SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Using Parametric Test to determine the Significance of Banded patterns in N-dimensional 0-1 dataset

Fatimah Binta Abdullahi[1], Salihu Idi Dishing[2], Salisu Aliyu[3], Jamilu Falgore[4]
[1,2,3]Department of Computer Science, Faculty of Physical Sciences, Ahmadu Bello University, Zaria
[4]Department of Statistics, Faculty of Physical Sciences, Ahmadu Bello University, Zaria
zeeh429@gmail.com, dishing@yahoo.com, aliyusalisu@abu.edu.ng, jamilu@yahoo.com

Abstract

Background*: The identification of banded patterns in N-dimensional (ND) 0-1 dataset is one where all the elements in the dimensions are arranged such that the one's entries are ordered about the center of dimensions. Reordering 0-1 datasets in order to determine banded patterns in data, allows for the identification of interesting pattern that are hidden in the data. The challenge is whether or not the identify banded patterns are significant. Previous work on the significance of banded patterns using parametric test was aimed at 2D and 3D banding algorithms, using a single banding algorithm in each case, which meant no work on the significance of banded patterns using different banding algorithms with different dimensions.* ***Aim****: To this end, this paper presents a comparison between ND banding algorithms for 2D and 3D.* ***Method****: The approach is to use parametric test on synthetic data, UCI and the real datasets taken from the cattle tracing system (CTS). The ND banding algorithms considered for 2D are:2D banding, barycentric (BC) and 2D sort, and for 3D are: exact-Euclidean, exact-Manhattan variations and the approximate.* ***Results****: The experimental results presented shows the significance of banded patterns with p value less than 0.05. However, the post-hoc test result shows a statistically significant difference between 2D banding and BC, 2D banding and 2D sort, exact-Euclidean and exact-Manhattan, exact-Euclidean and the approximate but no significant difference between BC and 2D sort as well as exact-Manhattan and the approximate.*

Keywords— *Banded pattern, N-dimension, 0-1 data, Parametric test.*

1. Introduction

The paper presents the techniques for determining the significance of banded patterns in ND 0-1 datasets, that consists of only 0's and 1's. For better understanding, the cell with a one value has a "dot" and zero value empty cell. 0-1 datasets commonly occur in a number of real-world application domains, these are gene and probe mapping in bio-informatics Alizadeh et al., (1995), Cheng and Church (2000), and Myllykangas et al., (2006). Sites and species co-occurrences in paleontology (Atkins et al., (1998) and Mannila et al., (2006)). The advantages of banded patterns include the following; first it gives an insight about the data by clustering records that co-exist. Secondly, it makes visualization of the dataset possible by given a better understanding of the datasets. Thirdly, it compresses the datasets in a way that improves the operation of algorithms that work with 0-1 datasets.

Several studies on Banded pattern identification in ND 0-1 datasets, where (N=2) exists; however, the concept of ND banded patterns as presented in this paper started from the work of Garriga et al., (2008), Garriga et al., (2001) and Makinen and Sirtola (2005) (similarly the work of Abdullahi et al., (2014), Abdullahi et al., (2015), Abdullahi et al., (2018), Abdullahi (2019a), Abdullahi (2018a), Abdullahi (2018b), Abdullahi and Coenen (2018a), Abdullahi and Coenen (2020) in the context of 2D and 3D. The early work on banded patterns identification was based on permutations and was considered NP-complete problem in the work of Papadimitriou (1976). The work presented in this paper is aimed at determining the significance of banded patterns on difference ND banding

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

algorithms using parametric tests. Given any ND 0-1 dataset with dimensions $\{\dim_1, \dim_2, ..., \dim_n\}$, we can achieve some banding by rearranging the elements in each dimension. Though in a real-life situation, a perfect banding may not be feasible; but there is a possibility of achieving a best banding. Therefore, the following are the objectives of this paper: (i) to investigate the significance of banded patterns on different ND banding algorithms for (2D and 3D) and (ii) to assess whether the banded patterns obtained from the ND banding algorithms have the same significance or differ in significance. The idea is to find out whether the banded patterns (b) obtained from different ND banding algorithms on datasets (d) after a number of iterations differs significantly or are the same. We let hundred (100) be the expected maximum number of iterations required to achieve banding (This is because the best bandings were all achieved within hundred iterations). We define the null hypothesis (H_0) by assuming that there is no significant difference between the banding algorithms and the alternative hypothesis (H_A) that there is a significant difference between the banding algorithms with a significance level of $p = 0.05$.

The structure of remaining part of the paper is as follows. The second section presents the materials and methods on different banding algorithms and a brief background review of some relevant work. The third section presents the evaluation results of the ND banding algorithms. The paper discusses the relevant results obtained in section four and finally, section five of the paper concludes with the main findings.

2. Materials and Methods

The work presented in this paper is to determine the significance of banded patterns on different banding algorithms that operate in ND 0-1 data sets using parametric test. Banding applications include identifying entities that co-occur in co-occurrence analysis Atkins et al., (1998), Mannila et al., (2006). An example is the paleontological application that combine Neolithic sites with its fossil species. Similarly, Very-Large Scale Integration (VLSI) chip design by Koebe and Knochel (1990), for block alignment problems aimed at using banding concept to identify channels between the circuit component blocks. Finally, in graph drawing, the idea is minimizing the number of edge cross overs, this was achieved using banding concept in Sugiyama et al (1981). The idea of banding emanates from numerical analysis according to Atkinson (2008), where it was used in the context of resolution of linear equations.

The banding concept relates to reorderable matrices, reorderable patterns and bandwidth minimization. Reorderable matrices simply refers to data visualization in 2D. This concept has helped in achieving a better understanding of data as proposed in Bertin (1981), Bertin (1999) and Makinen and Siirtola (2000). In 19th century, an English Egyptologist known as Petrie, used manual reordering approach to understand archaeological datasets as mentioned in Liiv (2010). For reorderable pattern, the idea is to arrange a 2D matrix column wise and row wise so that hidden pattern of interests can be identified (Aykanat et al., (2004) and Deutsch and Martin (1971)).

The approach finds a pattern P, revealed by arranging a given 2D matrix column wise and row wise, this however are not necessarily 0-1 datasets as such the pattern P may be of any form. The problem with this concept is how to rearrange 2D datasets to find the pattern P. Similarly, minimization of Bandwidth according to Cheng (1973a), Cheng (1973b) is the process of minimizing the bandwidth of 2D sparse matrix by arranging it row wise and column wise such that the one's entries are as close as possible along the main diagonal. However, it is worth noting that there are several studies on methods of data reordering in different applications domain. Such example is the Barycentric (BC) algorithm proposed by Makinen and Sirtola (2005), which is directed at 2D data, originally the concept support the drawing of graphs.

The algorithm finds permutation of 2D matrix row wise and column wise such that the one's entries converge about the diagonal using a barycentric measure. The BC algorithm finds the average position of each one's entries column wise/row wise. In this paper, the BC algorithm was considered one of the 2D algorithms used to determine the significance of banded patterns. Similarly, the 2D sort method, is an approach used to reorder 2D

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

matrix in order to detect patterns along the main diagonal of the matrix Bertin (1999), Harri and Eekki (2005). The 2D sort method works by computing the sum of each given 2D matrix column wise and row wise and then sorted the matrix accordingly. For the 3D banding algorithms, the approximate algorithm finds an approximate banding using dimension pairing; while the exact algorithm finds an exact banding for the entire data space.

The two metrics considered for exact banding algorithms are the Euclidean and Manhattan distance proposed by Abdullahi et al., (2016). There is limited work on the significance of banded patterns, for instance the work of Abdullahi and Coenen (2018b), Abdullahi and Hassan (2020), proposed the statistical significance testing for banded patterns using Gaussian distribution. The idea was to apply Gaussian distribution on randomly generated 2D datasets without banding and then used it confirmed whether the obtained banded patterns are significant or not with respect to their distance from mean.

The experimental results indicate the statistical significance of banding with at least one to two standard deviations from the mean. Additionally, the findings also establish the possibility of generating Gaussian distribution curves with ranges of one density values. Although, the results were promising however, the limitation was that it requires generating the Gaussian distribution curve for each of the data sets. Abdullahi (2019b) also proposed the use of student t-statistic and chi-square test to determine the statistical significance of 3D banded patterns. The experimental results presented shows 5% level of significant, and the paper only consider the exact-Euclidean metric of the 3D banding algorithm. Similarly, Abdullahi et al (2021), proposed the used of three parametric tests: student t-test, paired student t-test and ANOVA test to determine the significance of 2D banding Algorithms. The results from the experimental analysis shows a significant difference between the banding Algorithms.

The identification of banded patterns in ND 0-1 datasets requires arranging all elements in the dimensions to identify a best banding (or an approximate if no best exist). The idea presented by Abdullahi et al., (2015), use the "banding scores" (BS) concept. BS is a score value obtained by summing the one's entries associated with each index in the dimension and divide by the maximum number of the one's entries. The datasets comprise of set of dimensions (DIM) that consist of n number of dimensions, where DIM = $\{dim_1, dim_2, \ldots, dim_n\}$. Note that the dimension sizes are not equal. The dimension i represents a sequence of k index values $\{e_{i1}, e_{i2}, \ldots, e_{ik}\}$, i the dimension identifier and e_{ij} represents item e in dimension i with respect to index j. Every location has 0's and 1's. The distribution of the data entries is based on the application domain. Every one entry represents a set of coordinates (c1, c2, . . . , cn), where n is the number of dimensions and $c1 \in dim_1$, $c2 \in dim_2$ and so on.

For 2D, DIM = $\{dim_i, dim_j\}$, meaning computing the banding scores for dimension i with respect to dimension j and vice versa, by simply adding the index values for y (x) and dividing by the maximum indexes. For 2D datasets, the banding score (bs_{iq},) is computed as shown in Eq. 1.

$$bs_{iq} = \frac{\sum_{i=1}^{i=|M|} m_i \in M \times n_i \in N}{\sum_{i=1}^{i=|M|} \left(k_p - i + 1\right) \times n_i \in N'} \qquad (1)$$

Where: $M = \{m_1, m_2, \cdots m_k\}$ list of dimension j indexes for the one's entry that features index q in dimension i, $N = \{n_1, n_2, \cdots n_k\}$ list of the number of one's entries at each location corresponding to list l, N′ represents list similar to N but rearranged in descending order to the number of one's entry at each location and k_j is the size of dimension $_j$. Converting Eq. 1 to address ND data, Eq. 2 is used to calculate the banding scores for ND.

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

$$bs_{iq} = \frac{\sum_{i=1}^{i=|L|} dis\tan ce(l_i \in L) \times n_i \in N}{\sum_{i=1}^{i=|MAXDIST|} (\max Dist_p \in MAXDIST) \times n_i \in N'} \quad (2)$$

Where: L represents list of locations in the data space of size n less than 1, and $n = |DIM|$), $MAXDIST$ represents list of maximum distances in descending order where $|MAXDIST|$ is equivalent to $|L|$, N and N' as previously defined. The function distance () returns a distance to the origin of the datasets with one less than the original data space presented as coordinates set $(s_1, s_2, s_3 \cdots, s_{n-1})$.

For the 3D approximate banding algorithm, the banding score for index k in dimension i with respect to dimension j, denoted as bs_{ijk}, is calculated using Eq. 3.

$$bs_{ijk} = \frac{\sum_{k=1}^{k=|N|} n_k}{\sum_{l=1}^{l=|N|} (|dim_l| - p + 1)} \quad (3)$$

Where N is the set of dimension j indexes representing one's entry whose dimension i coordinate is equivalent to k ($N = \{n_1, n_2, \ldots\}$). The Overall Score (OS) for the whole configuration is computed using Eq. 4, where OS_i is the overall scores for dimension *i*.

$$OS = \frac{\sum_{i=1}^{i=|\dim|} OS_i}{|\dim|} \quad (4)$$

The OS_i value is calculated using Eq. 5. Here the banding is weighed with respect to its index location, as the aim is to arrange the one's entries from the top corner of the data space.

$$OS_i = \frac{\sum_{j=1}^{j=|p_i|} bs_{ij} \times (p_i - k + 1)}{\frac{p_i(p_i+1)}{2}} \quad (5)$$

2.1 The ND Banding Algorithm

The pseudocode for the ND banding algorithm is presented in Listing 1. The inputs in (lines 1 and 2) are: 0-1 data set D, consisting set of tuples of the form (c1, c2, …) indicating the location of each one's entry in the datasets; and $DIM = \{\dim_1, \dim_2, \cdots, \dim_n\}$ set of dimensions in D. The output is an arranged data set D with minimum OS value. The algorithm iterates over the datasets. For every iteration, the algorithm rearranges the indexes in DIM, using the scoring concept in Eq. 1, 2 and 3 to identify a revised DIM (Step 3). This revised DIM is then used to arrange D to give D′ (Step 9). A new value of OS is computed (using Eq. 3 and 4). If the new OS value (OS_{new}) is less than the current OS values. The algorithm terminates with the previous data configuration and the value OS. Hence, D, DIM and OS are updated (Steps 9 to 11) and the algorithm continues.

INPUT: 0-1 dataset D with DIM, where DIM = {dim_1, dim_2, …, dim_n}
OUTPUT**:** N rearrange to minimize OS
Begin:
Step 1: OS = 1.0
Step 2: LOOP

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Step 3: DIM′ = The rearranged DIM using Eq. 1, 2 and 3
Step 4: D′ = The rearranged D according to DIM′
Step 5: OS_{new} = calculated OS values for D using Eq. 4 and 5
Step 6: IF ($OS_{new} \geq OS$) THEN
Step 7: break
Step 8: ELSE
Step 9: D = D′
Step 10: DIM = DIM′
Step 11: OS = OS_{new}
Step 12: END IF
Step 13: END LOOP
Step 14: Terminate with D and OS
End

Listing 1. Generic ND banding Algorithm

The approximate variation, of the ND banding algorithm pairs the dimensions to compute the banding scores as presented in Eq. 3. We replace Steps 3 and 4 in Listing 1, with the pseudocode in Listing 2. The Approximate ND algorithm operates by considering all pairings ij in the dimension. For each pairing, the banding score bs_{Sijk} for e index k in dimension i is computed with respect to dimension j (Step 4), and the indexes in dimension i are arranged according to the computed banding score values (Step 6) and then D (Step 8).

Step 1: FOR $i = 1$ to $i = |DIM| - 1$ DO
Step 2: FOR $j = i + 1$ to $j = |DIM|$ and $j \neq i$ DO
Step 3: FOR $k = 1$ to $p = |p_i|$ DO
Step 4: bs_{ijk} = calculate the banding score for index k in dimension i with respect to dimension j using Eq. 3
Step 5: END FOR
Step 6: DIM = rearrange dimension i according to their bs_{ijk} values
Step 7: D = rearrange D according to DIM_i
Step 8: END FOR
Step 9: END FOR

Listing 2. The Approximate ND banding Algorithm

For the exact ND variation, we replace Steps 3 and 4 in Listing 1, with pseudocode given in Listing 3. The exact ND banding algorithm iterates over the entire datasets computing the banding scores for index q in dimension i using Eq. 2. For each dimension, it rearranges the indexes in the dimension according to the bs_{ij} values (Step 5), then reconfigure D (Step 6).

Step 1: FOR $i = 1$ to $i = |DIM|$ DO
Step 2: FOR $j = 1$ to $j = ki$ DO
Step 3: bs_{ij} = calculate banding score for index q in dimension i using Eq. 2
Step 4: END FOR

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Step 5: DIM = rearrange DIM_i according to bs_{ij} values
Step 6: D = rearrange D according to DIM_i'
Step 7: END FOR

Listing 3. The exact ND banding Algorithm

The 2D banding algorithm iterate over the 2D datasets by calculating the banding scores for index *q* in dimension i using Eq. 1 (Steps 3 to 5) as presented in Listing 4. The calculated banding scores are used to rearrange the indexes in the dimension i (Step 6). The banding scores for each index *q* in dimension j is then calculated using Eq. 1 (Steps 7 to 9). The calculated banding scores are used to rearrange the indexes in the dimension j (Step 10). The new OS values is calculated using Eq. 4 and 5 (Step 11), if the new OS value is greater than the previous OS value (Step 12), then algorithm exits and return M and OS (Steps 15 and 16). Otherwise, the algorithm repeats and exits if no further change.

INPUT: n x m (0-1) matrix M
OUTPUT: Reordered matrix row wise and column wise
BEGIN:
Step 1: OS = 1.0
Step 2: LOOP
Step 3: FOR ALL $j \in \{1 \cdots \dim_i\}$ DO
Step 4: Bs_{ij} = calculate banding score for dimension i using Eq. 1
Step 5: END FOR
Step 6: M = Rearranged dimension i according to the banding score
Step 7: FOR ALL $i \in \{1 \cdots \dim_j\}$ DO
Step 8: bs_{ij} = calculate banding score for dimension j using Eq. 1
Step 9: END FOR
Step 10: M′ = Rearranged dimension j according to the banding score
Step 11: OS_{new} = calculated OS value for M using Eq. 4 and 5
Step 12: IF ($OS_{new} \geq OS$) THEN
Step 13: break
Step 14: ELSE
Step 15: M = M′
Step 16: OS = OS_{new}
Step 17: END IF
Step 18: END LOOP
Step 19: EXIT with M and OS
END

Listing 4. The 2D Banding Algorithm

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

The BC algorithm first calculates the BC for each row *i,* and then arranged the rows in ascending order of BC values. The matrix then transposes and recomputes the BC values for the columns and arrange it in ascending order of values. The process repeats until no more changes. This BC compute the Average Moment (AM) as shown in Eq. 6. Where b_{ij} is the j^{th} entry for the row /column *i,* and *n* number of columns / rows.

$$AM = \frac{\sum_{j=1}^{n} j b_{ij}}{\sum_{j=1}^{n} b_{ij}} \qquad (6)$$

Listing 5 presents the pseudocode for the BC algorithm. The input is a 0-1 data set M. The output is a reordered matrix *M*. The BC computes the AM for all rows in M (Steps 2 to 4), then reorders the rows in ascending order of the AM value (Step 5). The algorithm then transposes the matrix M^T (Step 6) and iterates until no more change (Step 7). It exits with M (Step 10).

```
INPUT: An n x m (0-1) matrix M
OUTPUT: Reordered matrix row wise and column wise
BEGIN:
Step 1: LOOP
Step 2: FOR each row i ϵ M DO
Step 3: Calculate AM values for row i using Eq. 6
Step 4: END FOR
Step 5: M′ = rearrange M with row wise according to AM values in ascending order
Step 6: M^T = M′ transpose
Step 7: Repeat process on M′ until no more change
Step 8: M′ = M
Step 9: END LOOP
Step 10: EXIT with M
END
```

Listing 5. The Barycentric Algorithm

The 2D sort algorithm iteratively reorder a 2D matrix so as to identify banded patterns, the one's entries known as "black areas" in Bertin (1981) are arranged about the main diagonal. The 2D sort algorithm works by calculating the sum of rows and columns and then arranges them row wise / column wise in ascending order of the row /column sum. The 2D sort algorithm then calculate the row / column sum as shown in Eq. 7. Where N_{ij} is the position j of the cell for row/column *i,* and *n* is the number of columns /rows.

$$S = \sum_{j=1}^{j=n} N_{ij} \qquad (7)$$

The pseudocode for the 2D sort algorithm is presented in Listing 6. The input is a 0-1 data set M. The output is a rearranged matrix M. The 2D sort algorithm calculates the sum for all rows in M (Steps 2 to 4), then reorders them row wise in ascending order of their sum (Step 5). The algorithm then calculates the sum for all columns in M′ (Steps 6 to 8), then reorders them in column wise in ascending order of their sum (Step 9). The process is repeated until no more changes (Step 10). It then terminates with M (Step 12).

```
INPUT: An n x m (0-1) matrix M
OUTPUT: Rearranged matrix M row wise and column wise
BEGIN:
Step 1: LOOP
```

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Step 2: FOR EACH row i ϵ M DO
Step 3: Calculate row sum for row i using Eq. 7
Step 4: END FOR
Step 5: M′ = rearrange M row wise in ascending order of the row sum
Step 6: FOR EACH column j ϵ M DO
Step 7: Calculate column sum for column j using Eq. 7
Step 8: END FOR
Step 9: M″ = rearrange M column wise in ascending order of the column sum
Step 10: Repeat process on M″ until no more change
Step 11: END LOOP
Step 12: EXIT with M
END

Listing 6. The 2D sort Algorithm

3. Results

This section presents the experimental results obtained from the ND banding algorithms. We conduct the experiments using: (i) 2D randomly generated datasets in Coenen (2003), (ii) 2D datasets taken from University of California Irvine data repository Blake and Merz (1998), (iii) 3D randomly generated datasets in Coenen (2003) and (iv) 3D datasets taken from the Cattle Tracing System repository in Nohuddin et al., (2010). The evaluation objectives are to compare the statistically significant difference between the 2D banding algorithms: (i) 2D banding, (ii) BC, (iii) 2D sort and the 3D banding algorithms: (i) exact-Euclidean, (ii) exact-Manhattan and (iii) the approximate. Finally, to determine whether these algorithms are similar in significance and which differs. The parametric test considered are: (i) student t-test, (ii) Paired student t-test and (iii) ANOVA Parametric test is a statistical test that assumes parameters with known distributions about population. The parametric test was considered in the paper because it uses mean value to measure the central tendency, which relies on statistical distributions of data, and also has more chances of accuracy.

After the significant difference between the algorithms are tested, the post-hoc test is performed to determine which of these algorithms differs from the other. In this paper, Benjamin-Hochberg post hoc test proposed by Gupta in (2013) was adopted and used to determine the algorithms that differs in significance. The post-hoc test method works by initially testing the hypothesis under consideration and compute the test-statistic, which is then use to obtain a p-value from the distribution. Table 1 presents the statistics test and their respective p-values in bold fonts for (i) 2D banding with BC, (ii) 2D banding with 2D sort, (iii) BC with 2D sort, as well as (iv) 2D banding, BC with 2D sort. Similarly, Table 2 presents the statistic test and the p-values in bold fonts for (i) exact-Euclidean with exact-Manhattan, (ii) exact-Euclidean with the approximate, (iii) exact-Manhattan with the approximate as well as (iv) exact-Euclidean, exact-Manhattan with the approximate. Table 3 then presents the post-hoc test results for p = 0.05 in bold fonts respectively. Figures 1 and 2 is the bar chart representation of 2D random datasets with the number of iterations and the UCI datasets with the number of iterations. Similarly Figures 3 and 4 is the bar chart representation of the 3D random datasets with the number of iterations and the CTS datasets with number of iterations respectively.

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Table 1. 2D Parametric Statistic test results (Source: Abdullahi et al., 2021)

Parametric test	**2D Banding algorithms**	**Random datasets**		**UCI datasets**	
		Statistic test	**p-value <0.05**	**Statistic test**	**p-value < 0.05**
Student t-test	2D banding with BC	2.4056	**0.0271**	2.4839	**0.0230**
	BC with 2-D Sort	3.1716	**0.0052**	4.4385	**0.0003**
	2D banding with 2D sort	2.4251	**0.0260**	2.5310	**0.0209**
table-value		1.8330	0.05	1.8330	0.05
Paired Student t-test	2D banding with BC	3.0937	**0.0128**	2.3551	**0.0429**
	BC with 2D Sort	3.3797	**0.0081**	5.0825	**0.0006**
	2D banding with 2D sort	3.2299	**0.0103**	2.7664	**0.0218**
		2.2620	0.05	2.2620	0.05
table-value					
ANOVA	2D banding, BC with 2D sort	7.7799	**0.0021**	11.336	**0.0002**
table-value		4.4100	0.05	4.4100	0.05

Table 2. 3D Parametric Statistic test result

Parametric test	**3D banding algorithms**	**3D Random datasets**		**CTS datasets**	
		Statistic test	**p-value <0.05**	**Statistic test**	**p-value < 0.05**
Student t-test	exact-Euclidean with exact-Manhattan	2.9858	**0.0079**	2.6060	**0.0178**
	exact-Euclidean with the Approximate	6.8963	**$1.89^{e-0.6}$**	5.3033	**$4.83^{e-0.5}$**
	exact-Manhattan with the Approximate	2.7021	**0.0145**	2.9395	**0.0087**
table-value		1.8330	0.05	1.8330	0.05
Paired Student t-test	exact-Euclidean with exact-Manhattan	2.4046	**0.0395**	2.7688	**0.0217**
	exact-Euclidean with the Approximate	7,0601	**5.91^{e-05}**	4.3528	**0.0018**
	exact-Manhattan with the Approximate	2.8900	**0.0178**	2.7618	**0.0220**
table-value		2.2620	0.05	2.2620	0.05
ANOVA	Exact-Euclidean, exact-Manhattan with the Approximate	18.239	**9.72^{e-06}**	14.373	**5.616^{e-05}**
table-value		4.4100	0.05	4.4100	0.05

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Table 3. Post-hoc test result for 2D and 3D

2D banding algorithms	P=0.05	Reject (H_0)	3D banding algorithms	P=0.05	Reject (H_0)
2D banding with BC	-0.9157	True	exact-Euclidean with exact-Manhattan	-0.4437	True
2D banding with 2D sort	-2.8157	True	exact-Euclidean with the approximate	-2.1437	True
BC with 2D sort	0.0843	False	exact-Manhattan with the approximate	0.1563	False

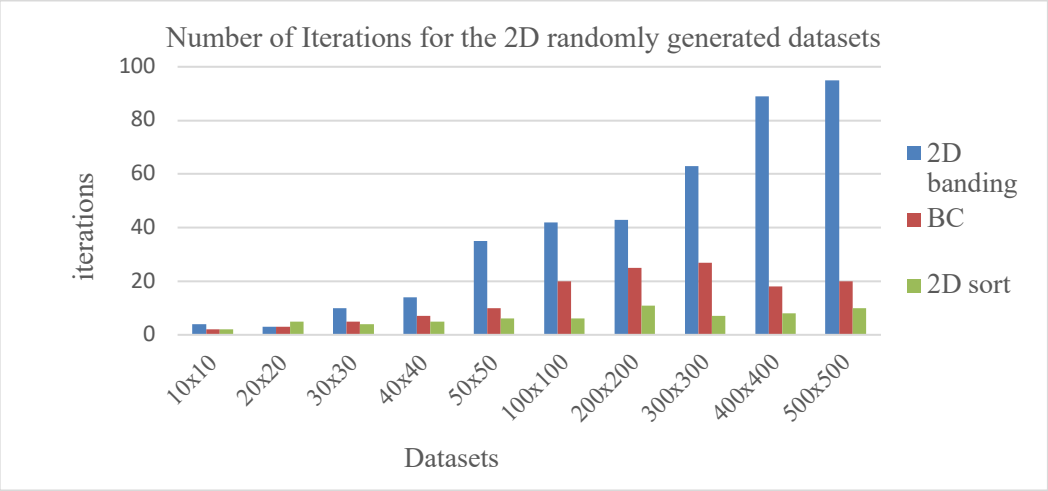

Figure 1. 2D random data with Number of iterations

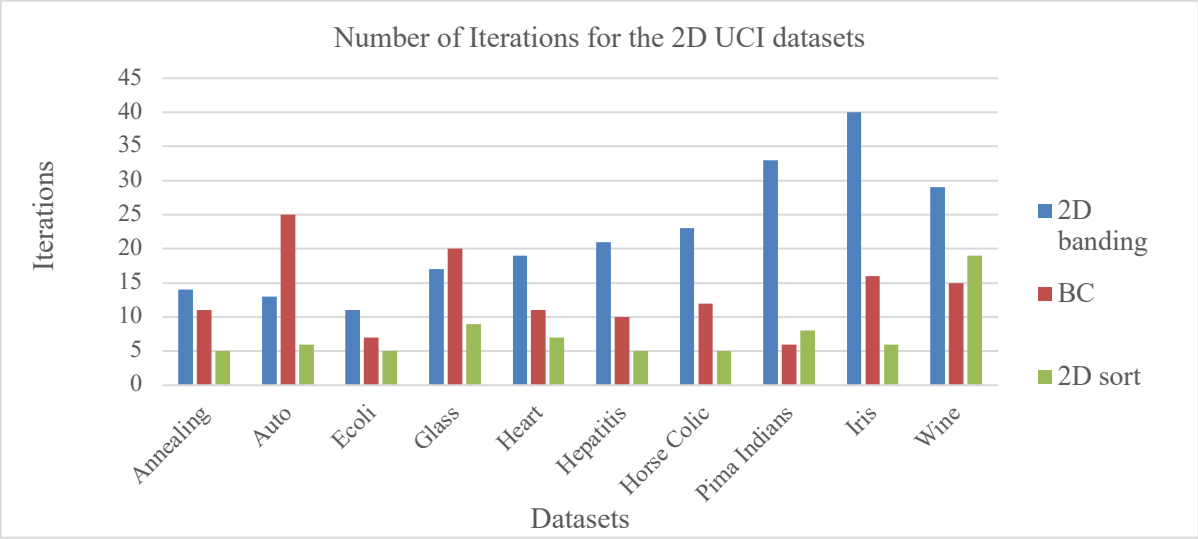

Figure 2. 2D UCI datasets with Number of iterations

Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

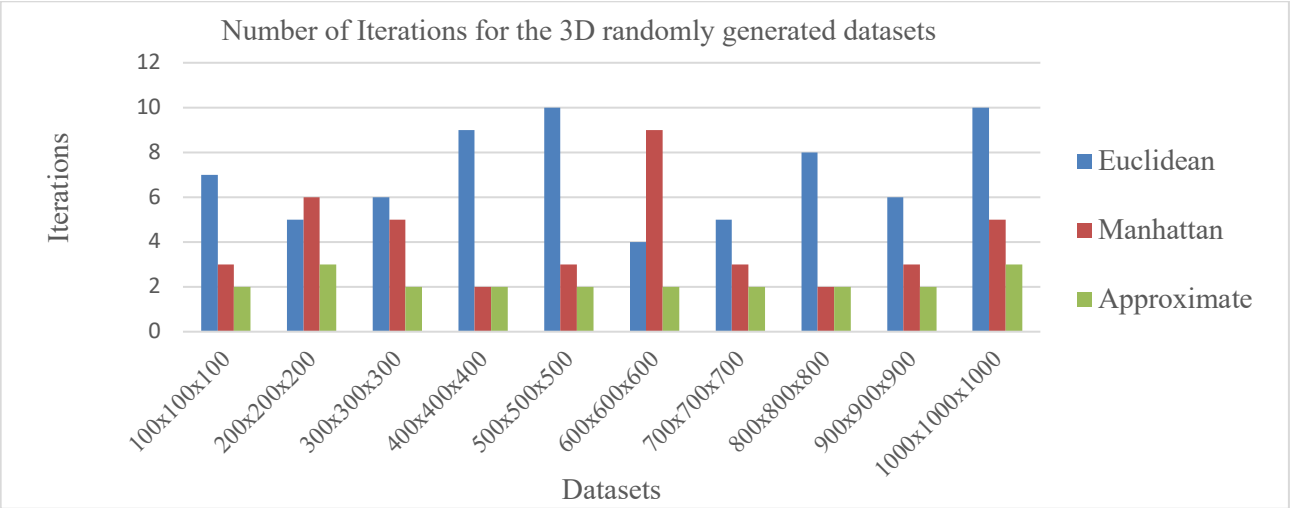

Figure 3. 3D random datasets with Number of iterations

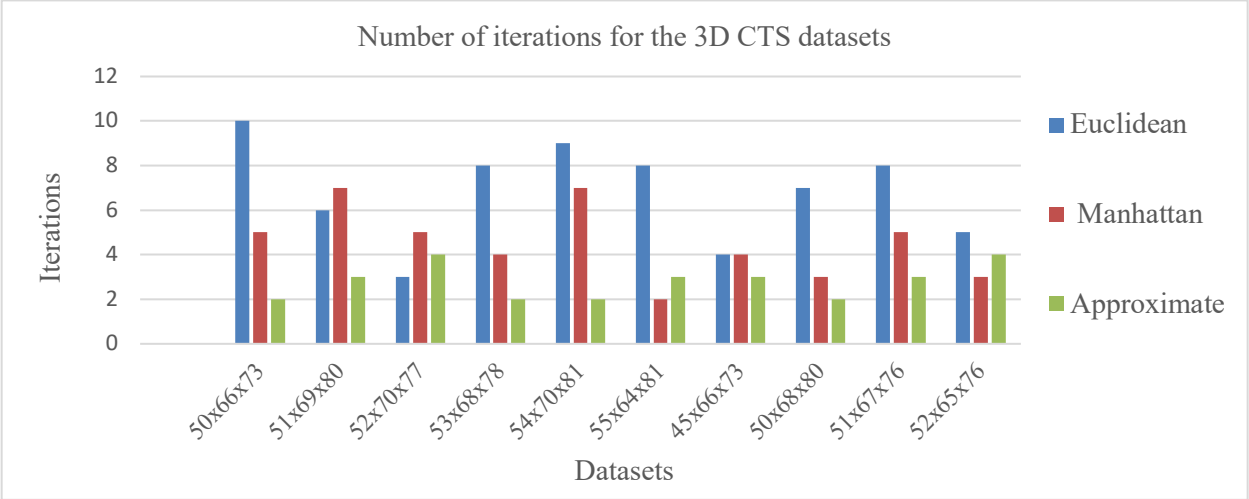

Figure 4. 3D CTS datasets with Number of iterations.

4. Discussion

The experimental results presented in Tables 1 and 2 above shows the significance of banded patterns with respect to 0.05 p value for the 2D and 3D respectively. From table 1, the result for the 2D random datasets and the UCI datasets are presented. The p value results for the 2D banding with BC, 2D banding with 2D sort and BC with 2D sort are: **0.0271, 0.0052, 0.026, 0.0230, 0.0003** and **0.0209** less than 0.05 using the student t-test we thereby reject H_0 that state there is no significant difference between the banding algorithms and accept H_A that assume significant difference between the banding algorithms. Similarly, for the 2D random datasets and the UCI datasets using paired

SLUJST
Sule Lamido University Journal of Science & Technology Vol. 4 No. 1&2 [July, 2022], pp. 147-160
https://doi.org/10.56471/slujst.v4i.278
Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

student t-test, the p value results were **0.0128, 0.0081, 0.0103, 0.0429**, **0.0006** and **0.0218** less than 0.05, as such we reject H_0 and accept H_A. Furthermore, the results for ANOVA test are **0.0021** and **0.0002** respectively still less than 0.05, thus H_0 is rejected and H_A accepted. Table 2 presents the p values results for 3D random and CTS datasets: the student t -test results are **0.0079, 1.89e-0.5, 0.0145, 0.0178, 4.83e-0.5, 0.0087** which are less than 0.05, hence H_0 is rejected, H_A accepted. The Paired student t-test results are: **0.0395, 5.91e-0.5, 0.0178, 0.0217, 0.0018 and 0.0220** still less than 0.05, thus we reject H_0 and accept H_A. Similarly, the ANOVA test results are **9.72e-0.6 and 5.616e-0.5** less than 0.05, H_0 rejected but H_A accepted. Table 3 present the post-hoc test, from the table, the experimental results for 2D banding with BC is **-0.9157** less than 0.05 H_0 rejected is True, 2D banding with 2D sort is **-2.8157** less than 0.05 H_0 rejected is True, BC with 2D sort is **0.0843** greater than 0.05 H_0 rejected is False. The results for exact-Euclidean with exact-Manhattan is **-0.4437** less than 0.05 H_0 rejected is True, exact-Euclidean with the approximate is **-2.1437** less than 0.05 H_0 rejected is True and finally, exact-Manhattan with the approximate is **0.1563** greater than 0.05 H_0 rejected is False.

5. Conclusion

The ND banding algorithms have been presented. Specifically, the 2D banding algorithms are: 2D banding, BC and 2D sort and the 3D banding algorithms are: exact-Euclidean, exact-Manhattan and the approximate respectively. The experimental results presented confirmed the following findings: (i) All the banding algorithms tested were statistically significant with p values less than 0.05 for randomly generated, UCI and the CTS datasets respectively for the three parametric test considered, (ii) The post-hoc test result shows significant difference between: (i) 2D banding with BC, (ii) 2D banding with 2D sort, (iii) exact-Euclidean with exact-Manhattan, (d) exact-Euclidean with the approximate with their p values less than 0.05. However, it fails to show a significant difference between (i) BC with 2D sort and (ii) exact-Manhattan with the approximate as their p values were greater than 0.05 respectively. For future work, the authors intend to investigate visualization of the ND banded patterns.

References

Alizadeh, F., Karp, R.M., Newberg, L. A. and Weisser, D.K. (1995). Physical Mapping of Chromosomes: A combinatorial problem in Molecular Biology. *Algorithmica*, pp. 52-76.

Cheng, Y. and Church, G.M. (2000). Biclustering of expression data. In: ISMB, vol. 8, pp. 93–103.

Myllykangas, S., Himberg, J., Bohling, T., Nagy, B., Hollman, J., & Knuttila, S. (2006). DNA copy number amplification profiling of human neoplasms. Oncogene 25, pp. 7324–7332

Atkins, J.E., Boman, E.G. and Hendrickson, B. (1998). Spectral algorithm for seriation and the consecutive one's problem. J. Computing. SIAM 28, pp. 297–310

Fortelius, M., Kai Puolamaki, M.F. & Mannila, H. (2006). Seriation in paleontological data using Markov chain monte Carlo methods. PLoS. Computing. Biol., 2, pp. 2.

Garriga, G.C., Junttila, E. and Mannila, H. (2008). Banded structures in binary matrices. In: Proceedings Knowledge Discovery in Data Mining (KDD 2008), pp.292–300.

Garriga, G.C., Junttila, E. and Mannila, H. (2011). Banded structure in binary matrices. Knowledge Information System 28(1), pp.197–226.

Makinen, E. and Siirtola, H. (2005). The barycenter heuristic and the reorderable matrix. Informatica 29, pp.357–363.

Abdullahi, F.B., Coenen, F. and Martin, R. (2014). A novel approach for identifying banded patterns in zero-one

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

data using column and row banding scores. In: Perner, P. (ed.) MLDM 2014. LNCS (LNAI), Springer, vol. 8556, pp.58–72.

Abdullahi, F.B., Coenen, F. and Martin, R. (2015). Finding banded patterns in data: the banded pattern mining algorithm. In: Madria, S., Hara, T.(eds.) DaWaK LNCS, Springer, vol. 9263, pp. 95–107.

Papadimitriou, Ch. H. (1976). The np-completeness of the bandwidth minimization problem. Computing, vol. 16, pp.263-270.

Koebe, M., & Knochel, J. (1990). On the block alignment problem. Elektronische Informationsverarbeitung und Kybernetik 26(7), pp. 377–387.

Sugiyama, K., Tagawa, S. and Toda, M. (1981). Methods for visual understanding of hierarchical system structures. IEEE Trans. Syst. Man Cybernet, vol.11, pp.109–125.

Atkinson, K.E. (2008). An Introduction to Numerical Analysis. John Wiley & Sons, New York, pp. 64

Bertin, J. (1981). Graphics and Graphic Information Processing. Walter de Gruyter, New York.

Bertin, J. (1999). Graphics and graphic information processing. In: Readings in Information Visualization, Morgan Kaufmann Publishers Inc. pp.62–65.

Liiv, I. (2010). Seriation and matrix reordering methods: an historical overview. Stat. Anal. Data Min, 3(2), pp.70–91.

Makinen, E. and Siirtola, H. (2000). Reordering the reorderable matrix as an algorithmic problem. In: Anderson, M., Cheng, P., Haarslev, V. (eds.) Diagrams 2000. LNCS (LNAI), vol. 1889, pp. 453–468. Springer, Heidelberghttps://doi.org/10. 1007/3-540-44590-0 37.

Aykanat, C., Pinar, A. and Catalyurek, U. (2004). Permuting sparse rectangular matrices into block-diagonal form. SIAM Journal of Science Computing. 25, pp.1860–1879.

Deutsch, S.B. and Martin, J.J. (1971). An ordering algorithm for analysis of data arrays. Operation Research 19(6), pp.1350–1362.

Cheng, K.Y. (1973a). Minimizing the bandwidth of sparse symmetric matrices. Computing 11(2), pp. 103–110.

Cheng, K.-Y. (1973b). Note on minimizing the bandwidth of sparse, symmetric matrices. Computing 11(1), pp. 27–30.

Harri, S & Eekki, M. (2005). Constructing and reconstructing of reorderable matrix. *Information Visualisation*, vol.4, pp.32-48

Abdullahi, F.B., Coenen, F. and Martin, R. (2016). Banded pattern mining algorithms in multi-dimensional zero-one data. In: Hameurlain, A., Kung, J., Wagner, R., Bellatreche, L., Mohania, M. (eds.) Transactions on Large-Scale Data- and Knowledge Centered Systems XXVI. LNCS, vol. 9670, pp. 1–31. Springer, Heidelberg. https://doi.org/10.1007/978-3-662-49784-5 1

Abdullahi, F.B. and Frans Coenen (2018a). Statistical Significance Testing for Banded Patterns Using Gaussian distribution. In Proceedings of 1st International Conference on Information Technology *in Education and Development (ITED),* Baze University, Abuja 25th-27th April, *pp. 209-219*.

Abdullahi, F.B. and Frans Coenen (2018b). Multi-Dimensional Banded Pattern Mining. In Proceedings of the Pacific Rim Knowledge Acquisition Workshop (PKAW2018), in Conjunction with the 15th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2018), Nanjing, China, 28th-29th, August, pp. 154 – 169.

Abdullahi, F.B (2018a). Finding Banded Patterns in 2-Dimrnsional Zero-Data Using Weighted Column and Row scores Nigerian Journal of Scientific Research (NJSR), 17(2), pp. 186-196.

Abdullahi, F.B (2018b). Three-Dimensional Banded Pattern Mining Algorithms Nigerian Journal of Scientific Research (NJSR), 17(3), pp. 292-302.

Abdullahi, F.B (2019a). Reordering 2-D Binary matrices without permutation generation Nigerian Journal of Scientific Research (NJSR), 18(3), pp. 195-203.

Abdullahi, F.B (2019b). Statistical Significant Testing For 3-Dimensional Banded Patterns. Katsina Journal of Natural and Applied Sciences (KAJONAS), 8(2), pp. 144-153.

Online ISSN: 2736-0903
Print ISSN: 2736-089X
www.slujst.com.ng

Abdullahi, F.B. and Hassan, T., (2020). Testing the Significance of 2D banded data using statistical methods Bayero Journal of Pure and Applied Sciences (BAJOPAS), 13(1), pp. 113-120.

Abdullahi, F.B. and Coenen, F., (2020). Finding banded patterns in large data set using Segmentation Bayero Journal of Pure and Applied Sciences (BAJOPAS), 13(1), pp. 90-96.

Gupta, S. P. (2013). Statistical Methods. Publisher Sultan Chand & Sons ISBN: 978-81-8054-931-1, 882-1000.

Coenen, F., (2003). LUCS-KDD data generator software. http://www.csc.liv.ac.uk/~frans/ Department of Computer Science, The University of Liverpool, UK.

Blake, C. I. and Merz, C. J., (1998). UCI Repository of Machine Learning Databases. http://www.ics.uci.edu/mlearn/MLRepository.htm

Nohuddin, P. N. E., Christley, R., Coenen, F & Setzkorn, C. (2010). Trend mining in social networks: A study using a large cattle movement database. Advances in Data mining, Applications and Theoretical Aspects LNAI 6171, pp. 464–475.

Abdullahi, F.B., Salihu, I. and Falgore, J. (2021). Determining the Significant of Banded Patterns in 2-Dimensional Binary-valued Dataset Using Parametric test. Proceedings of International Conference on Computing and Advances in Information Technology (ICCAIT), ABU, Zaria,15th-17th November, pp. 231-237