BAJOPAS Volume 13 Number 1, June, 2020 http://dx.doi.org/10.4314/bajopas.v13i1.16

Bayero Journal of Pure and Applied Sciences, 13(1): 113 - 120
Received: April, 2020
Accepted: June, 2020
ISSN 2006 – 6996

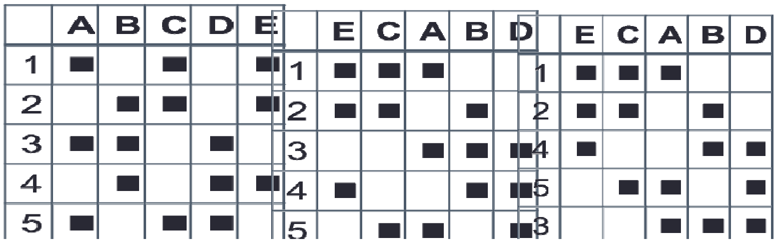
TESTING THE SIGNIFICANCE OF 2D BANDED DATA USINGSTATISTICAL METHODS

Abdullahi[1], F. B and Hassan[2], T
[1]Department of Computer Science, Ahmadu Bello University Zaria Kaduna-State
Email: zeeh429@gmail.com Phone: 08039235984
[2]Department of Statistics, University of Abuja, F.C.T
Email:hasmoten@gmail.com

ABSTRACT
This paper presents statistical methods for testing the significance of 2D banded data. A 2D zero-one dataset is said to be banded when the column (attributes) and rows (records) are arranged in such a ways that the nonzero (1s)-entries converges along the leading diagonal. The challenge with respect to banding in 2D is whether the identified bandings are significant or not. To address this issue, this paper propose statistical methods; the Student t-Distribution, Chi-square test and Normal Distribution to test the significance of bandings in 2D data. This paper also presents a 2D banding algorithm that incorporate a score mechanism: the dimension score (DS).We conduct evaluation using artificial and UCI data sets. The evaluation results shows that in the case of t-distribution and Chi-square test, the calculated statistic test exceeds the critical value in the table, while the normal distribution result shows significance of banding with regards to either one or two standard deviation (1SD or 2SD) from the mean.
Keyword: 2Dimension, Banded Data, Significance Test, Statistical Methods

INTRODUCTION

This paper presents techniques for testing the significance of 2D banding using statistical methods. Given any 2D datasets, bandings can be identified when the columns and rows are arranged to obtain a pattern along the main diagonal. The challenge however is whether the generated banding is significance or not. An example of a 2D banding is presented in Figure 1, where the columns (dim_x) and rows (dim_y) are rearranged to form a banding.



	A	B	C	D	E
1	■		■		■
2		■	■		■
3	■	■		■	
4		■		■	■
5	■		■	■	

(a) raw data

	E	C	A	B	D
1	■	■	■		
2	■	■		■	
3			■	■	■
4	■			■	■
5		■	■		■

(b) column reorder

	E	C	A	B	D
1	■	■	■		
2	■	■		■	
4	■			■	■
5		■	■		■
3			■	■	■

(c) row reorder

Figure 1: 2D banding: (a) raw data, (b) column reordered and (c) row reordered

Matrix reordering in 2D dataset has a long history. The idea of banding in 2D dataset as adopted in this paper was first proposed by (Makinen et al (2005), Mannila (2007) and Gemma et al (2008)), where a binary valued data is said to feature banding, if the columns and rows indexes can be rearrange so that the one entries are presented about the leading diagonal. Natural interpretations of banded structures include; patterns of species occurring in spatially correlated locations (Mannila et al (2007)), overlapping roles of genes in various diseases (Gemma et al (2008)) and overlapping communities in social networks (Puolamki et al (2006)). An alternative way of identifying banded pattern was later proposed in (Abdullahi et al (2014a),(2014b), (2015a), (2015b), (2016a), (2016b)) using the concept of scoring mechanism. A number of researchers have investigated the zero-one banding problem. Notable algorithms include:

1) Minimum Banded Augmentation (MBA) by Gemma et al (2008): The proposed algorithm that consider rows and columns permutations for non-zero entries in a given 2D matrices, where each column permutation is considered to be fixed whilst row permutations are considered. The algorithm commences by flipping zeros entries to ones and one entries to zero so that the rows feature a Consecutive-Ones Property (C1P).
2) Barycentric (BC), Makinen et al (2005): The approach is based on the "Barycentric" measure used to identify promising rearrangement of rows/columns. BC operates by calculating the average of location indexesof dots within each row (column).
3) Nestedness and Segmented Nestedness proposed by Mannila et al (2007): They introduced the concept of nestedness, where each row of a given 0-1 dataset is a subset of the column where the row has a one. A nested dataset is a dataset where all pairs of rows is either a superset or subset of the other. Similarly, the concept of k-nestedness dataset state that the set of columns can be partition into k parts so that each part is almost nested.

The contribution of this paper are (i) mechanism for detecting bandings in 2D datasets, (ii) techniques for determining the significance of 2D datasets using statistical methods, (iii) application of the techniques to artificially generated and real datasets.

The 2D Banding algorithm iteratively rearranges the columns and rows for a 2D data with respect to their banding score until no more (positive) changes can be made (Abdullahi (2016b), Abdullahi and Coenen (2018a) and Abdullahi and Coenen (2018b)) .

A Dimension Scores (DS) for an individual column (dim_x) and rows (dim_y) is calculated using Equation 1:

$$DS = \frac{\sum_{k=1}^{k=|C_j|} Dim - c_k - 1}{\sum_{k=1}^{k=|Ci|} Dim - k + 1} \quad (1)$$

Where Dim is the size of columns (rows), C_j the Transaction ID list for the column (row) and C_i the column (row) index at position k in C_j

Equation 2 calculates the Global Score (GS) for the entire data configuration, by adding up all the dimension scores and dividing by the size of the dimension.

$$GS = \frac{\sum_{i=1}^{i=Dim} \frac{\sum_{j=1}^{j=k_i} DS_j * \left(k_i - j + 1\right)}{\frac{k_i * \left(k_i + 1\right)}{2}}}{Dim} \quad (2)$$

The 2D Banding pseudo code is shown in Algorithm 1. The algorithm takes binary matrix M as input (Line1 and 2) and DIM (Dim and Dim). The output is datasets M rearrange to maximise GS. On each iteration, algorithm 1 loops over M to calculate Dimension score (DS) for columns (rows) (Line 6 to 8) using Equation 1. The indexes in the dimension are arrange in descending of BS (Line 7). The score for the entire configuration GS is then calculated (Line 10) using Equation 2. If the new GS is worse than the current GS, the algorithm exit otherwise M, DIM and GS are updated (Line 14). Also if no changes the algorithm exit (Line 17).

MATERIALS AND METHODS

Algorithm 1: The 2D Banding Algorithm

```
1. Input: Binary matrix M
2. DIM = (dim1 x dim2)
3. Output: M arrange to maximise GS
4. GS = 0
5. Loop
6. For all index in DIM do
7. BS = calculate column (row) scores in DIM using Equation 1
8. End For
9. M'= M arranged in descending order according BS
10. GS' = Overall GS using Equation 2
11. If(GS' < GS) Then
12. break
13. Else
14. M=M', GS = GS'
15. End If
16. End loop
17. Exit with M and GS
```

Overview of Statistical Methods

The student's t-distribution test is a parametric statistics, which is equivalent to Mann Whitney U-test in a non-parametric statistics. T-distribution used is to test the significance of the mean of a random sample in order to determine whether the sample mean from the normal distribution move away from value of the population mean. The t-distribution works with the small sample size and with unknown population standard deviation. The t-distribution test the significance sample drawn from a normal distribution deviates from the stated value of the population mean (Grupta, 2013). When using t-distribution, if the calculated, |t| value is more than the table value at any given level of significant then there is significance difference between $\overline{x}$ and μ, otherwise, there is no significant difference between $\overline{x}$ and μ. The t-Distribution is defined in Equation 3.

$$t = \frac{(\overline{x} - \mu)}{s \big/ \sqrt{n}} t(n-1) \quad (3) \quad \text{and } s = \frac{\sqrt{\sum (x - \overline{x})^2}}{n-1}$$

(4)

Where: $\overline{x}$ = sample mean

μ = population mean

n = sample size

s = sample standard deviation and t(n-1) the degree of freedom of student t-distribution. The null hypothesis is defined as H_0: $\mu = \mu_0$against any possible alternatives

(a) H_1: $\mu \neq \mu_0$

(b) H_1: $\mu > \mu_0$

(c) H_1: $\mu < \mu_0$

Where$\mu_0$is the some hypothesis value forμ.

Chi-square test is a statistical method used to test hypothesis. In chi- square test (χ^2-test), sometimes we need to consider data from population that are classified with respect to two or more different attributes. Our interest may be in the number of outputs, objects or responses which fall in various categories, The chi-square test is also called "Goodness of fit test", since it is used to test whether a significant difference exist between an observed number of subject or responses in each category and the expected number obtained under the null hypothesis. They show relationship between categorical variables. In addition, chi-square test use a single number to represent the difference between the observed values and the expected values (Grupta,2013). After calculating the chi-square values, then they are compare with the table/critical value on the chi-square table. If the calculated chi-square value is greater than the tabulated value, then there is significance difference, otherwise no difference exist. Equation 5 defines the chi-square hypothesis.

$$\chi^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right] (5)$$

Where: Subscript O_i is the observed frequency, E_i the corresponding expected frequency of the i^{th} class with v = (k-1) and v = (c-1) x (r-1) value of chi square degree of freedom (*d.f.*).Where c and r are the number of columns and rows respectively.

The normal distribution also known as Gaussian distribution is a probability function that describes how the values of variables are distributed (Jagadish (1996), Lukac and Edgar (2004), Feller (1971)). It is symmetric in nature where most observations are cluster around the central peak. The empirical rules in Normal distribution shows the percentage of data that fall within a certain number of standard deviation from the mean. A normal distribution represents bell shaped density curve defined as the mean and standard deviation. A standard normal curve comprise of a mean of zero (0) and a standard deviation of one (1). A dataset that follow a normal distribution, has 68% observations within one standard deviation from the mean, 95% observations within two standard deviation from the mean and 99.7% observations within three standard deviation from the mean (Pukelsheim, (1994), Abdullahi and Coenen (2018a)).

Thus, the standard normal distribution is shown in Equation 6.

$$Z = \left(\frac{X - \mu}{\sigma} \right) (6)$$

Where X is a random variable for normal distribution with μ as the sample mean and σ standard deviation.

RESULTS AND DISCUSSION

This section presents the evaluation and discussion of results. Two experiments were conducted, using artificially generated datasets of varying sizes with a 10% density and a real data from UCI data repository (Blake and Merz, (1998))using t-Distribution, Chi-square test and normal distribution.

All the synthetic data sets were generated using the LUCS-KDD generator (Coenen, (2003)). For the first experiment ten (10) datasets measuring: (i) (10×10), (ii) (15×10), (iii) (15×15), (iv) (20×10), (v) (20×15), (vi) (20×20), (vii) (25×10), (viii) (25×15), (ix) (25×20), (x) (25×25), were used for the t-Distribution statistic test. Similarly, ten (10) datasets measuring (i) (10×10), (ii) (20×20), (iii) (30×30), (iv) (40×40), (v) (50×50), (vi) (100×100), (vii) (200× 200), (viii) (300×300), (ix) (400×400), (x) (500×500), were used for Chi-square test. In the case of normal distribution, we generated five (5) datasets measuring: (i) (100×100), (ii) (200× 200), (iii) (300×300), (iv) (400×400), (v) (500×500), each for 50 times. For the second experiment, we used ten (10) datasets from UCI data repository. In this paper, we test the significance of 2D banding using the GS, a value defined between 0 and 1.However, in significance testing the idea was to deem a value significant or a mere occurrence of random chance. To determine whether a banding (b)from a given 2D datasets(d)is significant after a number of iterations. We let ten (10)be the expected number of iterations from d to b (equivalent to the assumed population mean (μ)). We define the null hypothesis (H_0) by assuming banding (b)does not exist in2D datasets (d) after 10 iterations and the alternative hypothesis (H_A) that banding (b) exist in 2D datasets (d) after 10 iterations. To test this assumption, we calculate the global scores (GS) values for each datasets (d) and recorded their respective iterations required to arrive at banding(b). We chose a significance level of p = 0.01 and 0.05. Using t-statistics test, we obtained the following values shown in Table1. From the table, our calculated t-statistics test result was **4.5927,** with a degree of freedom(9), and the table value at $t_{0.01}$and $t_{0.05}$are = 2.821 and 1.833 respectively. From the table, since the t-statistics test value exceeds the table values, were eject the null hypothesis (H_0.) and we conclude the hypothesis that state banding (b) exist on 2D datasets (d) after x iterations is right at 1% and 5% level of significance.

Table 1: Mean and Standard deviation calculation in t-Distribution

Data sets	GS	x	$(x - \bar{x})$	$(x - \bar{x})^2$
1	0.69	3	-3.2	10.24
2	0.74	2	-4.2	17.64
3	0.68	9	2.8	7.84
4	0.76	8	1.8	3.24
5	0.72	7	0.8	0.64
6	0.70	4	-2.2	4.84
7	0.75	9	2.8	7.84
8	0.74	5	-1.2	1.44
9	0.73	9	2.8	7.84
10	0.71	6	-0.2	0.04
		$\sum x= 62$		$\sum(x - \bar{x})^2 = 61.6$

$$\bar{x} = \frac{\sum x}{n} = \frac{62}{10} = 6.2 \quad s = \sqrt{\frac{61.6}{9}} = \sqrt{6.8444} = 2.6162$$

$$t = \frac{|6.2 - 10|}{2.6162/3.162} = \frac{3.8}{0.8274} = \mathbf{4.5927}$$

Using the chi-square test, we obtained the following value as shown in Table 2. From the table, the calculated chi-square (χ^2_{cal}) value **23.30** was obtained and compared with the exact critical value for the chi-square degree of freedom(9) at$\chi^2_{0.01}$(0.01)= 21.666 and $\chi^2_{0.05}$(0.05)= 16.919respectively. Since the calculated chi square value is more than the exact chi-square critical value in the table, the result is highly significant and we reject the null hypothesis at 1% and 5% level of significance. We now conclude that, banding does exist on 2D datasets after a number of iteration.

Table 2: Observed and Expected values calculations in Chi-Square

Data sets	Observed frequency (o)	Expected frequency (e)	Global Score (GS)	(O-E)	$(O-E)^2$	$(O-E)^2/E$
1	6	10	0.69	-4	16	1.6
2	5	10	0.68	-5	25	2.5
3	7	10	0.68	-3	9	0.9
4	3	10	0.68	-7	49	4.9
5	2	10	0.67	-8	64	6.4
6	5	10	0.65	-5	25	2.5
7	8	10	0.62	-2	4	0.4
8	8	10	0.60	-2	4	0.4
9	9	10	0.59	-1	1	0.1
10	4	10	0.59	-6	36	3.6
						$\sum(O-E)^2/E = 23.30$

χ^2_{cal} = **23.30**and chi-square degree of freedom (d.f.) = (c-1)(r-1) = (2-1)(10-1) = 9. The exact critical value from the table at d.f. (9) for the level of significance $\chi^2_{0.01}(0.01)$= 21.666 and $\chi^2_{0.05}(0.05)$= 16.919respectively.

Using normal distribution, the idea was to design normal distribution curves, from random data to which we have not apply bandings and then use it to test whether the obtained bandings are significant or a mere random occurrence, defined as the distance away from the mean. The results from Normal distribution is presented in Table 3, we define five distributions, for each data configuration. Table3, lists the GS occurrence counts for each of the data configuration without applying banding, Table 2 presents the mean (μ), standard deviation (σ), 1SD,and 2SD. Figure 2 presents the distribution curves associated with the data distributions. From the figure and tables, similar distribution curves were obtained regardless of data set size. Note that the significance of the distribution curves was that will be used to compare with the GS values obtained from similar data sets after applying banding. We generate 10 additional random data sets for each data configuration used for the distribution curves. The GS results produced after banding have been applied and compared with the normal distributions. The result presented in Table 5, for each datasets in the table, the column features (i) datasets, (ii) Average GS after banding, (iii) Average GS distance from the mean, (vi) significant with respect to 1SD and (v) significant with respect to 2SD. From the table the generated average GS after bandings were applied were located at least 1SD or 2SD of the mean. Therefore, we can state that the banding generated in the 2D datasets are statistically significant.

Table 3: Occurrence GS counts for each data configuration

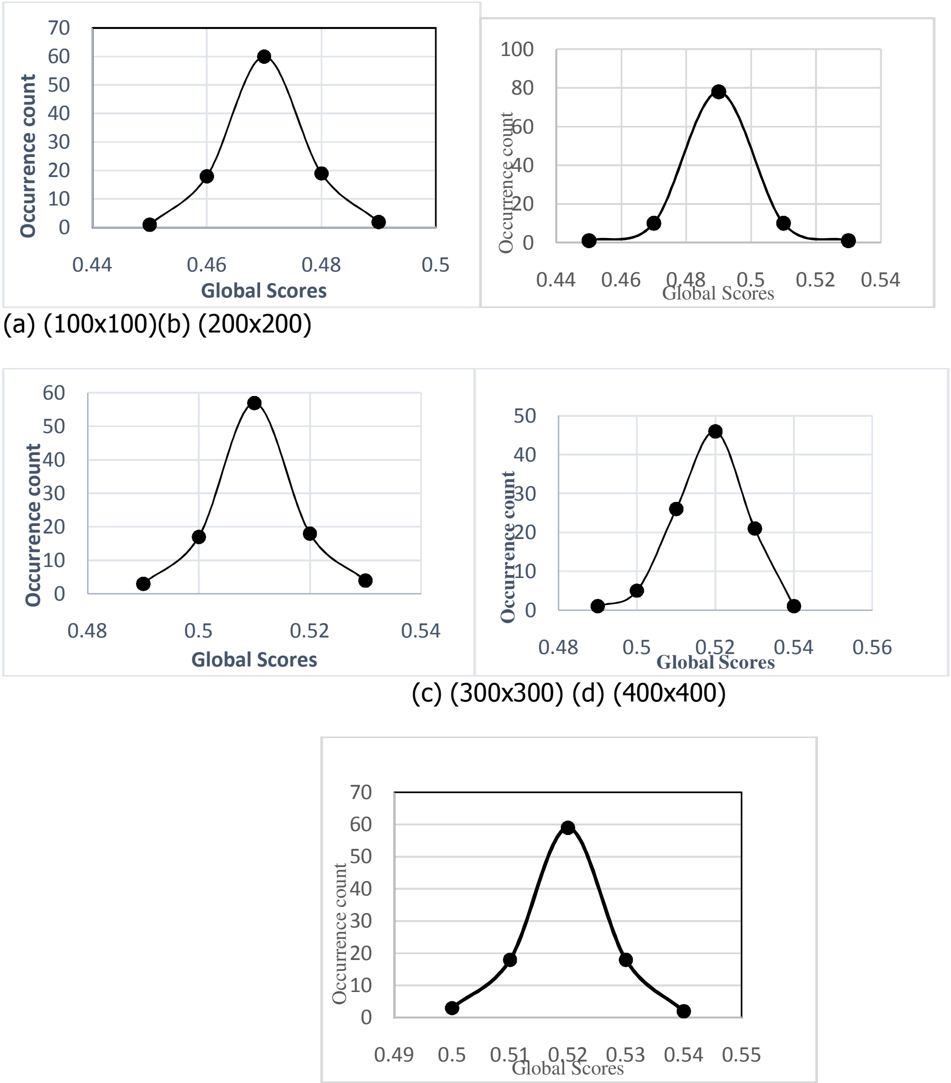
	Data sets				
GS	100x100	200x200	300x300	400x400	500x500
0.45	1	1	-	-	-
0.46	18	-	-	-	-
0.47	60	10	-	-	-
0.48	19	-	-	-	-
0.49	2	78	3	1	-
0.50	-	-	17	5	3
0.51	-	10	57	26	18
0.52	-	-	18	46	59
0.53	-	1	4	21	18
0.54	-	-	-	1	2

Table 4: Calculation of the mean and standard deviation values from Table 3

		Data sets				
		100x100	200x200	300x300	400x400	500x500
	μ	0.47	0.49	0.51	0.52	0.52
	σ	0.01	0.01	0.01	0.02	0.02
1SD	μ - σ	0.46	0.48	0.50	0.50	0.50
	μ+ σ	0.48	0.50	0.51	0.54	0.54
2SD	μ - 2σ	0.45	0.47	0.49	-	-
	μ + 2σ	0.49	0.51	0.52	-	-

Table 5:GS Normal Distribution Results

Rows x Columns	GS mean	SD	1SD (yes/no)	2SD (yes/no)
100 x 100	0.65	0.01	yes	yes
200 x 200	0.62	0.01	yes	yes
300 x 300	0.61	0.01	yes	no
400 x 400	0.60	0.02	yes	no
500 x 500	0.59	0.02	yes	no



(a) (100x100)(b) (200x200)

(c) (300x300) (d) (400x400)

(e) (500x500)

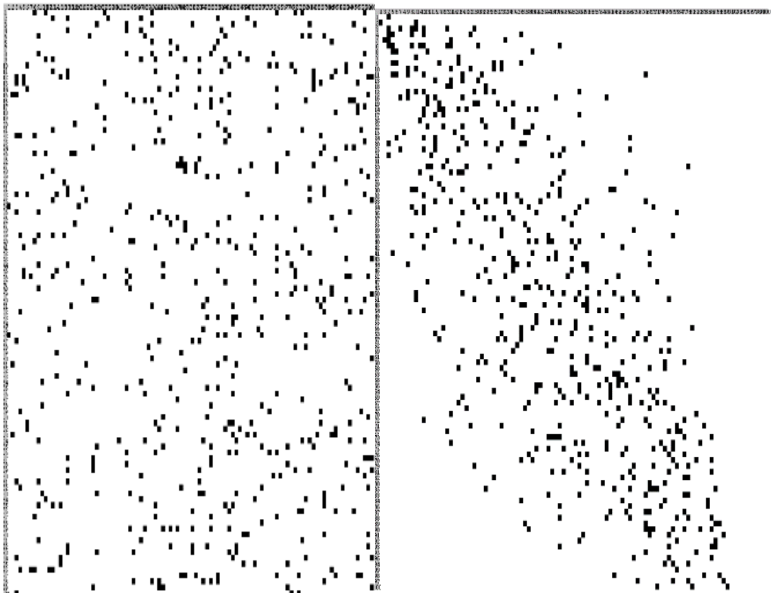
Figure 2: Normal distribution curves for Table 3

The second set of experiment on UCI datasets using the chi-square testis presented in Table 6. The table records the datasets, the number of records, number of attributes, the observed frequency, expected frequency. The difference between the observed and expected frequency and the sum of their difference. The calculated chi-square (χ^2_{cal}) test =**22.90**was compared with the exact critical value for the chi square degree of freedom (9) at $\chi^2_{0.01}$ (0.01)= 21.666and $\chi^2_{0.05}$(0.05)=16.919respectively. The result shows that the calculated chi square result is more than the exact critical value in the table at 1% and 5% level of significance. Thus the result is significant and the difference between the observed and expected frequency is significant and not a mere random chance, therefore we reject the null hypothesis (H_0) at 1% and 5% level of significance. Figure 3 shows a (100x100) datasets before and after rearranging the columns (dim_x) and rows (dim_y).

Table 6: Observed and Expected Frequencies calculation of UCI datasets

Datasets	# Recs	# Attr.	Observe frequency (o)	Expected frequency (e)	GS values	(O-E)	$(O-E)^2$	$(O-E)^2$ /E
Heart	302	52	6	10	0.80	-4	16	1.6
Annealing	898	73	8	10	0.80	-2	4	0.4
Car	1728	25	5	10	0.83	-5	25	2.5
Glass	214	48	7	10	0.79	-3	9	0.9
Hepatitis	155	56	3	10	0.84	-7	49	4.9
Horse Colic	368	85	7	10	0.81	-3	9	0.9
Iris	150	19	2	10	0.83	-8	64	6.4
Wine	178	68	6	10	0.79	-4	16	1.6
Zoo	101	42	9	10	0.86	-1	1	0.1
Lymph-ography	148	59	4	10	0.83	-6	36	3.6
								$\Sigma(O-E)^2/E$ =**22.90**

χ^2_{cal} = **22.90**and chi square degree of freedom (d.f.) = (c-1)(r-1) = (2-1)(10-1) = 9. The exact critical value from the table at d.f. (9) for $\chi^2_{0.01}$= 21.666 and $\chi^2_{0.05}$= 16.919respectively.



(a) (100x100) matrix (b) (100x100) banded matrix

Figure 3: 100x100 2D matrix (a) Before banding and (b) after banding

CONCLUSION

This paper has presented statistical methods for testing the significance of bandings in 2D datasets defined by the GS values. Two sets of experiments conducted using: (i) Artificially generated data sets and (ii) UCI data repository. The evaluation results presented shows the significance of 2D banding using statistical methods. In the case of t-distribution, the calculated statistic test exceeds the critical value in the table at 1% and 5% level of significance. While in the case of Chi-square test, we compared the calculated chi square result with the exact critical values from the chi square table, and the χ^2_{cal} value obtained was more than the exact critical value in the table at $\chi^2_{0.01}$ (1%)and$\chi^2_{0.05}$(5%) level of significance respectively. Similarly, in the case of normal distribution test, the results show significance of banding with respect to either one or two standard deviation (1SD or 2SD) from the mean. However, the limitation of the normal distribution approach was that the normal distribution curve for the datasets must have been derive in each case. The experiments has clearly shown the usefulness of the proposed statistical methods in testing the significant of 2D bandings. For future research, the authors intends to investigate the statistical significance testing for 3D bandings.

REFERENCES

Mannila, H and Terzi,E. (2007). Nestedness and Segmented Nestedness, Proceedings of the 13h ACM SIGKDD international conference on knowledge discovery and data mining, New York, NY, USA, 2007, 2007, pp. 480–489.

Puolamki K, Fortelius M, and Mannila H (2006). Seriation in Paleontological data using Markov Chain Monte Carlo methods. PLoS Computational Biology, 2, 2006.

Abdullahi, F.B., Coenen, F., & Martin, R. (2014). A Novel Approach for Identifying Banded Patterns in Zero-One Data using column and row banding scores," Proceedings of International Conference on Machine Learning and Data Mining in Pattern Recognition. Springer, 2014a, pp. 58–72.

Abdullahi, F.B., Coenen, F., & Martin, R. (2015). Finding Banded Patterns in Big Data using Sampling, in 2015 IEEE International Conference on Big Data (Big Data).IEEE, 2015a, pp. 2233–2242.

Abdullahi, F.B., Coenen, F., & Martin, R. (2014). A Scalable Algorithm forBanded Pattern Mining in Multi-dimensional Zero-One Data, Proceedings of International Conference on Data Warehousing and Knowledge Discovery. Springer, 2014b, pp. 345–356.

Abdullahi, F.B., Coenen, F., & Martin, R. (2015). Finding Banded Pattern in Data: The Banded Pattern Mining Algorithm, Proceedings of International Conference on Big Data Analytics and Knowledge Discovery. Springer, 2015b, pp. 95–107.

Abdullahi, F.B., Coenen, F., & Martin, R. (2016). Banded Pattern Mining Algorithms in Multi-dimensional Zero-One Data, in Transactions on Large-Scale Data and Knowledge-Cantered Systems XXVI. Springer, 2016a, pp. 1–31.

Abdullahi, F.B. (2016). Banded Pattern Mining for N-Dimensional Zero-One data PhD Thesis, University of Liverpool, United Kingdom, 2016b.

Abdullahi, F.B., and Coenen, F. (2018). Multi-Dimensional Banded Pattern Mining in Proceedings of 15th Pacific Rim Knowledge Acquisition Workshop (PKAW2018),Springer LNAI 11016, pages. 154–169, 2018a.

Abdullahi, F.B, and Coenen, F.(2018).Statistical Significance Testing for Banded Patterns Using Gaussian distribution Proceedings of 1st International Conference onInformationTechnology *in* Education and Development (ITED), 2018b, pages 209-219

Blake, C. I. and Merz, C. J. (1998). UCI repository of machine learning databases. http: //www.ics.uci.edu/mlearn/MLRepository.htm, 1998.

Coenen, F., (2003), LUCS-KDD Data generator Software. Department of Computer Science, The University of Liverpool UK.http://www.liv.ac.uk/ frans/KDD/Software/LUCS KDD DataGen Generator.html, 2003.

Gemma, G .C. Juntilla, E., & Manilla, H. (2008). Banded Structures in Binary Matrices, in Proceedings Knowledge Discovery in Data Mining(KDD08), 2008, pp.292–300.

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

Feller, W. (1971). Introduction of Probability Theory and Its Applications New York Vol 2 3rd ed pp.45 1971.

Makinen, E and Siirtola, H. (2005). The Barycenter Heuristic and the Reorderable Matrix, Informatica, vol. 29, pp. 357–363, 2005.

Jagadish, P. K, Read and Campbell, B. (1996). Handbook on Normal Distribution (2nd ed) CRC Press 1996.

Pukelsheim, P. (1994). The Three-sigma Rule. American Statistician 34: 477-495, 1994

Lukac Eugene and King Edgar., (2004). A Property of Normal Distribution. The Annals of Mathematics, 11, 2004.