A Visual-Interactive Idiom to Diagnose Missing Data Mechanism

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Missing values are a pervasive problem in most data collection processes. Several methods can deal with missing values, and choosing one depends on the diagnosis of the missing data mechanism—the way that missingness correlates with variables. One way of diagnosing the mechanism is by comparing pairs of variables using data visualizations. However, the standard visualizations for this task use only simple idioms—such as scatterplots and matrix plots—not designed to take on the diagnosis task upfront. Additionally, not only can they take too much screen space, but they also overlook vital pieces of information, consequently making the user actively pursue cues for diagnosis instead of explicitly presenting those cues. Thus, this paper proposes a visual-interactive idiom for diagnosing missing data mechanisms. The approach consists of a data derivation algorithm that quantifies two relevant cues for diagnosis: randomness similarity —how much the missing data differs from a perfect uniform sample— and randomness plausibility —how much the difference between those distributions resembles a random shape. The idiom uses a progressive-rendered interactive visualization, enabling exploration of those metrics during computation. We present the algorithm, its technical aspects, and the design choices for the idiom, showing how it supports the whole diagnosis task. We use a synthetic dataset to validate the capability of the idiom in depicting the missing data mechanisms, and then apply the idiom to real datasets to demonstrate how it can assist real exploratory data analysis scenarios.**

***Index Terms*—missing values, data preprocessing, exploratory data analysis**

I. Introduction

In almost every data collection processes, there is the risk of having missing values—the blank cells in data tables [1]. For example, in medicine, longitudinal studies face the chance of participant drop-out, which leaves the data incomplete [2]. In environmental sciences, where a range of sensors measure an extension of land, water or air, failures in the capture hardware or software might leave gaps in the measurements [3]. In biological and chemical domains, where the scale of observation can reach microscopic levels, certain phenomena might go unnoticed or even ignored [4]. When missing values occur, one option is to rerun the whole experiment, which is often implausible as it might require extensive resources and time [5]. Thus, when faced with this problem, data analysts usually drop out the missing values, conducting what is called a complete case analysis. However, although this is the default approach in many statistical packages [6], it might introduce severe biases and errors to the analysis [7] [8] [9]. The alternative is to use more advanced methods, such as multiple imputation [10] and likelihood estimation [11]. However, each method makes assumptions about the distribution of missing values [12]—known as the missing data mechanism—and understanding this distribution is a key task before applying those methods [13].

The missing data mechanisms describe how the distribution of missingness in a given data variable correlates to other variables [14]. There are three missing data mechanisms:

- Missing Completely at Random (MCAR): The probability of a value being missing does not correlate with other variables, either observed or non-observed (missing).
- Missing at Random (MAR): The probability of a value being missing correlates with observed variables.
- Missing Not at Random (MNAR): The probability of a value being missing correlates with unobserved variables.

Those mechanisms compose a fuzzy spectrum: almost no dataset is entirely MCAR, MAR, or MNAR, but instead has each of these mechanisms scattered to some degree across pairs of data variables. The most complex mechanism is the MNAR —also called non-ignorable— because it introduces an unknown uncertainty [15], so diagnosing it requires domain knowledge to create specific statical models to describe the missingness [16] [17]. In practice, without such domain knowledge, MAR and MNAR are indistinguishable, and some degree of MNAR should always be considered [18].

However, the MCAR and MAR mechanisms are the most common in real datasets, and identifying them is a possible data analysis task [19]. The standard way of visually exam-

Identify applicable funding agency here. If none, delete this.

ining the mechanism is with a matrix plot, which is a binary heatmap of the dataset that shows where the missing values occur [20]. Another strategy is to use a scatterplot matrix, with the missing values on the margin of the plot [20].

However, these plots have some caveats. If the data has many variables with missing values, the matrix plot requires many sorting interactions, and consequently, the analyst needs to remember several states of the visualization. High data dimensionality also affects the scatterplot matrix, as the number scatterplots displayed in a fixed area is finite.

Thus, this paper proposes an idiom to assist the diagnosis of missing data mechanisms when distinguishing MCAR from MAR mechanisms. We call it idiom because it comprises a data derivation algorithm, visual encodings, and interactions [21]. The data derivation extracts two quantitative metrics that summarize the resemblance to the mechanisms, acting as supporting cues for diagnosis. The visual-interactive representation assists the exploration of pairs of variables with coordinated selections, filtering, and details on-demand so that the analyst can overview, explore, and then diagnose data.

The remainder of this paper organizes as follows. Section II presents related works about visualization of missing values and missing mechanisms. Section III presents the idiom, describing the steps involved in missing data diagnosis and how we address them with data derivations, visual encodings, and interactions. Section IV presents a series of experiments with synthetic and real data, demonstrating how the idiom can assist in exploratory data analysis and mechanism diagnosis. Finally, Section V concludes with final remarks, limitations, and future works.

II. Related Works

Current and past studies in the visualization of missing values subdivides into two branches: one that deals with missing values at a visual encoding level —such as proposing encodings to visualize the missing values— and one that pursues depictions about the missingness itself, to assist analysts in choosing effective strategies to deal with them.

The first branch often overlooks the missing data mechanism problem. For instance, Twiddy et al. [22] addresses missing values in scientific visualization domain with a color-based idiom that fixes missing spots. Similarly, Popov [23] propose a idiom to visually place missing values at estimated positions. Song and Szafir [24] measured how visual depictions of missing values affect the perceived quality of datasets. While such studies contribute to clarify "what it is not there," they do not help as much in the task of identifying "why it is not there."

When looking for visualization as assistance to deal with missing values (i.e., to diagnose the missing data mechanism before an imputation or deletion method), the analyst needs to tap into visual idioms that support the diagnosis of the mechanism [25]. Without those methods, analysts are led to assume mechanisms [26], which might result in non-optimal or biased results.

To overcome this issue, the second branch of research deals with depictions for the missing value mechanism—a branch heavily represented by visualization toolkits and packages. The R Package VIM by Templ and Filzmoser [18] provide visualizations to both see the missing values, and to assist diagnosis. Python library Missingno by Bilogur [27] also present visualizations for both tasks. Recently, Tierney and Cook [28] acknowledge the little body of research focusing on this task, and introduce the R package Naniar to incorporate novel plots in the workflow of data workers.

However, these practical efforts are quite detached from the visualization theory and research. Although the design choice of those plots works, they are usually not justified with information visualization fundamentals. The low visibility of package papers and reports contribute to this problem: such plots receive less formal critique, the feedback that is crucial to advance any field in science. As a consequence, analysts end up using the available choices not necessarily because they are optimal, but because of the ready availability and limited alternatives.

The main source of this problem is that the existence of missing data is a source of uncertainty, and as such, different coping strategies apply [29]. Visualization literature is currently focusing on those who want to see the missing values (e.g., "what is not there"), leaving for practitioners the task of helping those who wish to minimize or exploit such uncertainty (e.g., "why it is not there") —which requires understanding the missing data mechanisms. In other words, few studies currently focus on understanding missing values, developing justified visual idioms to convey the information needed for diagnosing missing data mechanisms effectively.

Such effort is beneficial because the resulting methods can open the door for deeper visual analytics applications [30], to bundle the diagnosis and treatment methods in a single software.

Thus, this paper aims at initiating the conversation about how visualization theory can help to understand the mechanisms of missing data. We define the diagnosis task as a three steps process: overview, exploration, and diagnosis. From that definition, we propose an interactive-visual idiom for the exploration of the missing data mechanisms in general multivariate mixed data, comprised of data derivations and visual encodings that were specifically tailored to support the steps involved in diagnosing the missing values. The goal of this paper is not only to propose a solution for this visualization task, but also to bring awareness and visibility to the topic, which can lead to novel visualizations, taxonomies, and visual analytics systems that may assist data workers in dealing with such practical issues.

III. Idiom

To design an idiom that supports the diagnosis of missing values mechanism, it is important to define the steps involved in the diagnosis task.

A. The Mechanism Diagnosis Task

There is one main question that defines the missing data mechanism: Does the missingness of the data correlate to data variables? If so, then it has MAR characteristics, and if not, then it has MCAR characteristics. The goal of visualizations and statistical tests is to answer this question with the appropriated degree of uncertainty, as it is only an estimation [28]. For instance, the Little's MCAR test [19] verifies the whole data and provide a p-value for the hypothesis of not being MCAR.

However, the mechanism can be seen not only as a characteristic of a dataset but as a pairwise characteristic: a variable can simultaneously have a MAR relationship with some variables and an MCAR relationship with others. Thus, the question can be reframed in pairwise relations: Given a variable X and a variable Y, does the missingness of X correlates with Y? The consequence of this pairwise approach is that the number of relationships grows asymptotically by the square of the number of variables.

This quadratic growth is the main reason why current idioms, such as the matrix plot and scatterplot grid, have downsides. For the matrix plot, it means that the number of states that the analyst must remember grows to impracticable amounts, and for the scatterplot grid, it indicates that the number of grids can be unfeasible to the available space. Despite this scalability issue, those idioms can be useful even for large datasets, as long as the analyst already knows which pairs of variables to diagnose—i.e., when there are pre-defined variables of interest.

The problem is that when analyzing a dataset for the first time, those variables of interest are unknown. In that case, the analyst usually wants to find relevant variables first, then start diagnosing then. Hence, we characterize the diagnosis task as a three-step process as follows:

Overview: It is the need to analyze the data as a whole, to check if the data contains more MCAR or MAR relationships. It is the goal of the Little's MCAR test [19], and the step that current idioms lack support for larger datasets.

Exploration: It is the need to find which variables have enough evidence to be diagnosed as MCAR or MAR, and which are in the middle of the spectrum—being harder to diagnose. As the number of variables grows, the difficulty of this step increases, and current idioms support it using either too much screen space, or too much cognitive load.

Diagnosis: It is the need to evaluate if a pair of variables has evidence in favor (or against) a given mechanism, and the uncertainty of this evidence. Current idioms depict this uncertainty only intuitively, which increases the subjectivity of the diagnosis.

Thus, an improvement over the current idioms would address the three steps of the diagnosis task while using a constant screen space, avoiding excessive cognitive work on the analyst, and explicitly depicting the uncertainty inherent to diagnosis. This paper proposes an idiom that implements such improvements by quantifying the evidence of missing mechanisms into quantitative metrics and providing visual encodings and interactions that support the diagnosis steps. The following sections details this idiom.

B. Data Derivation

What standard approaches do is to map the data itself into visualizations, which might not necessarily be the best information for diagnosis. We propose the use of a data derivation step, extracting what we call evidence metrics—information directly relevant for diagnosis. The goal of those metrics is not to be the only source of data, but to augment the visual encodings and enable a novel design.

The proposed idiom derives two metrics: (1) A bootstrapped estimation of the probability that describes how similar the distribution of missing values is to a random distribution, and (2) a plausibility metric that describes how plausible is the hypothesis that the distribution is random. Both metrics are quantitative in the range of 0 and 1 and are computed for each pair of variables. This section shows the rationale behind those metrics and the algorithms that compute them.

1) Bootstrapped Randomness Score: This metric is an estimation of the probability that a given distribution of missing values is random when taking into account another variable. If the probability of randomness is high, the relation is more likely to be MCAR, and if not, is more likely to be MAR.

To compute this metric, we use the bootstrap resampling method [31]. Given a variable X with missing values, and a variable Y, the algorithm computes how much the values of Y that have missing values in X are similar to a random subsample from Y. The rationale is that if the distribution $Y|X_{miss}$ is too similar to a uniform sampling, then it is likely to be, in fact, a uniform random sampling. The opposite case is if the distribution $Y|X_{miss}$ deviates too much from a uniform sampling, and so it is likely that the missingness of X correlates with the values of Y. Figure 1 shows this difference between those distributions.

The algorithm quantifies those deviations in terms of the probability that they occurred at random. The output of running this algorithm is an estimated probability matrix of size $m \times n$, where m is the number of variables with missing values, n is the total number of variables in the dataset, and the elements a_{ij} being the probability that the missingness of the variable i correlates with variable j.

Since bootstrapping is an expensive operation, we used a progressive-rendering approach [32] in which the computation is fragmented in small steps. That way, each incremental step is rendered to the user as frames in an animation, while the algorithm is still running in the background. Each frame of the animation has a more refined probability, as the bootstrap converges into an estimation when more and more resamples are made.

The Algorithm 1 shows in detail how to compute this operation to create a matrix.

Line 5 samples the values of the variable Y with missing values at variable X. The values of Y are binned using a binning rule (e.g., sturges rule [33]), or simply counted if categorical or ordinal. An extra bin to keep track of the number

Algorithm 1: Bootstrapping the chance of the sampling distribution of the missing values being completely random (MCAR)

Result:
E: An estimation matrix;
Data:
B: resamples this iteration;
T: total resamples taken;
$previousE$: the last E computed;

1 $E \leftarrow Matrix[][]$;
2 **for** *each variable with missing values* x **do**
3 **for** *each other variable* y **do**
4 $sampleSize \leftarrow \frac{length(x_{miss})}{length(y)}$;
5 $sample \leftarrow y \cap x_{miss}$;
6 **if** y *is quantitative* **then**
7 $binWidth \leftarrow$ `BinningRule(`y`)`;
8 $j_{bins} \leftarrow$ `Histogram(`y, *binWidth*`)`;
9 $sample_{bins} \leftarrow$ `Histogram(`*sample, binWidth*`)`;
10 **else**
11 $j_{bins} \leftarrow$ `Count(`y`)`;
12 $sample_{bins} \leftarrow$ `Count(`*sample*`)`;
13 **end**
14 $expected_{bins} \leftarrow j_{bins} \cdot sampleSize$;
15 $errors \leftarrow (sample_{bins} - expected_{bins})^2$;
16 $RMSE \leftarrow \sqrt{\texttt{Mean}(errors)}$;
17 $estimation \leftarrow 0$;
18 **for** n *from* 1 *to* B **do**
 `// B is constant`
 `// Time Complexity` $\mathcal{O}(Bij) = \mathcal{O}(ij)$
19 $random \leftarrow \texttt{Resample}(y, sampleSize)$;
20 **if** y *is quantitative* **then**
21 $random_{bins} \leftarrow \texttt{Histogram}(random, binWidth)$;
22 **else**
23 $random_{bins} \leftarrow \texttt{Count}(random)$;
24 **end**
25 $errors \leftarrow (expected_{bins} - random_{bins})^2$;
26 $bRMSE \leftarrow \sqrt{mean(errors)}$;
27 **if** $bRMSE > RMSE$ **then**
28 $E \leftarrow estimation + \frac{1}{B}$;
29 **end**
30 **end**
31 $E[x][y] \leftarrow estimations$;
 `// Space Complexity` $\mathcal{O}(ij)$
32 **end**
33 **end**
34 **if** $T > 0$ **then**
35 $E \leftarrow \frac{B \cdot E + T \cdot previousE}{B + T}$;
36 **end**
37 $T \leftarrow T + B$;
38 $previousE \leftarrow estimations$;

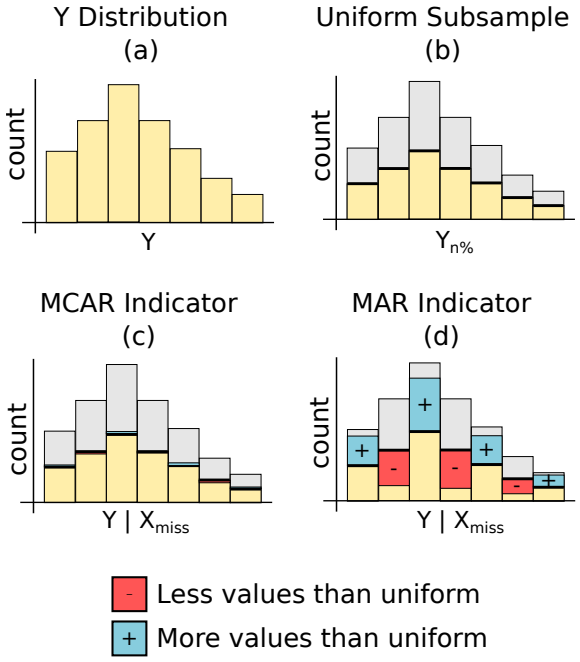

Fig. 1. Given the distribution of a variable (a), the distribution a perfect uniform subsample of n% values is equal to n% of its bins (b). When the missing values of a variable X dictates the sampling, a MCAR relationship (c) is likely to be closer to the uniform, while a MAR relationship (d) is likely to deviate.

of missing values of Y is added, similar to the approach of Kang and Schneiderman [34].

Line 14 computes the histogram of a perfect uniform sampling of Y —which we call the expected bins. As previously shown in Figure 1, the difference between the expected bins and the actual sample bins indicates how far the values of $Y|X_{miss}$ are from the uniform distribution. This difference is computed using a RMSE of the deviations of the bins in line 16.

The bootstrap starts at line 18, where the same process is repeated B times, but now comparing the perfect uniform subsample with random subsamples of Y—the bootstrapped errors are called bRMSE. Due to the central limit theorem, the distribution of those different bRMSE scores represents the probability that a given RMSE appears in uniformly random situations.

If the computed RMSE of the subsample of $Y|X_{miss}$ is low enough that it occurs too often in the sampled RMSE, then it is likely that the subsample $Y|X_{miss}$ is no different from a uniform sample, and thus an evidence in favor of MCAR hypothesis. If, in contrast, the RMSE occurs too rarely in the sampled RMSE, then it is likely that is not uniform, and thus an evidence in favor of the MAR hypothesis.

Figure 2 synthetizes the core of the derivated metric: the estimation of the MCAR probability for a given pair of variables. The bootstrapped probability is the number of times that the random subsamples achieved lower RMSE than the actual subsample, divided by the total number of random subsamples taken. Thus, if the RMSE is lower than 100%

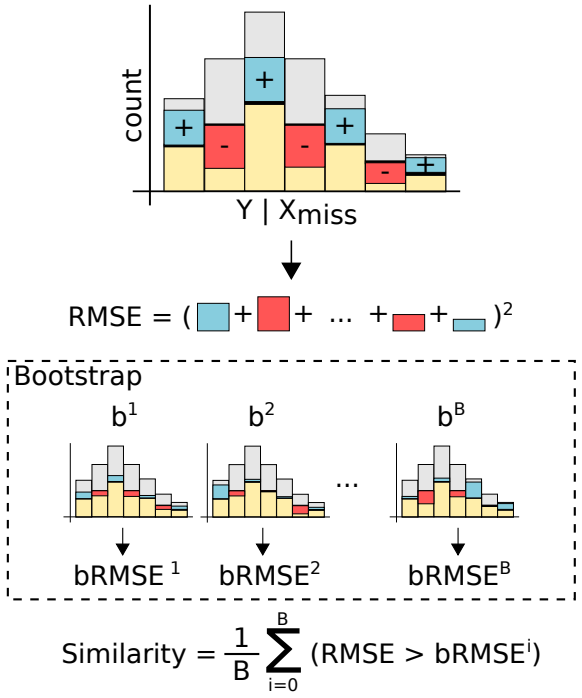

Fig. 2. The algorithm to estimate the MCAR similarity for . First, the RMSE of the deviations is computed, followed by several bRMSE for random distributions. The estimation is the number of bootstrapped bRMSE that are smaller than the actual RMSE

of the random subsamples, the bootstrapped estimation is 1, which favours MCAR. Conversely, if the RMSE is higher than 100% of the random subsamples, the bootstrapped estimation is 0, which favours MAR.

Since the algorithm requires many B resamples to provide accurate estimations, it was designed to repeat this computation as animation frames, with small values of B at each frame. Line 35 shows how it refines the computed estimations with a previous one. Each time the algorithm runs, it increments the total number of bootstrap iterations T. When a new matrix is computed, it is merged to the previous one using the number of interactions as a weighted average. Line 36 updates the estimation that used B repetitions with the previous one that used T repetitions, thus refining the matrix and achieving the data for the next animation frame.

The space complexity is $\mathcal{O}(ij)$. The time complexity is $\mathcal{O}(Tij)$, meaning that is quadratic for every iteration of the bootstrap. To delimitate T, we stop the progressive computation when not a single pair had its probability changed by more than 0.5%, as further estimations are likely to be too small to affect the analysis. Because T is delimited, and the algorithm always converge, T is assintotically constant, so $\mathcal{O}(Tij) = \mathcal{O}(ij)$.

2) Plausibility of Randomness: The shape plausibility is a much simpler quantitative metric and it can be computed without bootstrapping, but it only works for ordered data. This metric is based on the assumption that, if there is a correlation between missingness and other variables, this correlation is

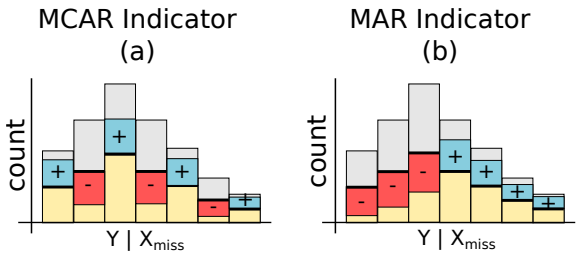

Fig. 3. Although the magnitude of deviations is similar, the relationship of (a) is more plausible to be MCAR than the one in (b) due to the inconsistence of the deviations.

likely to have a smooth shape. The assumption is not always true, but in most cases, it serves as an indicator of the mechanism.

To visualize, consider a histogram from a quantitative variable Y. If we sample 50% of in a perfectly uniform distribution Y, the expected histogram has 50% of each bin, as shown in Figure 3.

When a random uniform distribution is taken, the natural variance induces some random offsets of the bins. The consequence is that those offsets do not have any correlation with the value of Y, such that it is possible to have bins with less than expected values and bins with more than expected values closer to each other. (Figure 3a). Now, consider a random sampling with unknow distribution is taken from the same variable, and the result is the histogram of Figure 3b. Although this new sample has the same magnitude in the offsets as the random one (i.e., same bar heights), those offsets are clustered: bins with less and more expected values appear together (Figure 3b).

That suggests that the distribution of the second sample, although it does not deviate as much from the expected in terms of magnitude, has a more consistent shape that is unlikely to have occurred at random. The more consistent the shape, the more plausible is the MAR assumption, and the more inconsistent, the more plausible the MCAR assumption. To quantify the plausibility of the shape, we count the number of swaps in magnitude orientation (less or more than expected), and divide by the number of possible swaps. The missing values bin is not included, as it is not a part of the quantitative scale of the variable.

If, in a histogram with n bins, all clusters have only one bin, then the number of swaps is the maximum value $n - 1$, and thus the metric is 1 —high plausibility of randomness. If there are no swaps, then the metric is 0—low plausibility of randomness. The zero case occurs when all bins are less or more than expected, and the only one in the other orientation is the missing values bin: since it is not counted, there are zero swaps in the bins.

Because categorical data do not have an order, this metric makes no sense, and thus it is not used. The pair with categorical variables receive no plausibility score, but this absence is depicted in the visual encodings.

Thus, the derived metrics are a bootstrapped randomness

probability metric —based on the magnitude of deviation— and a randomness plausibility score —based on the consistency of deviations in ordered variables. Those metrics are not intended to be decisive for diagnosis, but rather to augment the visualization, providing visual encodings that are explicitly correlated to the missing data mechanisms. The next section shows how the proposed idiom map those metrics to visual encodings, and how they support the steps of a diagnosis task.

C. Visual Encodings

Because of the data derivation process, the proposed idiom has two additional information to plot for every pair of variables, therefore it has four data dimensions to plot:

- The list of variables with missing values;
- The list of all variables;
- The bootstrapped similarity score;
- The shape plausibility score;

There are many resources to assist the design of visualizations [35] [36] [37]. For the design choices based on empirical principles, we refer to the ranks of visual variables and guidelines in Munzner's book [21].

We start by acknowledging the level of importance of the dimensions, in order to choose proper visual depictions. Both list of variables are crucial, as they are the variables under analysis. For that reason, we allocate spatial position, the most powerful depiction [38], for them.

Using only two spatial depictions would lead to a grid-like idiom. However, the bootstrapped similarity score is also very important, as it is the metric that drives the computation and animation. Thus, we also allocate the spatial variable for this metric, opting for a idiom based on parallel axes, similar to the choice used in parallel coordinates [39]. As shown later, this also allows for easier interactions with spatial range selectors. To depict pairs of dimensions, the idiom uses a line that crosses the three axes at the associated values.

For the fourth dimension, the randomness plausibility score, since it is a metric based on an assumption, and it is not computed for all pairs, we chose the color. The color encoding allows for a visual depiction that acts more as a supportive metric, and the benefit that it is a visual mapping with an identifiable neutral position for the pairs that include categorical variables, and thus lack the score. We achieve that neutral position by using a sequential color scale for the pairs that have a score, but also allocating a color hue outside of that scale to represent the absence of the metric.

Figure 4 shows how this design choice looks. The parenthesis after the variable names indicates the absolute number of missing values in it. For aesthetic purposes, we used polycurves instead of polylines in order to apply a loose edge bundling effect [40], which makes groups of pairs with close similarity scores more distinguishable [41].

Each axis that holds variable names uses a categorical axis with small bands, to use the whole available space. The polylines cross the bands at a height that is proportional to where they cross the central similarity axis, so the bands acts like a miniature of the central axis. The bands also give the

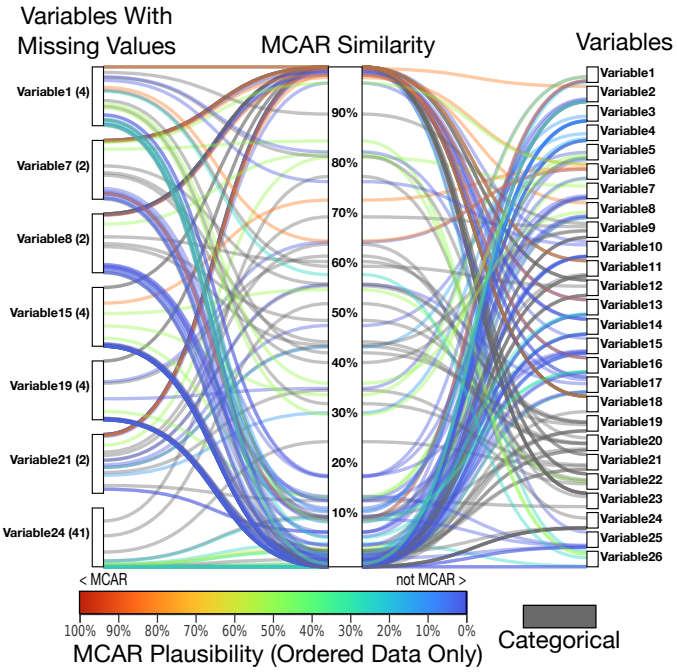

Fig. 4. The result of the design choices. The similarity metric is depicted by the central axis, and the plausibility metric is mapped in color. The list of variables are also axes, and the number of missing values appears in parenthesis.

option of partially filling then, in order to use the filling length as a visual encoding, which we make use of in the interactions.

Until now, the idiom is a static visualization that provides an overview of the dataset, because the analyst can see how many variables tends towards the MCAR or MAR side of the derived metrics. However, details about the distributions are still missing, and diagnosis is not possible yet. To complete the visual encodings, the idiom also includes a details on-demand view to fill that need.

The visual encodings for details on-demand are the main diagnosis view, in which the distributions are shown in a histogram-like encoding. Figure 5 shows the histogram for a given pair of dimensions. The histogram shows the difference between the perfectly uniformly sampled histogram and the values of Y that have missing values in X —the same histogram used to compute the scores. The border of the histogram has the same color of the associated polycurve, making it easier to change the look between views. The histograms are displayed on a scrollable list to the side of the main view.

Thus, the main encoding encompasses the overview step, and the details-on demand encompasses the diagnosis step. However, there is too much information on the screen, and the support for the intermediate step, exploration, is still missing. The next section shows how an interaction idiom can fill that gap, showing the relevant variables for analysis.

D. Interactions

The exploration step in the diagnosis task is represented by the question: "which variables are more likely to have the

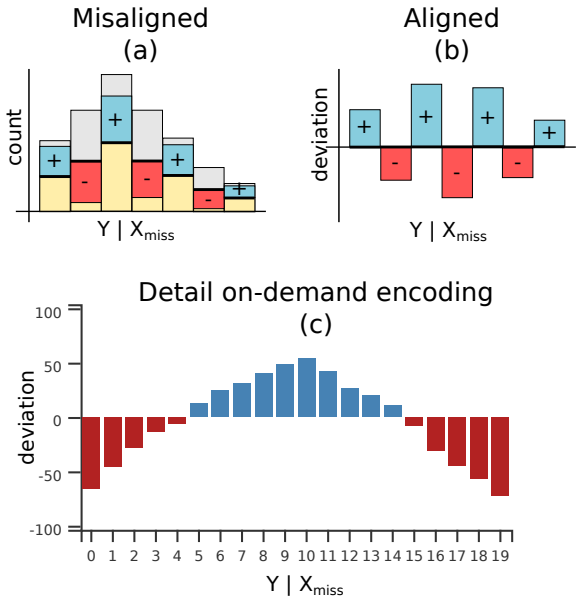

Fig. 5. Given the histogram of the distribution (a) the idiom aligns the expected values to count only the deviations of the bins (b). The result is the final visual representation displayed as details on-demand (c).

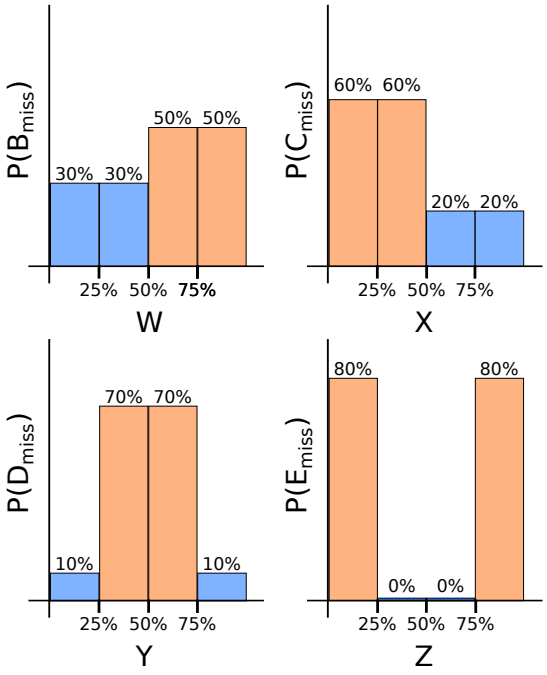

Fig. 6. The probability of missingness in variables B, C, D, and E correlate with the quartile in which the associated value of W, X, Y, or Z lies. Variable A is MCAR and thus missingness do not correlate with any other variable

mechanism I'm looking for?" The problem of exploration can then be further subdivided in three subquestions:

- Which pairs of variables are likely to have a MCAR/MAR relationship ?
- Which variable(s) are likely to have a MCAR/MAR relationship, given a specific variable with missing values?
- Which variable(s) with missing values are likely to have a MCAR/MAR relationship, given a specific variable?

To assist the anwser to those questions, each axis has a filter selection. By selecting a variable from the variable axes, or a range of values from the randomness similarity axis, the visualization dims the polycurves outside the selected items, highlighting the selected ones. The three filters compose using the logical operator AND.

Thus, to answer the first question, the analyst can select either the higher values from the similarity axis (selecting that tend to MCAR hypothesis) or the lower values (pairs that tend to MAR hypothesis). To answer the other questions, the filter can be composed with a selection of variables, showing only the pairs that include a specific one. The filter interaction also reduces the list of details-on demand: only the selected items appear on the histogram list.

Another interaction is the hover. Hovering a polycurve instantly makes the associated histogram appear at the top of the list, even if it was outside the scroll area, making it quicker to see histograms of specific pairs. When hovering either a histogram or polycurve, a synchronized glowing animation runs over the color, to reinforce the association between curves and histograms. The glows alternate between a highlight color outside the used scheme and the encoded color, so the color information is not totally overwritten by the animation.

To summarize, the proposed idiom derives two assistance metrics, and uses then in a visual encoding that supports overview, exploration, and diagnosis. The next section demonstrates the idiom with a synthetic dataset, to check how customized missing mechanisms appears in the idiom, and if their diagnosis is possible.

IV. Experiments

This section showcases the idiom in use with four datasets: a synthetic one—to validate if mechanisms are depicted as they should—and three real ones—to demonstrate usage in practical scenarios.

A. Synthetic Data

We used a data generator [42] to design the synthetic dataset. The intention is to include different levels of correlation in the missingness of variables, to assess if they are identifiable in the idiom when no noise is present.

The dataset has five variables with 25% of missing values (A, B, C, D, E), and four variables without missing values (W, X, Y, Z).

All variables are quantitative. The variables with missing values have a gaussian distribution ($\mu = 0$, $\sigma = 1$) and the others have a uniform distribution ($min = 0$, $max = 1$).

Variable A will have an MCAR mechanism with all other variables, that is, its missing values are randomly chosen without interference of any underlying model. Variables B, C, D, and E has missingness correlated with the values of W, X, Y, and Z, respectively. Figure Figure **??** describes how these values influence the probability of missingness.

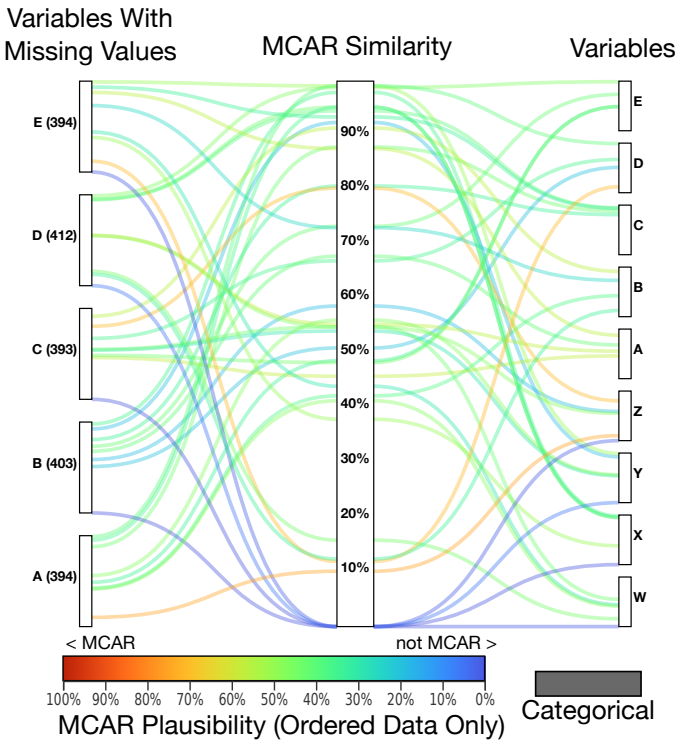

Fig. 7. The synthetic dataset depicted in the proposed idiom. Notice how the MAR relations are the only ones assigned to the extreme lower side of both derived metrics

The result is the visualization of Figure 7. Notice how the only lines that cross the similarity axis in 0% are the ones with the MAR mechanisms. Those lines are also the ones with the bluest color in the scale. All other pairs are, by data definition, MCAR. Notice how they appear closer to the 100% side of the metrics and do not take extreme values. The variance occurs because even uniform distributions have some chance of deviating from the expected.

This scenario shows that the idiom is able to depict missing mechanisms in synthetic data, without noise. The next datasets are from the UCI repository [43], and aims at evaluating the idiom in real analysis scenarios, where diagnosis is more complex.

B. Automobile Dataset

We chose this dataset due to its high usage in visualization research, and because of the different relationships with the missingness of the ”normalized-losses” variable. Figure 8 shows the view of this dataset, already selecting the normalized losses attribute on the left axis (as the others do not have enough missing values to estimate the mechanism). Notice how only a handful of variables without a strong MAR similarity evidence, showing the missingness correlates to many other variables of the dataset.

The line with the bluest color in Figure 8 is horsepower, which gives the hint that this variable has a high plausibility of being MAR. Hovering that line shows the histogram of Figure 9, which allow the analyst to inspect the distribution.

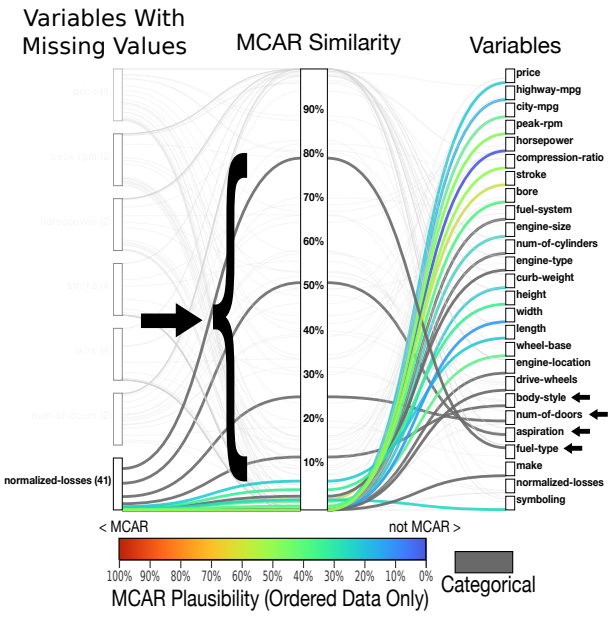

Fig. 8. The automobile dataset visualized in the proposed idiom. Notice how the variables body-style, num-of-doors, aspiration, and fuel-type are the only ones had MCAR Similarity above 10% with normalized-losses

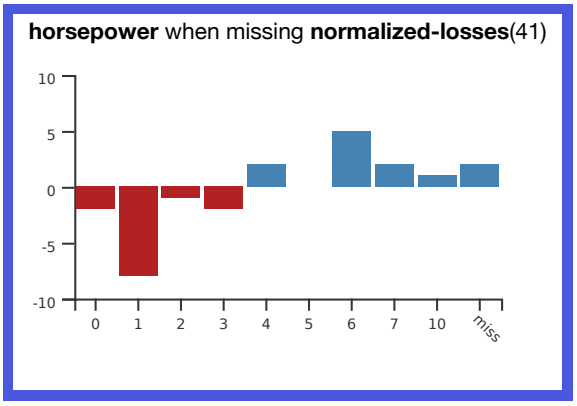

Fig. 9. The histogram view of the horsepower variable when missing normalized losses. The consistence makes clear that lower values of horsepower are less likely to have missing normalized-losses.

Notice how small values in horsepower are less likely than expected to have missing normalized-losses.

C. Cylinder Bands Dataset

We chose this dataset due to its high occurrence of missing values and high dimensionality, and because it has a good example of the ambiguity and uncertainty inherent to diagnosis.

Figure 10 shows the view of this dataset, already use a range selector in the similarity axes towards low scores. Notice how the variable Solvent_pct has more than half of its pairs in the selected range (Figure 10a), while the variable with most missing values, Blade_pressure, has less than half of pairs inside the same selection (Figure 10b). The filter dims out some pairs, but the length of the bar still gives an notion of how much is selected.

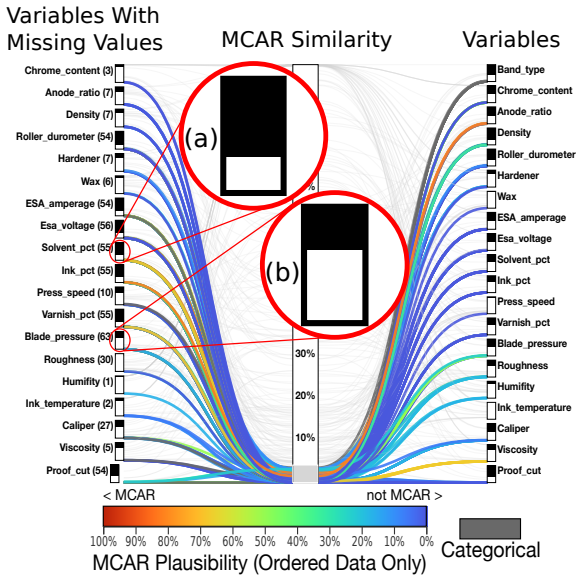

Fig. 10. The cylinder bands dataset depicted in the proposed idiom, selecting the lower scores of the MCAR similarity axis. Notice how the variable Solvent_pct (a) has more pairs in the selection than Blade_pressure (b).

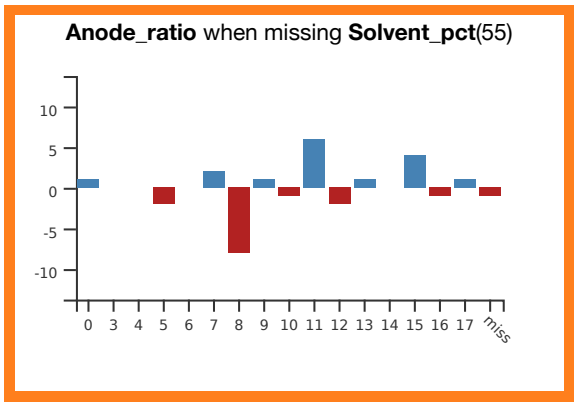

Fig. 11. The histogram view of the Anode_ratio variable when missing Solvent_pct. The inconsistence cast undertainty to diagnosis: it can be a MAR with a uncommon underlying model, or a MCAR with high deviation due to chance.

Figure 11 shows the histogram of how the missingness in Solvent_pct correlates with values in Anode_ratio—one of the pairs with high MCAR score in the color scale. Notice how this pair has one of the smallest randomness similarity score —which supports the MAR evidence— while also having a high randomness plausibility —which supports the MCAR evidence.

Such conflicting scores makes explicit an uncertainty that is inherent to diagnosis, and depicting it to the analyst is vital to support reasoning and diagnosis. In this case, the diagnosis consists of judging wherever the magnitude of deviation (similarity) is big enough to assume a MAR relationship with an underlying model that indeed has an unplausible shape, or if the randomness of the shape (plausibility) is big enough to assume this is just a MCAR distribution whose high deviation was only a matter of chance.

V. Final Remarks

This paper presented a visual-interactive idiom that uses derived metrics and progressive rendering to assist the diagnosis of missing data mechanisms. Each metric is taken pairwise, providing more detailed insights about the relationships in the data.

The derived metrics are two quantitative scores: a bootstrapped similarity between the missingness distribution and a random one (the more similar, the more likely the MCAR assumption), and a plausibility score that describes the consistency of the shape of the distribution (the more consistent, the more likely the MAR assumption). Those metrics act as supporting evidence, assisting the analyst in overviewing, exploring, and diagnosing the types of missing data.

The visual encoding consists of a parallel-axes approach, using also a list of histograms as a details on-demand view. The histograms and main views are linked through the interactions, including filtering and two-way hovering.

We evaluated the idiom using a synthetic dataset to demonstrate how specific mechanisms are depicted, and then used three public real datasets with missing values —Automobile, Cylinder Bands, and Thyroid Disease— to demonstrate the practical capabilities of the idiom.

There is an argument that the practical value of statistical tests to diagnose between mechanisms is unclear [44]. We argue that, with visualization, the value of diagnosis is clear: it provides an understanding of the underlying model that is, in itself, a piece of information. Rather than just giving a probability of a mechanism, visualization helps to understand why this mechanism happens, and what are its consequences on distributions. To name a few examples, specifically to our pairwise approach, understanding how the missingness correlates to other variables can: (1) give insight on how to fix the missing data in the collection process itself, (2) help choosing meaningful variables for imputation models, (3) check if simple methods that are only possible in MCAR are viable, and in which variables can use them, and (4) assist the creation of synthetic datasets with tailored mechanisms.

A. Limitations

There is a conflict between the constraints of screen space, cognitive overload, and computational power. Our idiom deliberately lays the load on the computer using the bootstrapping method, and as a consequence, it can be unsuitable for big data. However, the core contribution of the idiom is the extraction, depiction, and interaction with mechanism quantifications, it can be used with other metrics that may arise. Thus, research in optimizations, algorithms, and statistical tests—either to enhance or be an alternative to the bootstrap approach—could support not only this method, but also future ones.

An idiom based on parallel-axes is susceptible to the inherent problems of this encoding [45], especially visual clutter. This also means that methods that deal with this issue might

be applicable in the proposed idiom, such as adding supportive and coordinated views of the same data [46], or metadata navigation [47]. Such investigation could

The major limitation is the current lack of a user study to validate our approach. Although this paper focuses on describing the idiom and demonstrating it is practical applicability in both synthetic and real datasets, conducting a proper evaluation to empirically measure this efficacy is crucial for advancing visualization research [48]. Thus, we recognize that the first future work that arrives from this research must be an empirical evaluation, which could also compare variations of the idiom to achieve a higher level of efficacy.

B. Future Works

The most prominent direction of research is to develop novel idioms for missing data mechanism diagnosis, filling this gap in the literature and bringing it closer to the needs of data workers that deal with this uncertainty.

Additionally, a formal taxonomy of what means to diagnose a missing data mechanism and of the elements involved in this task could provide a foundation for such research. In this paper, we only briefly divide the task into overview, exploration, and diagnosis, and only derive two metrics to support mechanisms. Expanding and solidifying these concepts would make it easier to visualize the key information. Inspiration can be taken for the literature of visualization for medical diagnosis [49].

Finally, an ambitious direction is the development of a visual analytics pipeline that allows analysts to reason, explore, and deal with data that is possibly MNAR. Although MNAR is impossible to diagnose using the data at hand, a system that allowed domain experts to translate mental models into the data and explore their impacts —asking "what if?"— could lead to a better estimation of the "unknow unknows" of the studied phenomena. Adding the human intuition into the loop might be the missing link, and developing such a system—that consequently drives advances in general human-computer interaction, as well as in techniques for quantifying and visualizing uncertainties associated with estimations and subjective thinking—requires attention towards the mechanisms of missing data.

REFERENCES

[1] E. R. Buhi, P. Goodson, and T. B. Neilands, "Out of sight, not out of mind: Strategies for handling missing data," *American journal of health behavior*, vol. 32, no. 1, pp. 83–92, 2008.

[2] N. M. Laird, "Missing data in longitudinal studies," *Statistics in medicine*, vol. 7, no. 1-2, pp. 305–315, 1988.

[3] L. Kong, M. Xia, X.-Y. Liu, M.-Y. Wu, and X. Liu, "Data loss and reconstruction in sensor networks," in *2013 Proceedings IEEE INFOCOM*, pp. 1654–1662, IEEE, 2013.

[4] B. Roure, D. Baurain, and H. Philippe, "Impact of missing data on phylogenies inferred from empirical phylogenomic data sets," *Molecular biology and evolution*, vol. 30, no. 1, pp. 197–214, 2013.

[5] T. D. Pigott, "A review of methods for missing data," *Educational research and evaluation*, vol. 7, no. 4, pp. 353–383, 2001.

[6] A. Briggs, T. Clark, J. Wolstenholme, and P. Clarke, "Missing.... presumed at random: cost-analysis of incomplete data," *Health economics*, vol. 12, no. 5, pp. 377–392, 2003.

[7] G. J. Van der Heijden, A. R. T. Donders, T. Stijnen, and K. G. Moons, "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example," *Journal of clinical epidemiology*, vol. 59, no. 10, pp. 1102–1109, 2006.

[8] M. J. Knol, K. J. Janssen, A. R. T. Donders, A. C. Egberts, E. R. Heerdink, D. E. Grobbee, K. G. Moons, and M. I. Geerlings, "Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example," *Journal of clinical epidemiology*, vol. 63, no. 7, pp. 728–736, 2010.

[9] S. Demissie, M. P. LaValley, N. J. Horton, R. J. Glynn, and L. A. Cupples, "Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model," *Statistics in medicine*, vol. 22, no. 4, pp. 545–557, 2003.

[10] J. L. Schafer, "Multiple imputation: a primer," *Statistical methods in medical research*, vol. 8, no. 1, pp. 3–15, 1999.

[11] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica: Journal of the Econometric Society*, pp. 1–25, 1982.

[12] M. Soley-Bori, "Dealing with missing data: Key assumptions and methods for applied analysis," *Boston University*, vol. 4, pp. 1–19, 2013.

[13] S. Nakagawa and R. P. Freckleton, "Missing inaction: the dangers of ignoring missing data," *Trends in ecology & evolution*, vol. 23, no. 11, pp. 592–596, 2008.

[14] P. E. McKnight, K. M. McKnight, S. Sidani, and A. J. Figueredo, *Missing data: A gentle introduction*. Guilford Press, 2007.

[15] M. Skeels, B. Lee, G. Smith, and G. G. Robertson, "Revealing uncertainty for information visualization," *Information Visualization*, vol. 9, no. 1, pp. 70–81, 2010.

[16] J. K. Kim and C. L. Yu, "A semiparametric estimation of mean functionals with nonignorable missing data," *Journal of the American Statistical Association*, vol. 106, no. 493, pp. 157–165, 2011.

[17] Y. Yuan and G. Yin, "Bayesian quantile regression for longitudinal studies with nonignorable missing data," *Biometrics*, vol. 66, no. 1, pp. 105–114, 2010.

[18] M. Templ and P. Filzmoser, "Visualization of missing values using the r-package vim," *Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology*, 2008.

[19] R. J. Little, "A test of missing completely at random for multivariate data with missing values," *Journal of the American statistical Association*, vol. 83, no. 404, pp. 1198–1202, 1988.

[20] Z. Zhang, "Missing data exploration: highlighting graphical presentation of missing pattern," *Annals of translational medicine*, vol. 3, no. 22, 2015.

[21] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[22] R. Twiddy, J. Cavallo, and S. M. Shiri, "Restorer: A visualization technique for handling missing data," in *Proceedings Visualization'94*, pp. 212–216, IEEE, 1994.

[23] S. Popov, "Large-scale data visualization with missing values," *Technological and Economic Development of Economy*, vol. 12, no. 1, pp. 44–49, 2006.

[24] H. Song and D. A. Szafir, "Where's my data? evaluating visualizations with missing data," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 914–924, 2018.

[25] M. Templ, A. Alfons, and P. Filzmoser, "Exploring incomplete data using visualization techniques," *Advances in Data Analysis and Classification*, vol. 6, no. 1, pp. 29–47, 2012.

[26] K. Lakshminarayan, S. A. Harp, and T. Samad, "Imputation of missing data in industrial databases," *Applied intelligence*, vol. 11, no. 3, pp. 259–275, 1999.

[27] A. Bilogur, "Missingno: a missing data visualization suite," *Journal of Open Source Software*, vol. 3, no. 22, p. 547, 2018.

[28] N. J. Tierney and D. H. Cook, "Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations," *arXiv preprint arXiv:1809.02264*, 2018.

[29] N. Boukhelifa, M.-E. Perrin, S. Huron, and J. Eagan, "How data workers cope with uncertainty: A task characterisation study," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 3645–3656, 2017.

[30] J. J. Thomas and K. A. Cook, "A visual analytics agenda," *IEEE computer graphics and applications*, vol. 26, no. 1, pp. 10–13, 2006.

[31] M. R. Chernick, "Boostrap methods: a practitioner´ s guide/por michael r. chernick.," tech. rep.

[32] M. Angelini, G. Santucci, H. Schumann, and H.-J. Schulz, "A review and characterization of progressive visual analytics," in *Informatics*, vol. 5, p. 31, Multidisciplinary Digital Publishing Institute, 2018.

[33] D. W. Scott, "Sturges' rule," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 3, pp. 303–306, 2009.

[34] H. Kang and B. Shneiderman, "Visualization methods for personal photo collections: Browsing and searching in the photofinder," in *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, vol. 3, pp. 1539–1542, IEEE, 2000.

[35] C. Ware, *Information visualization: perception for design*. Morgan Kaufmann, 2019.

[36] A. Kirk, *Data Visualization: a successful design process*. Packt Publishing Ltd, 2012.

[37] S. K. Card and J. Mackinlay, "The structure of the information visualization design space," in *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, pp. 92–99, IEEE, 1997.

[38] M. S. T. Carpendale, "Considering visual variables as a basis for information visualisation," 2003.

[39] A. Inselberg, "The plane with parallel coordinates," *The visual computer*, vol. 1, no. 2, pp. 69–91, 1985.

[40] R. S. D. A. D. Lima, C. G. R. Dos Santos, and B. S. Meiguins, "A visual representation of clusters characteristics using edge bundling for parallel coordinates," in *2017 21st International Conference Information Visualisation (IV)*, pp. 90–95, IEEE, 2017.

[41] J. Heinrich, Y. Luo, A. E. Kirkpatrick, H. Zhang, and D. Weiskopf, "Evaluation of a bundling technique for parallel coordinates," *arXiv preprint arXiv:1109.6073*, 2011.

[42] Y. P. dos Santos Brito, C. G. R. dos Santos, S. de Paula Mendonça, T. D. Aráujo, A. A. de Freitas, and B. S. Meiguins, "A prototype application to generate synthetic datasets for information visualization evaluations," in *2018 22nd International Conference Information Visualisation (IV)*, pp. 153–158, IEEE, 2018.

[43] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[44] S. Van Buuren, *Flexible imputation of missing data*. CRC press, 2018.

[45] A. Dasgupta, M. Chen, and R. Kosara, "Conceptualizing visual uncertainty in parallel coordinates," in *Computer Graphics Forum*, vol. 31, pp. 1015–1024, Wiley Online Library, 2012.

[46] E. Bertini, L. Dell'Aquila, and G. Santucci, "Springview: Cooperation of radviz and parallel coordinates for view optimization and clutter reduction," in *Coordinated and Multiple Views in Exploratory Visualization (CMV'05)*, pp. 22–29, IEEE, 2005.

[47] A. Dasgupta, R. Kosara, and L. Gosink, "Meta parallel coordinates for visualizing features in large, high-dimensional, time-varying data," in *IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 85–89, IEEE, 2012.

[48] C. Forsell, "A guide to scientific evaluation in information visualization," in *2010 14th International Conference Information Visualisation*, pp. 162–169, IEEE, 2010.

[49] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister, "Evaluation of artery visualizations for heart disease diagnosis," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2479–2488, 2011.