e) and Figure 7 (f)). These histograms are asymmetric and have mixtures of positive, negative and zero data values. These histograms show how wide data the range, shape and central location of the data are. The peaks of the heights in the histograms show that the frequency of data values of dataset. But both sides of the peaks do not have an equal number of data frequencies, which would make them look bell shaped.

Therefore, on the evidence of distribution and on the basis of normal quantile-quantile plots (not shown), we can infer that these three variates are not normally distributed. Whereas methods, Range (Figure 6f) and SDFS (Figure 7b), three dimensional plots illustrate that most dispersed cloud produced, and we can even see by naked eye. Moreover, the cloud produced by SDFS standardization has very sparse and Most interestingly, there are no outliers after this standardization. Therefore, we can say that SDFS has better data distribution over other existing standardization methods. In next sub section we will be confirmed it by clustering outputs as well.



(a) L

(b) L

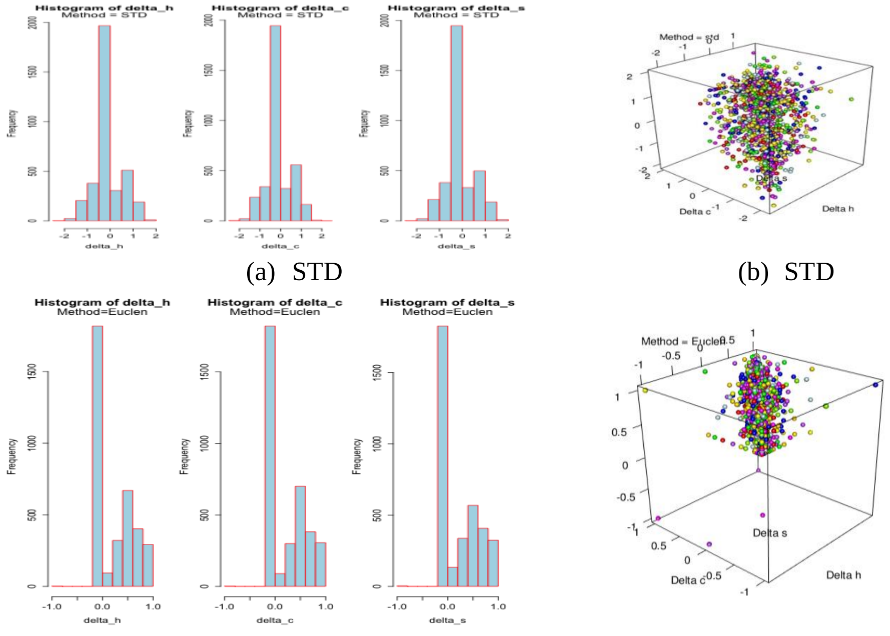
(c) Mean

(d) Mean

(e) Median

(f) Median

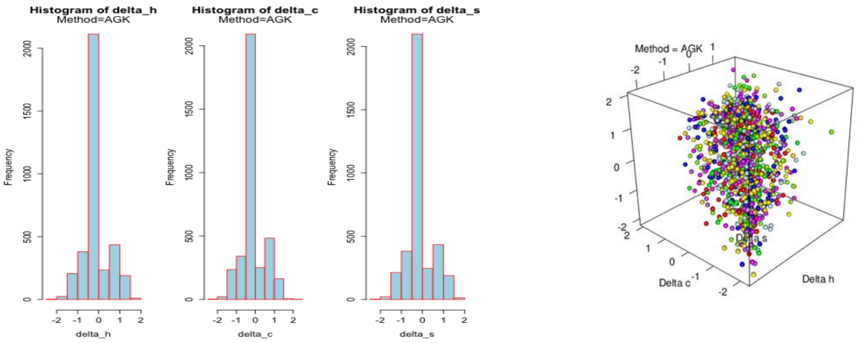
Figure. 2. Distribution of Standardized Data



(a) STD

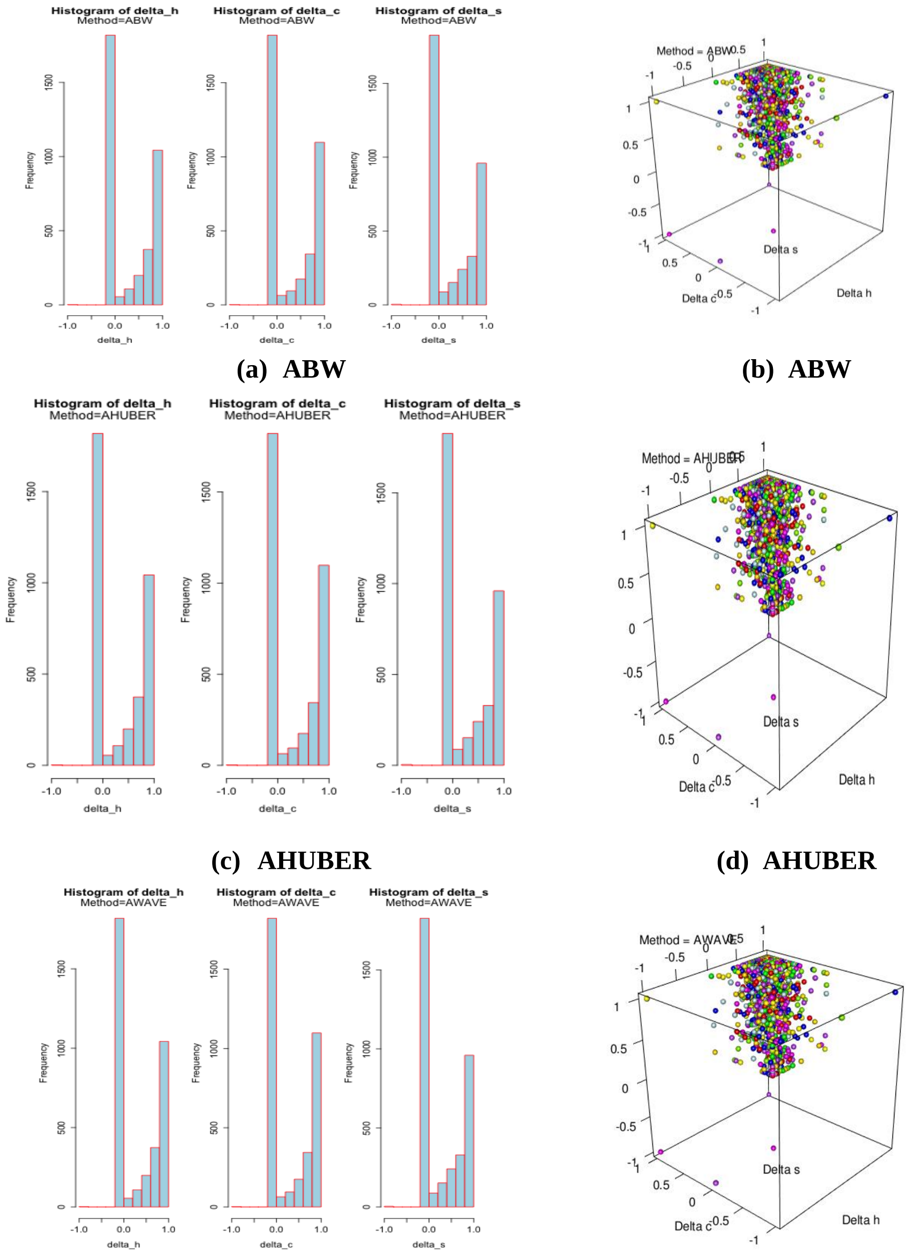
(b) STD

(c) Euclen (d) Euclen



(e) AGK (f) AGK

Figure. 3. Distribution of Standardized Data

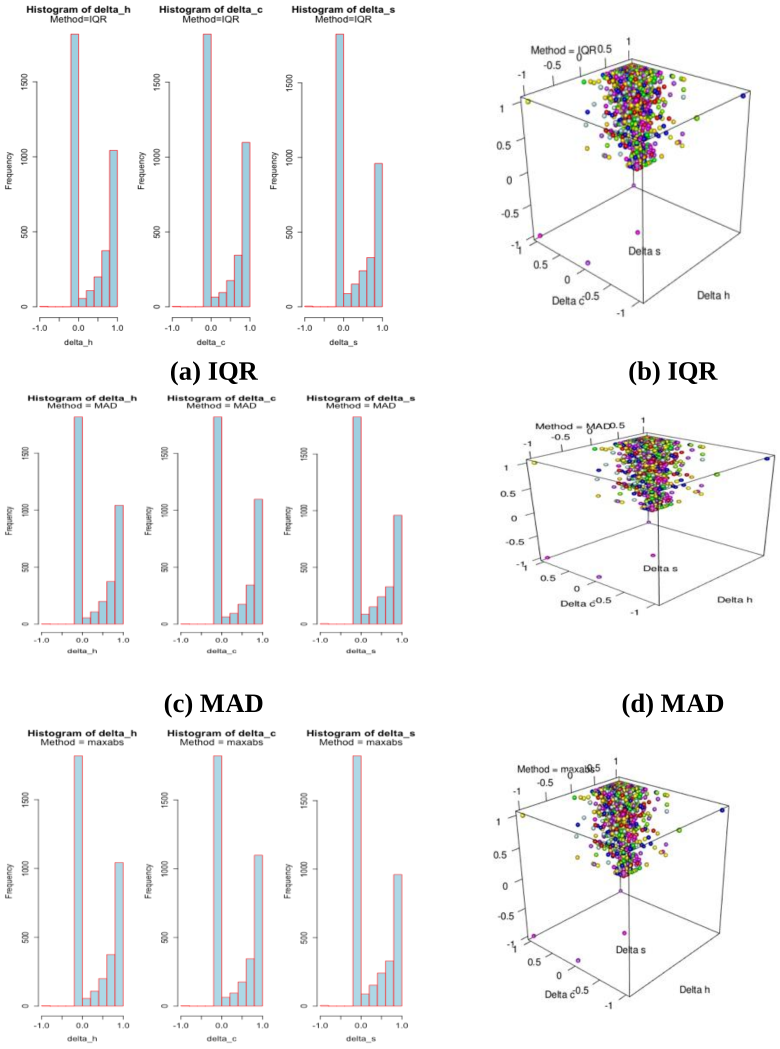


(a) ABW **(b) ABW**

(c) AHUBER **(d) AHUBER**

(e) AWAVE **(f) AWAVE**

Figure. 4. Distribution of Standardized Data

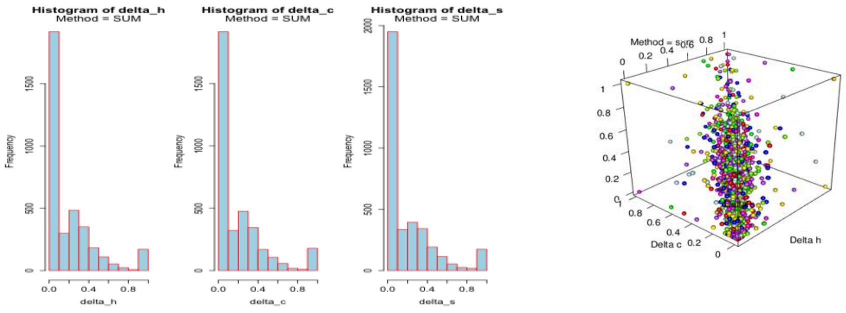


(a) IQR **(b) IQR**

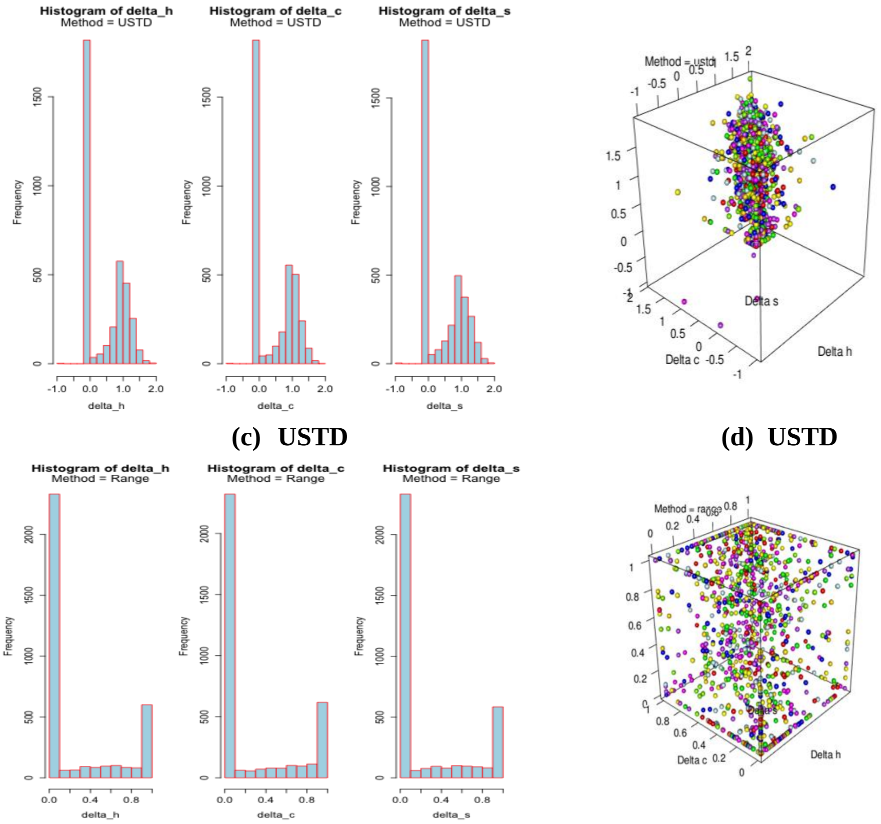
(c) MAD **(d) MAD**

(e) Maxabs **(f) Maxabs**

Figure. 5. Distribution of Standardized Data



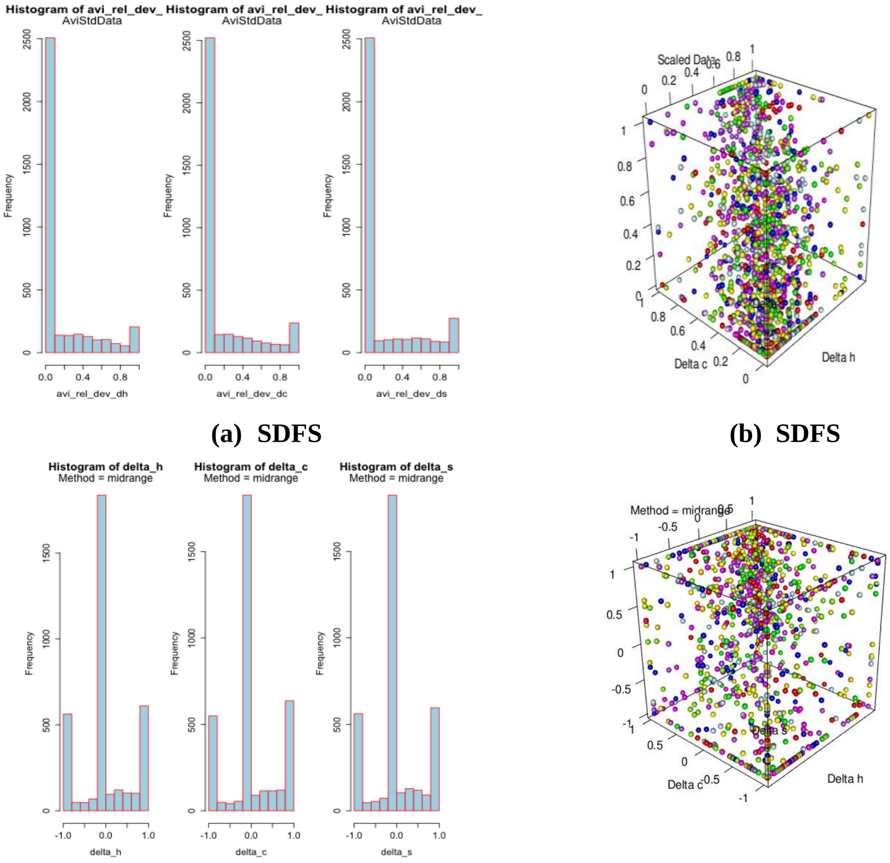
(a) SUM **(b) SUM**



(c) USTD **(d) USTD**

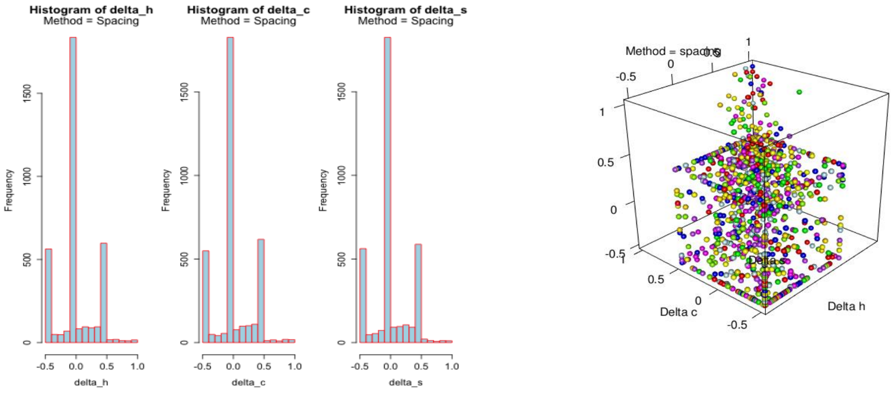
(e) Range **(f) Range**

Figure. 6. Distribution of Standardized Data



(a) SDFS **(b) SDFS**

(c) Midrange **(d) Midrange**



(e) Spacing **(f) Spacing**

Figure. 7. Distribution of Standardized Data

Effect of SDFS on Clustering

Since clustering is a very useful and popular data characterization method. To observe the effect of this characterization technique we used MODECLUS, a nonparametric clustering algorithm. The clustering performance is also used to evaluate the effect of the standardization technique. This standardization method helps MODECLUS in finding intuitive and natural clusters.

Figure 8, Figure 9, and Figure 10 represent plots of MODECLUS clustering output of standardized values of data points of ∆h, ∆c, and ∆s, which are calculated using standardization methods L, Mean, Median, STD, Euclen, AGK, ABW, AHUBER, AWAVE, IQR, MAD, Maxabs, Sum, USTD, Range, SDFS, Spacing and Midrange. The color of data points shows the cluster number in three dimensional plots.

All plots of clusters shows most of the data points are assigned in only one cluster and cloud are either compact diagonally or in top corner of three dimensional plots except for Range (Figure 10 (c)), SDFS (Figure 10 (d)), Spacing (Figure 10 (e)) and Midrange (Figure 10 (f)). The proposed novel standardization method (SDFS) clustering output (Figure 10 (d)) gives us clusters data points that are well separated with compact boundaries of clusters, as seen by eye, and distributed in the range [0,1]. Contrasting the most comparable plot of represented by range method but its clusters data points are sparse, and number of clusters are greater than this method as well.

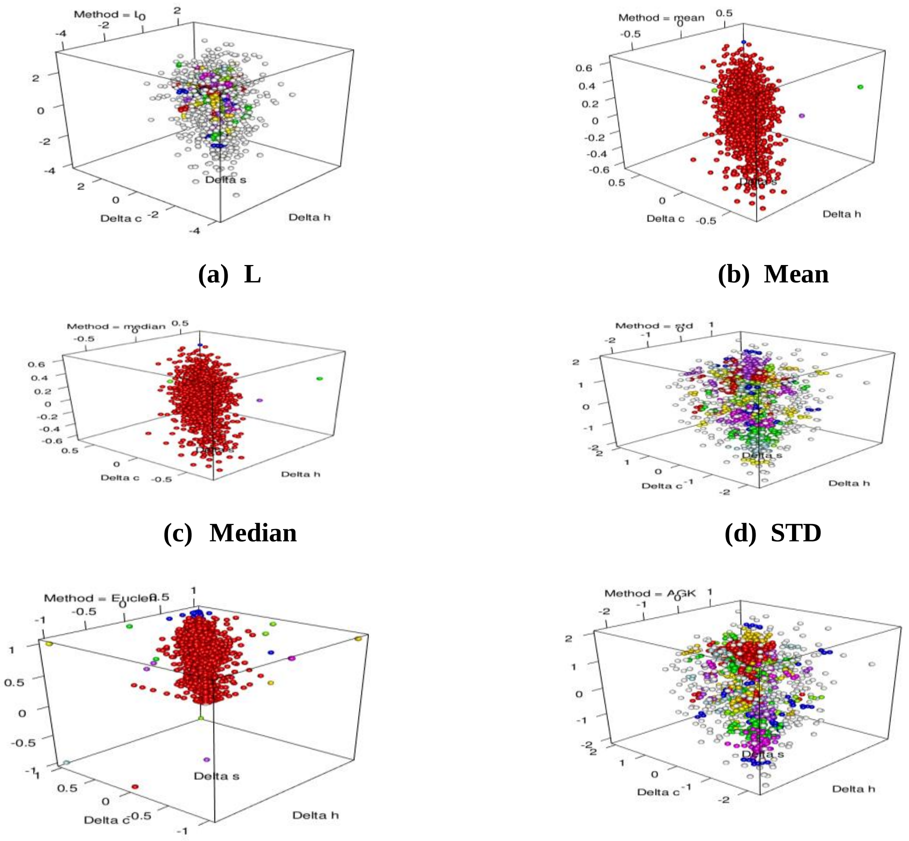
Therefore, the effectiveness of SDFS can be inferred that data points of clusters are well spread in three-dimensional space, showing non-linear relationships among them. Therefore, we can conclude that proposed standardization method, SDFS, is better standardization for nonparametric analysis for sparse dataset. On other hand this standardization is not showing good results for non-sparse data for parametric analysis.

Finally, the number of clusters and number of unclassified points with radius 0.20 at threshold 0.25 is shown in Table 1 using all existing and proposed standardization techniques (SDFS). Moreover, the clusters produced by range standardization is comparable, but it is very sparse around the space and number of clusters is ten with less density boundaries around the cluster's boundary. In contrast, the SDFS has six number of high frequency data boundary clusters with 100% classification of data points of the dataset.

Table 1. Summary of Number of Clusters Using Different Standardization Techniques

Methods	Number of Clusters	Number of Unclassified Points
Mean	5	0
Median	5	0
SDFS	6	0
Range	10	0
Sum	12	0
Spacing	15	0
Euclen	16	0
ABW	19	0

AHUBER	19	0
AWAVE	19	0
IQR	19	0
MAD	19	0
Maxabs	19	0
USTD	66	0
STD	99	336
AGK	100	341
L	100	579
Mid Range	100	59

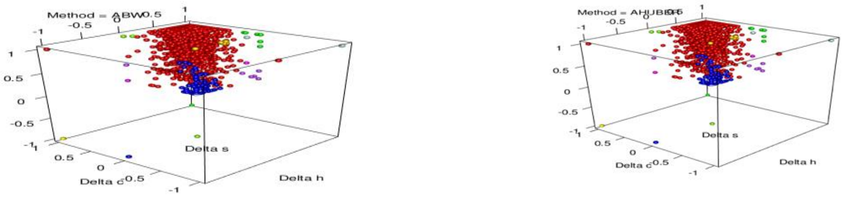


(a) L **(b) Mean**

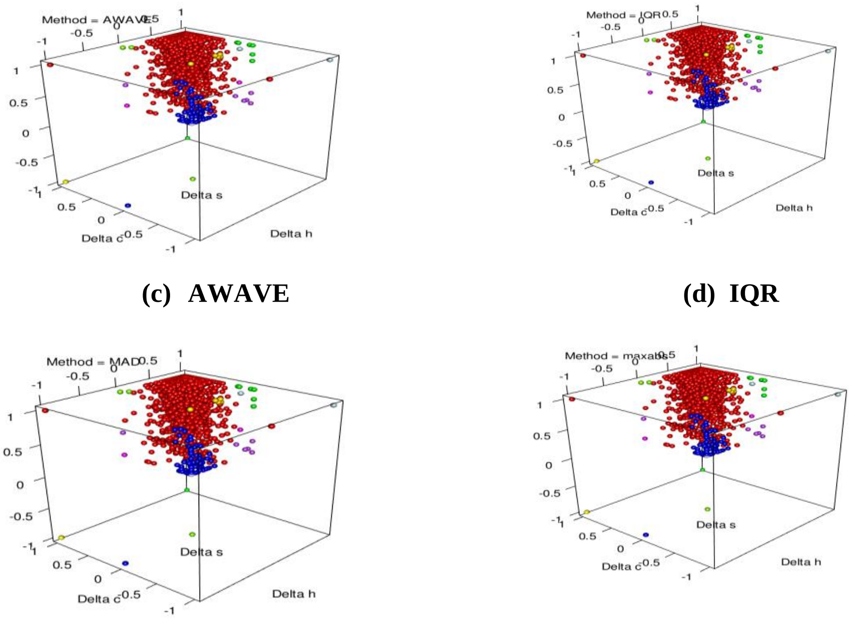
(c) Median **(d) STD**

(e) Euclen **(f) AGK**

Figure. 8. Clustering Output of Standardized Data



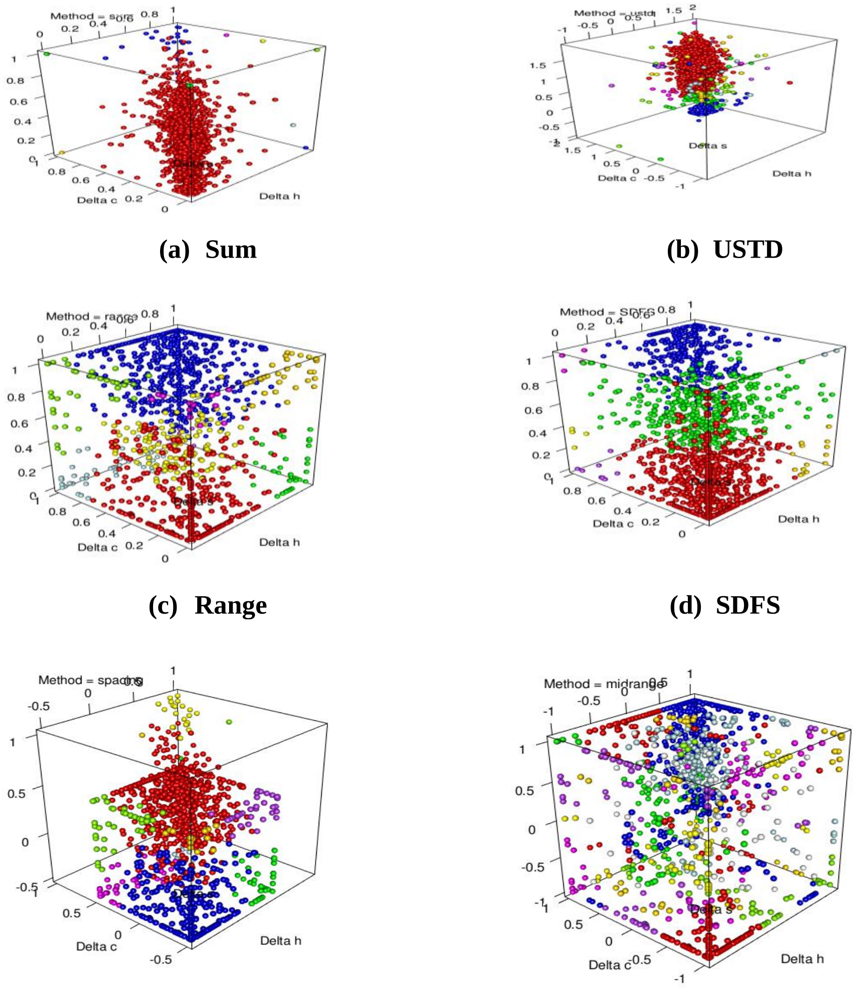
(a) ABW **(b) AHUBER**



(c) AWAVE **(d) IQR**

(e) MAD **(f) Maxabs**

Figure. 9. Clustering Output of Standardized Data



(a) Sum **(b) USTD**

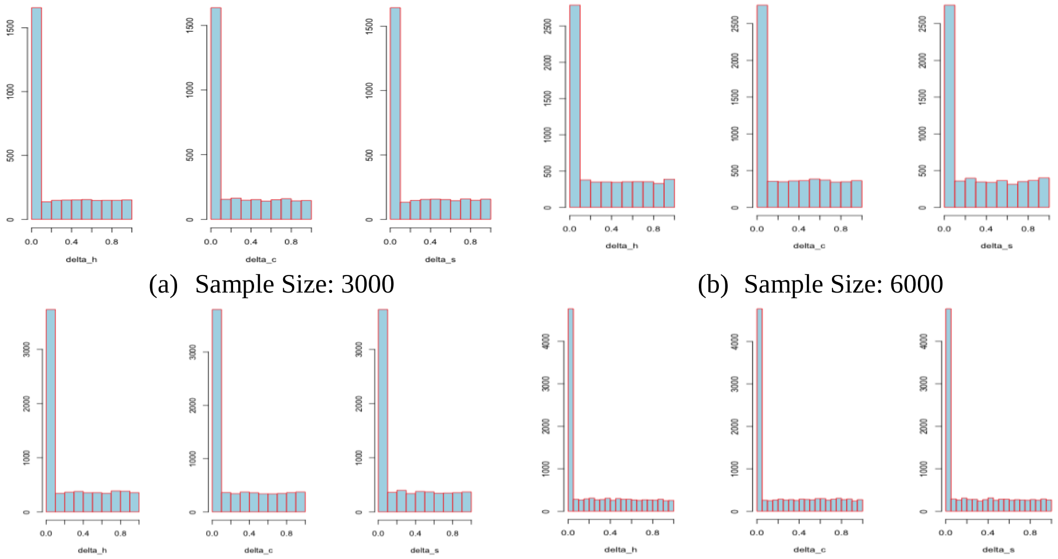
(c) Range **(d) SDFS**

(e) Spacing **(f) Mid-Range**

Figure. 10. Clustering Output of Standardized Data

Scalable Feature of SDFS on Clustering

Since SDFS is a standardization technique of non-parametric analysis, there is need to test scalability feature for sparse datasets. Therefore, four synthetic dataset of samples – 3000, 6000, 7000, and 10000 – have been created using NumPy library of Python (NumPy 2020). The distribution and sparseness of synthetic dataset are approximately similar the raw dataset of greenhouse experimental dataset. The distribution of synthetic dataset is represented by histogram in figure 11. The histograms of sample size 3000 (Figure 11 (a)), 6000 (Figure 11 (b)), 7000 (Figure 11 (c)), and 10000 (Figure 11 (d)) are illustrated and it shows that the data samples are sparse, and range of data points are between 0 and 1. It also represent the shape and central location of data which are similar (approximately) to experimental dataset.

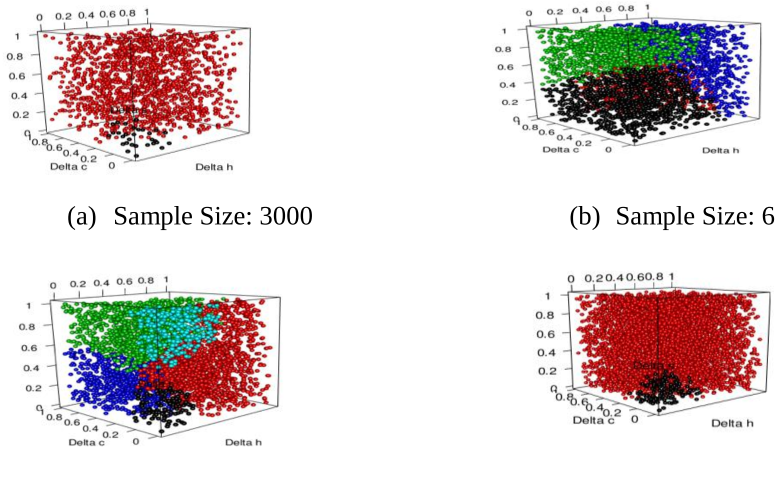


(a) Sample Size: 3000 (b) Sample Size: 6000

(c) Sample Size: 7000 (d) Sample Size: 10000

Figure. 11 Distribution of Synthetic Data

In order to find number of clusters in synthetic datasets, the data samples are standardized using SDFS. Thereafter, the standardized datasets are processed using MODECLUS clustering (radius = 0.20 and threshold = 0.25) method. The clustering output of sample size 3000 (Figure 12 a), 6000 (Figure 12 b), 7000 (Figure 12 c), and 10000 (Figure 12 d) are shown in figure 12. The color of data points in figure 12 shows the number of clusters. The figure 12 (a) and figure 12 (d) show only two clusters and most of data points belong to cluster one (red in color) and a few data points belong to cluster number two. Whereas figure 12 (b) and figure 12 (c) show four and five clusters respectively, these clusters are partially overlapping and representing natural clusters (arbitrary shape and size). Therefore, based on the result (figure 12), SDFS works perfectly if sample sizes are in the range of 3000 to 10000 data points.



(a) Sample Size: 3000 (b) Sample Size: 6000

(c) Sample Size: 7000 **(d)** Sample Size: 10000

Figure. 12 Clustering Output of Standardized (SDFS) Synthetic Data

Conclusion

Our experimental study show that the proposed method performs completely incomparable to other evaluated techniques. It is superior over existing techniques are due to fact that they assumed that data has some distribution (parametric analysis), while proposed techniques is very useful for data that has not any kind of distribution (nonparametric analysis) and sparse dataset with mixture of positive and negative data values. In contrast to existing standardization techniques, proposed standardization technique (SDFS) does not only aim at characterizing data but also will be used in finding optimization of methods in low dimensional data analysis process. Several questions remain to investigate in our future work like it outperform for univariate data, but we will have to make this for multivariate nonparametric analysis.

Acknowledgment

I sincerely appreciate Dr. Toni Kazic for advice and Dr. Ann Steplton for providing raw dataset of maize phenotype in order to test proposed method. A part of this work was done while I was a graduate student at University of Missouri – Columbia.

References

Hopkins, S., Dettori, J. R., & Chapman, J. R. (2018). Parametric and Nonparametric Tests in Spine Research: Why Do They Matter? Global Spine J. issue 8(6), pp 652-654, Doi: 10.1177/2192568218782679.

Vatsa, A. (2017). An Approach of Clustering Biological Phenotypes (Doctoral dissertation), University of Missouri – Columbia: https://mospace.umsystem.edu/xmlui/handle/10355/62341?show=full.

Vatsa, A. (2015). Characterizing Low-Dimensional Phenotypes by Clustering (Master's thesis), University of Missouri – Columbia: https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/47047/research.pdf?sequence=2&isAllowed=y.

Kelly, D., Vatsa, A., Mayham, W., Ngoˆ, L., Thompson, A., & Kazic, T. (2015). An opinion on imaging challenges in phenotyping field crops, *Mach. Vision Appl*.

Kelly, D., Vatsa, A., Mayham, W. & Kazic, T. (2015). Extracting complex phenotypes from images, *Mach. Vision Appl*.

Han, J., Kamber, M. & Pei, J. (2012). Data Mining Concepts and Techniques: *Cluster Analysis*. Morgan Kaufmann, NewYork. third edition.

Standardization, BioMedWare (2015–present). *methods for standardization.* http://www.biomedware.com/-Methods_for_data_standardization.htm: BioMedWare,.

Jajuga, K. & Walesiak, M. (2000). Standardization of data set under different measurement scales, *Chapter Classification and Information Processing at the Turn of the Millennium Part of the series Studies in Classification, Data Analysis, and Knowledge Organization*, vol. 1, pp. 105–112.

Milligan, G. W. & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis, *J. Classification*, vol. 5, pp. 181–204.

Stdize, SAS Proc., (2015–present). Http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#stdize_toc.htm: SAS(R) 9.4 Functions and CALL Routines: Reference, third ed.

Owen, M. (2010). *Tukey's Biweight Correlation and the Breakdown (*Master's thesis*)*. Pomona College, California.

Goodall, C. (1983). M-estimators of location: An outline of theory, in D. C. Hoaglin and Tukey [14], pp. 339–403.

Kafadar, K. (1982). "The efficiency of the biweight as a robust estimator of location," *J. Research of the National Bureau of Standards*, vol. 88, pp. 105–116.

Bickel, P. J. (1975). One-step huber estimates in the linear model. *J. Am. Stat.* location, in D. C. Hoaglin and Tukey [14], pp. 405–431. *Assoc.*, vol. 70, pp. 428–434.

Iglewicz, B. (1986). Robust scale estimators and confidence intervals for location, in D. C. Hoaglin and Tukey [14], pp. 405–431.

*Wikipedia, Welcome to Wikipedia (*2001–present*). the Free Encyclopedia that Anyone Can Edit*. http://en.wikipedia.org/wiki/Main_Page: WikiMedia Foundation.

Hoaglin, F. M. D. C., & Tukey, J. W. (1983). eds., *Understanding Robust and Exploratory Data Analysis*, (New York), John Wiley and Sons.

NumPy, (Access on 2020). Python - Random Numbers in NumPy: https://www.w3schools.com/python/numpy_random.asp