Supplemental Material to Missing Data Patterns Paper, published at SSDBM 2021

Michal Bechný, Lisa Ehrlinger*

May 2021

In 2021, we published our work on "Missing Data Patterns: From Theory to an Application in the Steel Industry" at the 33rd International Conference on Scientific and Statistical Database Management [Bec+21], available online here: `https://dl.acm.org/doi/proceedings/10.1145/3468791`.

Due to page limitations, we were not able to present our research in full length and therefore, would like to provide supplemental material in this document to facilitate repeatability and further evaluations.

This document consists of three sections: Section 1 covers an extend version of the related work, which supports further research in this direction, Section 2 repeats our approach on detecting missing data (MD) patterns and details on the two algorithms iBBiG and LBM, and Section 3 outlines the full original evaluation of our approach. All definitions of the formalism refer to the original paper and are not available in this document. For any further inquiries, please do not hesitate to contact us.

1 Overview on Missing Data Patterns

An observed data set $\mathbf{D}^{n \times m}$ is considered to be tabular and consists of n rows (observations) and m columns (variables). Unless mentioned otherwise, we assume rows to be independent and identically distributed (i.i.d.). The missingness of values in $\mathbf{D}$ is identified by binary matrix $\mathbf{M}^{n \times m}$, where $m_{ij} = 1$ if d_{ij} is missing, and $m_{ij} = 0$ otherwise. The index i refers to rows from the row set $\mathcal{I}$, and the indices j and l refer to columns from the column set $\mathcal{J}$. The $\mathbf{D}$ can be partitioned into $\mathbf{D} = (\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}^{n \times l} = (\mathbf{X}_1, ..., \mathbf{X}_l)$ stands for the set of fully observed variables and $\mathbf{Y}^{n \times k} = (\mathbf{Y}_1, ..., \mathbf{Y}_k)$ stands for the set of variables containing MD, such $m = l + k$. The $\mathbf{M}$ can be partitioned accordingly, $\mathbf{M} = (\mathbf{M}^{\mathbf{X}}, \mathbf{M}^{\mathbf{Y}})$, with zero-valued $\mathbf{M}^{\mathbf{X}}$ and binary $\mathbf{M}^{\mathbf{Y}}$, whose columns refer to Bernoulli processes (cf. [Cox17]).

There is no standardized list or agreement on the meaning of MD patterns in literature (cf. [BH14; LHH03; IS08; LR02; SG02]). Van Buuren [Van07]

*Corresponding author: lisa.ehrlinger@scch.at

introduces a MD pattern briefly as $\mathbf{M}$ without any further explanation. Enders [End10] describes a MD pattern as *"a configuration of observed and missing values in a data set"*. Independently of the author, at least an agreement that the MD patterns should be checked and considered, holds for all. In a recent work, Fernstad [Fer19] argues that a proper understanding of MD patterns and the distribution of $\mathbf{M}$ can improve the quality of conducted analysis. She describes three general patterns of missingness [Fer19]:

- *Amount Missingness* indicating the proportion of missing observations per variable,
- *Joint Missingness* related to the co-occurrence of MD among two or more variables, and
- *Conditional Missingness* denoting an existing dependency between some components of $\mathbf{M}$ and $\mathbf{X}$.

According to the recognized literature (cf. [Rub76; Van07]), the last mentioned type is rather related to the MD mechanism, which is defined as the dependence between data and their missingness. In the following, we discuss MD patterns that commonly appear in related work as well as three patterns that are of special interest for the industrial domain: *file-matching pattern (FMP)*, *line pattern (LP)*, and *multi-rate pattern*.

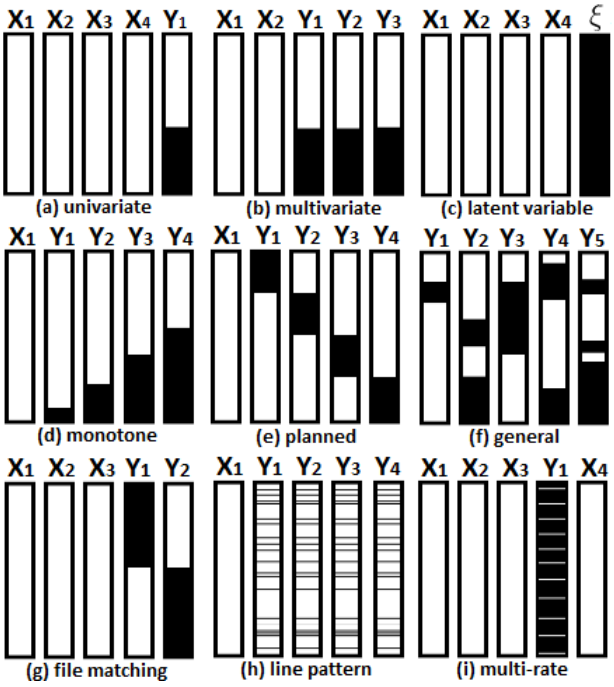

Figure 1: Missing data patterns

1.1 Common Missing Data Patterns

1.1.1 Univariate and multivariate pattern

The univariate pattern in Figure 1(a) is the simplest case, where exactly one variable in **D** contains MD [IS08; LR02; SG02]. Although Little & Rubin [LR02] described the multivariate pattern in Figure 1(b) separately, we claim that it can be seen as a multivariate generalization of the univariate case. An example for these patterns from the industrial domain is a failure of one or several sensors, which jointly stop recording measurements after a specific point in time.

1.1.2 Latent variable pattern

The latent variable pattern in Figure 1(c) is typical for generative models, where the latent variable ξ is missing for all observations, because its existence is only assumed. It is not necessary to understand this MD pattern as a problem, because it is used intentionally [End10]. Little & Rubin [LR02] mention that it can be useful to treat certain problems with completely unobserved variables to estimate the parameters of generative models using the methods of statistical inference.

1.1.3 Monotone pattern

The monotone pattern in Figure 1(d) is frequently observed in the social sciences, where participants tend to leave a study that is conducted over time [BH14; LR02]. This pattern has no significant relevance in industrial data [Ehr+18].

1.1.4 Planned pattern

The planned pattern in Figure 1(e) is common in designed experiments, where recording of all variables jointly is impossible, too expensive or causes burden [End10]. The missingness is in this case under the control of investigator.

1.1.5 General pattern

The general pattern in Figure 1(e) is the default case that is usually found in practice and is also denoted as "arbitrary pattern" (cf. [SG02; LHH03]) or "generalized pattern" (cf. [BH14]). It is also the most difficult pattern to handle, because it is typically a combination of more MD patterns together with non-systematically missing values.

1.2 Industry-Specific MD Patterns

In addition to MD patterns listed above, we want to highlight patterns that are of special importance in the industrial domain: FMP, LP, and multi-rate pattern.

1.2.1 File matching pattern

The FMP in Figure 1(g) was introduced by Little & Rubin [LR02] as a case when two variables are never observed jointly. We argue that FMP is of special importance nowadays, when data is queried from multiple information systems. The FMP is typically caused by joining data from heterogeneous sources with different column dimensions. This yields blocks of MD over several variables, which can be viewed as a generalization of how FMP was originally introduced in [LR02].

1.2.2 Line pattern

The term "line pattern" was coined in [Ehr+18] to describe the previously mentioned "sensor breakdown" pattern from [IS08] more generally. The LP, which can be after suitable rearrangements of columns seen as a "line" of MD, is typically caused by a breakdown of multiple sensors (cf. [IS08]) due to extreme physical state in the process environment. Based on our experience gained from analyzing data from the steel industry by voestalpine Stahl GmbH, we distinguish between three LP subtypes, depending on the occurrence of MD, as illustrated in Figure 2.

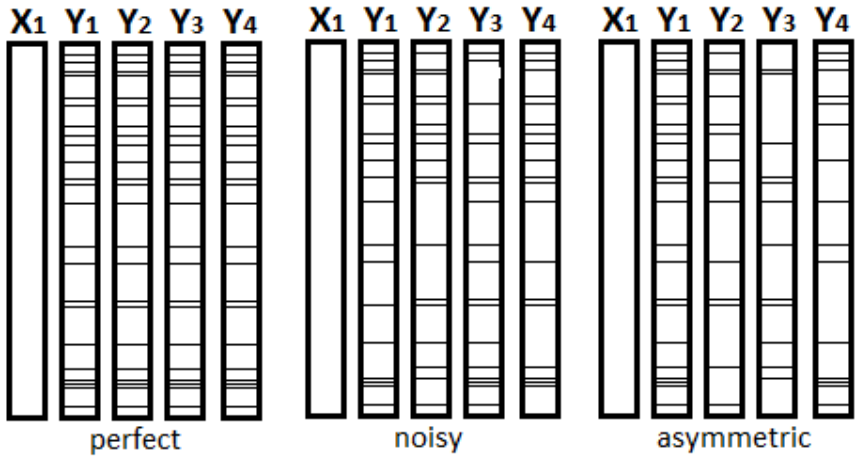

Figure 2: Different types of line pattern

Based on the sensitivity of communicating sensors, the LP can be either perfect, noisy, or asymmetric. A perfect LP refers to the joint missingness of a set of observations over a set of variables, typically communicating sensors measuring the same physical instance. The variables affected by a perfect LP are "perfectly" associated in terms of their (jointly) missingness on a subset of observations. In case of a noisy LP, the affected variables are highly but not perfectly associated, meaning that a missing value in one sensor is not necessarily accompanied with the missingness of all other related variables. The asymmetric

LP refers to a set of physically related variables (communicating sensors), when some of them are missing more often than others, which corresponds to different sensitivity of the respective sensors.

1.2.3 Multi-rate pattern

Another common MD pattern from the industrial domain is the multi-rate pattern introduced by [IS08] (cf. Figure 1(f)), where the variable $\mathbf{Y}_1$ is measured less often and consequently, regularly reoccurring MD are produced. In accordance to [IS08], we agree that it doesn't necessarily represent an error, since the values are not observed intentionally.

1.3 Summary on MD Patterns

Although some MD patterns in Figure 1 have similar appearance after aligning the rows, we argue that it is still necessary to distinguish them. Specifically, the following properties add understanding for their distinction: *amount of MD* (e.g., 1–10 % in LP vs. ˜90 % in multi-rate pattern), *reason for MD* (e.g., intention for planned pattern, participant's drop-out for multivariate pattern, DB query for FMP, and extreme temperature for LP), as well as the *type of association* between affected variables. Since the detection of MD patterns is a preceding step to deleting or imputing MD, it is important to distinguish between the types of patterns to select an appropriate action.

The fact that several MD patterns in Figure 1 have a similar appearance also illustrates the inconsistency in MD research due to the absence of a common formalization.

2 Approach to Detect MD Patterns

Traditional strategies for detecting MD patterns are often not appropriate for data of higher dimensionality, which is a rule rather than an exception in automatically collected industry data. Examples for such strategies are the visual inspection of an aggregation plot of $\mathbf{M}^{\mathbf{Y}}$ or the manual verification of different combinations between missing and observed components of the data [LR02; Van07]. Even the investigation of $\mathbf{M}^{\mathbf{Y}}$ with "only" 20 variables can yield up to 2^{20} combinations, making such manual strategies for MD pattern detection hardly feasible.

Therefore, we propose a new approach to automate detection of any MD pattern, with a particular focus on patterns appearing in industrial applications. The core idea is to discover significant regularities in the structure of $\mathbf{M}^{\mathbf{Y}}$. Figure 3 illustrates an iterative procedure with the following four steps:

Step (1) As a prerequisite for MD pattern detection, $\mathbf{D}$ is uniquely transformed to $\mathbf{M} = (\mathbf{M}^{\mathbf{X}}, \mathbf{M}^{\mathbf{Y}})$ using the knowledge about the encoding of MD.

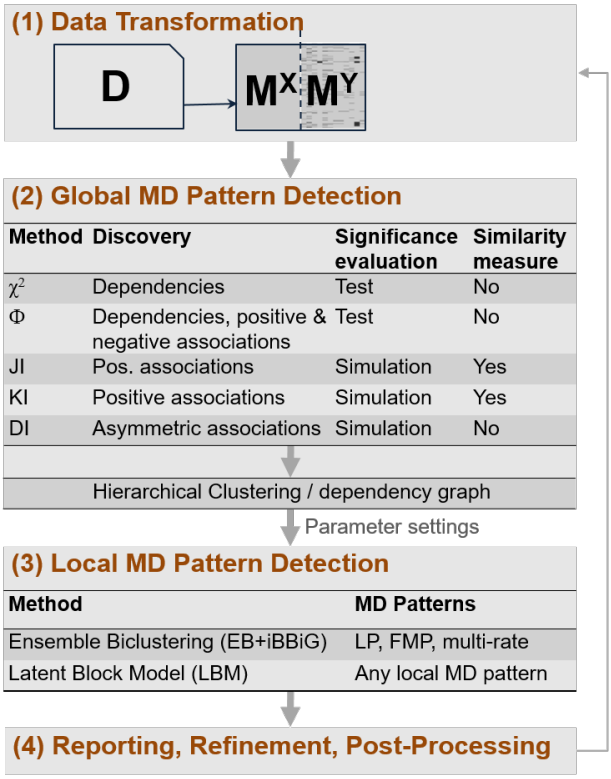

Figure 3: Approach to detect MD patterns in the steel industry

Step (2) To detect global MD patterns according to Def. 3.1 in [Bec+21], associations and dependencies among $\mathbf{M^Y}$ columns are investigated. Significant relations found are used as an input for hierarchical clustering or a dependency graph to effectively identify subset(s) of variables in accordance with Def. 3.1 in [Bec+21]. While an overview on the measures with their properties is provided in Figure 3, they are introduced in Section 2.1 and evaluated in Section 3.3.2.

Step (3) To detect local MD patterns according to Def. 3.5 in [Bec+21], two biclustering methods are introduced in Section 2.2: *Ensemble Biclustering using iterative Binary Biclustering for gene sets* (EB+iBBiG) and the *Latent Block Model* (LBM). Biclustering finds associations simultaneously in rows and columns of $\mathbf{M^Y}$, which makes it suitable to detect local MD patterns. Both methods are evaluated and compared in Section 3.3.3. The results from step (2) can be used to select proper setting for the biclustering.

Step (4) The reported results are used to refine the data set, to reformulate

database queries, or to impute MD.

While the entire approach is presented in Figure 3, the focus in this paper is on step (2) and (3). Step (1) is implemented according to the requirements for voestalpine Stahl GmbH (details in Section 3.3.1) and step (4) is non-trivial and therefore planned as follow-up research to this work.

2.1 Global MD Pattern Detection

To identify global MD patterns according to Def. 3.1 in [Bec+21] and step (2), the dependencies and associations among the $\mathbf{M^Y}$ variables need to be evaluated. In the following subsections, we discuss the following five measures: chi-square test (χ^2), Φ coefficient, Jaccard index (JI), Kulczynski index (KI), and Dice index (DI).

The two well-known dependency measures χ^2 and Φ were selected to evaluate Def. 3.1 in [Bec+21] ($\rightarrow \chi^2$) and to interpret the properties of the MD pattern ($\rightarrow \Phi$). As outlined in Figure 3, the latter three measures (JI, KI, DI) focus on the detection of associations. JI and KI are used to detect the co-occurrence of MD, which refers to industry-specific MD patterns or is an indicator to consider joint imputation for $\mathbf{M^Y}$. DI evaluates the symmetry of MD patterns, which could be either used for their interpretation, or the order in which the variables in $\mathbf{M^Y}$ should be imputed. Since JI, KI, and DI do not follow any well described probability distribution, there is no tabulated test to evaluate their statistical significance. To resolve this issue, we use a bootstrap resampling method introduced in Appendix A.1.

Since all five measures are intended for pairwise comparison, we use them in combination with hierarchical clustering or a dependency graph to identify subset J (cf. Def. 3.1 in [Bec+21]). Hierarchical clustering groups a pair of variables with the minimum distance in each clustering step [JMF00] and can also be used for the automated analysis of the dependency graph. The dependency graph provides a more intuitive visual presentation of the results (e.g., for discussion and verification with a customer) and provides further insights into the missingness structure. Both methods (i.e., hierarchical clustering and the dependency graph) allow the automated detection of global MD patterns by "cutting" the graph or dendrogram with a threshold for the strength of the given association.

The following subsections discuss the properties (completeness from Def. 3.2 in [Bec+21], minimality from Def. 3.3 in [Bec+21], and symmetry from Def. 3.4 in [Bec+21]) and suitability of five measures for our use case: χ^2, Φ, JI, KI, and DI. To achieve a common terminology, we derived a functional form for each measure from the occurrence of MD for an arbitrary pair of $\mathbf{M^Y}$ variables (summarized in Table 1).

2.1.1 χ^2-test

The χ^2-test [Coc52] can be applied to determine whether two binary variables are dependent. Thus, Def. 3.1 in [Bec+21] can be directly tested by applying

Table 1: Frequency table, where n_{11}/n_{00} indicate the number of cases when both variables j, l are missing/observed jointly and n_{10}/n_{01} indicate the number of cases when one variable (j or l) is missing and the other is observed

	$\mathbf{M}_{.l} = 1$	$\mathbf{M}_{.l} = 0$	
$\mathbf{M}_{.j} = 1$	n_{11}	n_{10}	$n_{1+} = n_{10} + n_{11}$
$\mathbf{M}_{.j} = 0$	n_{01}	n_{00}	$n_{0+} = n_{00} + n_{01}$
	$n_{+1} = n_{01} + n_{11}$	$n_{+0} = n_{10} + n_{01}$	

the χ^2-test to all pairs of $\mathbf{M}^{\mathbf{Y}}$ variables. For testing the null hypothesis of independence ($= H_0$), the χ^2-statistics is used:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}},$$

where $\mu_{ij} \overset{H_0}{=} \frac{n_{i+}n_{+j}}{n} = E(n_{ij})$ under the assumption that H_0 is true [Coc52]. The statistics is under H_0 approximately $\chi^2_{df=1}$ distributed, and the H_0 is rejected in favour of alternative that these variables are associated if p-value $= P(\chi^2 \geq \chi^2_{df=1,1-\alpha} | H_0 = \text{true}) \leq \alpha$, where α is a predefined significance level, typically $\alpha = 0.05$.

2.1.2 Coefficient Φ

For any pair of $\mathbf{M}^{\mathbf{Y}}$ variables the coefficient Φ is calculated as:

$$\Phi(\mathbf{M}_{.j}, \mathbf{M}_{.l}) = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1+}n_{0+}n_{+1}n_{+0}}}$$

and its value as well as interpretation is the same as by Pearson's correlation coefficient [Cra99]. Coefficient Φ takes values from $\langle -1, +1 \rangle$ and evaluates both the strength and direction of the linear dependency between variables. Pairwise independence of variables implies $\Phi = 0$, but observing $\Phi = 0$ does not generally imply independence. Coefficient Φ is functionally related to the χ^2-statistics, $\Phi^2 = \frac{\chi^2}{n}$, thus its significance can be evaluated in the same way. The Coefficient Φ is used in our approach to detect arbitrary global MD patterns. In addition to χ^2-test, Φ evaluates the direction and the strength of the association, which supports the interpretation of the detected patterns.

2.1.3 Jaccard index

Jaccard index (JI) is a similarity measure defined as the ratio between the cardinalities of the variables' intersection and the union [Jac12]:

$$\text{JI}(\mathbf{M}_{.j}, \mathbf{M}_{.l}) = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}.$$

If these cardinalities are equal, there is a perfect overlap between events of MD of the two variables and JI = 1. Contrary, JI is minimal and equals 0, if the

variables never had MD jointly. The JI is therefore suitable for the identification of co-occurrence between MD events, which refer to a positive association, as it is the case with LP, FMP, or multivariate pattern. However, JI is not suitable for the detection of disagreement (negative) relationships, such as depicted in Figure 1(e).

2.1.4 Kulczynski index

The Kulczynski index (KI) is a similarity measure defined as an average of two conditional probability estimates [Des13]:

$$\begin{aligned}\mathrm{KI}(\mathbf{M}_{.j}, \mathbf{M}_{.l}) &= \frac{\widehat{P}(\mathbf{M}_{.j} = 1|\mathbf{M}_{.l} = 1) + \widehat{P}(\mathbf{M}_{.l} = 1|\mathbf{M}_{.j} = 1)}{2} \\ &= \frac{1}{2}(\frac{n_{11}}{n_{11} + n_{01}} + \frac{n_{11}}{n_{11} + n_{10}}).\end{aligned}$$

Similarly to JI, the KI evaluates pairwise similarity based on the co-occurrence and synergy of MD and has the advantage of probabilistic interpretation.

2.1.5 Dice indices

The Dice1 and Dice2 indices are related to the two terms of KI, denoting the respective estimates of conditional probability:

$$\mathrm{Dice1}(\mathbf{M}_{.j}, \mathbf{M}_{.l}) = \widehat{P}(\mathbf{M}_{.j} = 1|\mathbf{M}_{.l} = 1) = \frac{n_{11}}{n_{11} + n_{01}},$$

$$\mathrm{Dice2}(\mathbf{M}_{.j}, \mathbf{M}_{.l}) = \widehat{P}(\mathbf{M}_{.l} = 1|\mathbf{M}_{.j} = 1) = \frac{n_{11}}{n_{11} + n_{10}}.$$

Since the triangular inequality does not hold for any of these indices, they cannot be considered as a proper similarity measures. Comparing Dice1 and Dice2 indicates whether pairs of related variables are associated symmetrically according to Def. 3.4 in [Bec+21]. The monotone and asymmetric LP are examples of non-symmetric patterns for which Dice indices are particularly useful.

2.2 Local MD Pattern Detection

In this section, we discuss two biclustering methods for the detection of local MD patterns according to Def. 3.5 in [Bec+21] and step (3) in Figure 3. Biclustering is an unsupervised method simultaneously grouping rows and columns of a data set [Kas+16], which make it suitable to detect local MD patterns. A bicluster $\mathbf{B}(\mathcal{I}, \mathcal{J})$ is a submatrix of data defined by a subset of rows, $\mathcal{I} \subseteq \{1, ..., n\}$, which satisfy certain similarity constrains on a subset of columns $\mathcal{J} \subseteq \{1, ..., m\}$.

In our application, we found biclustering highly effective to identify data cells d_{ij} missing due to either a LP or a FMP (see Section 3). The parts of $\mathbf{M}^{\mathbf{Y}}$ affected by a FMP typically built a perfect concentrated bicluster consisting of only 1's. Further, biclusters corresponding to a LP are dense (having a large proportion of 1's) with possibly containing a 0-valued noise, which depends on

the type of the LP. Both FMP and LP yield biclusters with a strong positive association between the respective subsets $\mathcal{I}$ and $\mathcal{J}$. The biclusters consisting of mostly 0's are of little importance for MD pattern detection, since they correspond to the observed part of the data.

Although several biclustering approaches have been proposed, little of them exist for binary data. One of the most widely used binary biclustering algorithm is BIMAX (cf. [Pre+06]), where the objective is to identify perfect biclusters of the maximal size. BIMAX is not useful for our application since zero-valued cells might appear within biclusters referring to local MD patterns. Therefore, we compared and evaluated the following two biclustering techniques: *iBBiG* [Gus+12] and *Latent Block Model* [GN03]. The iBBiG algorithm is suitable for the fast extraction of dense biclusters from a sparse binary matrix, which makes it particularly useful for the detection of positively associated MD patterns. We applied iBBiG within the framework of *ensemble biclustering* (EB) to increase the robustness of the result and refer to the combination of both techniques in the following as EB+iBBiG. Second, we evaluated the LBM – which aims to identify the distribution of $\mathbf{M}^{\mathbf{Y}}$ – to detect arbitrary types of MD pattern.

2.2.1 Ensemble Biclustering using iBBiG

The iBBiG (iterative Binary Biclustering of Gene sets) was proposed as a meta-analytic tool for discovering associations in sparse binarized genetical data sets [Gus+12]. iBBiG has several features that make it suitable for our work: (a) it extracts dense biclusters from the data, (b) it allows noise within biclusters, and (c) it allows overlapping biclusters [Gus+12]. In contrast to the majority of biclustering methods, a priori knowledge of the number of biclusters is not required. The iBBiG algorithm is an heuristic iterative method that consists of 3 main steps: (i) *bicluster fitness score*, (ii) *heuristic search* based on a genetic algorithm (GA) to effectively identify the biclustering solution in a high dimensional space, and (iii) *iterative extraction* to mask the signal from already detected biclusters [Gus+12].

The algorithm starts by randomly selecting two columns of $\mathbf{M}^{\mathbf{Y}}$, which are then used to select highly associated rows to form an initial bicluster. Following, a fitness score S_B (based on the size and homogeneity) is calculated for each inital bicluster to evaluate its quality. To illustrate the calculation of the fitness score, we assume that $\mathbf{B}^*(\mathcal{I}_r, \mathcal{J}_c)$ is an initial bicluster, having $r = |\mathcal{I}_r|$ rows and $c = |\mathcal{J}_c|$ columns, where $2 \leq c \leq k$. The following quantities are calculated for all rows to obtain S_{B^*}:

- The estimate of the probability of association between the i-th row and columns of $\mathbf{B}^*(\mathcal{I}_r, \mathcal{J}_c)$:
$$\hat{p}_i = \frac{1}{c} \sum_{j \in \mathcal{J}_c} m_{ij}. \tag{1}$$
- The entropy of $\hat{p}_i$:
$$H_i = -\hat{p}_i \log_2 \hat{p}_i - (1 - \hat{p}_i) \log_2 (1 - \hat{p}_i). \tag{2}$$

- The fitness-score of the i-th row:

$$S_i = \begin{cases} W_{iB^*}(1 - H_i)^\alpha & \text{if } \hat{p}_i > 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The S_{B^*} is then evaluated as: $S_{B^*} = \sum_i S_i$. The parameter $\alpha \in \langle 0.1, 1 \rangle$ from Eq. 3 scales the score and by default $\alpha = 0.3$, which is according to simulation studies by the authors optimal [Gus+12]. The weight term W_{iB^*} can be expressed as $W_{iB^*} = \sum_{j \in \mathcal{J}_c} w_{ij}$, where the entire weight matrix $\mathbf{W}$ is initialized by $\mathbf{M}^{\mathbf{Y}}$.

The procedure continues with the GA to optimize the growth or reduction of the biclusters with respect to S_{B^*}. The algorithm alternates between the GA and recalculation of S_{B^*} until the score stagnates for a specific number of iterations and the final bicluster is identified. The iterative extraction allows to find more than one bicluster in $\mathbf{M}^{\mathbf{Y}}$. Once a bicluster $\mathbf{B}^*(\mathcal{I}_r, \mathcal{J}_c)$ is identified, the respective weights of $\mathbf{W}$ are subtracted according to

$$W_{ij} \leftarrow W_{ij}(1 - (1 - H_i)^\alpha) \quad \forall j \in \mathcal{J}_c \tag{4}$$

and the entire procedure is applied again. Gusenleitner et al. [Gus+12] suggest to specify an upper bound for the expected number of biclusters since iBBiG terminates automatically when no more bicluster can be found [Gus+12]. We use the number of global MD patterns from step (2) as upper bound of biclusters to be identified.

Since the results of iBBiG are influenced by random initialization as well as by a heuristic GA, we applied the algorithm within the framework of EB [Shi+10], to increase the robustness of the solution. The core idea of ensembling is to run the desired algorithm multiple times under different starting seeds and/or parameter settings and subsequently to identify a hierarchy of solutions based on their pairwise similarity. Finally, robust biclusters are identified by intersecting the subsets of the solutions that satisfy certain homogeneity constrains. For our application, we used a similar procedure as proposed in [Kha13] and applied EB with iBBiG as follows:

- iBBiG is executed N-times to identify K biclusters in each run.
- The results are combined into a *bicluster set*.
- All identified biclusters are pairwise compared with the JI, yielding a similarity matrix.
- Based on the JI-matrix, a dendrogram of solutions is created using hierarchical clustering with complete linkage.
- Groups (= clusters, branches) of similar biclustering solutions are identified using a suitable similarity threshold.
- Robust biclusters are identified as an intersection of groups with sufficient cardinality C.

2.2.2 Latent Block Model

As a second approach to detect MD patterns from $\mathbf{M^Y}$, we evaluated binary LBM, originally introduced in [GN03]. In contrast to iBBiG, the objective of LBM is to identify the underlying distribution of $\mathbf{M^Y}$ by applying a mixture model jointly on its rows and columns. Therefore, LBM is very suitable for our application, since knowledge of the $\mathbf{M^Y}$ probability distribution is key for MD pattern recognition. As a result of LBM, $\mathbf{M^Y}$ can be segmented into several exclusive row- and column-classes which jointly partition $\mathbf{M^Y}$ into homogeneous blocks (= biclusters, co-clusters). More specifically, the LBM assumes existence of partitioning of both index sets $\mathcal{I}$ and $\mathcal{J}$ into Q-row and L-column classes, which jointly express a block-structure that can be viewed as a compressed information about the distribution of $\mathbf{M^Y}$. Thus, it can be used to identify arbitrary local MD pattern according to Def. 3.5 in [Bec+21]. For example, dense blocks can be used for the detection of rows and columns affected by multivariate pattern, FMP or different types of LP. Further, resulting blocks consisting of (mostly) 0's can be interpreted as parts of complete data which are unlikely to be a part of any local MD pattern. The analysis of dependencies between individual column classes can be used to describe global MD patterns as well, which makes LBM universally applicable in terms of MD pattern detection.

For the LBM specification, it is key to determine reasonable values for Q and L. For this purpose, we suggest to fit the LBM for each combination from the grid based on the number of global MD found in step (2), and to select the combination with the largest Bayesian information criterion (BIC):

$$\begin{aligned} BIC(Q,L) = \max_{\boldsymbol{\theta}} \log f(\mathbf{M^Y};\boldsymbol{\theta}) - \frac{Q-1}{2}\log(n) \\ - \frac{L-1}{2}\log(k) - \frac{QL}{2}\log(kn), \end{aligned} \tag{5}$$

where the first term is substituted by the result from the fitted model (cf. [Ker+15]). Since the details about statistical estimation of the LBM are beyond the scope of this paper, we refer to the literature [BIG14; GN13; NH98].

3 Implementation and Evaluation

In this section, we describe the application scenario and the employed data sets in Section 3.1, the prototypical implementation in Section 3.2, and the conducted experiments in Section 3.3. We conclude with a summary of the findings.

3.1 Application Scenario with Steel Data

In the steel mill of voestalpine Stahl GmbH, raw steel is casted into slabs. Those slabs are about 12 meters long and 200 millimeters thick. The slabs are heated and rolled in the hot rolling mill to receive steel coils with the desired thickness

for further processing. Such a steel coil can be rolled out up to 0.7 mm thickness and an average length of 1,600 meters. After the hot rolling, the steel is cooled down with water and wound into coils [Ehr+18].

The data sets used for our empirical evaluation contain sensor data from the hot rolling mill, which includes mainly temperature measures and information about the water cooling system. We faced the following challenges and characteristics of the data:

- The provided data sets can be attributed as *Big Data* since our evaluation data set, which represents only an excerpt of the entire hot rolling mill measurement DB, contains 193,700 records and 557 variables. Such large amounts of data highlight the need for automated approaches and novel ways to display the results.
- The data is automatically collected *sensor data*, which is why a lot of suggestion from MD research from social sciences (e.g., [LHH03]) cannot be applied.
- Since we mostly deal with *numerical data*, approaches for survey MD that include suggestions how to handle textual or discrete data are not relevant.

Voestalpine Stahl GmbH provided 3 data sets for our research, which are summarized in Table 2. DS1 contains process data from the steel mill with no knowledge about the causes of the missingness or the type of MD patterns. It was used mainly for the investigation of suitable techniques for the analysis of MD in large industrial data sets.

Table 2: Evaluation data sets, where k is the number of columns in $\mathbf{M}^{\mathbf{Y}}$ and Sim.=Simulated

DS	**Rows**	**Columns**	k	**Sim.**	**Patterns**	**Noise**
DS1	193,700	557	22	No	Unknown	No
DS2	2,641	1,125	73	Yes	3 FMP, 4 LP	5 %
DS3	2,500	1,000	108	Yes	18 LP	5 %

DS2 and DS3 contain several noisy LPs and FMPs simulated by domain experts from voestalpine Stahl GmbH to evaluate our approach. Since knowledge of the "underlying truth" about MD patterns is usually not available in practice, we decided to conduct a simulation study to accurately evaluate our results. The amount of the noise in the simulated LPs varied from 0 to 80 %. $\mathbf{M}^{\mathbf{Y}}$ of DS2 and DS3 contained 49 and 72 signal variables respectively, which follow one of the simulated patterns, and additionally 24 and 36 noisy variables, respectively, which do not follow any specific MD pattern and consist of 5 % binary noise only. 5 % background noise was added to the signal variables to assess the extent to which individual algorithms extract the "patterns" from the "noise". According to the domain knowledge of experts from voestalpine Stahl GmbH, there is usually no or a rather small proportion (0–3 %) of noise in the variables that follow an industry-specific MD pattern. Therefore, the 5 % noise used for evaluating the robustness of our method, reflects the worst case scenario for our application domain.

For both DS2 and DS3, the description of the true MD patterns was provided to us *post hoc* within the reference matrices $\mathbf{M}^{\mathbf{Y}}_{\text{ref}}$, whose cell (i,j) is equal to 1, if the corresponding cell of $\mathbf{M}^{\mathbf{Y}}$ belongs to some of the LP/FMP simulated, and equals 0 otherwise, i.e., corresponding cell is either complete or noisy, i.e., missing non-systematically. For a clear presentation of the results, the names of variables following some MD patterns begin with prefix "x" whereas the noisy variables begin with "n".

3.2 Implementation Details

Our approach has been implemented in `R`. The most important packages are: `biclust` [Kai+18] as the basis package to apply biclustering in `R`, `iBBiG` [GC19] to apply iBBiG, `superbiclust` [Kha14] since it implements the concept of Ensemble Biclustering [Kha14], and the package `blockcluster` [SIG17] for the application of the LBM. Further, we used `pheatmap` [Kol19] to visualize the missingness matrices and `qgraph` [Eps+12] to plot the dependency graphs. The implementation has been successfully deployed and tested at the statistical department of voestalpine Stahl GmbH.

3.3 Experiments and Evaluations

In this section, we discuss the application of our MD pattern detection approach with a specific focus on detection of variables and observations affected by LP and FMP. However, most of the steps of our analysis are applicable to identify any kind of MD pattern. All experiments were conducted with an Intel core i7-8700 CPU, 3.2 GHz, and 32 GB RAM. The analysis of DS1, DS2, and DS3 follows the steps described in Figure 3.

3.3.1 Data Transformation

For step (1), the $\mathbf{M}^{\mathbf{Y}}$ was obtained for each data set using knowledge about the encoding of MD by voestalpine (here: -99). Although existing metrics (cf. [Hin02]) to measure missing data are often restricted to the detection of `null` values, in practice, also default values such as "01/01/2000" or "NaN" values need to be considered. Our real-world evaluation yielded an interesting insight, which underpins this statement: during the transformation of DS2, we discovered (initially unknown) `NA` values, which were all related to FMPs (B2.1, B2.2, and B2.7 in Table 6). This step deserves more attention in future research.

3.3.2 Detection of global MD patterns

For step (2), we investigated the global MD patterns. Since no specific MD patterns were expected for DS1, we first evaluated the correlation, which is the most widely applicable measure from Section 2.1. The only significant value of $\Phi = 0.35$ was found for {x224, x245} and refers to a positively associated global MD pattern. Further, we investigated whether some variables tend to have MD

jointly using KI, which has the advantage of probabilistic interpretation. Figure 4 shows the dependency graph of the corresponding similarity matrix. The variables are depicted as the nodes and the edges refer to significant associations between them. A strong association corresponds to a short distance between the nodes and a saturated color of the edge. The most interesting finding is that KI(x224, x245) = 0.37, which refers to a high probability that values in x224 and x245 are missing jointly and confirms the pattern detected by correlation. Further, rather weak associations were identified for {x196, x555}, {x508, x90}, {x80, x245}, and {x23, x244}. All pairs of associated variables correspond to global MD patterns which are symmetric and minimal. Only {x196, x555} and {x23, x244} are also complete.

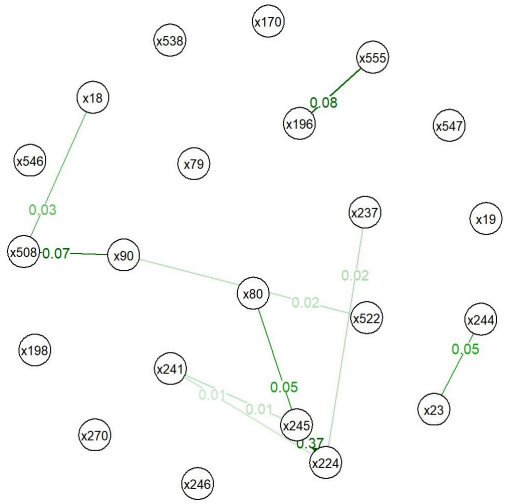

Figure 4: Dependency graph based on Kulczynski similarity for DS1

For DS2, we used KI and DI, since both are suitable to detect LPs and FMPs, which were expected to be present in this data set. Figure 5 shows the directed dependency graph based on DI. Table 3 summarizes the identified global MD patterns in DS2, which are all minimal and not complete. We further determined that GP1, GP2, and GP3 contain pairs of variables with significant difference in Dice indices (cf. Figure 5), which means that they are not symmetric.

With an increasing number of variables, DS3 (having 108 variables in $\mathbf{M}^{\mathbf{Y}}$) clearly showed the limitations of the dependency graph visualization, which is why hierarchical clustering is better suited for such use cases. Figure 6 shows the cluster dendrogram based on KI matrix, which yields 17 global MD patterns when cutting the height at a similarity level of 0.8, meaning that respective pairs of variables likely have MD jointly. All of these patterns are minimal. The pair {x632, x86} on the absolute left side of the dendrogram is the only complete MD pattern we found, since its variables are completely dissimilar to the remaining ones. No asymmetric patterns were identified using Dice indices.

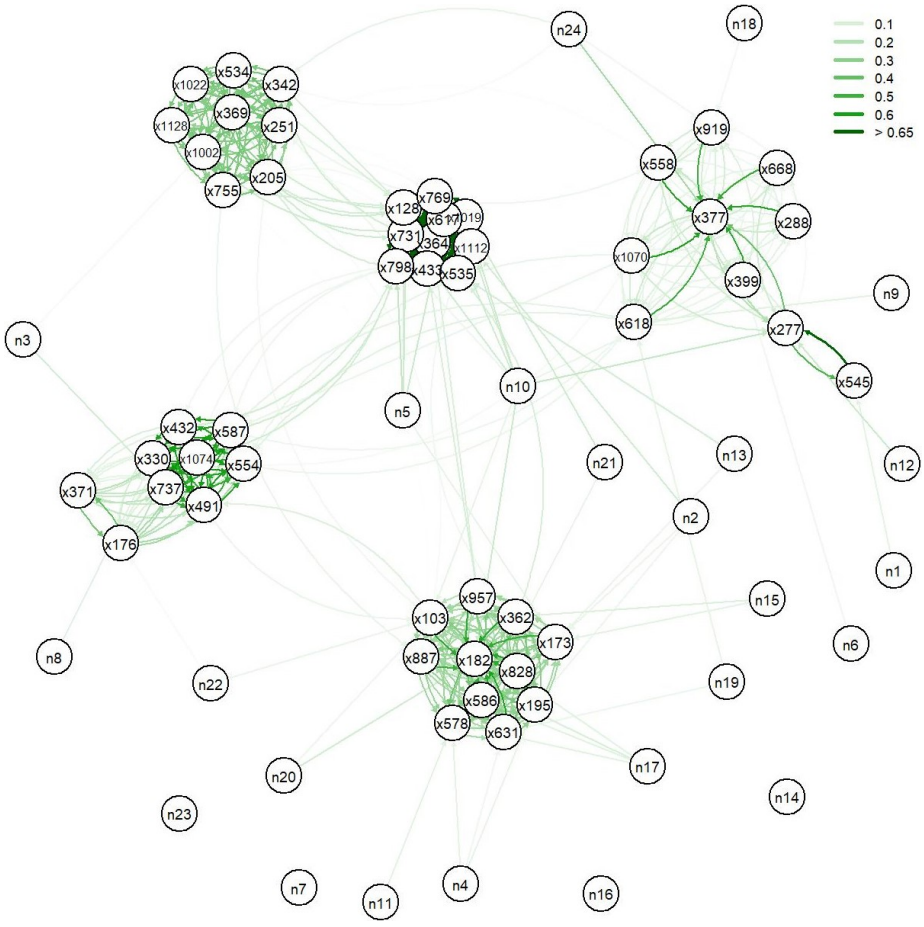

Figure 5: Directed dependency graph based on Dice indices for DS2

3.3.3 Detection of local MD patterns

Our procedure continues with step (3), the detection of local MD patterns using EB+iBBiG and LBM. The evaluation focuses on the detection of simulated LPs and FMPs, which were of primary interest for our company partner. LPs and FMPs can also be described as positively associated local patterns or "dense" biclusters.

Evaluation setup. We use the knowledge about the number of global patterns from step (2) to specify: (a) an upper bound for the expected number of biclusters for iBBiG, as well as (b) a range for the column classes for the LBM. Table 4 summarizes the setup of the EB+iBBiG parameters, which were introduced in Section 2.2.1, for all three data sets and includes the number of

Table 3: Global MD patterns found in DS2

Global Pattern	Variables	Symmetry
GP1	x182, x887, x103, x957, x362, x173, x828, x195, x631, x586, x579	No
GP2	x377, x1070, x558, x919, x668, x288, x399, x277, x618	No
GP3	x277, x545	No
GP4	x535, x1112, x731, x128, x1019, x769, x617, x798, x364, x433	Yes
GP5	x534, x369, x1022, x205, x1128, x342, x1002, x251, x755	Yes
GP6	x330, x432, x491, x554, x587, x737, x1074	Yes
GP7	x371, x176	Yes

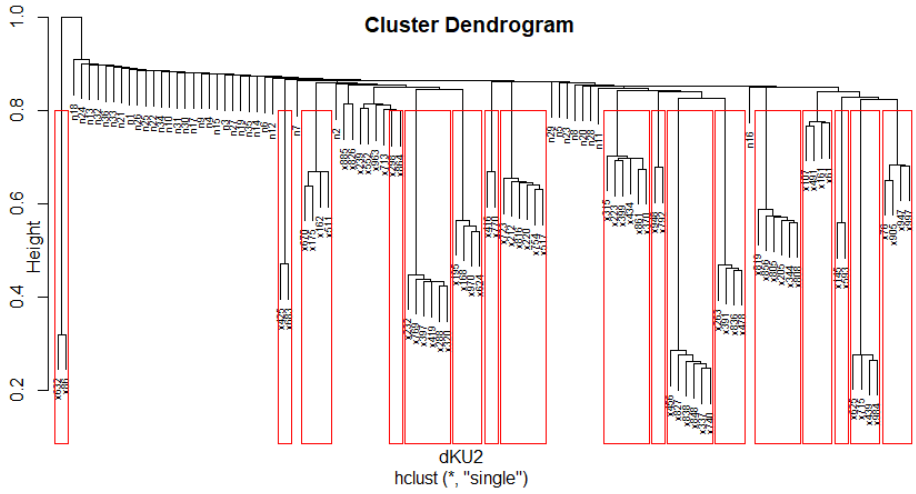

Figure 6: Hierarchical clustering based on Kulczynski similarity for DS3

identified robust biclusters. Since none or only little local MD patterns were expected to be present in DS1, both, the number of iBBiG repetitions (N) and the number of biclusters to store from each iBBiG run (K) were chosen rather small. Since both, DS2 and DS3, were expected to contain a larger number of (simulated) local MD patterns, the values of N and K were set accordingly higher. The cardinality threshold C equal to K and a strict JI threshold of 0.8 guarantee that each resulting bicluster is robust since it needs to occur in each iBBiG run.

Table 4: Parametrization of EB+iBBiG based on the identified global MD patterns

DS	N	K	**JI Threshold**	C	**Identified Biclusters**
DS1	5	5	0.8	5	4
DS2	30	15	0.8	15	7
DS3	30	20	0.8	20	18

Similarly, using knowledge about the number of global MD patterns found, we setup the ranges for the row- and column-classes Q and L for which the LBM should be estimated (cf. Table 4). The optimal LBM for each data set was identified according to maximal BIC (cf. Eq. 5).

Table 5: Parametrization of LBM based on the identified global MD patterns

DS	Q	L	**Optimum**	BIC
DS1	3:7	3:7	-	-
DS2	7:16	7:16	Q = 8, L = 10	- 44 771.91
DS3	15:25	15:25	Q = 20, L = 21	- 64 558.54

Results. Table 6 shows the result of our local MD pattern detection by listing all robust biclusters (Bicl.) discovered with EB+iBBiG. The set of associated variables, the number of rows ($|\mathcal{I}|$), and the mean to evaluate the homogeneity is provided for each bicluster. It can be seen that all simulated MD patterns (7 for DS2 and 18 for DS3) have been identified correctly as robust biclusters, without any false positives. In DS1, we identified 4 non-simulated MD patterns.

Biclusters with mean=1 represent perfect homogeneous biclusters (e.g., B1.1, B2.1, or B3.1), where variables are missing always jointly. Biclusters with mean<1 represent dense biclusters, which refer to strong positive local patterns with co-occurring MD.

While the execution time of EB+iBBiG was in the order of a few minutes for each data set, the LBM for DS1 did not converge for any of the combinations of Q and L. A possible reason is the very low incidence of MD ($\approx$ 2.3 %), which does not form a homogeneous block structure even though several global MD patterns had been detected. However, LBM aims to detect exactly such block structures. For DS2 and DS3, the optimal block structure was identified according to the maximal value of BIC (-44,771.91 for DS2, and -64,558.54 for DS3). Tuning all combinations of Q and L to find the optimal model required approx. 10 hours (i.e., 20-30 minutes per combination) per data set. Combinations with smaller values for Q and L required exponentially shorter computation time than combinations with larger values. However, if a suitable model for the value of (Q, L) is already provided, e.g., due to the analysis of global MD pattern, or sub-optimal solutions are allowed, the estimation takes only a few minutes on a standard PC. We refer to the beginning of Section 3.3 for the used hardware settings.

To illustrate identified local MD pattern, Figure 7 shows a heatmap of the two perfectly homogeneous biclusters B2.1 (which corresponds to a LP) and

Table 6: Robust biclusters found in DS1, DS2, and DS3 using EB+iBBiG

DS	Bicl.	Variables	|I|	Mean
DS1	B1.1	x80, x244	924	1.00
	B1.2	x224, x245	490	1.00
	B1.3	x19, x80, x508	935	0.67
	B1.4	x18, x244, x508	813	0.67
DS2	B2.1	x1112, x1019, x731, x364, x798, x433, x535, x128, x769, x617	258	1.00
	B2.2	x1074, x432, x330, x737, x587, x491, x554	162	1.00
	B2.3	x182, x578, x103, x828, x957, x195, x362, x887, x586, x173, x631	180	0.66
	B2.4	x1022, x205, x342, x369, x755, x1002, x534, x1128, x251	72	0.96
	B2.5	x545, x377, x277	318	0.72
	B2.6	x558, x377, x399	201	0.73
	B2.7	x176, x371	79	1.00
DS3	B3.1	x337, x456, x740, x827, x838, x848	274	1.00
	B3.2	x439, x625, x715, x984	258	1.00
	B3.3	x232, x320, x397, x419, x769	125	1.00
	B3.4	x205, x344, x805, x808, x819, x856	162	0.77
	B3.5	x263, x391, x478, x836	118	1.00
	B3.6	x632, x86	234	1.00
	B3.7	x168, x195, x624, x970	160	0.84
	B3.8	x212, x220, x517, x723, x754, x816	53	0.99
	B3.9	x223, x315, x370, x399, x434, x861	82	0.78
	B3.10	x425, x683	123	1.00
	B3.11	x145, x593	114	1.00
	B3.12	x162, x175, x511, x670	50	1.00
	B3.13	x76, x905, x947, x997	77	0.83
	B3.14	x416, x770,	61	1.00
	B3.15	x792, x948,	61	1.00
	B3.16	x161, x491, x61	58	0.80
	B3.17	x239, x298, x552, x713, x864	42	0.72
	B3.18	x826, x885	33	1.00

B2.2 (which corresponds to a FMP), respectively. Figure 8 shows a comparison between the heatmaps of the original and the rearranged (= co-clustered) $\mathbf{M^Y}$ according to the optimal block-structure found for DS2 and DS3.

Result validation and method comparison. To evaluate the performance of EB+iBBiG, we constructed a prediction matrix $\widehat{\mathbf{M}}^{\mathbf{Y}}_{\text{pred}}$, whose cell (i,j) equals 1, if the corresponding cell of $\mathbf{M^Y}$ belongs to a detected robust bicluster, and equals 0 otherwise. In case of LBM, the (i,j) cell of $\widehat{\mathbf{M}}^{\mathbf{Y}}_{\text{pred}}$ equals 1, if the respective cell of $\mathbf{M^Y}$ belongs to a block having a mean$>$ 0.5, and equals 0 otherwise. The selection of such a threshold allows to compare the results of the LBM to EB+iBBiG, since iBBiG forms biclusters by selecting rows with $\hat{p}_i > 0.5$

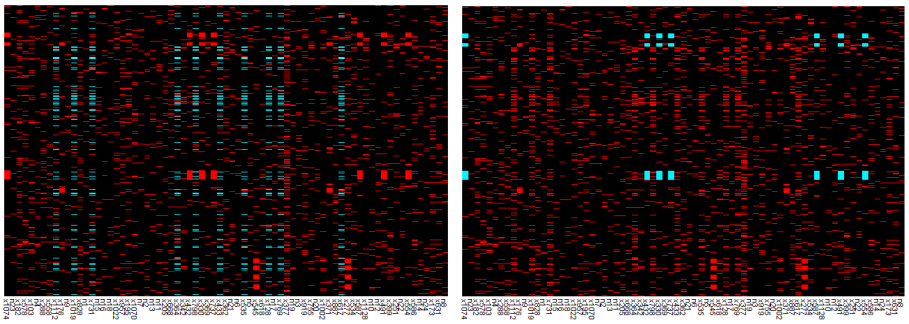

Figure 7: Robust biclusters B2.1 (left) and B2.2 (right) from DS2. The blue color represents $m_{ij} = 1$ cells being detected as a part of a bicluster, red represents $m_{ij} = 1$ not being a part of a bicluster, and black is $m_{ij} = 0$

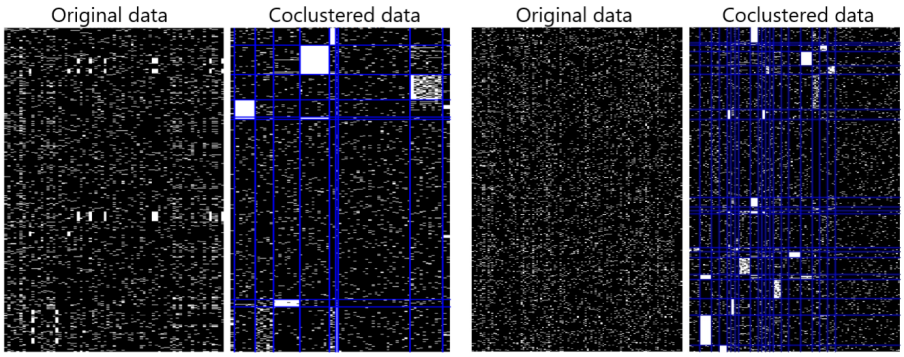

Figure 8: Visualization of the best LBM for DS2 (left) and DS3 (right). The white color represents $m_{ij} = 1$ cells and black color represents $m_{ij} = 0$.

(cf. Eq. 3), which guarantees that the detected biclusters have a mean> 0.5. We compared the $\widehat{\mathbf{M}}^{\mathbf{Y}}_{\text{pred}}$ of both algorithms to the reference matrix $\mathbf{M}^{\mathbf{Y}}_{\text{ref}}$, which was provided by voestalpine Stahl GmbH post hoc to the evaluations and indicates all "true" (i.e., simulated) MD patterns. Table 7 shows the confusion matrices, which were obtained by comparing $\widehat{\mathbf{M}}^{\mathbf{Y}}_{\text{pred}}$ to $\mathbf{M}^{\mathbf{Y}}_{\text{ref}}$ for each data set. Based on the confusion matrix, we calculated BACC[1], TPR[2], and TNR[3] for the final evaluation, where the results are summarized in Table 8.

[1] balanced accuracy$= \frac{\text{sensitivity + specificity}}{2}$

[2] true positive rate=sensitivity$= \frac{TP}{TP+FN}$

[3] true negative rate=specificity$= \frac{TN}{TN+FP}$

Table 7: Confusion matrices obtained by comparing detected local MD patterns ($\widehat{\mathbf{M}}^{\mathbf{Y}}_{\text{pred}}$) with the reference ($\mathbf{M}^{\mathbf{Y}}_{\text{ref}}$)

Method	**DS**			**Reference**	
				$m_{ij} = 1$	$m_{ij} = 0$
EB+iBBiG	**DS2**	**Predicted**	$m_{ij} = 1$	180 143	4 633
			$m_{ij} = 0$	375	7 642
	DS3		$m_{ij} = 1$	259 919	1 742
			$m_{ij} = 0$	235	8 104
LBM	**DS2**		$m_{ij} = 1$	180 470	5 148
			$m_{ij} = 0$	48	7 127
	DS3		$m_{ij} = 1$	259 902	3 530
			$m_{ij} = 0$	252	6 316

Table 8 shows that EB+iBBiG reached better BACC and TPR. On the other hand, slightly higher and almost perfect specificity was obtained with the LBM, which produced only very few false positive cases. The lower sensitivity of both approaches on DS2 was caused by an asymmetric LP, which contained 80 % noise (which made it difficult to detect) and was very large in contrast to the other patterns (affecting 20 % observations over {x277, x288, x377, x399, x558, x618, x668, x919, x1070}). This LP was only detected by EB+iBBiG (see biclusters B2.5, B2.6 in Table 6) and not with the LBM since the mean of the corresponding block cannot exceed the threshold of 0.5 due to the 80 % of noise in the underlying pattern. The higher TPR of EB+iBBiG refers to its better capacity in terms of industry-specific MD pattern detection.

Table 8: Comparison of biclustering algorithms

Data	**Algorithm**	**BACC[1]**	**TPR[2]**	**TNR[3]**
DS2	EB+iBBiG	81.02 %	62.25 %	99.79 %
	LBM	79.02 %	58.06 %	99.97 %
DS3	EB+iBBiG	91.12 %	82.31 %	99.91 %
	LBM	82.03 %	64.15 %	99.90 %

3.3.4 Discussion of the Experiments

For DS1, we identified several global and local MD patterns, which indicate the co-occurrence of MD and might be valuable information for a joint imputation of missing values. We plan to investigate the extent to which the runtime could be reduced for the overall imputation process in future research.

The evaluations with DS2 and DS3 showed that all identified global patterns correctly grouped sets of variables forming one of the simulated patterns. Another desirable result was that our approach assigned only signal variables "x" and none of the noisy variables "n" (with only random MD) to a pattern.

Further, we conclude that biclustering is suitable to detect local MD patterns, since in our application focused on the detection of LPs and FMPs, all

for DS3, and all but one for DS2 of the simulated biclusters were detected reliably. The single not detected bicluster corresponds to an asymmetric LP, which consisted of 80 % noise and was much larger than the other biclusters. Thus, lower TPR on DS2 was obtained for both approaches. While the TNR was similar and greater than 99 % in all cases, EB+iBBiG outperformed the LBM in terms of much higher sensitivity (81 % vs 79 % for DS2, and 82 % vs. 64 % for DS3). We conclude, that EB+iBBiG is the preferable approach to detect industry-specific MD pattern due to (i) comparable specificity, (ii) much better sensitivity, (iii) short computational time, and (iv) because it is not necessary to tune the parameters in contrast to LBM.

A Appendix

A.1 Significance of Association Measures

The p-value of any statistics $T^{obs}(\mathbf{D})$ observed on data $\mathbf{D}$ is defined as $P(T(\mathbf{D}) \geq T^{obs}(\mathbf{D})|H_0$ is valid) [Hal98], where $T()$ denotes hypothetical true value of that statistics. In our application, we test H_0 of no association, $H_0 : T^{obs}() = 0$, against H_1 that $T^{obs}()$ significantly differs from 0 using columns $\mathbf{M}_{.j}$ and $\mathbf{M}_{.l}$ as data. The approximate p-value can be calculated using the following procedure inspired by the *trial-shuffling test* [Alb+15]:

1. Calculate $T^{obs} = T(\mathbf{M}_{.j}, \mathbf{M}_{.l})$
2. For $b = 1, ..., B$:
   - I Sample with replacement from $\mathbf{M}_{.j}$ vector $\mathbf{M}^b_{.j} = (m^b_{1j}, \ldots, m^b_{nj})'$ of the same length.
   - II Sample with replacement from $\mathbf{M}_{.l}$ vector $\mathbf{M}^b_{.l} = (m^b_{1l}, \ldots, m^b_{nl})'$ of the same length.
   - III Calculate statistics on resampled data $\mathbf{M}^b_{.j}, \mathbf{M}^b_{.l}$: $T^b = T(\mathbf{M}^b_{.j}, \mathbf{M}^b_{.l})$
3. Calculate approximate p-value $= \frac{\sum_b \mathbb{I}(T^b \geq T^{obs})}{B}$, where $\mathbb{I}$ stands for the indicator function.

References

[Jac12] Paul Jaccard. "The distribution of the flora in the alpine zone. 1". In: *New phytologist* 11.2 (1912), pp. 37–50.

[Coc52] William G Cochran. "The χ2 test of goodness of fit". In: *The Annals of Mathematical Statistics* 23.3 (1952), pp. 315–345.

[Rub76] Donald B. Rubin. "Inference and Missing Data". In: *Biometrika* 63.3 (1976), pp. 581–592.

[Hal98] Anders Hald. *A History of Mathematical Statistics from 1750 to 1930*. Vol. 2. New York, NY, USA: John Wiley & Sons, Inc., 1998.

[NH98] Radford M Neal and Geoffrey E Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants". In: *Learning in Graphical Models*. Dordrecht: Springer, 1998, pp. 355–368.

[Cra99] Harald Cramér. *Mathematical methods of statistics*. Vol. 43. Princeton, NJ, USA: Princeton university press, 1999.

[JMF00] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. "Data Clustering: A Review". In: *ACM Computing Surveys (CSUR)* 31.3 (2000), pp. 264–323.

[Hin02] Holger Hinrichs. "Datenqualitätsmanagement in Data Warehouse-Systemen [Data Quality Management in Data Warehouse Systems]". PhD thesis. Universität Oldenburg, 2002.

[LR02] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data, Second Edition*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[SG02] Joseph L. Schafer and John W. Graham. "Missing Data: Our View of the State of the Art". In: *Psychological Methods* 7.2 (2002), pp. 147–177.

[GN03] Gerard Govaert and Mohamed Nadif. "Clustering with block mixture models". In: *Pattern Recognition* 36 (Feb. 2003), pp. 463–473. DOI: 10.1016/S0031-3203(02)00074-2.

[LHH03] Edith D. de Leeuw, Joop Hox, and Mark Huisman. "Prevention and Treatment of Item Nonresponse". In: *Journal of Official Statistics* 19.2 (2003), pp. 153–176.

[Pre+06] Amela Prelić et al. "A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data". In: *Bioinformatics* 22.9 (2006), pp. 1122–1129.

[Van07] Stef Van Buuren. "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification". In: *Statistical Methods in Medical Research* 16.3 (2007), pp. 219–242.

[IS08] Syed A. Imtiaz and Sirish L. Shah. "Treatment of Missing Values in Process Data Analysis". In: *The Canadian Journal of Chemical Engineering* 86.5 (2008), pp. 838–858.

[End10] Craig K Enders. *Applied Missing Data Analysis*. New York, NY, USA: Guilford Press, 2010.

[Shi+10] Fan Shi et al. "A bi-ordering approach to linking gene expression with clinical annotations in gastric cancer". In: *BMC Bioinformatics* 11.1 (2010), p. 477.

[Eps+12] Sacha Epskamp et al. "qgraph: Network Visualizations of Relationships in Psychometric Data". In: *Journal of Statistical Software* 48.4 (2012), pp. 1–18. URL: http://www.jstatsoft.org/v48/i04/.

[Gus+12] Daniel Gusenleitner et al. "iBBiG: iterative binary bi-clustering of gene sets". In: *Bioinformatics* 28.19 (July 2012), pp. 2484–2492. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts438. URL: https://doi.org/10.1093/bioinformatics/bts438.

[Des13] Bernard Desgraupes. "Clustering indices". In: *University of Paris Ouest-Lab Modal'X* 1 (2013), p. 34.

[GN13] Gérard Govaert and Mohamed Nadif. *Co-Clustering: Models, Algorithms and Applications*. New York, NY, USA: John Wiley & Sons, Inc., 2013.

[Kha13] Tatsiana Khamiakova. "Statistical Methods for Analysis of High Throughput Experiments in Early Drug Development". PhD thesis. Hasselt University, 2013.

[BH14] Patricia Berglund and Steven G. Heeringa. *Multiple Imputation of Missing Data Using SAS*. Cary, North Carolina, USA: SAS Institute Inc., 2014.

[BIG14] Parmeet Bhatia, Serge Iovleff, and Gérard Govaert. "Blockcluster: an R package for model based co-clustering". In: *Journal of Statistical Software* 76.9 (2014), pp. 1–24.

[Kha14] Tatsiana Khamiakova. *Superbiclust: Generating Robust Biclusters from a Bicluster Set (Ensemble Biclustering)*. R package version 1.1. R Foundation for Statistical Computing. 2014. URL: https://CRAN.R-project.org/package=superbiclust.

[Alb+15] Mélisande Albert et al. "Bootstrap and permutation tests of independence for point processes". In: *The Annals of Statistics* 43.6 (2015), pp. 2537–2564.

[Ker+15] Christine Keribin et al. "Estimation and selection for the latent block model on categorical data". In: *Statistics and Computing* 25.6 (2015), pp. 1201–1216.

[Kas+16] Adetayo Kasim et al. *Applied Biclustering Methods for Big and High-Dimensional Data using R*. Boca Raton, Florida: CRC Press, 2016.

[Cox17] David Roxbee Cox. *The theory of stochastic processes*. Abingdon-on-Thames, England, UK: Routledge, 2017.

[SIG17] Parmeet Singh Bhatia, Serge Iovleff, and Gérard Govaert. "blockcluster: An R Package for Model-Based Co-Clustering". In: *Journal of Statistical Software* 76.9 (2017), pp. 1–24. DOI: 10.18637/jss.v076.i09.

[Ehr+18] Lisa Ehrlinger et al. "Treating Missing Data in Industrial Data Analytics". In: *Proceedings of the 13th International Conference on Digital Information Management (ICDIM)*. Berlin, Germany: IEEE, 2018, pp. 148–155.

[Kai+18] Sebastian Kaiser et al. *biclust: BiCluster Algorithms*. R package version 2.0.1. R Foundation for Statistical Computing. 2018. URL: https://CRAN.R-project.org/package=biclust.

[Fer19] Sara Johansson Fernstad. "To Identify What is Not There: A Definition of Missingness Patterns and Evaluation of Missing Value Visualization". In: *Information Visualization* 18.2 (2019), pp. 230–250. DOI: 10.1177/1473871618785387. URL: https://doi.org/10.1177/1473871618785387.

[GC19] Daniel Gusenleitner and Aedin Culhane. *iBBiG: Iterative Binary Biclustering of Genesets*. R package version 1.28.0. R Foundation for Statistical Computing. 2019.

[Kol19] Raivo Kolde. *pheatmap: Pretty Heatmaps*. R package version 1.0.12. R Foundation for Statistical Computing. 2019. URL: https://CRAN.R-project.org/package=pheatmap.

[Bec+21] Michal Bechny et al. "Missing Data Patterns: From Theory to an Application in the Steel Industry". In: *33rd International Conference on Scientific and Statistical Database Management*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 214–219. ISBN: 9781450384131. URL: https://doi.org/10.1145/3468791.3468841.