Statistical Significance Testing for Banded Patterns Using Gaussian distribution

Fatimah Binta Abdullahi
Department of Computer Science
Ahmadu Bello University
Zaria Nigeria
zeeh429@gmail.com

Frans Coenen
Department of Computer Science
University of Liverpool
United Kingdom
coenen@liverpool.ac.uk

Abstract — *This paper, presents a technique for determining the statistical significance of banded patterns in 2-Dimensional (2-D) zero-one data. Given a 2D data set, some form of banding can be identified by rearranging the columns and rows, but the question is whether these bandings are significant or not. The approach advocated in this paper is to use Gaussian distribution mechanism on randomly generated 2D datasets to which banding had not been applied which is then used to established whether the generated banding is significant or not in terms of the distance from the mean. In this paper, a column and row scoring mechanism incorporated into the 2D Banded Pattern Mining (BPM) algorithm is presented. Evaluations were conducted using two sets of experiments: experiments using a collection of data sets, using a static dot density of 10% and experiments using collection of data sets using ranges of dot density values from 10% to 50% increasing in steps of 10%. The evaluation results presented indicate the significance of bandings with respect to either one standard deviation (1SD) or two standard deviation (2SD). The results also show that it is possible to generate generic normal distribution curves using range of dot density values.*

KEYWORDS
Banded Patterns, Zero-One data,
Statistical Significance, Gaussian distribution

I INTRODUCTION

The work presented in this paper is concerned with technique for determining the statistical significance of banded patterns in 2-Dimensional (2-D) zero-one data, using the Gaussian distribution. The objective of the paper is that, given any 2D data set, some form of banding can be identified by rearranging the columns and rows, but the question is whether the obtained banding is significant or not. A trivial example of 2D banding is given in Figure 1. The figure shows a 2D banding example with columns and rows rearranged to reveal banding.

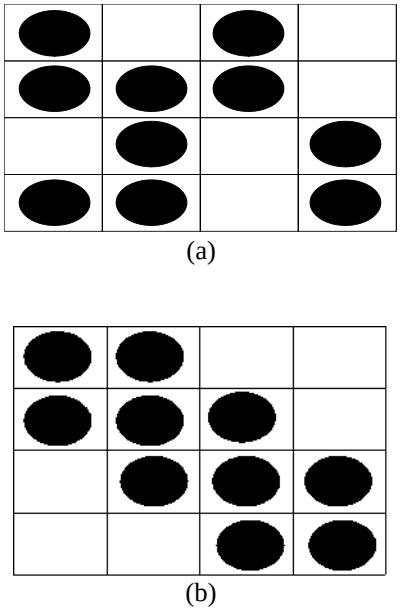


Fig 1: 2D banding example (a) original matrix and (b) original matrix with the columns and rows reordered to reveal a banding

The rest of the chapter is organised as follows; Section 2 presents related work. Section 3 presents the methodology of the paper. Section 4 presents evaluation and result. Finally, Section 5 concludes the paper.

II RELATED WORK

The identification of banding in 2-D zero-one data has a long history, although the ideas of 2D Banded Pattern Mining (2D-BPM) as adopted in this paper was first proposed in [1], [2], [3], [17] and [18]. Existing work on identifying bandings in zero-one data as proposed in [19] and [20]

concentrated on the generation and testing of permutations, whilst [21] used barycentric values to identify bandings. The main issue with the identification of banded patterns in this manner is the large number of permutations to be

considered makes the identification of banding in 2D data a resource intensive enterprise. To address this issue according to [1], [2], [3], [17] and [18], an alternative solution to the permutation generation and test approach that does not require the generation of permutations but instead operates using the concept of a Banding Score (BS). The proposed solution is to iteratively reorder the items in each column and row according to their individual BS until a "best" banding is arrived at defined in terms of a Global Banding Score (GBS).

III. METHODOLOGY

This paper presents mechanism for determining the statistical significant of bandings. The basic idea presented is that if we had "n" randomly generated data sets, all featuring the same dimensions and approximately the same density, each of these data sets would have a Global Banding Score (GBS) value associated with it. The assumption here is that these GBS values would be distributed following the normal (Gaussian) distribution. However, it is expected that the GBS value generated after banding had been applied would be located away from the median of the distribution by a distance of at least one standard deviation. Note that the normal (Gaussian) distribution mechanism was selected in this paper because it was assumed that the data sets to which banding will be applied are likely to follow this distribution. Further reason was that the Gaussian distribution is easy to work with and many statistical tests can be derived from it. This paper explores this idea and demonstrates that normal distributions can be usefully employed to establish the statistical significance of banding.

A. *Overview of statistical significance testing*

The normal distribution is concerned with the operation of a continuous probability distribution [4, 6, 7, 12 and 14] that represents a real-valued random variable. The normal distribution is described by the probability density function Φ(x) given in Equation 1, where x is an observation of some kind. Note that the factor $\sqrt{2\pi}$ ensures the total area under curve Φ(x) is one [4, 5, 7 and 8] and that the distribution has a unit variance (unit standard deviation). $\phi(x) = \frac{e^{1/2}x^2}{\sqrt{2\pi}}$. Though, authors differ on which normal distribution should be called the "standard" one, Gauss [16] defined standard normal distribution as having variance $\sigma^2 = 1/2$ and a probability density function of: $\phi(x) = \frac{e^{x^2}}{\sqrt{x}}$. While Stigler [12, 13] define standard normal distribution as having a variance $\sigma^2 = 1/2$ and a probability density function of: Using the probability density function Φ(x) given above, for a range of values of x, a "bell curve" [11] describes a mean μ, a standard deviation σ and a variance σ2. Figure 2, taken from [10] presents many examples of bell curves associated with the normal (or Gaussian) distribution. In the figure the x-axis indicates a range of values for the variable x

from -5 to 5 and the y-axis represents the frequency or probability of the occurrence count. The red curve in the figure is the standard normal curve with (μ = 0, σ = 1), the blue and green curves represents the normal curves with (μ= 0, σ = 0:2) and (μ = -2, σ = 0:5), whilst the purple curve is a normal curve with (μ = 0, σ = 5:0). Thus, the normal distribution is symmetric about its mean μ. The normal distribution value tends to zero when the value x lies more than a few standard deviations away from the mean.

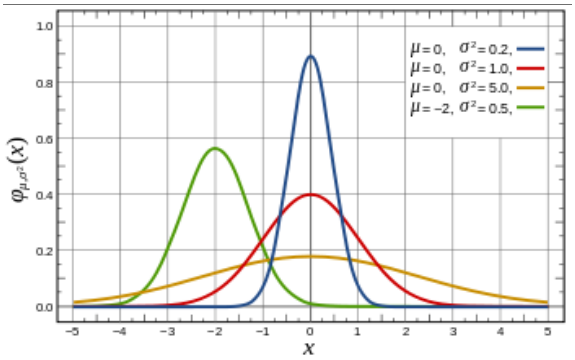


Fig 2: Gaussian or Normal Probability Curve [10]

In the normal distribution, the three-sigma rule is used to show the percentage of values that lie within a band around the mean width of "one", "two" and "three" standard deviations; this means that; 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations from the mean. In other words, for the normal distribution, values of less than one standard deviation away from the mean accounts for 68.27% of the values, two standard deviation from the mean accounts for 95.45% of the values and three standard deviation accounts for 99.73% of the values. Figure 3 taken from [9], illustrates the three-sigma rule for the normal distribution. With respect to the work presented in this paper, the normal (Gaussian) distribution was used to test the statistical significance of bandings.

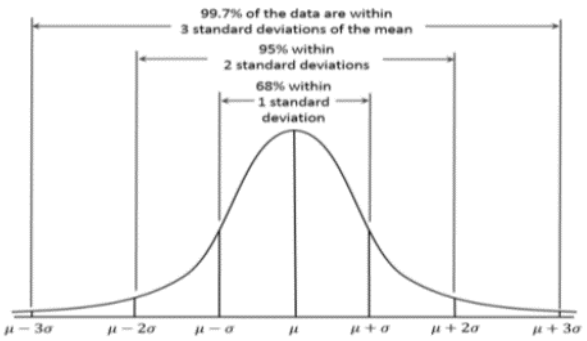


Fig 3: three-sigma rule for the normal distribution [9]

B. The Column and Row Score Mechanism

The fundamental idea presented in [1], [2], [3], [17] and [18] was that the "bandedness" of a 2D dataset can be expressed in terms of a Global Banding Score (GBS), a number between "0" and "1". Depending on how the GBS is calculated, "1" indicates a perfect banding and "0" the other extreme the "most imperfect" banding. The GBS is calculated by summing and normalizing individual column (CS) and row scores (RS) associated with individual columns and rows in each 2D matrix (data set). Individual CS (RS) is expressed as a number between 0 and 1. The idea is that given a set of columns (rows) scores, these can be used to reorder the columns (rows) to reveal a banding. Once the rows and columns have been reordered the individual CS (RS) values will need to be recalculated, as it is likely that they will have changed because of the reordering and a new GBS generated. The expectation is that the new GBS will be better than the initial GBS calculated prior to the reordering. It was anticipated that the reordering would have to be undertaken over several iterations until the GBS value "stabilised". However, the important point to note is that the time complexity of this approach is linear according to the number of columns/rows.

C. Column and Row Score Calculation

The fundamental idea underpinning the 2D-BPM algorithms considered in [1], [2],[3],[17] and [18] was the concept of a banding score (column and row score mechanism). In 2-D data space, DIM = {Dim_x, Dim_y}, this implies calculating the banding sores for Dim_x with respect to Dim_y. Given a set of dots ($Dots_{xj}$) associated with index j in Dim_x and index j in Dim_y, a CS and RS can be calculated using Equation 1, 2.

$$CS_{xj} = \frac{\sum_{n=1}^{n=|C_y|} C_n}{\sum_{n=1}^{n=|C_y|} k_2 - n + 1} \quad (1)$$

$$RS_{yj} = \frac{\sum_{n=1}^{n=|C_x|} C_n}{\sum_{n=1}^{n=|C_x|} k_1 - n + 1} \quad (2)$$

where C_y is the set of y-coordinates associated with $Dots_{xj}$ ($|Dots_{xj}| \equiv |C_y|$). Similarly, where C_x is the set of x-coordinates associated with $Dots_{yj}$ ($|Dots_{yj}| \equiv |C_x|$). And k_1 is the maximum size for dimension Dim_x and k_2 the maximum size for dimension Dim_y. Note also that k_1 and k_2 equate to the maximum indexes for Dim_x and Dim_y respectively.

$$CS_{x1} = \frac{1+2}{3+4} = \frac{3}{7} = 0.428$$

$$\frac{(0.428x4)+(0.667x3)+1.000x2)+(1.000x1)}{4(4+1)} =$$

Following on from this CS_{x2} = 0.667, CS_{x3} = 1.000 and CS_{x4} = 1.000. The same scores would be obtained for the y dimension in Fig 1 because the banding is symmetrical about the leading diagonal. The idea is then to reveal a banding by arranging the indexes, in ascending order from the origin, according to their associated column scores. The column and row score mechanism can also be used to calculate a Global Banding Score (GBS) for an entire banding configuration using Equation 2 where GBS_x and GBS_y are the GBS for dimension x and y.

$$GBS = \frac{GBS_x + GBS_y}{2} \quad (3)$$

The value for GBS_x and GBS_y are then calculated using Equation 3 and 4.

$$GBS_x = \frac{\sum_{j=1}^{j=k_1} CS_{xj} \times (k_1 - j + 1)}{\frac{k_1(k_1+1)}{2}} \quad (4)$$

$$GBS_y = \frac{\sum_{j=1}^{j=k_2} RS_{yj} \times (k_2 - j + 1)}{\frac{k_2(k_2+1)}{2}} \quad (5)$$

Thus, returning to the configuration given in Figure 1, using Equation 3, the value for GBS_x will be calculated as follow:

$$\frac{1.712 + 2.001 + 2.000 + 1.000}{10} = \frac{6.713}{10} = 0.671$$

Because the configuration is symmetrical about the leading diagonal GBS_y will also equal 0:671. The GBS value for the entire configuration will then, using Equation 3, be:

$$GBS = \frac{0.671 + 0.671}{2} = 0.671$$

over the data space. On each iteration, the column score for each index j in Dim_x is calculated (Line 7). The index in Dim_x is then rearranged in ascending order of the CS_{xi} to produce D′ (Line 9). The GBS for the x-dimension is calculated using Equation 3 (line 10). The same process is then followed for Dim_y so as to produce D′′ (lines 11-14). The GBS_y value calculated for the y-dimension is calculated using Equation 5 (line 15). A new GBS value is then calculated using Equations 3 (line 16). Then, if the new GBS is worse than the current GBS (GBS_{sofar}) we exit with the previously stored configuration and GBS (lines 17-18). Otherwise D, Dim_x, Dim_y and the value for GBS are updated (lines 20) and the algorithm repeats. If no changes (line 23) are made the algorithm also exit.

Algorithm 1: The 2D-BPM Algorithm

1. **Input: D** = Zero-One data matrix measuring k_1 x k_2
2. $Dim_x = \{0, 1, \ldots, k_1\}$, $Dim_y = \{0, 1, \ldots, k_2\}$
3. **Output**: Rearranged data space D that minimise GBS
4. $GBS_{sofar} = 1.0$
5. **Loop**
6. **for** all index in Dim_x **do**
7. CS_{index} = column score for $index_j$ in Dim_x is calculated using Equation 1
8. **end for**
9. D′ = The data set D rearranged according to column score for Dim_x
10. GBS_x = Global banding score for Dim'_x is calculated using Equation 4
11. **for** all index in Dim_y **do**
12. RS_{index} = row score for $index_j$ in Dim_y is calculated using Equation 2
13. **end for**
14. D′′ = The data set D′ rearranged according to the row score for Dim_y
15. GBS_y = Global banding score for Dim'_y using Equation 5
16. GBS_{new} = Overall Global Banding Score calculated using Equation 3
17. **if** ($GBS_{new} \geq GBS_{sofar}$) **then**
18. break
19. **else**
20. D = D′′, $Dim_x = Dim_x'$, $Dim_y = Dim_y'$, $GBS_{sofar} = GBS_{new}$
21. **end if**
22. **end loop**
23. Exit with D and GBS

D. The 2D Banded Pattern Mining (2D-BPM) Algorithm

This section presents the 2D-BPM algorithm proposed in [1],[2][3],[17] and [18] for identifying bandings in 2D data sets. The algorithm operates by iteratively rearranging the column and row indexes until the GBS is minimised.

The pseudo code for the 2D-BPM algorithm is presented in Algorithm 1. The inputs (lines 1 to 2) are: (i) a binary data set D and (ii) k_1 ($Dim_x = \{0, 1, \ldots, k_1\}$), k_2 ($Dim_y = \{0, 1, \ldots, k_2\}$. The output is a rearranged data set D that minimises the GBS value. The algorithm iteratively loops

V EVALUATION AND RESULTS

In this paper a 2D-BPM algorithm has been considered. The reported evaluations indicated that in all cases a better GBS value was produced after banding than existed prior to banding. The question remained as to whether the detected bandings were indeed statistically significant or not. This section considers a process that can determine whether an obtained banding is statistically significant or not. The idea was to create a bank of normal distribution curves, from randomly generated data sets to which banding had not been applied, which could then be used to establish whether a generated banding was significant or not in terms of distance from the mean. To demonstrate this approach, two sets of experiments were conducted, each involving a collection of 1000 data sets grouped into batches of 100 according to row/column size. More specifically the row and column dimensions used were:

i. 100 x 100,
ii. 141 x 141
iii. 173 x 173,
iv. 200 x 200,
v. 224 x 224,
vi. 245 x 245,
vii. 265 x 265,
viii. 285 x 285,

ix. 300 x 300 and
x. 316 x 316.

distinction between the two sets of experiments was the density used:

1) Static Dot Density value: Experiments using a collection of data sets, using a static density of 10%.
2) Range of Dot Density values: Experiments using density values ranged from 10% to 50%
3) increasing in steps of 10% (each data set size featured five different dot densities distributed evenly).

The rational for the second set of experiments was to determine the more general applicability of the approach. The data sets were generated using the LUCS-KDD generator [15]. The results were then used to define ten normal distributions, one for each data set configuration. The normal distributions associated with the first set of experiments are discussed in further detail in Subsection 6.1 while that associated with the second set is discussed in Subsection 6.2. The banded pattern significance testing is discussed in further detail in Subsection 6.3.

The effect was to have data sets ranging from 10,000 to 100,000 locations in steps of 1,000. The

*A. **Static Dot Density***

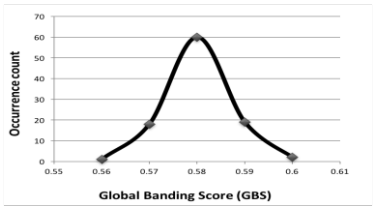
In this subsection, the experimental result using a static dot density of 10% is presented. Table 1 lists the natural GBS occurrence counts for each data set configuration (without banding), whilst Table 2 lists the accompanying μ, σ and one and two standard deviation limits. Fig 4 shows the normal distribution curves associated with the distributions (and the information in Tables 1 and 2). Inspection of the figure (and tables) indicates that similar distribution curves result regardless of data set size. The significance of these distribution curves is that they can now be used to compare GBS values obtained from similar data sets (same size and density) after banding has taken place. This is illustrated in the following section.

Table 1: Mean and Standard Deviation values extracted from data presented in Table 1 (static dot density)

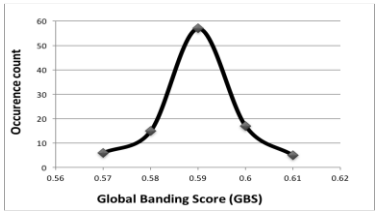
	Data sets									
GBS	**100 x 100**	**141 x 141**	**173 x 173**	**200 x 200**	**224 x 224**	**245 x 245**	**265 x 265**	**283 x 283**	**300 x 300**	**316 x 316**
0.56	1	-	-	-	-	-	-	-	-	-
0.57	18	6	1	1	-	-	-	-	-	-
0.58	60	15	5	10	-	-	-	-	-	-
0.59	19	57	26	78	1	1	-	1	-	-
0.60	2	17	46	10	20	15	2	14	5	3
0.61	-	5	21	1	58	65	18	35	20	18
0.62	-	-	1	-	20	18	61	34	53	59
0.63	-	-	-	-	1	1	17	15	18	18
0.64	-	-	-	-	-	-	2	1	4	2
Total	100	100	100	100	100	100	100	100	100	100

Table 2 Mean and Standard Deviation values extracted from (ranged of dot density)

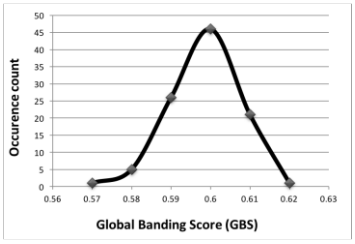
		Data sets									
		100 x 100	**141 x 141**	**173 x 173**	**200 x 200**	**224 x 224**	**245 x 245**	**265 x 265**	**283 x 283**	**300 x 300**	**316 x 316**
	μ	0.58	0.59	0.60	0.61	0.61	0.61	0.62	0.615	0.62	0.62
	σ	0.01	0.01	0.02	0.01	0.02	0.01	0.01	0.02	0.01	0.01
1SD	**μ-σ**	0.57	0.58	0.58	0.60	0.59	0.60	0.61	0.595	0.61	0.61
	μ+σ	0.59	0.60	0.62	0.62	0.63	0.62	0.63	0.635	0.63	0.63
2SD	**μ-2σ**	0.56	0.57	-	0.59	-	0.59	0.60	-	0.60	0.60
	μ+2σ	0.60	0.61	-	0.63	-	0.63	0.64	-	0.64	0.64



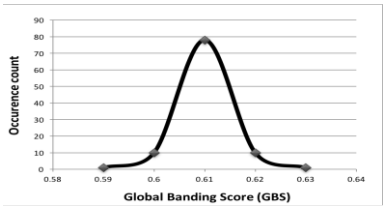
(a) 100 x 100



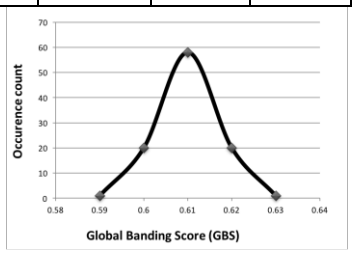
(b) 141 x 141



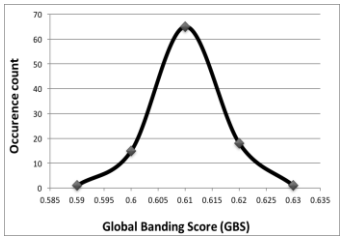
(c) 200 x 200



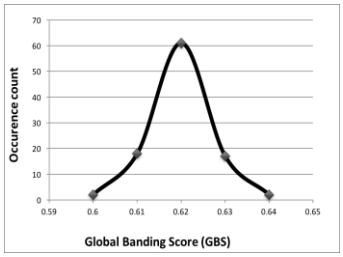
(d) 173 x 173



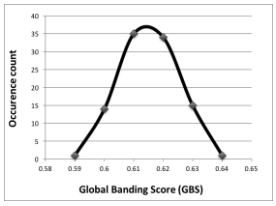
(e) 224 x 224



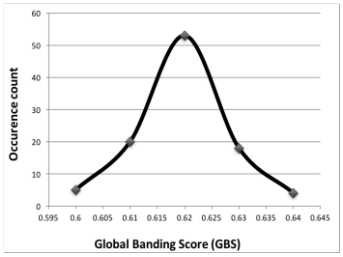
(f) 245 x 245



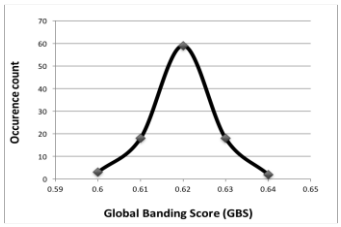
(g) 265 x 265



(h) 283 x 283



(i) 300 x 300



(j) 316 x 316

Fig 4 Standard distribution curves for data presented in Table 1 (static dot density)

B. Range of Dot Density

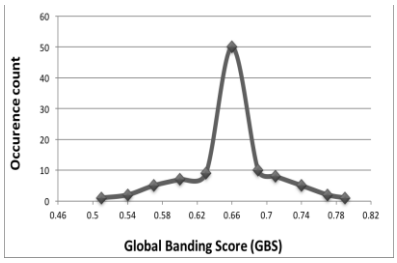
This subsection considers the normal distributions that result with respect to data set generated using a range of dot density values instead of a static dot density value. More specifically density values ranging from 10% to 50% increasing in steps of 10%. In the same manner, as in the previous subsection. Table 3 lists the natural GBS occurrence counts for each data set configuration (without banding), whilst Table 4 lists the accompanying μ, and one and two standard deviation limits. The associated normal distribution curves are given in Fig 5. Inspection of the figure indicates that similar distributions are produced; however, comparison with the distribution curves presented previously in Figure 4 indicates a marked difference in shape indicating that it is not a "one size fits all" situation. The significance of the distribution curves, as already noted was that they can be used to compare the GBS values obtained from data sets (same size but different densities) after banding has taken place to determine if the resulting banding is statistically significant or not.

Table 3 (ranged of dot density)

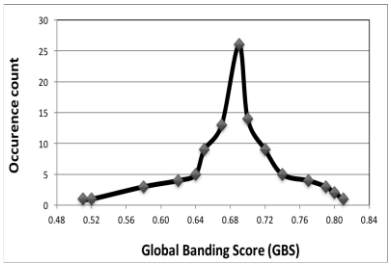
	Data sets									
GBS	**100 x 100**	**141 x 141**	**173 x 173**	**200 x 200**	**224 x 224**	**245 x 245**	**265 x 265**	**283 x 283**	**300 x 300**	**316 x 316**
0.51	1	1	-	-	-	-	-	-	-	-
0.52	-	1	1	1	1	2	-	1	-	-
0.53	-	-	-	-	-	3	1	-	1	1
0.54	2	-	5	3	3	-	1	3	-	-
0.55	-	-	-	-	-	5	-	-	3	5
0.56	-	-	6	-	5	-	2	-	-	-
0.57	5	-	-	-	-	7	-	-	7	7
0.58	-	3	8	4	7	-	-	5	-	-
0.59	-	-	-	-	-	-	-	-	9	-
0.60	7	-	14	5	10	9	4	7	-	10
0.61	-	-	-	-	-	-	-	-	-	-
0.62	-	4	-	9	12	12	10	-	-	-
0.63	9	-	-	-	-	-	-	10	14	12
0.64	-	5	-	-	25	27	15	-	-	-
0.65	-	9	27	-	-	-	-	-	-	-
0.66	50	-	-	15	11	-	-	-	-	32
0.67	-	13	-	-	-	-	-	12	31	-
0.68	-	-	-	23	10	10	35	-	-	-
0.69	10	26	-	-	-	-	-	26	-	-
0.70	-	14	15	15	-	-	-	-	-	-
0.71	8	-	-	-	-	-	-	-	-	-
0.72	-	9	9	10	7	9	14	11	15	11
0.73	-	-	-	-	-	-	-	-	-	-
0.74	5	5	8	6	5	7	9	10	9	9
0.75	-	-	-	-	-	-	-	-	-	-
0.76	-	-	-	5	-	-	5	-	-	-
0.77	2	4	-	-	-	-	-	-	-	-
0.78	-	-	6	3	3	5	2	7	7	7
0.79	1	3	-	1	-	-	-	5	-	-
0.80	-	2	1	-	1	3	1	2	3	5
0.81	-	1	-	-	-	1	1	1	1	1
Total	100	100	100	100	100	100	100	100	100	100

Table 4 Mean and Standard Deviation values extracted from (range of dot density)

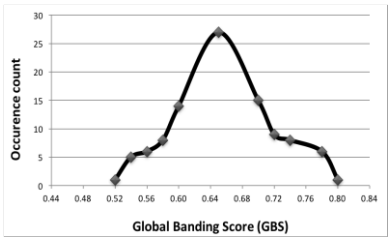
		Data sets									
		100 x 100	**141 x 141**	**173 x 173**	**200 x 200**	**224 x 224**	**245 x 245**	**265 x 265**	**283 x 283**	**300 x 300**	**316 x 316**
	μ	0.66	0.68	0.65	0.68	0.64	0.64	0.68	0.68	0.67	0.66
	σ	0.06	0.07	0.06	0.09	0.07	0.09	0.07	0.09	0.07	0.09
1SD	**μ- σ**	0.60	0.61	0.57	0.59	0.57	0.55	0.61	0.59	0.60	0.57
	μ+ σ	0.72	0.75	0.71	0.77	0.71	0.73	0.75	0.77	0.74	0.75
2SD	**μ-2σ**	0.54	-	0.53	-	0.50	-	0.54	-	0.53	-
	μ+2σ	0.78	-	0.77	-	0.78	-	0.82	-	0.81	-



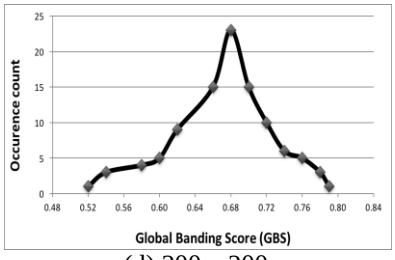
(a) 100 x 100



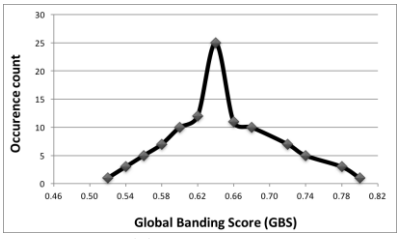
(b) 141 x 141



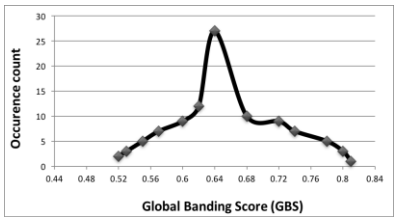
(c) 173 x 173



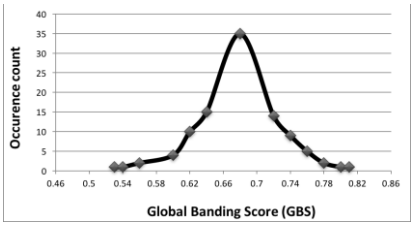
(d) 200 x 200



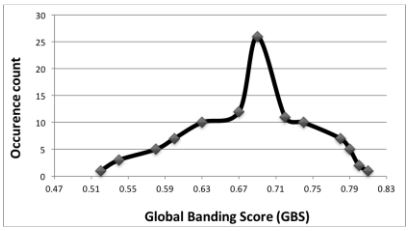
(e) 224 x 224



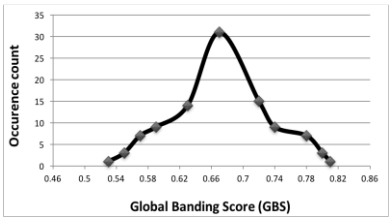
(f) 283 x 283



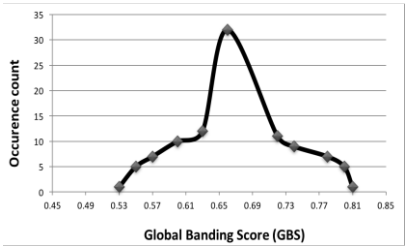
(g) 300 x 300



(h) 245 x 245



(i)265 x 265



(j) 316 x 316

Fig 5 Standard distribution curves for data presented in Table 3 (ranged of dot density)

C. Banded Pattern Significance Testing

To evaluate the proposed approach to significance testing of generated banded patterns two set of experiments were

conducted using: (i) a static dot density value of 10% and (ii) a range of dot density values (the same range as used to generate the distribution curves described above). In each case, several additional synthetics data sets were generated, 10 for each of the data set configuration used above to generate distribution curves. The resulting GBS values produced because of applying banding were then compared with the normal distributions. Note that for this purpose the 2D-BPM banding algorithm was used. The results are presented in Tables 5 and 6. In the tables, for each data set configuration, the columns indicate: (i) the average GBS value obtained after banding, (ii) the
distance of the average GBS value from the corresponding (μ) value shown in Tables 2 and 4 as appropriate, (iii) whether the results were significant or not (yes/no) with respect to one standard deviation (1SD) and (iv) whether the results were significant or not (yes/no) with respect to two standard deviation (2SD). From the tables, the generated average GBS values after banding had been applied in every case was found to be located at least one or two standard deviations away from the median. It is therefore argued that these bandings are statistically significant. The results also show that the proposed mechanism; the normal (Gaussian) distribution for determining the statistical significant of bandings is a viable approach and can be effectively used to determine the statistical significance of banding.

Table 5: GBS results with Normal Distribution (static dot density)

Data sets	Mean GBS	Distance from μ	Significant w.r.t 1SD (yes/no)	Significant w.r.t 2SD (yes/no)
100 x 100	0.57	0.10	no	**yes**
141 x 141	0.57	0.09	no	**yes**
173 x 173	0.58	0.09	no	**yes**
200 x 200	0.59	0.09	**yes**	no
224 x 224	0.59	0.09	no	**yes**
245 x 245	0.59	0.09	**yes**	no
265 x 265	0.59	0.09	no	**yes**
283 x 283	0.59	0.09	**yes**	no
300 x 300	0.60	0.09	no	**yes**
316 x 316	0.60	0.09	**yes**	no

Table 6: GBS results with Normal Distribution (range of dot density)

Data sets	Mean GBS	Distance from μ	Significant w.r.t 1SD (yes/no)	Significant w.r.t 2SD (yes/no)
100 x 100	0.41	0.02	no	**yes**
141 x 141	0.42	0.01	**yes**	no
173 x 173	0.43	0.01	**yes**	no
200 x 200	0.44	0.01	**yes**	no
224 x 224	0.46	0.02	no	**yes**
245 x 245	0.45	0.02	no	**yes**
265 x 265	0.45	0.03	no	**yes**
283 x 283	0.46	0.02	**yes**	no
300 x 300	0.46	0.03	no	**yes**
316 x 316	0.46	0.03	no	**yes**

V CONCLUSIONS AND FUTURE WORK

This paper has presented some ideas on how to determine whether the generated bandings are statistically significant or not. Two set of experiments were conducted using: (i) a static dot density value and (ii) a range of dot density values. The idea was that any data set irrespective of the density used and size, will feature some form of banding defined by a GBS value and these values will form a normal distribution. Whether, after column and rows have been reordered using the banding score concept, the resulting banding is significant or not can then be determined by how far the new GBS value is away from the mean of the associated normal distribution (μ). To analyse this approach twenty normal distributions were derived using ten 2D data set configurations. The usage of these distributions was then evaluated by using them to determine the significance of several further bandings. The evaluation results presented indicated the significance of bandings with respect to either 1SD or 2SD. A criticism of the approach is that the normal distribution for a data set under consideration must to be derived in each case; however, the results show that it is possible to generate generic normal distribution curves using ranges of dot density values (but a fixed size). The experiments had clearly indicated a useful mechanism for determining whether a banding is statistically significant or not. For Future work, the authors intend to investigate other ways of assessing the statistical significance of banded patterns.

References

[1] F. B Abdullahi, F. Coenen, and R. Martin. A novel approach for identifying banded pattern mining in zero-one data using column and row banding score. In Proc. Machine Learning and Data Mining in Pattern Recognition (MLDM'14), Springer LNAI 7376, pages 336 - 379, 2014.

[2] F. B Abdullahi, F. Coenen, and R. Martin. A scalable algorithm for banded pattern mining in multi-dimensional zero-one data. In Proc. Data Warehousing and Knowledge Discovery (DaWaK'14). Springer, LNAI, pages 391- 404, 2014.

[3] F. B Abdullahi, F. Coenen, and R. Martin. Finding banded patterns in data: The banded pattern mining algorithm. In Proc. 17th International Conference on BIG data Analytics and Knowledge Discovery (DaWaK'15), Springer LNAI 9263, pages 95-107, 2015.

[4] Veronica Czitrom and Patrick D. Spagon. Statistical case studies for industrial process improvement. SIAM, pages 342-345, 1997.

[5] Lukac Eugene and King Edgar. A property of normal distribution. The Annals of Mathematics, 11, 2004.

[6] W. Feller. Introduction to Probability Theory and Its Applications. New York Willey vol 1 3rd ed, 1968.

[7] W. Feller. Introduction to Probability Theory and Its Applications. New York Willey vol 2 3rd ed p.45, 1971.

[8] Krishnamoorty Kalimuthu. Handbook of Statistical Distribution with Applications. Chapmall and Hall / CRC Press: ISBN 1-58488-635-8, 2006.

[9] F. Pukelsheim. The three-sigma rule. American Statistician, 34:477- 495, 1994.

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

[11] Herrnstein J. Richard and Murray Charles. The Bell Curve: Intelligence and Class Structure in American Life. Free Press ISBN 0-02-914673-9, 1994.

[12] Stigler M. Stephen. Mathematical statistics in early states. The annals of the Statistics, 6:239-265, 1978.

[13] Stigler M. Stephen. Statistics on Table. Harvard University Press, 1999.

[14.] Wesstein Eric W. Normal distribution. http:// math-world.wolfram.com/NormalDistribution.html, 2015.

[15] F. Coenen, "Lucs-kdd data generator software." http://www.csc.liv.ac.uk/_frans/KDD/Software/LUCS KDD DataGen˙Generator.html, Department of Computer Science, The University of Liverpool, UK, 2003.

[16] Carl Friedrich Guass. Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections. Little Brown and Company, 1857.

[17] F. B. Abdullahi, F. Coenen, and R. Martin, "Finding banded patterns in big data using sampling," in 2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015, pp. 2233–2242.

[18] F. B. Abdullahi, F. Coenen, and R. Martin, "Banded pattern mining algorithms in multi-dimensional zero-one data," in Transactions on Large-Scale Data-and Knowledge-Centered Systems XXVI. Springer, 2016, pp. 1–31.

[19] Gemma C. Garriga, Esa Junttila, and Heikki Mannila. Banded structures in binary matrices. Proceedings Knowledge Discovery in Data Mining (KDD08), pages 292-300, 2008.

[20] Gemma C. Garriga, Esa Junttila, and Heikki Mannila. Banded structures in binary matrices. Knowledge Discovery and Information System, 28:197- 226, 2011.

[21] Erkki Makinen and Harri Siirtola. The barycenter heuristic and the reorderable matrix. Informatica, 29:357{363, 2005.