Research Article

Similarity Statistics for Clusterability Analysis with the Application of Cell Formation Problem

Yingyu Zhu and Simon Li

Department of Mechanical and Manufacturing Engineering, University of Calgary, Alberta, Canada

Correspondence should be addressed to Simon Li; simoli@ucalgary.ca

Received 9 August 2018; Accepted 17 October 2018; Published 2 December 2018

Academic Editor: Luis A. Gil-Alana

This paper proposes the use of the statistics of similarity values to evaluate the clusterability or structuredness associated with a cell formation (CF) problem. Typically, the structuredness of a CF solution cannot be known until the CF problem is solved. In this context, this paper investigates the similarity statistics of machine pairs to estimate the potential structuredness of a given CF problem without solving it. One key observation is that a well-structured CF solution matrix has a relatively high percentage of high-similarity machine pairs. Then, histograms are used as a statistical tool to study the statistical distributions of similarity values. This study leads to the development of the U-shape criteria and the criterion based on the Kolmogorov-Smirnov test. Accordingly, a procedure is developed to classify whether an input CF problem can potentially lead to a well-structured or ill-structured CF matrix. In the numerical study, 20 matrices were initially used to determine the threshold values of the criteria, and 40 additional matrices were used to verify the results. Further, these matrix examples show that genetic algorithm cannot effectively improve the well-structured CF solutions (of high grouping efficacy values) that are obtained by hierarchical clustering (as one type of heuristics). This result supports the relevance of similarity statistics to preexamine an input CF problem instance and suggest a proper solution approach for problem solving.

1. Introduction

The research of this paper is like a crossroad of manufacturing systems and computer science. Based on our disciplinary background, we initially study the cell formation (CF) problem that seeks for the clustering of similar machines and parts to support mass customization in [1]. In other words, a CF problem is a two-mode clustering problem [2]. Due to the NP-hard nature of the CF problem [3], many algorithms, including exact, metaheuristic, and heuristic approaches, have been proposed (to be discussed in Section 2.2.3). In the study of hierarchical clustering (abbreviated as HC, classified as a greedy-based heuristic approach), although HC is not the most powerful in searching for near-optimal solutions, it can yield satisfactory results comparable to some powerful metaheuristic approaches (e.g., genetic algorithms) for "well-structured" solutions. In this context, this research investigates the conditions based on the statistics of similarity values to estimate the potential structuredness of a given CF problem without solving it.

In the domain of computer science, the notion of structuredness somehow corresponds to the clusterability concept [4]. Intuitively, clusterability can be interpreted as a measure of an "intrinsic structure" of a dataset to be clustered [5]. Computer scientists have observed that a dataset of good clusterability can be clustered quite effectively (i.e., less impact from the NP-hard nature of the clustering problem). This observation has been summarized in a statement that "clustering is difficult only when it does not matter" (abbreviated as the CDNM thesis) [4, 6].

Notably, the measure of clusterability remains an open topic in computer science. Ackerman and Ben-David [4] have surveyed different definitions of clusterability and shown their incompatibility in pairwise comparisons. Nowakowska et al. [5] argued that a clusterability measure should be partition-independent so that it does not depend on the clustering algorithms and the resulting solutions. Ackerman et al. [7] proposed the use of the statistical distributions of pairwise distances between any two objects to evaluate clusterability.

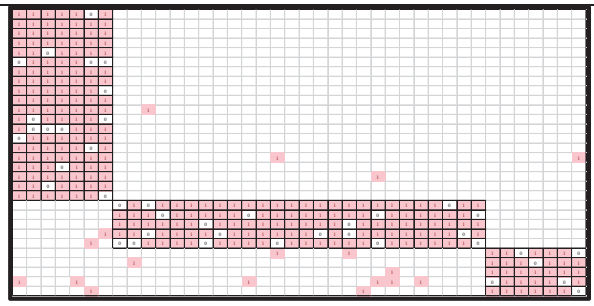
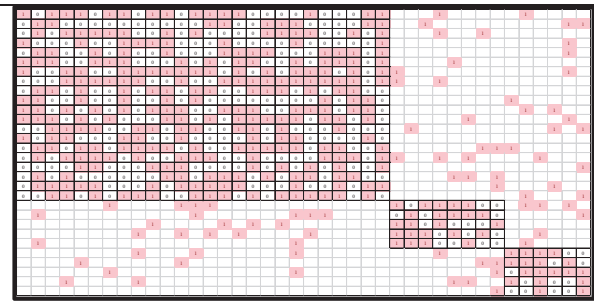
Example of a well-structured matrix	Example of an ill-structured matrix
	
Grouping efficacy = 0.8142	Grouping efficacy = 0.4780
81 (out of 435) machine pairs have similarity values higher than or equal to 0.80	4 (out of 435) machine pairs have similarity values higher than or equal to 0.50

Figure 1: Comparison of well-structured and ill-structured matrices.

Back to the context of the CF problem, in response to the CDNM thesis, we also observed that a heuristic approach (e.g., HC in our case) can yield satisfactory results. To further utilize this observation in practice, this research develops the criteria that assess the potential structuredness (corresponding to clusterability in computer science) of a given CF problem and suggest either using HC or genetic algorithm (GA) for problem solving. To verify the development, we have applied numerical examples to examine the results of the structuredness criteria and the quality of CF solutions via HC and GA.

Though developed independently, we want to acknowledge that our approach of evaluating the structuredness criteria is similar to the statistical approach by Ackerman et al. [7]. The difference lies in our application's focus on the CF problem, while Ackerman et al. [7] have focused on the relatively high-level development for clustering tasks. This difference explains our use of similarity measures (instead of distances) in statistical analysis since they are common for the CF problem and allow for some normalization in setting the structuredness criteria. Further, our work numerically checks the relations between structuredness criteria and the solution quality by two different clustering approaches (i.e., HC and GA).

Notably, this paper was extended from our conference paper [8] with the improvement of the techniques (e.g., the threshold setting and the normalization approach). Also, additional numerical examples have been used in the evaluation.

The rest of this paper is organized as follows. Section 2 will overview the CF problem and discuss the three properties of a well-structured CF solution in order to clarify the logical relation of similarity statistics. Section 3 will introduce the histogram analysis of similarity values and develop the U-shape criteria. Section 4 will introduce the Kolmogorov-Smirnov (K-S) test, which is used to develop another criterion to inform the matrix's structuredness. Section 5 will discuss the procedure that applies the developed criteria to classify well-structured and ill-structured matrices. Section 6 will examine the structuredness criteria via numerical examples, which are also used to check the effectiveness of metaheuristics via a two-stage solution process. Section 7 will conclude this paper.

2. Background: Cell Formation Problem

2.1. Problem Introduction. In the design of a cellular manufacturing system, one early and important decision is the formation of machine groups and part families, and it is often referred to as the cell formation (CF) problem. A simple CF problem can be compactly captured by a machine-part incidence matrix. Let $M = \{m_i\}$ (for $i = 1$ to m) be the set of machines and $P = \{p_j\}$ (for $j = 1$ to n) be the set of parts. Then, an incidence matrix, denoted as $B = [b_{ij}]$, indicates whether machine m_i is required to produce part p_j (if so, $b_{ij} = 1$; otherwise, $b_{ij} = 0$). After solving the CF problem, the matrix's rows and columns can be reordered to reveal which subset of machines (i.e., a machine group) is highly related to which subset of parts (i.e., a part family).

By using the incidence matrices to represent CF solutions (i.e., block-diagonal matrices), they can be roughly classified into two types: well-structured and ill-structured matrix [2, 9]. As illustrated in Figure 1, a well-structured matrix has few nonzero matrix entries outside the blocks (defined as exceptional elements) and few zero matrix entries inside the blocks (defined as voids). Precisely, exceptional elements are the matrix entries of $b_{ij} = 1$ with m_i and p_j in different cells, and voids are the matrix entries of $b_{ij} = 0$ with m_i and p_j in the same call. The opposite conditions apply for an ill-structured matrix (i.e., a matrix solution with many exceptional elements and voids). A well-structured matrix implies that part families can be produced quite exclusively by some machine groups so that the changes of few part families will not be adversely impacting the production of other parts. This is one desirable feature of cellular manufacturing systems [1].

To quantify the structuredness of a CF matrix solution, we use the traditional grouping efficacy (denoted as μ), which is formulated as follows [10].

$$\mu = \frac{n_e - n_{out}}{n_e + n_{in}} \quad (1)$$

where n_e, n_{out}, n_{in} are the total number of nonzero matrix entries, exceptional elements, and voids, respectively. In a perfect CF solution where $n_{out} = n_{in} = 0$, the grouping efficiency is equal to its maximum value, i.e., one. When there are more exceptional elements (n_{out}) and voids (n_{in}), the grouping efficacy value will become smaller.

Yet, not all incidence matrices can be converted to a well-structured matrix due to the original complex interdependency of the production requirements among machines and parts. This situation cannot be resolved by advanced optimization techniques as the root cause stems from the original inputs of the CF problem. However, we cannot practically know whether a given CF problem is going to have a well-structured matrix or not until we actually solve this problem. In this context, the purpose of this paper is to assess the structuredness of a given CF problem by analyzing the similarity of machines without actually solving it. In the traditional CF notion, two machines can be said similar if they are required mainly to produce a subset of common parts. In this work, the Jaccard similarity coefficient is applied [11, 12]. Let s_{xy} be the similarity value between machines m_x and m_y. The formulation of the Jaccard similarity coefficient is provided below.

$$s_{xy} = \frac{a_{xy}}{a_{xy} + b_{xy} + c_{xy}} \quad (2)$$

where a_{xy} is the number of parts that need both machine m_x and m_y; b_{xy} is the number of parts that need machine m_x but not machine m_y; c_{xy} is the number of parts that need machine m_y but not machine m_x. Conceptually, the Jaccard similarity coefficient focuses on the number of common features (e.g., a_{xy}) that is normalized by the total number of relevant features (e.g., a_{xy}, $b_{xy,}$ and c_{xy}). Notably, similarity is only evaluated for any two machines (i.e., a machine pair).

After specifying the notion of machine similarity, let us revisit the two examples in Figure 1. Each example has 30 machines, leading to 30×(30-1)/2 = 435 machine pairs. By examining the similarity of any two machines (or machine pairs), we find that the well-structured matrix has a higher number of machine pairs with high-similarity values. In the examples of Figure 1, we can get the following two statements concerning the statistics of the machine similarity values.

(i) Well-structured matrix: 81 (out of 435) machine pairs have similarity values higher than or equal to 0.80.

(ii) Ill-structured matrix: 4 (out of 435) machine pairs have similarity values higher than or equal to 0.50.

In this illustration, it is roughly identified that a well-structured matrix can have quite a different statistical distribution of machine similarity values as compared to an ill-structured matrix. This observation leads to an investigation question on the statistical conditions in which a well-structured matrix can be classified. This investigation is the focus of this paper. By knowing such statistical conditions, engineers in the design of cellular manufacturing systems can initially assess their production requirements via the statistics of machine similarity. If the statistical data shows unfavorable results (i.e., chance of getting a well-structured matrix is low), they can either modify the production requirements (e.g., buy more machines) or seek for other manufacturing systems. It can save the efforts to solve the CF problem with such initial assessment. Also, this paper will show that a well-structured matrix can be satisfactorily obtained by some less time-consuming heuristics (where complex optimization methods may not bring additional benefits).

2.2. Properties of a Well-Structured CF Solution. To investigate the statistical conditions of the structuredness of a CF solution, this section will discuss the three properties of a well-structured matrix. These three properties include (1) high grouping efficacy, (2) high percentage of high-similarity machine pairs, and (3) relative ease of obtaining satisfactory CF solutions. Afterward, a research plan will be discussed.

2.2.1. Property I: High Grouping Efficacy. The original formulation of the grouping efficacy (GE) in (1) can be found in Kumar and Chandrasekharan [10], and it is intended to replace a weighted sum function with a simple ratio to assess the goodness of a CF solution (in a block-diagonal form). Since then, the GE measure has become popular in the CF research (e.g., [9, 13]). Despite its popularity, some researchers have criticized its "built-in weights" [14], where a lower number of voids (i.e., n_{in}) tend to give a better GE measure (as compared to exceptional elements (i.e., n_{out})). Brusco [15] has commented that the nonlinearity of the GE measure has incurred a challenge for finding the exact solutions for the CF problems. As commented by Sarker and Mondal [16] in their survey paper, it is not easy to develop a standard measure that fits all CF problems. It is generally recognized that the GE measure is good to discern the structuredness of the matrix-based CF solutions [2]. Thus, we choose the GE measure in this study.

Based on its definition, a well-structured matrix should have few exceptional elements and voids, leading to a high value of GE. While GE is effective in indicating the structuredness of a CF solution (high value $\longrightarrow$ well-structured matrix), this value cannot be known until the CF problem is solved. Thus, in this research, GE is used as a verification measure to examine how well machine similarity can be related to the structuredness of a CF solution.

2.2.2. Property II: High Percentage of High-Similarity Machine Pairs. Compared to the property of high grouping efficacy, it is less obvious to know that a well-structured matrix has a high percentage of high-similarity machine pairs. In view of the Jaccard similarity coefficient in (2), there are two types of factors used to assess the machine similarity. While a_{xy} (i.e., the number of common parts) is taken as a commonality factor, both b_{xy} and c_{xy} (i.e., the number of parts processed in one machine but not another one) serve as differentiating factors to normalize the similarity measure. In turn, if the similarity value of both machines is high, a_{xy} cannot be zero

and the values of b_{xy} and c_{xy} should be small, implying not only commonality but also exclusiveness of these two machines to process their common parts. This feature can potentially lead to smaller numbers of voids and exceptional numbers, leading to a well-structured matrix.

In literature, the notion of similarity has been applied for many years to address the CF problem, and the Jaccard similarity coefficient is one of the early applications [11]. Since then, many similarity coefficients have been proposed, and the comparison study of similarity coefficients can be found in Sarker [17], Mosier et al. [18], and Yin and Yasuda [19]. Notably, similarity is a context-dependent concept, and it depends on the application and relevant information to assess how similar between two objects. In our investigation, we choose the Jaccard similarity coefficient because its notion on the commonality and differentiating factors is straightforward to the simple CF application.

While similarity coefficients have been studied extensively for CF problems, the statistical distribution of similarity values of a CF problem has not been investigated reasonably in our understanding. Notably, these similarity values can be found without solving the CF problems. Then, if we know the relation between the statistical distribution of similarity values and the GE measure, we can use the statistical distribution of similarity values to assess the potential of yielding a well-structured matrix for a CF problem. This is the major aim of this paper.

2.2.3. Property III: Relative Ease of Obtaining Satisfactory CF Solutions. At this point, we may wonder why it is important to know the potential of yielding a well-structured matrix before solving the CF problems. First of all, it has been recognized that a CF problem is a NP-hard problem [3] so that there will be less likely to find a practical algorithm that can guarantee an exact solution for a moderate-size problem. As a result, the effort required to solve a CF problem is not trivial. In literature, many metaheuristic algorithms have been proposed to solve the CF problems such as genetic algorithms [20, 21] and simulated annealing [22, 23]. Related comprehensive reviews can be found in Papaioannou and Wilson [24] and Renzi et al. [25]. While metaheuristic algorithms have capacities to yield high-quality solutions, they generally require users to have good mathematical skills to understand these algorithms [26] and good experiences to make some "implementation decisions" [15, p. 293] (e.g., terminating conditions in genetic algorithms).

In contrast to metaheuristic algorithms, heuristic algorithms are easier to implement but the quality of their solutions is often targeted [27, p. 159]; [24]. In a nutshell, a common feature of heuristic algorithms is their greedy or hill-climbing approaches that focus on best solutions at a stage without backtracking for other solution possibilities. This feature allows them to converge to some feasible solutions quickly with the trade-off of checking a smaller solution space (thus, potentially weaker solution quality). Hierarchical clustering (HC), which was one early approach for CF problems [11], is one example of heuristic algorithms since HC always groups the object pairs with the highest similarity values progressively without backtracking.

As its third property, it is observed that a well-structured matrix can be obtained relatively easily by a heuristic approach (referred to HC specifically in this paper), where the metaheuristic approach does not necessarily have an advantage for getting higher-quality solutions. Alternately, the advantage of the metaheuristic approach is observed more often in the case of ill-structured matrices. As discussed before, a well-structured matrix demonstrates sharp differences between similar and dissimilar machine pairs. This feature supports the "greedy" nature of the heuristic approach, which can easily distinguish high-similarity pairs in the progressive grouping process. In contrast, an ill-structured matrix has more machine pairs with middle-similarity values so that some borderline cases can potentially lead to solutions of lower quality. While this third property may not be obvious, more verifying examples will be reported later in Section 6.3 as part of the investigation effort of this paper.

Given this third property of a well-structured matrix, the statistical analysis of similarity values can then lead to another application, i.e., supporting the choice of the algorithmic approach for solving CF problems. If the statistical analysis shows a high potential to obtain a well-structured matrix, we can choose a heuristic approach to solve the CF problems. Alternately, if it indicates a high chance of getting an ill-structured matrix, we may consider revising the input incidence matrix (e.g., adding more machines or changing some part requirements). Also, we can prepare to use the metaheuristic approach to seek for high-quality solutions. In sum, the statistical analysis can preliminarily probe the structure of a given CF problem in order to determine the next problem solving step.

2.3. Research Plan. In view of the three properties of a well-structured matrix discussed above, the research and development questions are set as follows.

(i) What are the criteria related to the statistics of similarity values to assess the potential of getting a well-structured matrix?

(ii) How do we decide on whether using a metaheuristic or heuristic approach for solving a CF problem?

To address the first question, this paper will utilize two statistical tools: histogram and the Kolmogorov-Smirnov (K-S) test. Histogram will be used to analyze the distribution of machine similarity values of a given CF problem, and twenty CF solutions will be set to investigate the threshold values for informing the potential structuredness of a matrix. The K-S test will be used to assess the normality of the distribution of machine similarity values. That is, if the set of similarity values roughly follow the normal distribution, it means that many machine pairs have the average similarity value, implying a low proportion of high-similarity values (i.e., an ill-structured matrix).

Based on the investigation using the histogram and the K-S test, we will develop a procedure to probe the structure of a given CF matrix and suggest whether using a metaheuristic or heuristic for problem solving (i.e., address the second question). In this paper, we have implemented

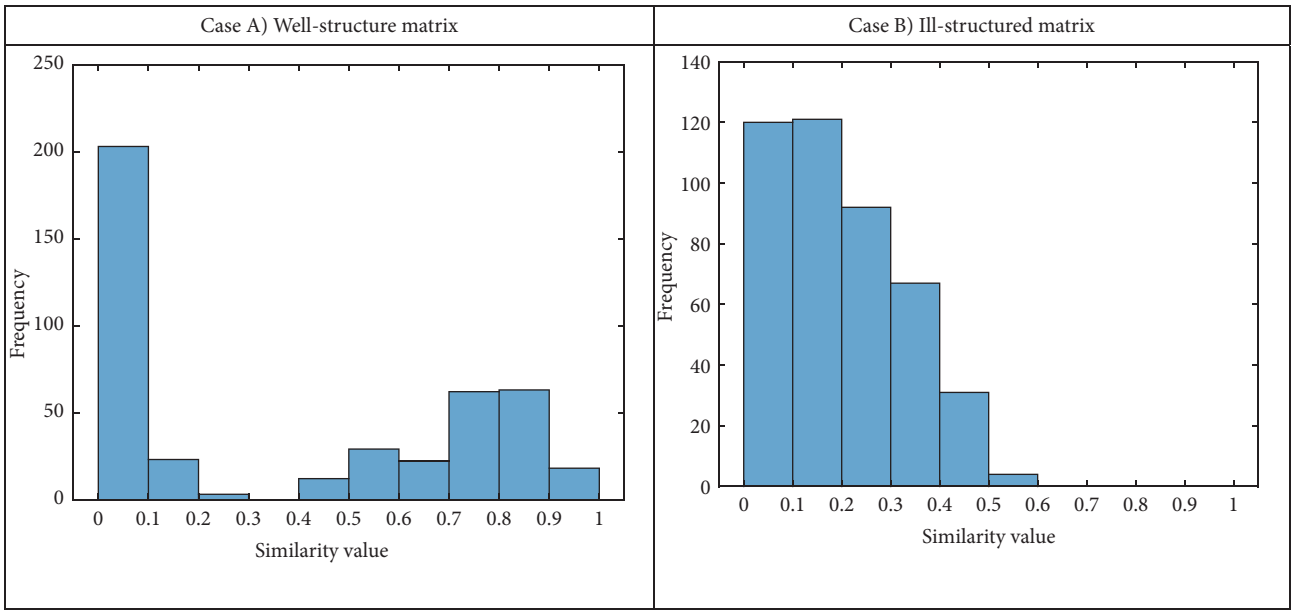


Figure 2: Histograms of well-structured and ill-structured matrices.

genetic algorithm (GA) and hierarchical clustering (HC) as the metaheuristic and heuristic approaches, respectively, for solving the CF problems. To verify the procedure, additional forty CF matrices will be set. These CF matrices will be solved by HC and then genetic algorithm to observe the relation between the matrix's structuredness and the utility of the metaheuristic approach for better CF solutions.

3. Histogram Analysis of Similarity Values

3.1. Histogram and the U-Shape. In this study, histograms are used to report the frequency distribution of machine similarity values with an increment of 0.1. Figure 2 shows two histograms for the well-structured and ill-structured matrices of Figure 1, respectively. In these histograms, the horizontal axis stands for the machine similarity values ranging from 0 to 1, and the vertical axis stands for the number of machine pairs within those ranges of similarity values. Notably, these histograms are independent of the orders of a matrix's rows and columns. That is, we can get these histograms of similarity values without solving the CF problem.

From these two histograms, it is observed that a well-structured matrix tends to yield an U-shape histogram, i.e., relatively high numbers of extreme similarity values. The right peak of the U-shape can be explained by the property of high percentage of high-similarity machine pairs discussed in Section 2.2.2. While the numbers of low-similarity machine pairs are high in both cases of well-structured and ill-structured matrices, a well-structured matrix has a low number of machine pairs of similarity values between 0.2 and 0.4. In contrast, an ill-structured matrix has a good number of those middle-similarity machine pairs, which cause a challenge of clear grouping in cell formation. Given this general U-shape observation, the next subsections will discuss the criteria that classify the structuredness of a matrix (i.e., well-structured or ill-structured) based on the histogram data.

3.2. Setup of 20 Benchmark Matrices. Since the frequency distribution of a histogram will not be altered by the orders of a matrix's rows and columns, we can set the CF solution matrices with known structuredness and then observe their histograms to develop the structuredness criteria. In this investigation, twenty 30×40 solution matrices (i.e., 30 machines and 40 parts) with three cells (or blocks) are set. These matrices are varied by two factors: (1) block sizes and (2) numbers of exceptional elements and voids. Concerning the block sizes, five cases are set as follows, where each bracket indicates the size of a block as (number of machines × number of parts).

(i) Case A ⟶ even case: (10×13) (10×13) (10×14)

(ii) Case B ⟶ uneven case with a large block: (20×26) (5×7) (5×7)

(iii) Case C ⟶ uneven numbers of machines and parts in two blocks: (20×7) (5×26) (5×7)

(iv) Case D ⟶ uneven numbers of machines and parts in three blocks: (20×5) (5×17) (5×18)

(v) Case E ⟶ uneven case with two large blocks: (14×18) (14×19) (2×3)

Besides, four cases are set below to characterize the structuredness of matrices via the control of the numbers of exceptional elements and voids.

(i) Case I (well-structured): few exceptional elements and no voids

(ii) Case II (well-structured): no exceptional elements and few voids

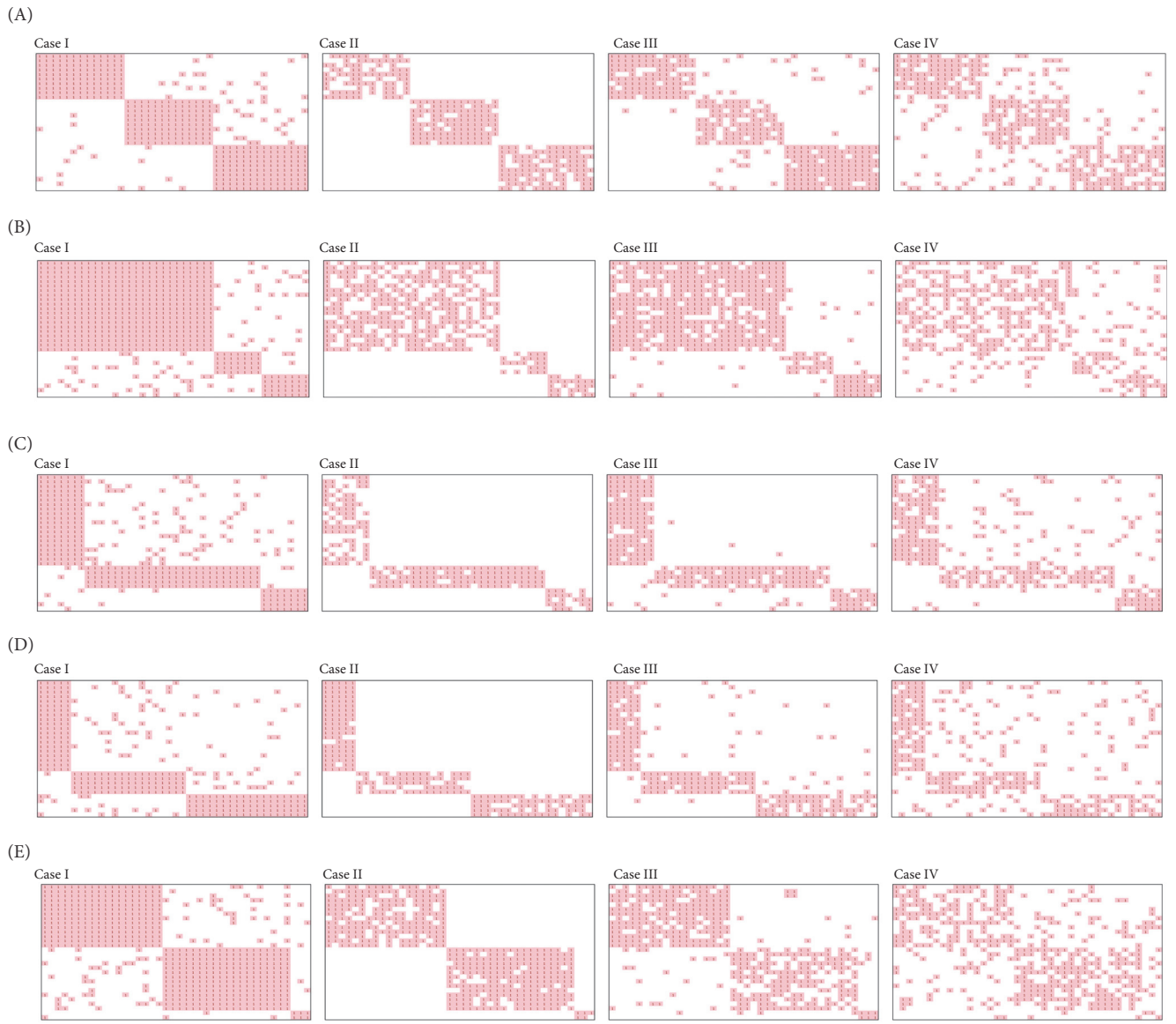


Figure 3: The resulting matrices of 20 benchmark cases.

(iii) Case III (well-structured): few exceptional elements and few voids

(iv) Case IV: (ill-structured): good numbers of exceptional elements and voids

The resulting 20 matrices are shown in Figure 3. As general inspections, the matrices in Cases I and II have clear boundaries of three cells. The matrices in Case III have more exceptional elements and voids but their structures are still quite discernible. In contrast, the structure of matrices in Case IV is messier with higher numbers of exceptional elements and voids. Based on these matrices, the next subsection will investigate their histograms and develop the U-shape criteria to classify the matrix's structuredness.

3.3. Histogram-Based U-Shape Criteria. To inform the matrix's structuredness, two conditions as the U-shape criteria are set toward the low and high-similarity values. Let $F_{left}(x)$ be the fraction of similarity values that are lower than x and $F_{right}(y)$ be the fraction of similarity values that are higher than y. Then, the general U-shape criteria can be expressed as follows.

$$F_{left}(x) \geq a \tag{3}$$

$$F_{right}(y) \geq b \tag{4}$$

where a and b are the thresholds of the minimum fractions of low and high-similarity values, respectively, to characterize the U-shape of a well-structured matrix. The setup of these parametric values (i.e., x, y, a, and b) will be based on the above 20 benchmark matrices.

Figure 4 shows the histograms of the 20 benchmark matrices. As the preliminary observations, the frequency

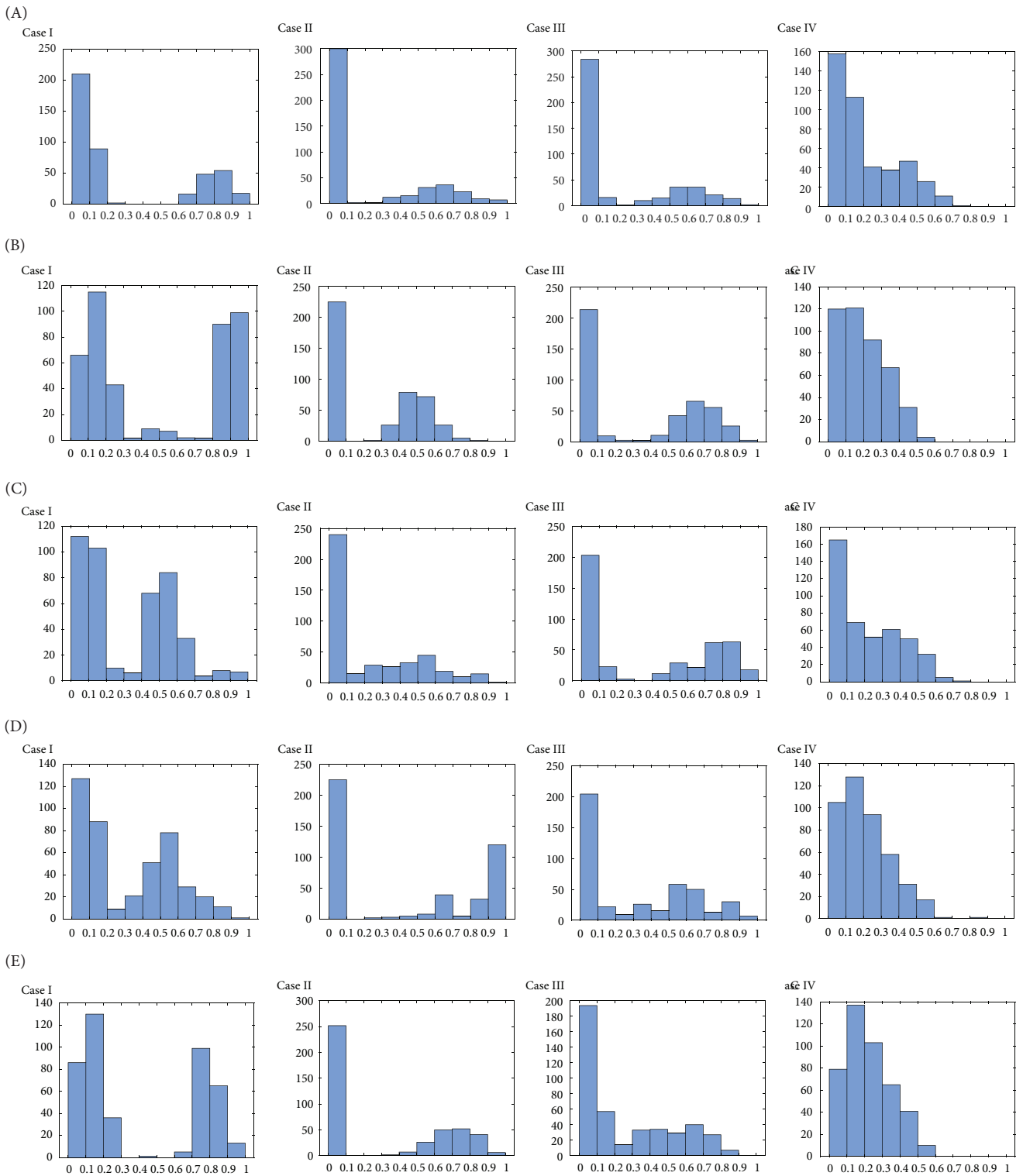


Figure 4: Histograms of 20 benchmark matrices.

distributions of these histograms are perceived quite different between the well-structured (i.e., Cases I, II, and III) and ill-structured matrices (i.e., Case IV). Yet, some U-shapes are not plainly obvious (e.g., Cases A-III and C-II), and the peaks of high-similarity values of the well-structured matrices are not located at the rightmost region (e.g., Cases C-I and D-I). The U-shape criteria will then be set based on these observations.

Concerning the region of low-similarity values (i.e., the left side of the U-shape), it is found that both well-structured (i.e., Cases I, II, and III) and ill-structured (i.e., Case IV)

Table 1: Number of machine pairs with similarity values equal to zero.

	Case I	Case II	Case III	Case IV
A	34	300	110	32
B	7	225	83	23
C	23	240	121	57
D	22	225	116	15
E	11	252	73	14

Table 2: Number of machine pairs with similarity values greater than or equal to 0.5.

	Case I	Case II	Case III	Case IV
A	135	106	109	38
B	200	104	194	4
C	136	91	194	38
D	139	203	158	19
E	182	175	103	10

matrices have high proportions because many machines, as long as they are not in the same cell, have less common parts to work with in both cases. As a result, the proportions of low-similarity values from a well-structured matrix can become less discernible statistically. Thus, we choose to investigate the extreme value when the similarity values equal to zero, i.e., $F_{left}(x{=}0)$. Table 1 records the number of machine pairs with the similarity values equal to zero. As observed, while the matrices of Cases II and III have low right-side peaks, they have high proportions of such zero-similarity machine pairs. As the U-shape criteria will be used for the early screening, we set this criterion rather strictly as follows.

$$F_{left}(0) \geq 0.5 \tag{5}$$

This criterion requires 50% of machine pairs to have zero-similarity values in order to qualify a well-structured matrix. By checking the benchmark matrices with 30 machines (i.e., 435 machine pairs), the threshold is 218 machine pairs, and the matrices in Case II pass this criterion.

Concerning the region of high-similarity values (i.e., the right side of the U-shape), as discussed earlier, not all well-structured matrices have high proportions of high-similarity values at the rightmost region. By inspecting the histograms in Figure 4, we identify a reasonable cut-off of high-similarity values should be 0.5, i.e., $F_{right}(y{=}0.5)$. Table 2 records the number of machine pairs with the similarity values greater than or equal to 0.5. As observed, the proportions of high-similarity values ($s_{xy} \geq 0.5$) in Case IV (i.e., ill-structured matrices) are relatively low. In contrast, Case C-II is the well-structured matrix with the lowest number of high-similarity values (i.e., 91), and the corresponding fraction is $91/435 \approx 0.21$. As a result, another U-shape criterion for the right-hand side is set as follows.

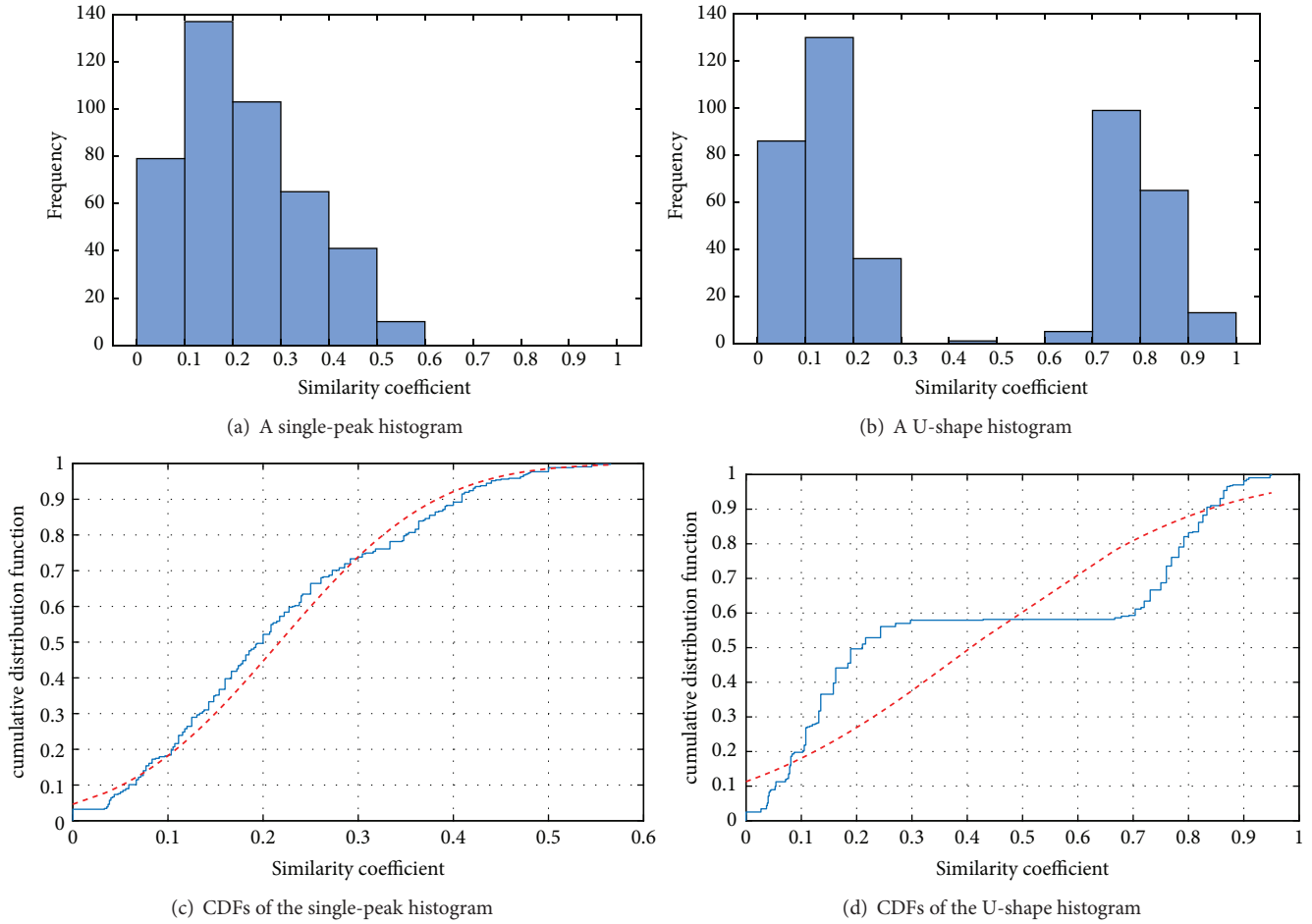
$$F_{right}(0.5) \geq 0.2 \tag{6}$$

In sum, if an input incidence matrix satisfies one of the two U-shape criteria formulated in (5) and (6), this matrix has a good chance to yield a well-structured CF solution. Notably, we treat the histogram-based U-shape criteria as a preliminary filter in this work. That is, if a matrix does not satisfy these criteria, it does not immediately imply that this matrix is ill-structured. In fact, other parameters of an input incidence matrix, such as the number of machines and the density of nonzero matrix entries, can impact the frequency distribution of a histogram. Thus, the next section will develop another criterion based on the K-S test.

4. Criterion Setting Based on the Kolmogorov-Smirnov (K-S) Test

4.1. Background. The Kolmogorov-Smirnov (K-S) test is one type of hypothesis testing in statistics (Corder and Foreman) [28]. As one of its applications, the K-S test is used in this paper to evaluate how well a dataset represents a normal distribution (i.e., the normality of the dataset). The use of the K-S test in this study is mainly motivated by the observation of the histograms in Figure 2 that a well-structure matrix will tend to give a U-shape. As the U-shape will generally exhibit two peaks in the histogram representation, the normality of the associated data (i.e., similarity values) will be weak in comparison to that of an ill-structured matrix.

Figure 5 illustrates the concept of the normality of similarity values with two cases: single-peak histogram and U-shape histogram. The K-S test essentially compares the curves of two cumulative distribution functions (CDFs) [29, 30]. While one CDF represents the empirical data points (i.e., empirical CDF, solid line), another CDF is based on the normal distribution curve fitted by the empirical data (i.e., hypothesized normal CDF, dashed line). As seen in Figures 5(c) and 5(d), the single-peak histogram has higher normality than the U-shape histogram since the single-peak histogram yields a closer match between the empirical and hypothesized normal CDFs. In contrast, the U-shape histogram yields its empirical CDF in Figure 5(d) with rapid increases at the beginning and the end, along with a relatively flat region in the middle, and this CDF curve significantly deviates from normality [31].



(a) A single-peak histogram

(b) A U-shape histogram

(c) CDFs of the single-peak histogram

(d) CDFs of the U-shape histogram

FIGURE 5: CDFs of similarity values (solid line: empirical CDF; dashed line: hypothesized normal CDF).

The P value is a common concept in hypothesis testing [32]. It can be interpreted as the smallest probability value associated with a given dataset to reject the null hypothesis (i.e., smaller P value $\longrightarrow$ more likely to reject the null hypothesis). In this work, we treat the P value of a K-S test as a proxy measure on the normality of a set of similarity values. That is, if the P value is smaller, the dataset tends to be *less*-normal [33]. Interpreted in our context, a *less*-normal condition implies a U-shape and thus a well-structured matrix. For example, the P value of the single-peak histogram in Figure 5(c) is 7.44×10^{-4}, and the P value of the U-shape histogram in Figure 5(d) is 9.27×10^{-22}.

Notably, the purpose of using the K-S test in this work is not about hypothesis testing, but only using its P value as a proxy measure to assess the normality of a set of similarity values and then inform the structuredness of a CF matrix. Yet, the P values in our applications tend to be very small. To conveniently handle this proxy measure, let P_{value} be the P value of a set of similarity values based on the K-S test, and an alternative proxy measure (denoted as L_p) is defined as follows:

$$L_p = -\log_{10} P_{value} \tag{7}$$

As L_p is the negative logarithm of the P value, a higher value of L_p implies a higher tendency of having a U-shape of the dataset. For example, the values of L_p for the single-peak histogram (i.e., Figure 5(c)) and the U-shape histogram (i.e., Figure 5(d)) are 3.13 and 21.03, respectively. In other words, if a CF matrix yields a higher value of L_p, it has a better chance to be solved as a well-structured CF solution.

By knowing the property of the trend associated with L_p, it leads to the next investigation question on setting the threshold value of L_p to classify ill-structured and well-structured matrices. To do so, it is recognized that the values of L_p can be sensitive to the number of machines and the density of nonzero entries of a given matrix. Thus, the next subsection will investigate the upper bound of L_p of a given matrix to normalize the value of L_p. Then, we will apply the 20 benchmark matrices in Figure 3 to determine the threshold.

4.2. Estimate the Upper Bound of L_p for Normalization. The upper bound of L_p can be estimated by a perfect block-diagonal matrix, where the numbers of exceptional elements (n_{out}) and voids (n_{in}) are zero (i.e., the grouping efficacy μ = 1). In this case, the machine pairs have similarity values equal to either one (when two machines belong to the same block)

Table 3: L_p, L_{bp} and their ratios of the benchmark matrices.

	Case I			Case II		
	L_p	L_{bp}	L_p/L_{bp}	L_p	L_{bp}	L_p/L_{bp}
Case A	43.44	62.39	0.70	69.58	85.40	0.81
Case B	22.03	31.35	0.70	44.35	72.93	0.61
Case C	14.80	70.55	0.21	42.57	97.73	0.44
Case D	13.13	75.90	**0.17**	43.58	96.69	0.45
Case E	23.06	39.67	0.58	53.56	65.65	0.82
	Case III			**Case IV**		
	L_p	L_{bp}	L_p/L_{bp}	L_p	L_{bp}	L_p/L_{bp}
Case A	39.67	76.05	0.52	10.77	73.53	**0.15**
Case B	26.84	54.38	0.49	3.77	72.19	0.05
Case C	23.06	88.52	0.26	7.69	86.15	0.09
Case D	18.77	91.49	0.21	3.13	83.92	0.04
Case E	17.56	67.29	0.26	1.97	67.14	0.03

or zero (when two machines are in different blocks). This kind of "bipolar" distribution can be viewed as a far extreme of the normal distribution, and the corresponding P value can be taken as the upper bound of L_p.

In the normalization process, we can first identify the size and the number of nonzero entries of a given matrix. Let m and n be the numbers of machines and parts, respectively, as the size of the matrix. The number of nonzero matrix entries has been denoted as n_e. Then, the density of nonzero entries of a matrix (denoted as D_s) can be determined as follows.

$$D_s = \frac{n_e}{m \times n} \tag{8}$$

Given an incidence matrix, its upper bound of L_p can be considered in a case when its nonzero entries can be freely moved to form a nearly perfect block-diagonal matrix. By fixing the values of m, n, and D_s, there can be a corresponding theoretical upper bound of L_p. Let L_{bp} denote such an upper bound of L_p of a given matrix. Then, for any given matrix, we can determine its L_p and L_{bp}, where L_{bp} is treated as a normalizing factor. Since this paper focuses on machine similarity, we drop the consideration of n to simplify the investigation. Then, the next step is to determine the following function.

$$L_{bp} = f\left(m, D_s\right) \tag{9}$$

To estimate the function of L_{bp}, our strategy is to systematically generate a good number of perfect block-diagonal matrices by varying the numbers of machines, parts, and even-size cells (note: the number of even-size cells will determine the number of nonzero entries). The ranges of these varying parameters in this work are listed as follows.

(i) Number of machines: from 10 to 50 machines

(ii) Number of parts: from 10 to 110 parts (with an increment of 10)

(iii) Number of even-size cells: from 2 to 14 cells (also restricted by the matrix's size to avoid extremely large and small cells)

Further details of the setup of these perfect matrices can be found in Zhu [34]. As a result, this work has generated 2519 perfect matrices. Then, the values of P value and L_p are determined for these matrices, giving 2519 points to approximate the function formulated in (9) via curve fitting techniques. The resulting regression equation is found as follows.

$$\begin{aligned} L_{bp}\left(m, D_s\right) = {} & -29.08 + 2.164m + 132.3D_s \\ & + 0.1049m^2 - 10.35mD_s \end{aligned} \tag{10}$$

In practice, we can determine the values of L_p via (7) and L_{bp} via (10) for a given matrix. Then, we can check its ratio of L_p to L_{bp} and examine the U-shapeness and then the possible structuredness of the matrix. The next subsection will discuss the criterion based on the ratio of L_p to L_{bp}.

4.3. Ratio Criterion Based on L_p and L_{bp}. The setting of the ratio threshold for L_p and L_{bp} is based on the 20 benchmark matrices in Figure 3. The values of L_p, L_{bp} and their ratios are recorded in Table 3. As a recall, Cases I, II, and III are set to represent the well-structured matrices, and Case IV represents ill-structured matrices. As an initial assessment, the average of the ratios of Cases I, II, and III (i.e., well-structured matrices) is 0.48, while the ratio average of Case IV is 0.07. This observation indicates that the ratio L_p/L_{bp} can make distinctions between well-structured and ill-structured matrices quite effectively from a statistical standpoint.

Yet, when we examine the extreme situations, the lowest ratio of the well-structured cases is 0.17 (i.e., Case D-I, bold in Table 3), and the highest ratio of the ill-structured cases is 0.15 (i.e., Case A-IV, also bold in Table 3). As observed, the gap between the two is close, and we intend to impose a tight criterion to classify well-structured matrices. As a result, we set the threshold value at 0.2, formulated as follows.

$$\frac{L_p}{L_{bp}} \geq 0.2 \tag{11}$$

At this point, Case D-I is the only well-structured matrix that does not satisfy this criterion. Yet, Case D-I satisfies one of

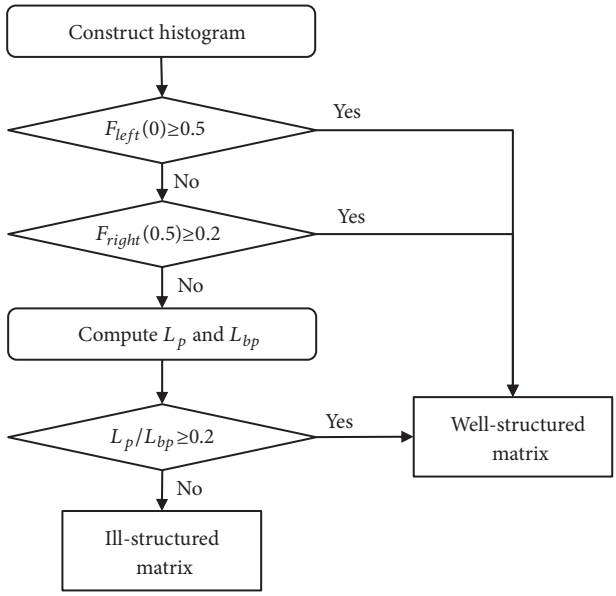


Figure 6: Procedure to assess the potential structuredness of an incidence matrix.

the earlier U-shape criteria. Thus, our next step is to combine the U-shape criteria and the ratio criterion in a procedure to examine the potential structuredness of an incidence matrix. That is, if a given matrix satisfies one of these criteria, it is indicated that this matrix has a high potential to yield a well-structured CF solution. The next section will discuss this procedure to apply these criteria to inform the potential structuredness of a given matrix.

5. Procedure

This section provides a four-step procedure below to assess the potential structuredness of an incidence matrix using the histogram-based U-shape criteria and the criterion based on the P value of the K-S test. Figure 6 illustrates the decision branches of this procedure.

Step 1 (construct histogram). By receiving an incidence matrix as an input, the similarity values of machine pairs are first determined based on (2). If there are m machines, there will be $m\times(m\text{-}1)/2$ machine pairs with their similarity values, forming the dataset of the statistical analysis. A histogram is then constructed to analyze these similarity values.

Step 2 (apply the histogram-based U-shape criteria). This represents the preliminary check based on the frequencies of having high and low-similarity values. If either one of the criteria $F_{left}(0) \geq 0.5$ or $F_{right}(0.5) \geq 0.2$ is satisfied, the incidence matrix is considered having a good potential to yield a well-structured CF solution. If none of these two criteria is satisfied, we will move on to the analysis based on the P value of the K-S test.

Step 3 (compute L_p and L_{bp}). The dataset of similarity values is treated as the input to determine the P value of the K-S test in view of assessing the normality of the dataset. This calculation can be performed via some statistics software tools. In this work, we have used the statistics functions from Matlab to compute the P value. Then, the value of L_p can be evaluated using (7). With the incidence matrix, the value of L_{bp} can be evaluated using (10) by identifying the number of machines (i.e., m) and the density of nonzero entries (i.e., D_s).

Step 4 (apply the ratio criterion L_p / L_{bp}). With the values of L_p and L_{bp}, we can check the criterion if L_p / L_{bp} $\geq$ 0.2. If this criterion is satisfied, the input matrix should have a good potential to yield a well-structured CF solution. If not, the input matrix would have a good chance to result in an ill-structured CF solution. The practitioners may consider modifying the input matrix by adding machines or revising the production requirements.

6. Application and Verification

To examine the statistical analysis of similarity values for CF problems in this paper, other 40 matrices (in addition to the earlier 20 benchmark matrices, making up a total of 60 matrices) will be generated and applied in this section. These 60 matrices will be used to examine the following two issues specifically.

(i) Given the three criteria for assessing the potential structuredness of a matrix, we are going to use these 60 matrices to examine their effectiveness to distinguish well-structured and ill-structured matrices.

(ii) While Property III (i.e., relative ease of obtaining satisfactory CF solutions) of a well-structured matrix has been discussed in Section 2.3, it will be verified via these 60 matrices by two stages of CF problem solving.

6.1. Setup of the 60 Incidence Matrices. The strategy to generate 60 matrices is based on the extension of getting the 20 benchmark matrices in Section 3.2. The additional varying factors include the following.

(i) In addition to the size of 30×40 matrix, another size of 40×100 matrix is set.

(ii) We add cases with more numbers of cells (from 3 to 6, 8, and 12 cells)

(iii) The evenness of cell sizes is also varied for each case.

Table 4 shows the setup of 60 matrices, where Cases A and E are repeated from Section 3.2 for comparison. Notably, the structuredness of matrices, which were classified as Cases I, II, III, and IV in Section 3.2, is also applied, leading to the study of 15×4 = 60 incidence matrices. As the intention of the setup, the matrices of Cases I and II have no voids and exceptional elements, respectively. Then, they should be classified as well-structured matrices. The matrices of Case III have only few exceptional elements and voids, and they should also be classified as well-structured matrices. In contrast, the matrices of Case IV have more exceptional elements and voids, and they should be classified as ill-structured matrices. The images and histograms of these 60

Table 4: Setup of incidence matrices.

	Matrix size	No. of cells	Cell sizes
Case A	30×40	3 cells	(10×13) (10×13) (10×14)
Case B	30×40	3 cells	(20×26) (5×7) (5×7)
Case C	30×40	3 cells	(20×7) (5×26) (5×7)
Case D	30×40	3 cells	(20×5) (5×17) (5×18)
Case E	30×40	3 cells	(14×18) (14×19) (2×3)
Case F	30×40	6 cells	(5×7) (5×7) (5×7) (5×7) (5×7) (5×5)
Case G	30×40	6 cells	(15×20) (3×4) (3×4) (3×4) (3×4) (3×4)
Case H	30×40	6 cells	(15×4) (3×20) (3×4) (3×4) (3×4) (3×4)
Case I	30×40	6 cells	(15×5) (3×7) (3×7) (3×7) (3×7) (3×7)
Case J	40×100	8 cells	(5×13) (5×13) (5×13) (5×13) (5×12) (5×12) (5×12) (5×12)
Case K	40×100	8 cells	(8×8) (8×8) (4×14) (4×14) (4×14) (4×14) (4×14) (4×14)
Case L	40×100	8 cells	(7×18) (7×18) (7×18) (7×17) (6×17) (2×4) (2×4) (2×4)
Case M	40×100	12 cells	(3×8) (3×8) (3×8) (3×8) (3×8) (3×8) (3×8) (3×8) (4×9) (4×9) (4×9) (4×9)
Case N	40×100	12 cells	(5×5) (5×5) (4×5) (4×5) (3×10) (3×10) (3×10) (3×10) (3×10) (3×10) (2×10) (2×9)
Case O	40×100	12 cells	(4×12) (4×12) (4×12) (4×12) (4×11) (4×11) (4×11) (4×11) (2×2) (2×2) (2×2) (2×2)

matrices are provided as supplementary materials (available here).

6.2. Examination of the Criteria. To evaluate the effectiveness of the criteria to assess the structuredness of the matrices, we have evaluated the criteria values for the 60 matrices. The results are provided in Table 5, where the values satisfying the criteria of well-structured matrices are bold. As observed in these results, the structuredness criteria can discern the well-structured matrices of Cases I, II, and III, where each matrix there satisfies at least one criterion. In contrast, no matrices of Case IV satisfy any criteria of well-structured matrices.

In view of the effectiveness of individual criteria, it is observed that $F_{left}(0)$ is effective in filtering the matrices of Case II (i.e., few voids and no exceptional elements). Due to the absence of exceptional elements in this case, any two machines of different blocks will have similarity values equal to zero. This explains the high values of $F_{left}(0)$ observed in Case II. In contrast, $F_{right}(0.5)$ is less effectiveness when the matrices have more cells (e.g., Cases H and I) and large sizes (e.g., Cases J to O). Notably, the values of $F_{right}(0.5)$ for Case IV are quite low (ranging from 0.00 to 0.09). In this view, the criterion of $F_{right}(0.5)$ is quite tight.

By comparison, the ratio criterion (i.e., L_p/L_{bp}) seems effective in distinguishing well-structured matrices, where Case D-I is the only case not identified as a well-structured matrix by this criterion only. Notably, the discernible gap of well-structured matrices (lowest at 0.17 in Case D-I) and ill-structured matrices (highest 0.16 in Case L-IV) is small. It explains the need of having $F_{left}(0)$ and $F_{right}(0.5)$, along with the ratio criterion, in the assessment of the structuredness of the matrices.

6.3. Examination of Property III via Optimization. As a recall from Section 2.2.3, Property III states that a well-structured matrix can be fairly obtained via a heuristic approach, where more complex metaheuristics may not bring in additional benefits. To verify this property, the sixty matrices were tested with a two-stage solution process. First, each matrix will be solved by a hierarchical clustering (HC) method as one heuristic to yield a CF solution. Then, we examine if we can further optimize the obtained CF solution via the genetic algorithm (GA), representing a metaheuristic method. In this way, we can check the correlation between grouping efficiency and the percentage of improvement of solution quality by GA. The algorithmic details of the HC method and the implementation details of GA applied in this study can be found in Zhu [34].

Table 6 lists the grouping efficacy (μ) results for the 60 matrices after running hierarchical clustering (HC) and then genetic algorithm (HC+GA). Also, the percentages of improvement in view of grouping efficacy by GA are reported for comparison. As observed, the matrix solutions in Cases I and II cannot be further improved by GA, while three matrix solutions in Case III can be improved by GA with small percentages (between 0.20% and 0.25%). In contrast, the ill-structured matrix solutions in Case IV can be improved by GA in the percentages of improvement between 0.63% and 22.69%. Overall, we consider that the numerical results generally follow Property III, given that the matrices in Case III are close to the boundary between well-structured and ill-structured matrices.

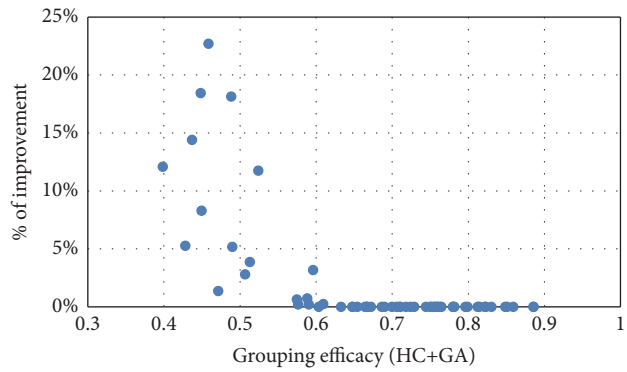


Figure 7: Percentage of solution improvement versus grouping efficacy.

Figure 7 shows the plots of the percentages of solution improvement versus the values of grouping efficacy based on HC+GA. Based on the 60 matrices studied in this paper, GA did not improve the quality of matrix solutions that have 0.60 or higher grouping efficacy. For the data points of grouping efficacy values less than 0.60, we find that these data points are negatively correlated, where the correlation value [32, p. 173] is -0.62. In the statistical interpretation, we can state that a lower value of grouping efficacy tends to allow a larger room of improvement by GA but its linearity is not strong. Notably, the capabilities of HC and GA to yield high-quality solutions can depend on other factors (e.g., density of nonzero entries in a matrix). Thus, it is not easy to observe a linear correlation just between the percentage of improvement and the grouping efficacy. More control factors and samples should be required for an in-depth investigation.

7. Conclusions

This paper has explored the statistics of similarity values to investigate the structuredness of cell formation (CF) matrix solutions. Using grouping efficacy (μ) as one recognized index to inform the quality of a CF matrix, it is found that a well-structured matrix has a high percentage of high-similarity machine pairs (i.e., Property II). Accordingly, this paper sets up 20 benchmark matrices, with varying structuredness, to develop the U-shape criteria and the criterion based on the Kolmogorov-Smirnov test. Then, a procedure is developed to assess the potential structuredness of a CF matrix without solving the CF problem. The criteria for assessing structuredness of matrices are examined via additional 40 matrices, and agreeable results are observed. Genetic algorithm (GA) is used to see if it can improve the CF solutions obtained by hierarchical clustering (as one type of heuristics). The results show that the matrix solutions with high grouping efficacy values (i.e., well-structured matrices) cannot be effectively improved by GA.

While the worst-case computational complexity of clustering problems (e.g., NP hardness) is well recognized, the

Table 5: Results of the criteria values.

	Case I			Case II			Case III			Case IV		
	F_{left}	F_{right}	L_p/L_{bp}	F_{left}	F_{right}	L_p/L_{bp}	F_{left}	F_{right}	L_p/L_{bp}	F_{left}	F_{right}	L_p/L_{bp}
A	0.08	**0.31**	**0.70**	**0.69**	**0.24**	**0.81**	0.25	**0.25**	**0.52**	0.08	0.09	0.15
B	0.02	**0.46**	**0.70**	**0.52**	**0.24**	**0.61**	0.19	**0.45**	**0.49**	0.05	0.01	0.05
C	0.05	**0.31**	**0.21**	**0.55**	**0.21**	**0.44**	0.28	**0.45**	**0.26**	0.13	0.09	0.09
D	0.05	**0.32**	0.17	**0.52**	**0.47**	**0.45**	0.27	**0.36**	**0.21**	0.03	0.04	0.04
E	0.03	**0.42**	**0.58**	**0.58**	**0.40**	**0.82**	0.17	**0.24**	**0.26**	0.03	0.02	0.03
F	**0.51**	0.14	**0.39**	**0.86**	0.11	**0.92**	0.43	0.10	**0.36**	0.26	0.03	0.14
G	0.18	**0.26**	**0.41**	**0.72**	0.14	**0.80**	0.26	**0.26**	**0.39**	0.18	0.07	0.14
H	0.34	0.14	**0.27**	**0.72**	**0.23**	**0.66**	0.43	0.18	**0.27**	0.23	0.03	0.08
I	0.18	0.11	**0.23**	**0.72**	**0.22**	**0.68**	0.34	0.15	**0.23**	0.19	0.01	0.06
J	0.29	0.10	**0.43**	**0.90**	0.09	**0.95**	**0.53**	0.06	**0.47**	0.14	0.00	0.10
K	0.21	0.08	**0.34**	**0.88**	0.08	**0.91**	0.17	0.03	**0.25**	0.08	0.00	0.07
L	0.17	0.13	**0.45**	**0.87**	0.12	**0.96**	0.30	0.04	**0.28**	0.10	0.02	0.16
M	0.37	0.04	**0.31**	**0.94**	0.05	**0.94**	0.41	0.02	**0.27**	0.27	0.00	0.13
N	0.40	0.04	**0.28**	**0.93**	0.06	**0.94**	0.42	0.02	**0.26**	0.24	0.00	0.08
O	0.30	0.06	**0.30**	**0.93**	0.06	**0.97**	0.48	0.04	**0.34**	0.14	0.00	0.08

Table 6: Grouping efficacy results with the two-stage solution process.

	Case I			Case II		
	HC	HC+GA	% Improve	HC	HC+GA	% Improve
Case A	0.7817	0.7817	0	0.7550	0.7550	0
Case B	0.8859	0.8859	0	0.6542	0.6542	0
Case C	0.7587	0.7587	0	0.7180	0.7180	0
Case D	0.7514	0.7514	0	0.8218	0.8218	0
Case E	0.8590	0.8590	0	0.8302	0.8302	0
Case F	0.8230	0.8230	0	0.7800	0.7800	0
Case G	0.8511	0.8511	0	0.6861	0.6861	0
Case H	0.7059	0.7059	0	0.7500	0.7500	0
Case I	0.6475	0.6475	0	0.7444	0.7444	0
Case J	0.7645	0.7645	0	0.7600	0.7600	0
Case K	0.7106	0.7106	0	0.7241	0.7241	0
Case L	0.7561	0.7561	0	0.8122	0.8122	0
Case M	0.6720	0.6720	0	0.7798	0.7798	0
Case N	0.6327	0.6327	0	0.8484	0.8484	0
Case O	0.6644	0.6644	0	0.8854	0.8854	0
	Case III			**Case IV**		
	HC	HC+GA	% Improve	HC	HC+GA	% Improve
Case A	0.7627	0.7627	0	0.5776	0.5959	3.17%
Case B	0.7961	0.7961	0	0.4650	0.4713	1.35%
Case C	0.8142	0.8142	0	0.5842	0.5884	0.72%
Case D	0.7290	0.7290	0	0.4938	0.5129	3.87%
Case E	0.6898	0.6898	0	0.4927	0.5065	2.80%
Case F	0.7097	0.7097	0	0.4689	0.5240	11.75%
Case G	0.7990	0.7990	0	0.5711	0.5747	0.63%
Case H	0.6667	0.6667	0	0.4134	0.4884	18.14%
Case I	0.6481	0.6481	0	0.4067	0.4281	5.26%
Case J	0.6995	0.6995	0	0.3785	0.4483	18.44%
Case K	0.6035	0.6035	0	0.3819	0.4369	14.40%
Case L	0.6079	0.6094	0.25%	0.4657	0.4898	5.18%
Case M	0.5753	0.5765	0.21%	0.3738	0.4586	22.69%
Case N	0.5891	0.5903	0.20%	0.3556	0.3986	12.09%
Case O	0.6673	0.6673	0	0.4150	0.4494	8.29%

CDNM thesis (discussed in Section 1) has implied that not all clustering problems in practice are difficult to solve. This research corresponds to the "clustering pipeline" proposed by Ackerman et al. [7], where clusterability (or structuredness in our context) can be evaluated to inform the selection of effective clustering algorithms. In this view, one intended contribution of this work is to implement this idea in the context of the CF problem. In future work, we will explore more applications in manufacturing systems that require grouping and combinatorial decisions (e.g., product and systems modularity). Also, we can explore more statistical and machine learning techniques such as multimodality tests and random forest to replace the K-S test for better predication performance.

Data Availability

The matrix data used to support the findings of this study are included within the supplementary information file (pictorial illustrations). Other data formats (e.g., Excel file) can be available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the NSERC Discovery Grants, Canada.

Supplementary Materials

One file of supplementary materials is included with the manuscript. This file contains the images of the matrix data for the problems presented in Section 6. *(Supplementary Materials)*

References

[1] N. L. Hyer and U. Wemmerlov, *Reorganizing the Factory: Competing through Cellular Manufacturing*, Productivity Press, Portland, OR, USA, 2002.

[2] M. J. Brusco, "An exact algorithm for maximizing grouping efficacy in part-machine clustering," *IIE Transactions*, vol. 47, no. 6, pp. 653–671, 2015.

[3] C. Liu, Y. Yin, K. Yasuda, and J. Lian, "A heuristic algorithm for cell formation problems with consideration of multiple production factors," *International Journal of Advanced Manufacturing Technology*, vol. 46, no. (9-12), pp. 1201–1213, 2010.

[4] M. Ackerman and S. Ben-David, "Clusterability: a theoretical study," in *Proceedings of the International Conference on Artificial Intelligence and Statistics, JMLR: Workshop and Conference Proceedings*, N. Lawrence, Ed., vol. 5, p. 18, 2009.

[5] E. Nowakowska, J. Koronacki, and S. Lipovetsky, "Clusterability assessment for Gaussian mixture models," *Applied Mathematics and Computation*, vol. 256, pp. 591–601, 2015.

[6] A. Daniely, N. Linial, and M. Saks, "Clustering is difficult only when it does not matter," *Computer Science (Machine Learning)*, 2012.

[7] M. Ackerman, A. Adolfsson, and N. Brownstein, "An effective and efficient approach for clusterability evaluation," *Computer Science (Machine Learning)*, 2016.

[8] Y. Zhu and S. Li, "Statistical Analysis of Similarity Measures for Solving Cell Formation Problems," *Procedia CIRP*, vol. 63, pp. 248–253, 2017.

[9] M. M. Paydar and M. Saidi-Mehrabad, "A hybrid genetic-variable neighborhood search algorithm for the cell formation problem based on grouping efficacy," *Computers & Operations Research*, vol. 40, no. 4, pp. 980–990, 2013.

[10] C. S. Kumar and M. P. Chandrasekharan, "Grouping efficacy: a quantitative criterion for goodness of block diagonal forms of binary matrices in group technology," *International Journal of Production Research*, vol. 28, no. 2, pp. 233–243, 1990.

[11] J. McAuley, "Machine grouping for efficient production," *Production Engineering Research and Development*, vol. 51, no. 2, pp. 53–57, 1972.

[12] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy: the Principles and Practice of Numerical Classification*, Freeman, San Francisco, Calif, USA, 1973.

[13] T.-H. Wu, C.-C. Chang, and J.-Y. Yeh, "A hybrid heuristic algorithm adopting both Boltzmann function and mutation operator for manufacturing cell formation problems," *International Journal of Production Economics*, vol. 120, no. 2, pp. 669–688, 2009.

[14] G. J. K. Nair and T. T. Narendran, "Grouping index: a new quantitative criterion for goodness of block-diagonal forms in group technology," *International Journal of Production Research*, vol. 34, no. 10, pp. 2767–2782, 1996.

[15] M. J. Brusco, "An iterated local search heuristic for cell formation," *Computers & Industrial Engineering*, vol. 90, pp. 292–304, 2015.

[16] B. R. Sarker and S. Mondal, "Grouping efficiency measures in cellular manufacturing: A survey and critical review," *International Journal of Production Research*, vol. 37, no. 2, pp. 285–314, 1999.

[17] B. R. Sarker, "The resemblance coefficients in group technology: A survey and comparative study of relational metrics," *Computers & Industrial Engineering*, vol. 30, no. 1, pp. 103–116, 1996.

[18] C. T. Mosier, J. Yelle, and G. Walker, "Survey of similarity coefficient based methods as applied to the group technology configuration problem," *Omega*, vol. 25, no. 1, pp. 65–79, 1997.

[19] Y. Yin and K. Yasuda, "Similarity coefficient methods applied to the cell formation problem: A taxonomy and review," *International Journal of Production Economics*, vol. 101, no. 2, pp. 329–352, 2006.

[20] C. Zhao and Z. Wu, "A genetic algorithm for manufacturing cell formation with multiple routes and multiple objectives," *International Journal of Production Research*, vol. 38, no. 2, pp. 385–395, 2000.

[21] F. M. Defersha and M. Chen, "A linear programming embedded genetic algorithm for an integrated cell formation and lot sizing considering product quality," *European Journal of Operational Research*, vol. 187, no. 1, pp. 46–69, 2008.

[22] S. Lee and H. P. Wang, "Manufacturing cell formation: A dual-objective simulated annealing approach," *The International Journal of Advanced Manufacturing Technology*, vol. 7, no. 5, pp. 314–320, 1992.

[23] R. Tavakkoli-Moghaddam, N. Safaei, and F. Sassani, "A new solution for a dynamic cell formation problem with alternative routing and machine costs using simulated annealing," *Journal of the Operational Research Society*, vol. 59, no. 4, pp. 443–454, 2008.

[24] G. Papaioannou and J. M. Wilson, "The evolution of cell formation problem methodologies based on recent studies (1997–2008): Review and directions for future research," *European Journal of Operational Research*, vol. 206, no. 3, pp. 509–521, 2010.

[25] C. Renzi, F. Leali, M. Cavazzuti, and A. O. Andrisano, "A review on artificial intelligence applications to the optimal design of dedicated and reconfigurable manufacturing systems," *International Journal of Advanced Manufacturing Technology*, vol. 72, no. (1-4), pp. 403–418, 2014.

[26] A. Stawowy, "Evolutionary strategy for manufacturing cell design," *Omega* , vol. 34, no. 1, pp. 1–18, 2006.

[27] G. J. Nair and T. T. Narendran, "CASE: A clustering algorithm for cell formation with sequence data," *International Journal of Production Research*, vol. 36, no. 1, pp. 157–180, 1998.

[28] G. W. Corder and D. I. Foreman, *Nonparametric Statistics: A Step-By-Step Approach*, John Wiley & Sons, Hoboken, NY, USA, 2014.

[29] A. P. Bradley, "ROC curve equivalence using the Kolmogorov-Smirnov test," *Pattern Recognition Letters*, vol. 34, no. 5, pp. 470–475, 2013.

[30] T. Arnold and J. Emerson, "Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions," *The R Journal*, vol. 3, no. 2, pp. 34–39, 2011.

[31] A. Justel, D. Peña, and R. Zamar, "A multivariate Kolmogorov-Smirnov test of goodness of fit," *Statistics & Probability Letters*, vol. 35, no. 3, pp. 251–259, 1997.

[32] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*, Wiley, Hoboken, NJ, USA, 5th edition, 2011.

[33] Z. Drezner and O. Turel, "Normalizing variables with too-frequent values using a Kolmogorov–Smirnov test: A practical approach," *Computers & Industrial Engineering*, vol. 61, no. 4, pp. 1240–1244, 2011.

[34] Y. J. Zhu, *Hierarchical Clustering and Similarity Statistics for Solving and Investigating Cell Formation Problems*, Department of Mechanical and Manufacturing Engineering, University of Calgary, 2017.