

To Explore What Isn't There — Glyph-based Visualization for Analysis of Missing Values

Sara Johansson Fernstad and Jimmy Johansson

Abstract—This paper contributes a novel visualization method, Missingness Glyph, for analysis and exploration of missing values in data. Missing values are a common challenge in most data generating domains and may cause a range of analysis issues. Missingness in data may indicate potential problems in data collection and pre-processing, or highlight important data characteristics. While the development and improvement of statistical methods for dealing with missing data is a research area in its own right, mainly focussing on replacing missing values with estimated values, considerably less focus has been put on visualization of missing values. Nonetheless, visualization and explorative analysis has great potential to support understanding of missingness in data, and to enable gaining of novel insights into patterns of missingness in a way that statistical methods are unable to. The Missingness Glyph supports identification of relevant missingness patterns in data, and is evaluated and compared to two other visualization methods in context of the missingness patterns. The results are promising and confirms that the Missingness Glyph in several cases perform better than the alternative visualization methods.

Index Terms—Missing data, information visualization, glyphs.

24 Nov 2020

arXiv:2011.12125v1 [cs.GR]

1 INTRODUCTION

DATA sets with missing values, commonly known as incomplete or missing data, are a frequent challenge in data analysis across a range of domains. They are known to cause issues such as biased, uncertain and unreliable results. A large number of statistical methods have been developed for dealing with missing values, which are mainly focused on replacing missing data with plausible values (known as imputation) [1]. Meanwhile, the visualization and visual analysis of missing data is a largely overlooked topic, even though visualization has great potential to support understanding and knowledge generation from incomplete data. The awareness of the existence of missing values and the patterns relating to these missing values can be improved by visualization, and through this many potential issues and data uncertainties can be highlighted. In addition to supporting identification of issues arising during data generation and pre-processing, visualization of missing data can reveal important patterns, such as patients missing appointments in medical studies where the missingness may highlight a health issue. Missing data visualization can also facilitate decision making as to how missing values can be most appropriately dealt with. The application of suitable statistical methods for imputation require understanding of the patterns of missingness, and questions such as if records are missing at random or if there are structured patterns needs to be taken into consideration.

Research by Johansson Fernstad [2] and Johansson Fernstad and Glen [3] suggested a set of missingness patterns of particular importance for analysis. Johansson Fernstad [2] provided context to the complexities of analysing incomplete data, and identified issues to address through evaluation of visualization methods. This paper presents the Missingness Glyph (MissiG), a novel visualization method for analysis of

missing values in multivariate data. MissiG represent both univariate and multivariate patterns of missingness as well as relationships between missing and recorded values, to support the identification and understanding of missingness patterns, and through this highlight uncertainty to aid decision making. To ensure its usability, MissiG is designed based on well established glyph design principles [4], [5], [6], utilizing simple and clearly separable visual channels (colour, height and shape) for category separation, magnitude representation and distribution comparison. It is designed to work equally well as a standalone visualization or as an enhancement to existing multivariate data visualization methods. To demonstrate its versatility, the paper presents two standalone layout options for MissiG, linear and radial, as well as examples of how it can be used as enhancement for Heatmap and Parallel Coordinates (PC). The usability of MissiG is demonstrated through two evaluations that compare MissiG with a Heatmap, which represents missing values using colour, and with PC where missing values are represented through location. The results indicate that MissiG is generally better or equally good as other visualization methods when it comes to identification of missingness patterns, and that the performance of other visualization methods can be improved by enhancing them with MissiG. The main contributions of this paper are:

- MissiG, a novel glyph visualization that supports analysis of missingness patterns in data;
- two layouts (linear and radial) with additional enhancements of MissiG;
- two evaluations comparing the performance of visualization methods in the context of identification of missingness patterns.

The paper is structured as follows: Section 2 provides background, the MissiG design is described in Section 3. Its usability is demonstrated by examples in Section 4 and evaluations in Section 5, with conclusions in Section 6.

- Sara Johansson Fernstad is with the School of Computing, Newcastle University, UK.
E-mail: sara.fernstad@newcastle.ac.uk
- Jimmy Johansson is with Linköping University, Sweden.

2 BACKGROUND

Missing records may occur in any type of data (numerical, categorical, text, relational networks etc.), and the most appropriate method for visualizing missing data will depend on the type of the recorded data. The focus of this paper lies mainly on missing values in multivariate (numerical) data. The visualization presented in the paper can, however, easily be adapted to categorical data and the missingness patterns they are based on are equally applicable to numerical and categorical data. The identification of missing values in data can be a challenging pre-processing step to data analysis, since missing values can be represented in a range of different ways during data collection. While the identification of missing values is an important challenge that can be facilitated by visualization, it is not within the scope of this paper. The contributed visualization method assume that missing values are explicitly marked in the data. This section will provide a brief overview of the analysis of data with missing values. More in depth discussions can be found in, for instance, Johansson Fernstad [2].

2.1 Analysis of Missing Data

The effect missing values have on analysis results depends both on the missingness mechanism, described in 2.2, and on how the missing values are handled. It may also be greatly affected by the degree of missingness and distribution of missing values across the data set. The two main approaches to dealing with missing values are removal and imputation. Removal is when data items with missing values are removed prior to analysis. This approach carries a considerable risk of biased results, unless the values are missing completely at random. Imputation is when missing values are replaced by estimated values. There exist a large number of imputation methods, ranging from replacement with arithmetic mean or random draws from representative distributions, to complex multiple imputation methods that combine several imputations following a set of rules [1]. Imputed values may bias and affect the analysis results, depending on the appropriateness of the imputation method.

2.2 Missingness Patterns

The missingness mechanism [7] is a model of how the probability of an observation being missing depends on its own value and on the values of other variables. There are three mechanisms defined: *Missing Completely at Random*, *Missing at Random* and *Missing Not at Random*. The missingness mechanisms are rarely known prior to analysis, they are fairly complex and may be difficult to apply to an exploratory analysis approach. More recent research suggests patterns that may be more straightforward for describing missingness in data. Wang and Wang [8] suggested three patterns in context of classification data, focussing on the distribution of missing values. In a later paper [9] they described concepts of relevance for understanding the impact of missing values on the analysis results, addressing both missing values and the relationship between missing and recorded. Based on previous research and interviews with data science practitioners, Johansson Fernstad [2] defined a set of three missingness patterns of relevance for analysing missingness in data, as described below.

Amount Missing (AM) refers to the relative amount of missing values in a variable or a data item, and supports

understanding of the distribution of missing values in the data set. Insight into AM in variables can, for example, support identification of variables where the missingness may be particularly difficult to deal with, or highlight subsets of data where conclusions drawn from the recorded values may be unreliable due to the large amount missing values. It can also be useful to investigate whether the missingness may be randomly distributed, since the amount missing would be relatively equal across the data set for random missingness.

Joint Missingness (JM) is a multivariate or pairwise pattern that refers to the amount of data items that have missing values in more than one variable at the same time. The pattern may, for example, occur in survey data where participants who refuse to answer a specific question also tend to not answer another specific question. Identification of JM can support discovery of issues in data collection or pre-processing that cause missingness to propagate across the data, as well as identification of data subsets where missingness may need to be dealt with differently to missingness in subsets with less JM.

Conditional Missingness (CM) is a pairwise pattern that describes the relationship between items that are missing in one variable and their recorded value in other variables. It aims to describe patterns where the probability of missingness is conditional upon recorded values, and as such supports understanding of relationships between missing and recorded. Investigation of CM can be useful to understand the cause of missingness, and can support decisions on how to deal with the missingness. For example, if missing values in variable A tend to have low recorded values in variable B , then imputation of missing values in A based on items with low values in B may be more valid than imputation based on all items.

As discussed in Johansson Fernstad [2], these three patterns bring together the main characteristics of the previously suggested missingness patterns. They also provide a more straightforward description of patterns than the missingness mechanism. The research presented in this paper address methods for exploration and identification of missingness in data based on the concepts of AM, JM and CM.

2.3 Missing Data Visualization

It can often be meaningful to consider missing values as information bearing signals, rather than issues that need to be removed, since they may provide valuable information and highlight potential issues in data gathering, pre-processing and analysis processes. Fielding et al. [10] and Djurcilov and Pang [11] provide examples from health-related surveys and meteorological studies where the absence of data is more informative than an estimated value, and emphasize the value of visualization for understanding of missingness in data.

Shape Coding [12] is an early example of representing missing values with colour to support identification of missingness related patterns in multivariate data. Twiddy et al. [13] adopted a similar approach where recorded and missing values were visually separated using a colour scheme. MANET [14], [15] was another early example where visual representations of missing values were incorporated in the visualization software. The XGobi [16] and gGobi [17] systems focussed on exploration of missingness, and represented missing data by imputed values which were linked to a separate view to keep track of missing values. Wang and Wang [8] presented

a visualization method for missing values in classification data, utilizing self-organizing maps [18] for clustering, with main focus on whether the missing values were randomly distributed, unevenly distributed or biased towards a particular class. Additionally, a number of R-packages support visualization of missing values, as described by for examples Unwin [19]. Some of the more recent packages include Naniar [20], which includes a range of table and barchart based visualization as well as an UpSet [21] style set visualization; and extracat [22] including the Visna visualization of missing values. VIM (Visualization and Imputation of Missing values) [23], utilize various visual attributes to highlight missingness in histograms, scatter plots, PC and other common visual representations. The miP (multiple imputation plots) package [24] use VIM to visualize imputation results from a range of packages. Cheng et al. [25] developed an R-package that distinguish imputed missing values from recorded values by colour. Some R-packages, such as AmeliaView [26], also provide graphical user interfaces for manipulation and control of imputation methods. While interesting, a large part of previous research in missing data visualization may not be able to efficiently deal with growing data sizes. Many methods are focussed on supporting imputation and representing imputed values, which is an important challenge that visualization can facilitate. Such methods may, however, be less appropriate for explorative analysis and knowledge generation.

Considering missing values as part of the broader issue of data quality, a range of tools have been designed to support data profiling and cleaning. Profiler [27] utilise inference and data mining approaches to identify quality issues, and use visualization to investigate them in context of the larger dataset. 'Know-your-enemy' [28] use a similar approach, with automated quality checks to support visual exploration of quality issues in time series data. Triana et al. [29] utilise data quality dimensions to enrich visualization with quality information. Schulz et al. [30] defined missing data as one of several data descriptors, and used PC with lines intersecting a point below the axis to represent missing values. Cedilnik and Rheingans [31] represented uncertainty using procedurally generated annotations, and represented missing data by a distance based probability value. Xie et al. [32] focussed on data quality and used imputation to obtain quality values for missing data. Arbesser et al. [33] presented a system where missing data is one of several quality classes represented by colour. These papers define missingness as one of several quality descriptors in the more general context of data quality, thus not focussing on visual analysis and representation of the specific features of missingness, as is the focus of the work presented here.

Although the importance of visualization of missing data has been emphasized [3], [34], [35], little research has investigated how to best represent missing values in data. Eaton et al. [36] discussed the impact of missing data on the interpretation of visualization, and evaluated three approaches to representing missing values in line graphs. Their results implied that poor indication of missing values has negative effect on interpretation, and suggest that visualization should be enhanced by dedicated visual attributes, annotation or animation to indicate the existence of missing data. They did, however, not suggest what visual attributes may be most appropriate. A later study by Andreasson and Riveiro [37] evaluated the impact on decision making of three techniques for representing missing

values visually (emptiness, fuzziness and emptiness plus explanation), and found that while emptiness plus explanation was the most preferred technique and rendered the highest decision confidence it also resulted in higher risk behaviour. Fernandes et al. [38] evaluated the impact on decision making of a set of uncertainty visualization on mobile displays, concluding that CDF plots and quantile dotplots can successfully improve decision making despite the limited display space. Song and Szafir [39] investigated four categories of missing value representation in line-graphs and bar charts, defined as highlight, downplay, annotation and information removal. They concluded that highlighting of missing values is usually perceived as higher quality, while downplay and information removal is perceived as lower quality. They also found that visualization using highlighting and annotation while preserving the continuity of the recorded data results in the highest perceived data quality and confidence in results.

Recent work by Johansson Fernstad [2] evaluated the performance of three visualization methods in the VIM package [23] (scatterplot matrix, heatmap and PC), which are enhanced by visual attributes for representation of missing values. The performance was investigated for tasks relating to the identification of the AM, JM and CM patterns. The results indicated that heatmap with missing values represented by colour generally performed best for tasks relating to AM and JM, while PC with missing values represented by lines intersecting a point above the axis, performed better for CM tasks. Conclusions from that study suggest that it is important to:

- 1) Include clear frequency representations through size, possibly combined with colour. This generally needs more attention, and in particular for visualization that lack a natural frequency representation.
- 2) Include features that connect missing and recorded, as well as missing and missing, across multiple variables. This is important for identification of multivariate patterns, and in particular need more attention in visualization with limited representation of connections.
- 3) Avoid separation of missing and recorded values in different sets of representations. This is particularly important for CM patterns, but also since missingness analysis would commonly be part of a more extensive analysis where overall data patterns are of interest.

It was also concluded that location rarely is suitable as sole representation of missing values, due to the intrinsic meaning of location in methods such as PC and scatterplot, and that additional features should be used to emphasize the missingness.

3 THE MISSINGNESS GLYPH

This section will describe MissiG, a novel visualization method designed to support exploration of missing values in data, based on the missingness patterns and results presented by Johansson Fernstad [2]. MissiG is designed to be usable both as a standalone visualization or as a glyph-style enhancement to be added to multivariate visualization methods, thus utilizing the strengths of common visualization methods while overcoming their limitations in terms of missing data analysis. While the examples provided in this paper mainly focus on numerical data sets, MissiG can easily be adapted to categorical data.

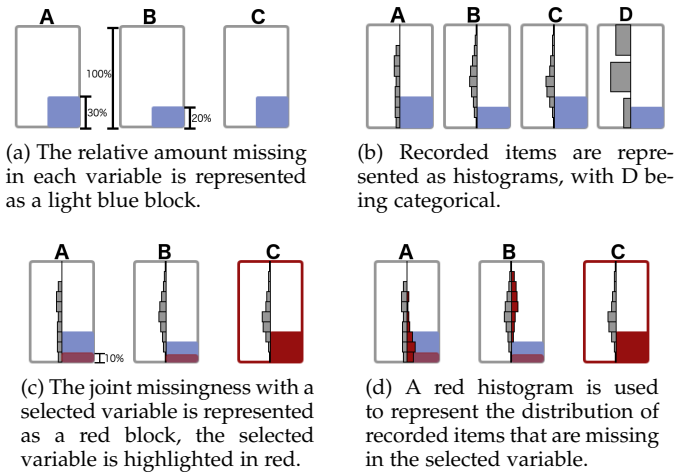


Fig. 1: The basic structure of MissiG for three or four variables. Variable C is selected in 1c and 1d

3.1 Visual Representation of Missingness Patterns

In its basic form, each MissiG glyph represent one variable in a data set. The glyph has a rectangular shape using blocks and histograms to represent the data and missingness patterns. As displayed in Fig. 1a, the relative amount missing (AM) is represented by the height of a light blue block, where the full length of the main rectangle represents 100%. In the figure, 30% of values are missing in the left and right variables, while 20% of values are missing in the centre variable. With block height being a straightforward representation of frequency, this approach provides an easily interpreted overview of the amount missing in multivariate data sets. The distribution of recorded items in the variable is represented by a grey histogram in the left half of the glyph, as shown in Fig. 1b, with low values represented at the bottom and high values at the top. The width of the histogram bins corresponds to the relative number of items with recorded values within the bin range. For a categorical variable, each bin would represent a unique category and the width would correspond to the relative frequency of that category (variable D , Fig. 1b)

The relative number of items that are subsequently missing in a pair of variables (i.e. the JM) and the relationship between missing in one variable and recorded in another (i.e. the CM) both operate on variable pairs and thus there exist a larger number of JM and CM relationships than there are variables. Furthermore, while the JM of a variable pair is non-directional ($A_{JM} \rightarrow B_{JM} = A_{JM} \leftarrow B_{JM}$), the CM is directional ($A_{CM} \rightarrow B_{CM} \neq A_{CM} \leftarrow B_{CM}$). Thus, there are twice as many unique CM relationships as JM relationships in a data set. To avoid visual clutter and increase scalability, MissiG in its basic form only displays representation of JM and CM for a selected variable. A red block is used to represent the JM, as displayed in Fig. 1c where variable C is selected and highlighted in red, and where the height of the red block in A indicate that 10% of items have concurrently missing values in A and C . CM is represented through a red histogram in the right half of the glyph, which display the distribution of recorded values in an unselected variable, for the subset of items that have missing values in the selected variable. In Fig. 1d variable C is selected, and the red histogram in A represent

the distribution in A of items that are missing in C but recorded in A . From Fig. 1d it is visible that items with missing values in C tend to have relatively low values in A while they have comparably high values in B .

When analysing missingness it is often relevant to understand if the values are missing at random or not. The visual features of MissiG can help out-ruling random missingness through multiple properties. Firstly, where missingness is completely random it can be expected that each variable has roughly the same amount missing; if the height of the blue blocks in the MissiG glyphs varies greatly, we can hence assume that the missingness is not completely random. Secondly, if the missingness is random we can expect certain levels of JM. This expected JM can be defined as $E(\vec{d}_j, \vec{d}_k) = P(\vec{d}_j) \cdot P(\vec{d}_k)$, where $P(\vec{d}_j)$ and $P(\vec{d}_k)$ are the probabilities that a value is missing in \vec{d}_j and \vec{d}_k respectively. If the JM deviates greatly from $E(\vec{d}_j, \vec{d}_k)$ we may assume that the missingness is not completely random. Thus, if 50% of values are missing in variable A and 50% are missing in B , we expect 25% of values to be concurrently missing in A and B if missingness is completely random. Thirdly, for CM, if the data is randomly missing in A , we would expect the overall distribution of recorded values in variable B to be similar to the distribution of recorded values in B for the subset of items that have missing values in A . Hence we would expect the grey and red histograms in the glyph for B to have a similar shape. If the shapes of the two histograms differ considerably, we can conclude that the missingness in A is not random and there may be a relationship between recorded values in B and missingness in A .

3.2 Glyph Design Considerations

Four pieces of information are represented in a glyph: 1) the relative amount missing in a variable; 2) the relative amount jointly missing with another, selected variable; 3) the overall distribution of recorded data; and 4) the distribution of recorded data for items that are missing in another, selected variable. MissiG was designed based on these definitions and a number of well established glyph design principles.

MissiG utilize three main visual channels: colour to distinguish between three categories of information (missing values in a variable, recorded values in a variable, and data relating to another, selected variable), size/height to represent magnitude, and shape for representation and comparison of data distributions. This is based on the principles of Typedness [4] and Semantic Relevance [5], which emphasize that visual channels should be appropriate for the semantics of the underlying data, as well as the guidelines suggested by Borgo et al. [6]. Orderability and Channel Capacity [4] has also been taken into account, with orderable data (magnitude) being represented by height which has a relatively high capacity, and non-orderable categories being represented by the lower capacity colour channel. Borgo et al. [6] also emphasize the importance of glyph property interaction normality as well as the use of perceptually uniform properties. Normality and uniformity is maintained in MissiG through magnitude values (AM and JM) being represented relative to the full glyph height, which is the same for all glyphs, and by distribution representations utilizing the full glyph height (this was not completely addressed in an earlier version of the glyph, as described in Section 5).

The importance of simplicity, the use of well known visual channels and well defined rules is highlighted by several design principles (Learnability [4], Complexity and Density, Simplicity and Symmetry [6]). The MissiG design addresses this through the use of basic visual channels representing a clearly defined type of information, with colour representing categories of information (missing values, recorded values, and data relating to a selected variable), height representing magnitudes and shape representing value distributions. The design principles of Separability, Searchability [4] and Channel Composition [5] emphasize the importance of clearly identifiable, non-conflicting visual channels. In the MissiG design this is achieved through the use of visually different and non-overlapping visual channels. Alternative designs that have been considered would generally increase the complexity of the glyph. For example, a pie-chart style representation could have been used for the magnitude (AM and JM), but this would have complicated the additional representation of distributions, as well as losing the intuitiveness of a minimum and maximum value and, to some extent, the ability to easily compare heights across glyphs. For numerical variables, a violin style representation could have been used instead of histograms, these would however not be viable for categorical variables.

Pop-out effect and saliency of visual channels are important to take into account in glyph design, such that more important information is made more salient. This is, for example, highlighted by the principles of Attention Balance [4], Pop-out Effects and Visual Hierarchy [5], and Importance-based mapping [6]. Colour is one of the most salient visual channels [40], and more intense colours, such as red, tend to have a stronger pop-out effect. In the MissiG glyph, colour is used to discriminate between the three categories 'missing data', 'recorded data' and 'data related to a selected variable'. Patterns related to a variable that is interactively selected by the user can be expected to be of more interest to the user, thus a more salient colour (red) is chosen to highlight these patterns, while the non-selected missing and recorded values have less salient colours (light blue and light grey). This is also related to the principle of Focus and Context [4].

3.3 Layouts and Enhancements

To provide an as flexible representation as possible, MissiG can be used as a standalone visualization technique with different layouts, or as a glyph-style enhancement to existing methods. In this section two standalone layouts are suggested, with added visual encodings to enhance representation of missingness patterns, but further layouts can also be considered. Furthermore, two examples are provided of how MissiG can be used to extend two common visualization methods.

Linear layout is the simplest multivariate representation, similar to PC, as displayed in Fig. 2. This layout facilitates comparison of in particular AM patterns, since the height of blocks are directly comparable across multiple variables. Upon selecting a variable by clicking on it, as displayed in Fig. 2b where x_5 is selected, the JM with the selected variable is enhanced through arcs linking variable pairs. The thickness of the arc represents the JM of the variable pair. Alternatively, the JM of all variable pairs can be represented by arc thickness, as in Fig. 2c. In Fig. 2b and 2c it is, for instance, visible from the thickness of the arcs that x_5 and x_3 have relatively high

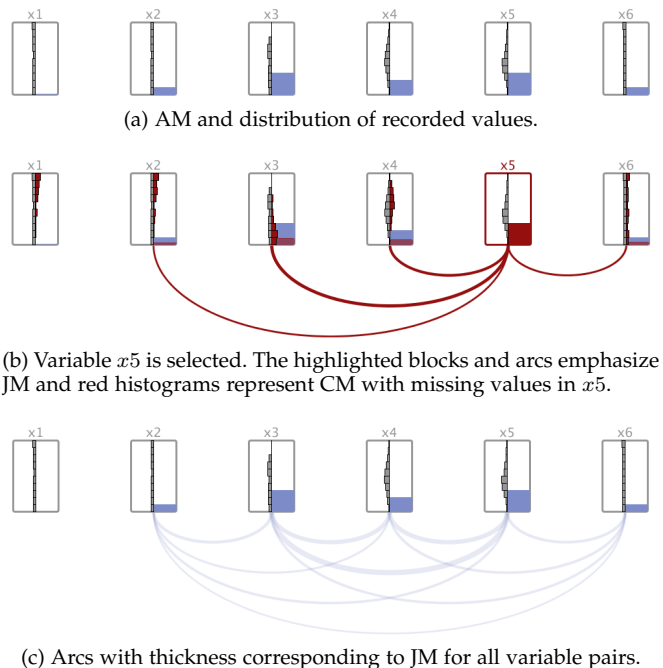


Fig. 2: MissiG with linear layout for a synthetic data set with 6 variables, where x_1 has no missing values and the remaining variables have 10% – 30% missing.

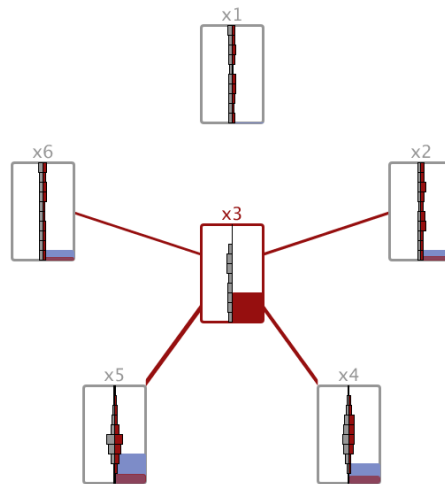
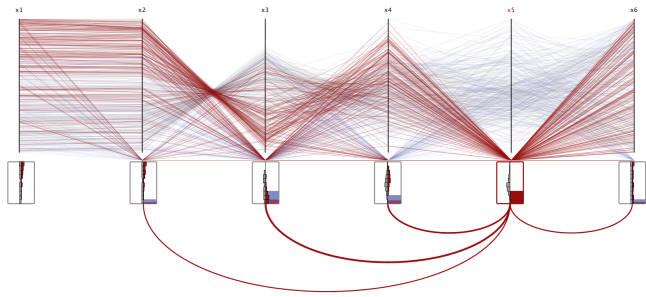


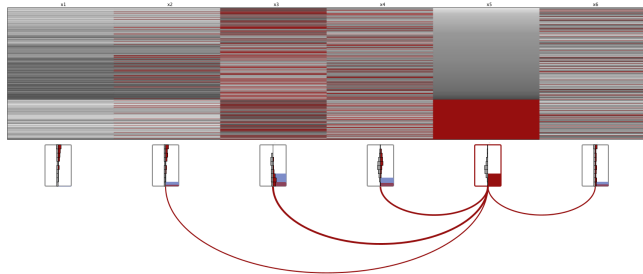
Fig. 3: MissiG with radial layout for the same data as in Fig. 2, with x_3 selected. Highlighted blocks and bands emphasize JM and red histograms represent CM with missing values in x_3 .

JM, while the JM of x_5 and x_2 is lower. The red histograms in the figure also reveal some CM patterns. For instance, the red histograms in the x_1 , x_2 and x_4 glyphs, which are denser for higher values than corresponding grey histogram, indicate a relationship between missing values in x_5 and high values in x_1 , x_2 and x_4 ; while there appear to be a relationship between missing in x_5 and low values in x_3 . For x_6 , the red and grey histogram have a similar shape, indicating that there is no relationship between missing in x_5 and recorded in x_6 .

Radial layout is a representation where the glyph of an interactively selected variable of particular interest is highlighted and positioned in the centre of a circle (Fig. 3). The other variables in the data set are positioned on the circumference



(a) PC where missing values are represented below the axes and items with missing values in x_5 is highlighted in red.



(b) Heatmap with missing values represented in red and recorded values represented using grey scale, with dark grey corresponding to low values and light grey corresponding to high values.

Fig. 4: PC and Heatmap enhanced with MissiG glyphs. Variable x_5 is selected in both figures.

of the circle, making it possible to analyse the relationships between a single variable of particular interest with respect to all other variables, inspired by the work of Johansson et al. [41]. The width of red bands between glyphs represents the JM with the selected variable, similar to the arcs in the linear layout. The pairwise relationships with the variable of interest may be more easily compared in the radial layout than in the linear, since the distance between the centred variable and other variables is equal and thus band length will not impact perception. The high JM of x_3 and x_5 is clearly identifiable in Fig. 3, and it is visible from the width of the red band that the JM of x_3 and x_4 is also relatively high. The red histograms are generally mirroring the shapes of corresponding grey histogram, which indicates that there are no strong CM patterns for items with missing values in x_3 .

Enhancements to existing techniques. As previously mentioned, the flexibility of the glyph design also makes it usable as an enhancement to existing visualization. Fig. 4 displays examples where MissiG is used as an enhancement of PC (Fig. 4a) and Heatmap (Fig. 4b). The PC and Heatmap, as implemented here, already include some representation of missing values. For the PC missing values are represented below the axis, with red highlighting of items with missing values for a selected variable. In the Heatmap, cells with missing values are represented by red colour, while recorded values are represented by grey scale. In the implementation described here, which is mainly focussed on exploration of missing values, MissiG is interactively linked to the other technique through variable selection. Selection of a variable of interest will highlight MissiG as described above. When linked to PC, data items with missing values in the selected variable will be coloured red in PC (see Fig. 4a), and when linked to Heatmap the rows will be ordered based on their

value in the selected variable (see Fig. 4b). The addition of MissiG to these visualization aims to overcome some of the limitations identified in [2]. In Fig. 4a it is for instance visible that the JM of x_5 and x_3 , as well as x_5 and x_4 is relatively high, something that likely would have been hard to spot in the PC alone. CM patterns are easier to identify in PC, as confirmed in [2], but the added MissiG provide a confirmation of patterns such as the relationship between missing values in x_5 and high values in x_1 , x_2 and x_4 ; particularly in situations where the data is dense with a large amount of missing values. The Heatmap in Fig. 4b displays the same data set and selection as the PC. While the Heatmap generally performed well in [2], particularly for identification of AM and JM patterns, it is still likely that its performance will decline with denser data and high amounts of missing values. For example, the CM relationship between missing in x_5 and high values in x_4 , and the relationship between missing in x_5 and high values in x_1 are equally visible in the MissiG representation, while the former is less perceivable than the latter in the Heatmap, due to x_4 having a relatively large number of missing values while x_1 have none. Alternative approaches to enhance visualization methods with MissiG can also be considered. The trade-off between increased understanding of missingness patterns and the potentially increased cognitive burden and possible interference with overall analysis has to be taken into account within the analytical context. In the suggested enhancements the glyph is not overlaying the other visualization, which reduces its visual interference while it requires additional screen-space compared to, for example, overlaying the glyphs on the PC axes. Furthermore, in an analytical context where missing data is not the main focus, it may be appropriate to use a less salient colour than red to highlight missing values and to provide the glyphs as an on-demand feature.

3.4 Scalability

Due to the relative simplicity of the glyph design and the representation of summary patterns rather than features of individual data items, the scalability of MissiG is comparable to or even better than the scalability of common multivariate visualization methods such as PC, Heatmap or Scatterplot matrix. The visual complexity of a single glyph will not grow with an increasing number of data items and, hence, from a perception and interpretation point of view it will not matter if the data set includes a few hundred items or a million items. On the other hand, with a multivariate MissiG layout, an increasing number of variables will require additional visual objects to be drawn and impact the usability of the visualization, similarly to how it impacts common multivariate visualization methods. The exact number of variables that can be visualized effectively depends not only on the available display space, but also on the missingness patterns in the data as well as which layout and visualization options are being used. A data set where only a small number of variables have missing values will likely result in a less cluttered display than a data set with missing values in most variables; and a linear display where all JM arcs are being displayed (as in Fig. 2c) will often result in a more cluttered display than if only JM arcs of a selected variable is displayed (as in Fig. 2b). Fig. 5, 6, 7 and 8 show the use of MissiG, PC and Heatmap for differently sized data sets with different missingness distributions and patterns. As with other

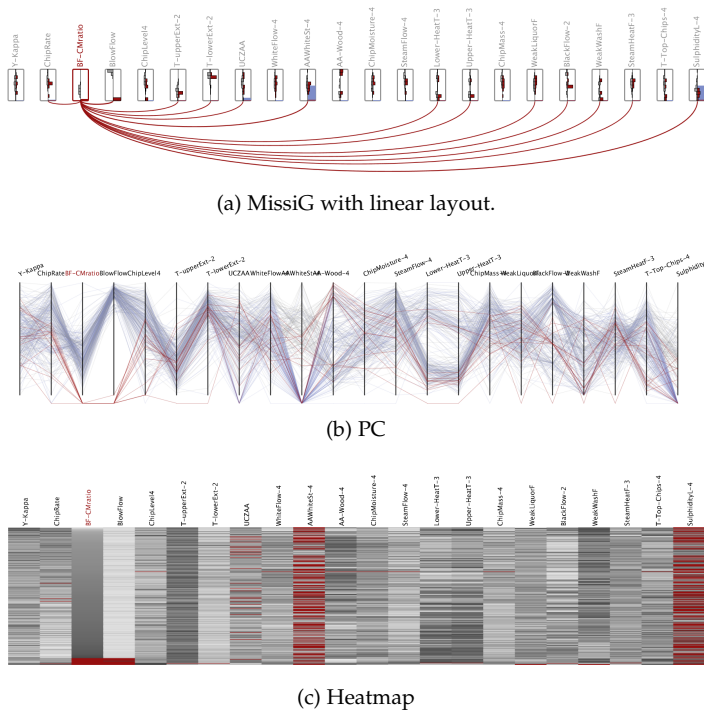


Fig. 5: Visualization of the *Kamyry Digester* data set with 22 variables and 301 items. *BF-CMratio* (third variable from the left) is selected and highlighted.

multivariate visualization techniques, approaches to deal with more complex and higher dimensional datasets may for example include simplified glyph design with details on demand, dimensionality reduction and quality metric approaches to aid identification of the most interesting missingness patterns.

4 EXAMPLES OF PATTERN IDENTIFICATION

This section provides examples of how MissiG can be useful for identifying missingness patterns. Larger versions of the figures (5 - 8) are provided in the supplemental material. The linear and radial MissiG layouts will be compared with PC and Heatmap. In PC, missing values are represented at a point below the axis, and items with missing values for a selected variable are highlighted in red. The Heatmap use red colour to represent missing values and grey scale to represent recorded values with light grey corresponding to high values. The rows in the Heatmap are ordered based on their values for a selected variable. Two public data sets with missing values are used for the examples, the *Kamyry Digester* data set [42] which contains 22 variables (observation date was removed from the data), 301 items and has 352 (5.6%) missing values, and the *Communities and Crime* data set [43], which in full contains 128 variables and 1994 data items. For the purpose of the examples here, 34 variables were randomly selected from the *Communities and Crime* data, with a total of 24126 (55.2%) missing values.

Fig. 5 and 6a displays the *Kamyry Digester* data set with the *BF-CMratio* variable selected. It is visible in the MissiG plots (5a and 6a) and the Heatmap (5c) that there is a strong JM between *BF-CMratio* and *BlowFlow* (third and fourth from left). This can to some extent also be identified in the PC (5b) through the red line below the third and fourth axis from the left, but it is

considerably harder to appreciate the amount of jointly missing values in PC. Looking at the histograms in the MissiG glyph of *SulphidityL-4* (rightmost glyph in 5a and upper left in 6a), it is visible that the shapes of the grey and red histograms differ, with the red histogram being denser towards lower values. This indicates a potential CM between missing values in the selected *BF-CMratio* and low recorded values in *SulphidityL-4*. The same pattern is considerably harder to identify in the PC (5b) and in the Heatmap (5c) due to the high number of missing values in *SulphidityL-4* which largely masks the patterns of recorded values. These JM and CM patterns are clearly not random patterns, and if not already known it may prompt the analyst to further examine the connection between these variables and investigate potential issues in the data collection or pre-processing steps.

Fig. 7 and 6b display the same data set as in Fig. 5 but with *AAWhiteSt-4* (tenth variable from the left) selected in all views. Starting with CM patterns, it is visible in Fig. 7a and 6b that the shapes of the red histograms are very similar to the shapes of corresponding grey histograms. This indicates that items with missing values in *AAWhiteSt-4* are randomly distributed across the recorded values of other variables, which may aid the choice of imputation method. These patterns are harder to visually verify in the PC and Heatmap (Fig. 7b and 7c), and in particular when patterns occur such as the dark grey blocks in variables *Lower-HeatT-3* and *Upper-HeatT-3* in the Heatmap (column four and five to the right of the selected *AAWhiteSt-4*), which are a result of the two density peaks in the variables (as visible from the histograms in 7a) combined with the ordering of items which in this figure is based on values in *AAWhiteSt-4*. While investigating JM, two potentially interesting relationships are identified in Fig. 7a and 6b. Firstly, from the red block in the rightmost variable (*SulphidityL-4*), which completely overlaps the blue block, it appears that all items that are missing in *AAWhiteSt-4* are also missing in *SulphidityL-4*. Both *AAWhiteSt-4* and *SulphidityL-4* have around 50% of values missing, which means that around 25% of items should be jointly missing if the values were missing completely at random, hence, the high JM indicates a non-random missingness pattern suggesting that the missingness in these variables and its cause should be investigated in conjunction. The pattern is easily identifiable also in the Heatmap (Fig. 7c) through the blocks of red rows, and in the PC (Fig. 7b where the items with missing values in *AAWhiteSt-4* (red lines) all intersects below the *AAWhiteSt-4* axis, although the percentage missing is hard to read from the PC. The second JM pattern that can be identified in the MissiG representations is that the JM between *AAWhiteSt-4* and *UCZAA* (two steps left of *AAWhiteSt-4* in 7a) is lower than expected from a random pattern. With half of values missing for *AAWhiteSt-4* it is expected that the red block in *UCZAA* would be half the size of the blue block in *UCZAA* if the missingness was random, but the red block is considerably smaller and hence indicates a non-random pattern. The same pattern can be spotted in the Heatmap, although less obvious, while it is considerably harder in PC since it does not clearly represent the frequency of missing values.

Fig. 8 and 6c displays 34 variables from the *Communities and Crime* data set. The *Community* variable (second from left) is selected in Fig. 8a and 8b, and some patterns related to distribution of missing values are easily identified in the linear MissiG layout. Only 15 variables have missing values. Of

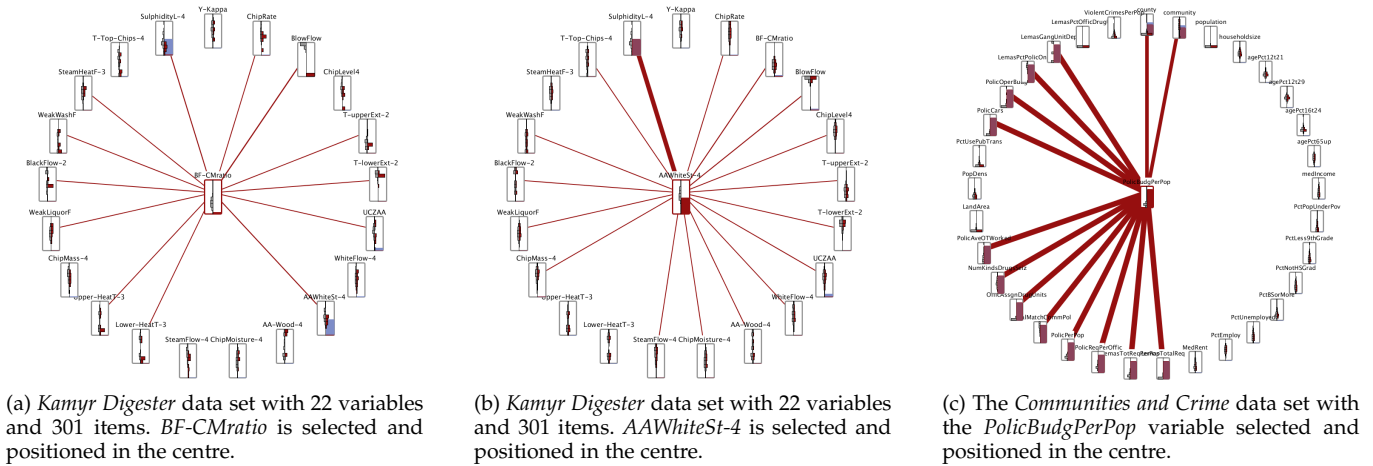


Fig. 6: Visualization of two data sets using MissiG with radial layout.

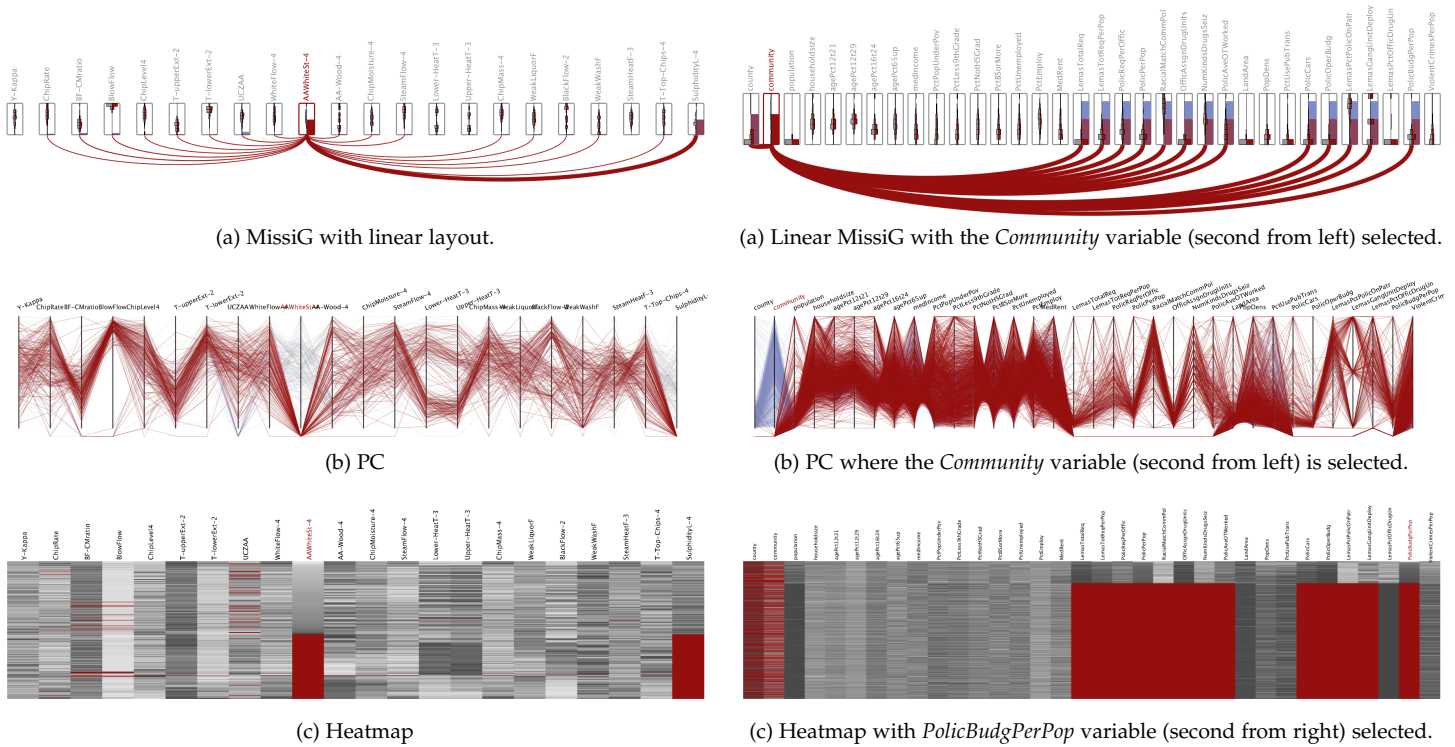


Fig. 7: Visualization of the *Kamyg Digester* data set with 22 variables and 301 items. *AAWhiteSt-4* (tenth variable from the left) is selected and highlighted.

Fig. 8: Visualization of 34 variables and 1994 data items from the *Communities and Crime* data set.

these, the two leftmost have around 50% of values missing, as visible from the blocks in the lower right part of corresponding glyphs, while the remaining 13 variables have nearly 85% missing, which can be seen from the blue blocks that are nearly as high as the full height of the glyphs. In the PC (Fig. 8b) it can be seen from the high number of red lines that the selected variable has a relatively high AM, but the AM in other variables is hard to appreciate. It can be seen from the two leftmost glyphs in Fig. 8a that nearly all items with missing values for *Community* also have missing values for *County*, which is also visible in the PC (8b) from the small number of red lines linking from missing values in *Community* (second

axis) to recorded values along the leftmost *County* axis. It is also clearly visible in the MissiG representation that approximately half of the items that have missing values in the 13 variables to the right, also have missing values for the *Community* variable, since the red blocks in the 13 variables is around half the height of the blue blocks, which can be expected given the AM of the individual variables. While it is visible in the PC that there is JM between *Community* and the 13 variables to the right, it is difficult to estimate the size of JM.

In Fig. 6c and 8c the *PolicBudgPerPop* variable is selected (second from right in Heatmap and centre in radial MissiG). It is clear from both figures that a majority of values (around 85%) are missing from this variable, and that more or less all of the

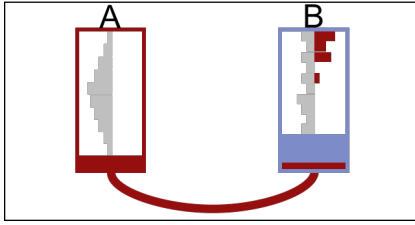


Fig. 9: The initial MissiG design, with amount missing blocks across the width of the glyph and histograms above the blocks.

items with missing values in *PolicBudgPerPop* also have missing values in 12 other variables, as visible from the red blocks in the MissiG completely covering the blue blocks, and from the red blocks in the Heatmap where rows are ordered by values in *PolicBudgPerPop*, this indicates a relationship between missing in *PolicBudgPerPop* and missing in the other 12 variables that is not occurring at random. If this JM pattern had been random, a JM rate of around 85% would be expected rather than 100%. In Fig. 6c it is however visible that the JM for the selected *PolicBudgPerPop* and the *County* and *Community* variables (top and top right glyph) is likely around 85%, since a part of the blue block is visible above the red block in the *County* and *Community* blocks, and that the grey and red histograms for other variables have similar shapes, which indicate a random relationship with missing values in *PolicBudgPerPop*. These patterns are not as easily detected in the Heatmap, where it for instance is hard to appreciate the amount of jointly missing values with *County* and *Community*. While the high AM suggest that any imputation would introduce too much bias, the knowledge of JM patterns can help understand data collection issues that may need to be addressed.

5 EVALUATION OF THE MISSINGNESS GLYPH

Two usability studies were carried out to establish the usability of MissiG. The first was designed using interactive visualization in the lab, while the second was an online study comparing static visual representations. The following section will describe the studies and their results in detail.

5.1 Visualization Methods

Six visualization methods were compared in the studies, including different MissiG layouts and extensions. There was a slight variation in the MissiG design between the first and second study, as the glyph was improved after the first study. In the initial design (Fig. 9) the amount missing blocks were stretched across the full width of the glyph, and histograms were squeezed into the space available above the blocks, resulting in different histogram heights for variables and difficulties comparing histograms across variables. As shown in Fig. 1 the amount missing blocks in the final design only cover half the width, to allow more space for the histograms, and use opacity and black borders to increase visibility of the part of the red CM histogram that overlap the blocks. The visualization methods compared were:

Linear MissiG (MissiG-L): The linear layout of MissiG with display of JM arcs only for the selected variable, as in Fig. 2b.

Radial MissiG (MissiG-R): The radial layout of MissiG, where the selected variable is positioned in the centre, as in Fig. 3. This layout was only included in the second study.

Heatmap (HM): Heatmap where missing values are represented by red cells, and recorded values are represented in grey scale with dark corresponding to low values and light corresponding to high values.

Heatmap with MissiG (HM+MissiG): Heatmap with the same colouring as above but enhanced with MissiG, as in Fig. 4b.

Parallel Coordinates (PC): PC where missing values are represented by polylines intersecting a point below the axis, and items with missing values for a selected variable is represented by red polylines.

Parallel Coordinates with MissiG (PC+MissiG): PC with missing value representation as above but enhanced with MissiG, as in Fig. 4a.

5.2 Hypotheses and Tasks

Based on the results in [2], the evaluations were designed to test the following hypotheses:

- H1** The MissiG glyphs will perform better than PC and better or equally well as Heatmap for AM tasks.
- H2** PC+MissiG will perform better than PC for AM tasks.
- H3** HM+MissiG will perform better or equally well as Heatmap for AM tasks.
- H4** The MissiG glyphs will perform better than PC and better or equally well as Heatmap for JM tasks.
- H5** PC+MissiG will perform better than PC for JM tasks.
- H6** HM+MissiG will perform better or equally well as Heatmap for JM tasks.
- H7** The MissiG glyphs will perform better than Heatmap and better or equally well as PC for CM tasks.
- H8** PC+MissiG will perform better or equally well as PC for CM tasks.
- H9** HM+MissiG will perform better than Heatmap for CM tasks.
- H10** Overall, MissiG will be the most preferred visualization method by participants.

Tasks were aimed to address different aspects of the three missingness patterns. This included approximation of percentage of missing values in a variable; identification of the variable with most missing values; identification of the variable pair that has highest joint missingness; comparison of difference in joint missingness between variable pairs; evaluation of trends in recorded data that possibly relates to missing values (such as: items with missing values in variable A tend to have high recorded values in variable B); and identification of differences between the general data distribution and the distribution of items that have missing values in some variable. To cover this as broadly as possible, two different questions were defined for each missingness patterns, with multiple choice style answers provided. The questions were defined as follows:

- AM1:** Approximately how much data is missing in variable X ?
- AM2:** Which of the following variables have the highest number of missing values?
- JM1:** With which of the following variable does variable X have the highest joint missingness?

- JM2:** Is the joint missingness of variable X and Y higher than the joint missingness of variable X and Z ?
- CM1:** Which of the following trends is most related to missing values in variable X ?
- CM2:** Which image displays the highest number of variables with a clear difference between the general data distribution and the distribution of items that are missing in the selected variable?

For both studies, the performance of the visualization methods was analysed in terms of accuracy and response time when completing the tasks. Based on the study hypothesis, three sets of analysis were relevant for each missingness pattern: 1) MissiG vs HM vs PC; 2) HM vs HM+MissiG; and 3) PC vs PC+MissiG. For the first set of significance testing, where more than two visualization methods were compared, one way ANOVA with repeated measures was used when the result data were normally distributed. If data were not normally distributed the Friedman test was used, followed by post-hoc tests for pairwise comparison to identify for which combinations of visualization methods the performance was significantly different, using Wilcoxon signed rank test with a Bonferroni correction applied resulting in a significance level set at $p < 0.017$ for the first study (comparing three methods), and at $p < 0.0125$ for the second study (comparing four methods). For the second and third set of significance testing, where pairs of visualization methods were compared, a dependent t-Test was used for normally distributed data, while Wilcoxon signed rank test was used for non-normally distributed data.

5.3 First Study

A study with 15 participants was conducted to evaluate the performance of an initial implementation of the MissiG visualization. This study compared MissiG-L with Heatmap and PC, as well as comparing standard Heatmap and PC with versions enhanced with MissiG glyphs, using an interactive environment allowing for highlighting of glyphs and polylines in PC, and sorting of rows in Heatmap (as described in Section 3.3). Three of the above questions (**AM1**, **JM1** and **CM1**) were used in the study.

5.3.1 Experimental Design and Procedure

The experiment was designed as a within-subject study with visualization method as factor. Each participant performed 45 tasks and equally many tasks were performed for each visualization method and pattern using the data sets described in 5.3.2. No data set was used more than once per participant. Performance was measured in terms of accuracy and response time when performing the tasks. Ethical approval was received prior to the study.

The study was conducted individually in a controlled setting using a 15-inch MacBook Pro and an external screen where the study interface was displayed as a fixed size 1700x920px window. An initial scripted presentation was used to ensure that all participants possessed the basic knowledge needed to interpret the visual representations and understand the missingness patterns and tasks. This was followed by a training period including a small number of test tasks using the different visualization methods. The training was used as a means for the participants to become familiar with the tasks, visualization methods and experimental environment. For the

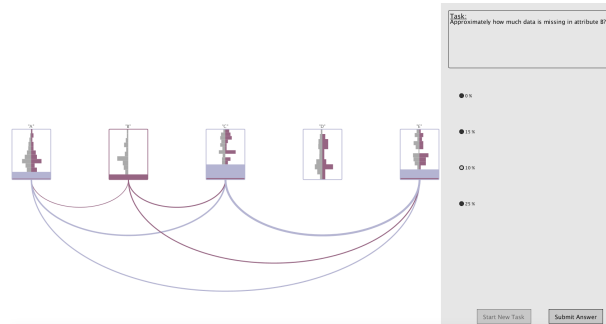


Fig. 10: The interface of the first study, with interactive visualization on the left and multiple choice options to the right.

experimental phase the tasks and visualization methods were counterbalanced using a Latin-square procedure [44], resulting in a unique ordering for each participant and, hence, reducing the potential learning impact on the results. The test environment (Fig. 10) consisted of two panels, one displaying one of the interactive visualization methods representing a study data sets, and the other displaying the task and multiple choice answers, along with buttons for submitting the answer and for displaying a new question. Response time was measured from when a question was displayed until the answer was submitted, allowing the participants to take a break before displaying a new question. The answers provided and response times were stored in log files, which were later used to analyse the results. The experimental phase was followed by a short questionnaire collecting information about the participants and their previous experience of data analysis, visualization and missing data, as well as information about which visualization method they found easiest, hardest and preferred to use.

5.3.2 Data

Three publicly available data sets were used and modified through controlled removal of values, to maintain realistic data structures while controlling the missingness patterns in the data. A total of 45 data sets were generated, 15 based on the *User Knowledge Modelling* (UKM) data set [45] with 5 variables and 403 items, 15 based on the *Concrete Compressive Strength* data set [46] with 9 variables and 1030 items, and 15 based on the *Parkinsons* data set [47] with 23 variables and 197 items. Varying levels of uniformly distributed noise, with a noise level between 1% and 15%, was randomly added, and the data sets were separated into three groups, one for each missingness pattern. Missingness patterns were created by replacing numerical values with a *NaN* string, with between 0% and 40% of values removed from each variable. The structure of missingness in variables was defined using a similar approach to Fernstad [2]. The variable names of the original data were replaced by letters to avoid impact of preconceptions based on variable names.

5.3.3 Results

15 participants finished the study, 2 were female and 13 male. The biggest age group among participants was 25-34 years (46.7%), followed by 45-54 (26.7%), 35-44 (20%) and 18-24 (6.7%). Participants were asked to rank their level of experience of 1) visualization methods, 2) data analysis, and 3) missing data, using 5 point likert scales ranging from No prior experience (1) to Professional (5). 46.7% ranked their experience of

TABLE 1: Pairwise results for MissiG, Heatmap and PC.

AM Wilcoxon	Accuracy		Response Time	
	Z	p	Z	p
HM vs MissiG	0.000	1.000	-0.284	0.776
PC vs MissiG	-2.505	0.012	-2.726	0.006
PC vs HM	-3.017	0.003	-3.181	0.001
JM Wilcoxon	Accuracy		Response Time	
	Z	p	Z	p
HM vs MissiG	-1.342	0.180	-2.897	0.004
PC vs MissiG	-3.373	0.001	-3.408	0.001
PC vs HM	-3.275	0.001	-3.294	0.001
CM Wilcoxon	Accuracy		Response Time	
	Z	p	Z	p
HM vs MissiG	-1.793	0.073	-0.568	0.570
PC vs MissiG	-1.303	0.193	-3.408	0.001
PC vs HM	-.905	0.366	-3.408	0.001

visualization methods high (4 or 5), while almost equally many (40%) ranked their visualization experience low (1 or 2). A clear majority (86.7%) ranked their experience of data analysis high, while only 6.7% ranked it low. The opposite was the case with missing data experience, with only 13.3% ranking their experience high and 60% ranking their experience as low. In addition to the results presented in this section, the descriptive statistics of the results are provided as supplemental material.

Amount Missing: The mean values with 95% confidence intervals for AM task results are displayed in Fig. 11. Statistical testing using Friedman test confirmed significant differences for both accuracy ($\chi^2(15) = 14.085, p = 0.001$) and response time ($\chi^2(15) = 12.113, p = 0.002$) when comparing MissiG with Heatmap and PC. The confidence intervals indicate worse performance for PC both for accuracy and response time. While there is a small overlap in response time between MissiG and PC, research by Cumming and Finch [48] conclude that when comparing groups using confidence intervals of individual group estimates, the p-value is near the significance value when the confidence limit of an interval reaches approximately the midpoint between the point estimate and the confidence limit of the other interval. These results were supported by the post-hoc analysis using Wilcoxon signed rank test (table 1, top) which confirmed significantly worse performance of PC compared to Heatmap and MissiG for AM tasks. For Heatmap compared to HM+MissiG, the confidence intervals overlap and no statistical significance was found for neither accuracy nor response time. PC+MissiG performed better than PC, although the confidence intervals for accuracy overlap. Using Wilcoxon signed rank test, there was no statistically significant difference for accuracy, while the difference in response time was significant ($Z = -2.669, p = 0.008$). These results all support **H1**, **H2** and, in part, **H3**.

Joint Missingness: Fig. 12 displays the mean values with 95% confidence intervals for JM task results. Comparing MissiG with Heatmap and PC, the confidence intervals indicate worst performance for PC and slightly better performance for MissiG compared to Heatmap. Friedman test confirm significant differences for both accuracy ($\chi^2(15) = 25.721, p < 0.001$) and response time ($\chi^2(15) = 24.400, p < 0.001$). Post-hoc tests using Wilcoxon signed rank test (table 1, centre) shows significantly worse result for PC compared to MissiG and Heatmap for both accuracy and response time, and also significantly worse response time for Heatmap compared to MissiG. Although HM+MissiG perform slightly better than

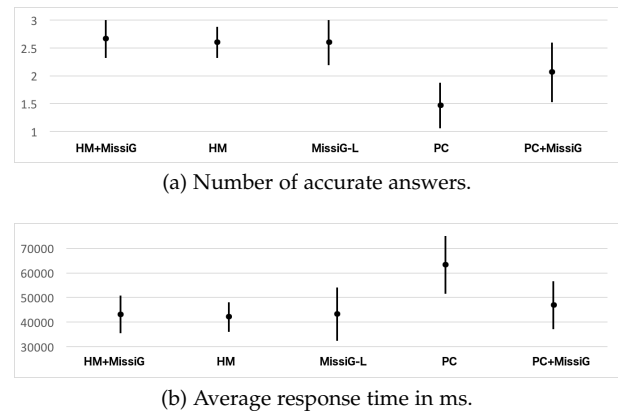


Fig. 11: Confidence intervals for AM tasks

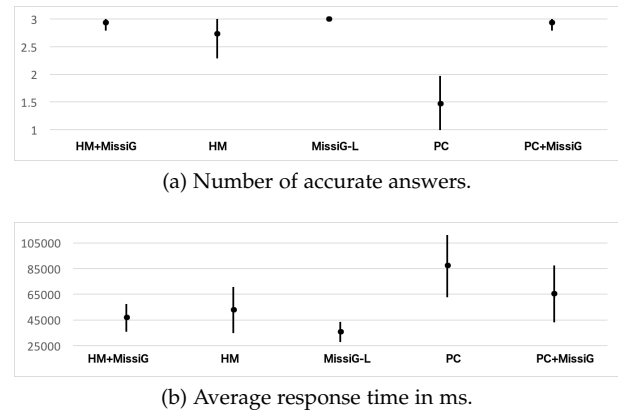


Fig. 12: Confidence intervals for JM tasks

Heatmap, the confidence intervals overlap for both accuracy and response time, and no significant differences were found. The performance of PC+MissiG for JM tasks is better than for PC, particularly in terms of accuracy where the two confidence intervals are clearly separated. Analysis using Wilcoxon signed rank test found statistically significant differences for accuracy ($Z = -3.244, p = 0.001$) as well as response time ($Z = -2.385, p = 0.017$). These results support **H4**, and, in part, **H5** and **H6**.

Conditional Missingness: The mean values with 95% confidence intervals for CM task results are displayed in Fig. 13. Comparing MissiG with Heatmap and PC the confidence intervals indicate slightly worse accuracy performance for MissiG and worse response time for PC, although the intervals largely overlap for both performance measures. Friedman test results indicate significant differences for both accuracy ($\chi^2(15) = 6.045, p = 0.049$) and response time ($\chi^2(15) = 22.800, p < 0.001$). The Wilcoxon signed rank test (table 1, bottom) however does not confirm any significant differences for accuracy when the Bonferroni correction has been applied, while the worse response time of PC is significant. The results when comparing HM+MissiG with Heatmap, and PC+MissiG with PC, were not significant. The results for CM tasks hence in part confirms **H7** (response time for PC is worse than for MissiG), but not **H8** or **H9**.

Further to the measured performance, information was gathered with regards to which visualization method the participants found easiest to use, hardest to use and which they preferred to use. A majority, 50%, found HM+MissiG the

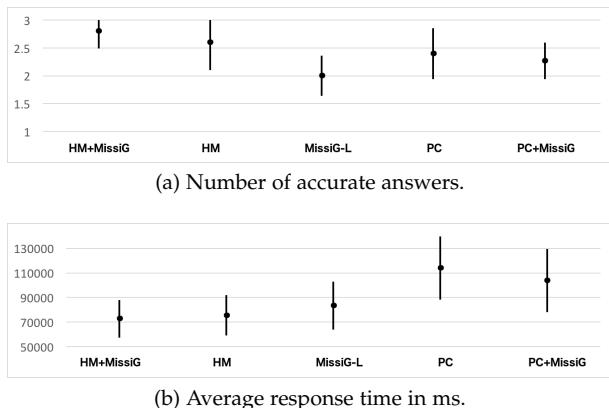


Fig. 13: Confidence intervals for CM tasks

easiest visualization to use, followed by 20% each for MissiG and Heatmap, and 10% for PC+MissiG. All participants stated that they found PC the hardest to use. The most preferred visualization to use was HM+MissiG, 64.7%, followed by 11.8% each for MissiG, Heatmap and PC+MissiG, and 0% for PC. This generally support H10 and indicates that the MissiG design is generally well received by the participants and in particular when combined with Heatmap.

5.4 Second Study

The first study was followed by an online study, aiming to further confirm the results and investigate a broader set of tasks related to the three missingness patterns (using all six questions in Section 5.2), as well as including the radial MissiG layout. Furthermore, the second study focused on the visual representation of data, using static images rather than an interactive environment to reduce the potential impact of variation in interactivity across methods. The study compared MissiG (both MissiG-L and MissiG-R) with Heatmap and PC, as well as Heatmap with HM+MissiG, and PC with PC+MissiG.

5.4.1 Experimental Design and Procedure

The experiment was designed as a within-subject study with visualization method as factor. Each participant performed 36 tasks, with one task per question, visualization method and missingness pattern, using the data sets described in 5.4.2. Performance was measured in terms of accuracy and response time. Ethical approval was received prior to the study.

The study was conducted online, and implemented using the Gorilla experiment builder [49]. The experiment was separated into three phases, one for each missingness pattern, each consisting of a training part and a test part. The training included descriptions of the visualization methods and how to interpret them, in context of the relevant missingness pattern, followed by training using the same type of questions as in the test phase but with feedback on whether the response was accurate to support understanding of the question and visualization. The test phase was similar to the training, with the difference of not including description of visualization and not providing feedback on whether responses were correct or not. The presentation order of visualization methods was randomized for both training and test phases, to reduce the impact of learning effects. The order of missingness patterns were fully counterbalanced, resulting in 6 different orders which

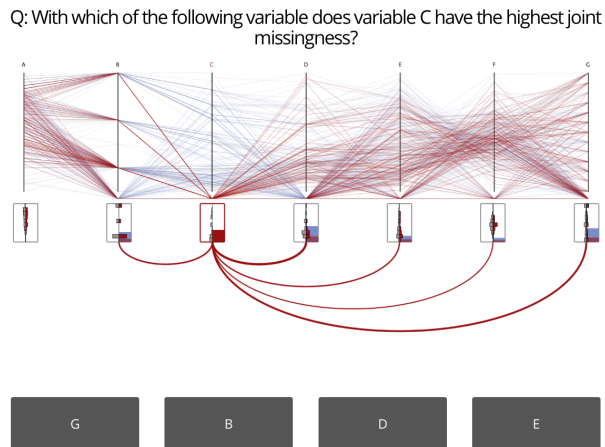


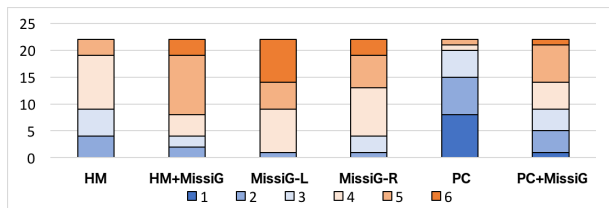
Fig. 14: The interface of the second study, with question and visualization at the top and answer buttons at the bottom.

were randomly balanced across the participants. The study interface (Fig. 14) consisted of a question and static image of the visualization method, with a set of multiple choice answers available through buttons. The study was restricted to only run on computers (not mobile phones and tablets) to reduce impact of screen size, and a range of anonymized study data was recorded through Gorilla, of which accuracy and response time was used for the analysis. Background information about participants were collected through a questionnaire, and at the end of each missingness pattern phase the participants were asked to rank their visualization preference for the task.

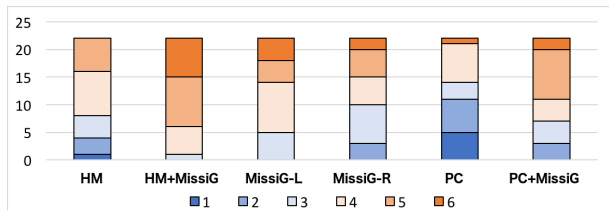
5.4.2 Data

As for the first study, a public data set was used and modified through controlled removal of values, to maintain realistic data structures while controlling the missingness patterns in the data. A total of 54 data sets were generated, using the 'cars' [50] data set, which contains real measurements of 392 cars. For each car, seven variables were collected (miles per gallon, number of cylinders, displacement, horsepower, weight, acceleration year, origin). The variable origin that describes where the cars were made (Europe, America and Asia) is categorical and was therefore removed, since some of the visualization methods used (PC in particular) does not perform well with categorical data. The variable names were anonymized and participants did not know which data set was used, thus limiting the impact of preconceptions based on variable names and any need of specific knowledge about cars.

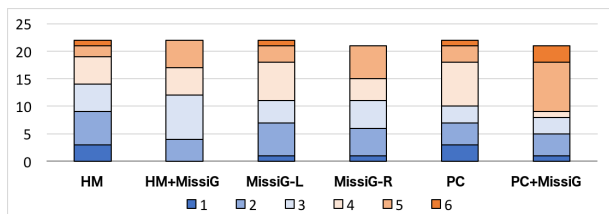
Missingness patterns were introduced by replacing numerical values with a NaN string. For AM and JM patterns, between 0% and 40% of values were randomly removed for each variable. This generated data sets where values are missing completely at random, and JM occur as a result of the random missingness in multiple variables, which works well for the tasks defined in Section 5.2 for AM and JM pattern identification. CM patterns required a more controlled removal, with slightly different approaches taken for CM1 and CM2 tasks. First, between 5% and 10% of values were removed randomly from all variables. Then two variables, X_1 and X_2 , were chosen and between 35% and 70% of values in X_1 were removed for data items with recorded values below the first quartile or above the third quartile in X_2 . Through this generating CM



(a) Preference of visualization methods for AM tasks.



(b) Preference of visualization methods for JM tasks.



(c) Preference of visualization methods for CM tasks. One participant did not provide answer for MissiG-R and PC+MissiG.

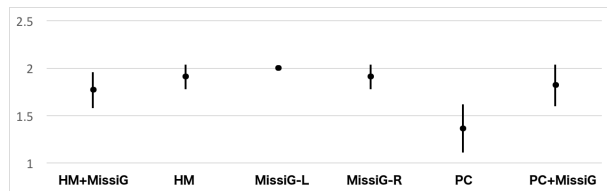
Fig. 15: Visualization method preference for different tasks, ranked using a 6 point likert scale ranging from Strongly Disliked (1) to Strongly Liked (6). Blue colour represents negative responses and red represent positive responses.

patterns between missing values in X_1 and low or high values in X_2 . For **CM2** tasks, which require more than one data set per task, each data set was separated into four subsets with four variables in each, of which one displayed more CM patterns.

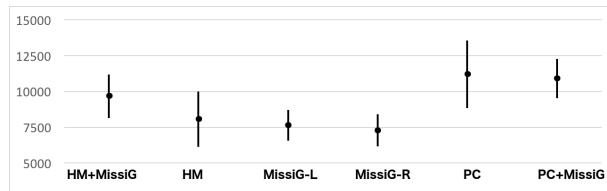
5.4.3 Results

24 participants initially finished the study. Of these two were removed from analysis due to data quality issues, one of them finishing the whole study in less than three minutes (only possible if not reading instructions and answering without trying to solve tasks), and the other due to the response time to a single question being more than 13 minutes (likely caused by disruption while responding). Of the 22 included participants (5 female, 16 male and 1 preferred not to disclose gender) the majority were between 25 and 44 years old (18-24: 4.5%, 25-34: 45.5%, 35-44: 31.8%, 45-54: 18.2%, 55 or older: 0%). Participants were asked to rank their level of experience of: 1) visualization methods, 2) data analysis, and 3) missing data, using 5 point Likert scales ranging from None (1) to Expert (5). 68.2% ranked their visualization experience as high (4 or 5), while only 9.1% ranked it as low (1 or 2); 81.8% ranked their experience of data analysis as high, while none ranked it as low; and 22.7% ranked their experience of dealing with missing data as high, while 31.8% ranked it as low.

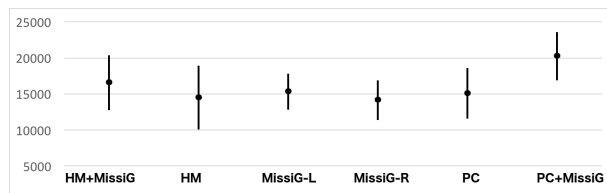
Amount Missing: Fig. 15a displays a summary of the preference of the participants for each visualization method for AM tasks, using a likert scale ranging from 1 (strongly disliked) to 6 (strongly liked). From the figure we can conclude that



(a) Number of accurate answers.



(b) Average response time in ms.



(c) Response time in ms for correct answers only.

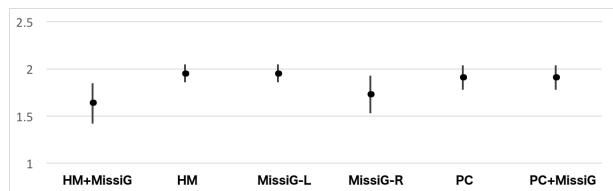
Fig. 16: Confidence intervals for AM tasks

MissiG-L was the most preferred visualization method for AM tasks, followed by HM+MissiG and MissiG-R, which supports **H10**. PC was the least liked method for AM tasks. Confidence intervals for the performance of the visualization methods are presented in Fig. 16. When comparing the two MissiG layouts with Heatmap and PC, Friedman tests confirmed statistically significant differences in accuracy ($\chi^2(22) = 29.110, p < 0.001$) and response time ($\chi^2(22) = 20.018, p < 0.001$), while differences for response time only for accurate answers were not significant. For accuracy and response time, Wilcoxon signed rank test (table 2) revealed significantly worse performance for PC compared to all other visualization methods. No other differences were significant. These results confirm that MissiG performs better than PC, and equally good or better than Heatmap for AM tasks (**H1**). Wilcoxon tests revealed significant difference in response time for Heatmap compared to HM+MissiG ($Z = -2.127, p = 0.033$) with better performance for Heatmap, while differences in accuracy and response time for correct answers were not significant. These results does not confirm **H3**. For PC compared to PC+MissiG, Wilcoxon tests showed significant differences for accuracy ($Z = -2.202, p = 0.028$) and response time for correct answers ($Z = -2.315, p = 0.021$), with higher accuracy but slower response time when PC is enhanced with MissiG. The overall difference in response time was however not significant. This support **H2** for accuracy, but not response time. An explanation to slower response times for visualization methods enhanced with MissiG can be the additional cognitive burden of combining two visualization methods instead of one, as well as the likely unfamiliarity of the MissiG visualization. The accuracy results, which generally would be more important than response time, however does support the benefit of enhancing PC with MissiG.

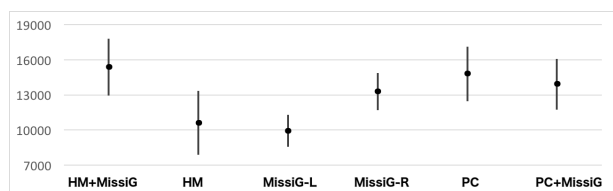
Joint Missingness: Fig. 15b displays a summary of the preference of the participants for each visualization method for JM

TABLE 2: Pairwise results for AM tasks for MissiG-L, MissiG-R, Heatmap (HM) and PC.

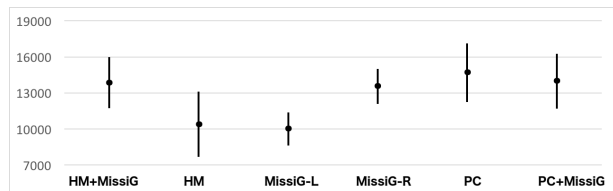
Wilcoxon	Accuracy		Response Time	
	Z	p	Z	p
MissiG-L vs HM	-1.414	0.157	-0.081	0.935
MissiG-R vs HM	0.000	1.000	-1.088	0.277
PC vs HM	-3.207	0.001	-2.646	0.008
MissiG-R vs MissiG-L	-1.414	0.157	-0.828	0.408
PC vs MissiG-L	-3.500	<0.001	-3.685	<0.001
PC vs MissiG-R	-3.207	0.001	-3.393	0.001



(a) Number of accurate answers.



(b) Average response time in ms.



(c) Response time in ms for correct answers only.

Fig. 17: Confidence intervals for JM tasks

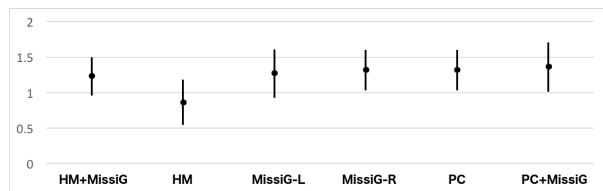
tasks. From the figure we can conclude that HM+MissiG was the most preferred visualization method for JM tasks, followed by MissiG-L, PC+MissiG and Heatmap, which generally confirms **H10** for JM tasks. PC was the least liked method also for JM tasks. Confidence intervals for the performance of the visualization methods for JM tasks are presented in Fig. 17. When comparing the two MissiG layouts with Heatmap and PC, Friedman tests confirmed statistically significant differences in response time ($\chi^2(22) = 32.073, p < 0.001$) and response time only for accurate answers ($\chi^2(22) = 30.491, p < 0.001$), while the differences for accuracy were not significant. For response time and response time of accurate answers, Wilcoxon signed rank post-hoc tests (table 3) revealed significantly better performance for Heatmap and MissiG-L compared to MissiG-R and PC, which support **H4** for response time, although not for accuracy. Wilcoxon tests revealed that the difference in performance for Heatmap compared to HM+MissiG was significant for accuracy ($Z = -2.646, p = 0.008$), response time ($Z = -3.360, p = 0.001$) and response time only for accurate answers ($Z = -3.295, p = 0.001$), with overall better performance for Heatmap. There were no significant performance differences for PC compared to PC+MissiG. Thus, **H5** and **H6** are not supported by these results.

Conditional Missingness: Fig. 15c displays a summary of the

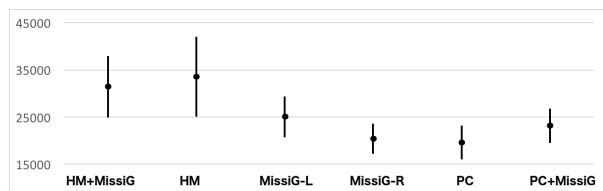
TABLE 3: Pairwise results for JM tasks for MissiG-L, MissiG-R, Heatmap (HM) and PC.

Wilcoxon	Response Time	
	Z	p
MissiG-L vs HM	-0.276	0.783
MissiG-R vs HM	-3.133	0.002
PC vs HM	-3.328	0.001
MissiG-R vs MissiG-L	-3.782	<0.001
PC vs MissiG-L	-3.782	<0.001
PC vs MissiG-R	-0.828	0.408

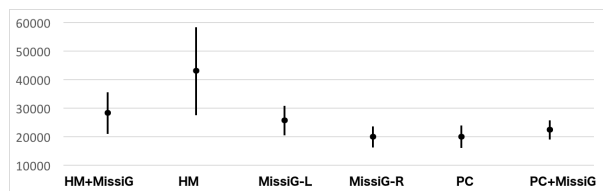
Wilcoxon	Response Time Correct	
	Z	p
MissiG-L vs HM	-0.893	0.372
MissiG-R vs HM	-3.198	0.001
PC vs HM	-3.100	0.002
MissiG-R vs MissiG-L	-3.912	<0.001
PC vs MissiG-L	-3.263	0.001
PC vs MissiG-R	-0.990	0.322



(a) Number of accurate answers.



(b) Average response time in ms.



(c) Response time in ms for correct answers only.

Fig. 18: Confidence intervals for CM tasks

preference of the participants for each visualization method for CM tasks. It can be concluded that PC+MissiG was the most preferred visualization method for CM tasks, followed by PC, MissiG-L, and MissiG-R, again confirming that the participants liked using MissiG as part of the analysis (**H10**). Heatmap was the least liked method. Confidence intervals for the performance of the visualization methods are presented in Fig. 18. When comparing the two MissiG layouts with Heatmap and PC, tests confirmed statistically significant difference in response time (ANOVA: $F(1.452, 30.499) = 11.906, p = 0.001$) and response time for accurate answers (Friedman: $\chi^2(14) = 20.229, p < 0.001$), but not for accuracy. Post-hoc analysis with ANOVA and Wilcoxon signed rank test (table 4) showed that MissiG-R and PC performed significantly better than Heatmap for both response time and response time for accurate answers, and PC performed significantly better than MissiG-L for both response times as well, which is also clear from the confidence intervals. Additionally, MissiG-L

TABLE 4: Pairwise results for CM tasks for MissiG-L, MissiG-R, Heatmap (HM) and PC.

ANOVA	Response Time	
	Z	p
MissiG-L vs HM	-2.731	0.006
MissiG-R vs HM	-2.953	0.003
PC vs HM	-2.897	0.004
MissiG-R vs MissiG-L	-2.959	0.003
PC vs MissiG-L	-3.243	0.001
PC vs MissiG-R	-0.121	0.904

performed significantly better than Heatmap, and MissiG-R performed significantly better than MissiG-L for response time for accurate answers. This supports **H7** in terms of response time, while accuracy results are inconclusive. Looking at the results in Fig. 18, HM+MissiG generally performs better than Heatmap, which supports **H9**, although Wilcoxon signed rank test show that the difference is only significant for response time for correct answers ($Z = -2.731, p = 0.006$). A dependent t-Test showed that the difference in response time for PC compared to PC+MissiG, with slightly better performance for PC, was significant ($t(21) = -2.2374, p = 0.027$), which does not support **H8**, while there was no difference for accuracy or response time for accurate answers.

6 DISCUSSION AND CONCLUSIONS

This paper presented MissiG, a novel glyph based method for visualization of missing values and missingness patterns in data. The method was designed based on established glyph design guidelines to focus on the representation of three patterns of importance for understanding missingness in data, namely the amount missing in variables, the joint missingness in pairs of variables, and the conditional missingness between missing and recorded values in variables. MissiG can be used both as a standalone multivariate visualization method, with two different layouts presented in the paper, and as an enhancement to existing visualization methods, here demonstrated through enhancement of Heatmap and PC. These four methods were evaluated against Heatmap and PC through two usability studies.

The results from the studies indicate that MissiG performs better than PC and equally well as Heatmap for amount missing and joint missingness tasks. The results for conditional missingness tasks indicate that MissiG is equally good as PC and may be better than Heatmap, although the results are not conclusive across the two studies. This may be due to the concept of conditional missingness being more complex than amount missing and joint missingness, which was indicated by questions made by participants in the first study, and thus can have resulted in a higher degree of uncertainty in the results. It is also worth noting that the lack of difference for accuracy in some cases may be the result of ceiling effects when a majority of participants answered correctly to most questions. The difference in results across the two evaluations

can furthermore be a result of the different study designs, with the first study using interactive visualization and the second study including additional questions. This difference was intentional to cover a broader range of analysis situations and tasks, but may have impacted results. In terms of using MissiG as an enhancement, PC with MissiG generally performed better than PC for amount missing in both studies and joint missingness in the first study, while it did not perform better for conditional missingness tasks. Conversely, Heatmap enhanced with MissiG did not perform as well as Heatmap for amount missing and joint missingness tasks, but did in part perform better for conditional missingness, although results were not conclusive across the two studies. Several aspects may have influenced these results. Firstly, the analysis of two coordinated views compared to a single view may have a higher cognitive burden, resulting in longer response times for the enhanced visualization methods. Furthermore, the majority of participants ranked their experience of visualization as high rather than low, which suggests that they may have had previous experience of Heatmap and PC, whereas MissiG is a novel visualization method. This can have impacted MissiG negatively, particularly in terms of response time. It is worth noting that enhancement with MissiG seems to be mostly beneficial for tasks where the basic visualization method is limited (i.e. amount missing and joint missingness for PC, and conditional missingness for Heatmap), whereas the benefit of enhancement is more questionable for tasks where the basic method already performs well. The studies confirmed that MissiG is the preferred choice by users as a method for visualizing missingness in data, which further confirms its usability and the potential of its utility as part of more complex visual analysis workflows. To summarize the main results:

- MissiG performs better than PC, and equally well as Heatmap, for amount missing and joint missingness.
- MissiG in part performs better than Heatmap, and equally well as PC, for conditional missingness.
- PC+MissiG generally performs better than PC for amount missing and joint missingness.
- Heatmap+MissiG performs equally well and in part better than Heatmap for conditional missingness.
- Visualization with MissiG was generally the preferred choice by users across all three missingness patterns.

These results are encouraging and strongly suggest that MissiG has potential to greatly improve analysis and understanding of missingness patterns in data, and through this support decision making as to how to deal with missing values, as well as to reveal important insights related to missing values. Future work includes the application and qualitative testing of MissiG with application domain experts in substantially more complex analysis settings. Of interest would be to apply the technique to IoT and sensor data. In domains such as Air traffic control and unmanned aerial management, as well as in medical digital mobility assessment, where missing values may have a pertinent meaning and where many parameters interplay with each other. Using this kind of tool to analyse relationships between missing data could potentially give a better understanding of how to improve such systems. Utilizing the flexibility of the glyph design to provide additional layout options and enhancement of more visualization methods is a topic for immediate future work.

REFERENCES

- [1] J. Carpenter and M. Kenward, *Multiple Imputation and its Application*. Wiley, 2013.
- [2] S. Johansson Fernstad, "To identify what isn't there: A definition of missingness patterns and evaluation of missing value visualization," *Information Visualization*, vol. 18, no. 2, pp. 230–250, 2019.
- [3] S. Johansson Fernstad and R. C. Glen, "Visual analysis of missing data – to see what isn't there," in *Poster Proceedings of IEEE Vis*. IEEE, November 2014.
- [4] D. H. Chung, P. A. Legg, M. L. Parry, R. Bown, I. W. Griffiths, R. S. Laramee, and M. Chen, "Glyph sorting: Interactive visualization for multi-dimensional data," *Information Visualization*, vol. 14, no. 1, pp. 76–90, 2015.
- [5] E. Maguire, P. Rocca-Serra, S.-A. Sansone, J. Davies, and M. Chen, "Taxonomy-based glyph design—with a case study on visualizing workflows of biological experiments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2603–2612, 2012.
- [6] R. Borgo, J. Kehrler, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen, "Glyph-based visualization: Foundations, design guidelines, techniques and applications." in *Eurographics (STARs)*, 2013, pp. 39–63.
- [7] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.
- [8] H. Wang and S. Wang, "Visualization of the critical patterns of missing values in classification data," in *International Conference on Advances in Visual Information Systems*. Springer, 2007, pp. 267–274.
- [9] H. Wang and S. Wang, "Data mining with incomplete data," in *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*. IGI Global, 2008, pp. 3027–3032.
- [10] S. Fielding, P. M. Fayers, and C. R. Ramsay, "Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches," *Health and Quality of Life Outcomes*, vol. 7, no. 1, p. 57, 2009.
- [11] S. Djurcilov and A. Pang, "Visualizing sparse gridded data sets," *IEEE Computer Graphics and Applications*, vol. 20, no. 5, pp. 52–57, 2000.
- [12] J. Beddow, "Shape coding of multidimensional data on a microcomputer display," in *Proceedings of the 1st Conference on Visualization'90*. IEEE Computer Society Press, 1990, pp. 238–246.
- [13] R. Twiddy, J. Cavallo, and S. M. Shiri, "Restorer: A visualization technique for handling missing data," in *Proceedings of the conference on Visualization'94*. IEEE Computer Society Press, 1994, pp. 212–216.
- [14] A. Unwin, G. Hawkins, H. Hofmann, and B. Siegl, "Interactive graphics for data sets with missing values — manet," *Journal of Computational and Graphical Statistics*, vol. 5, no. 2, pp. 113–122, 1996.
- [15] M. Theus, H. Hofmann, B. Siegl, and A. Unwin, "Manet extensions to interactive statistical graphics for missing values," in *In New Techniques and Technologies for Statistics II*. IOS Press, 1997, pp. 247–259.
- [16] D. F. Swayne and A. Buja, "Missing data in interactive high-dimensional data visualization," *Computational Statistics*, vol. 13, no. 1, pp. 15–26, 1998.
- [17] D. F. Swayne, D. T. Lang, A. Buja, and D. Cook, "Ggobi: Evolving from xgobi into an extensible framework for interactive data visualization," *Comput. Stat. Data Anal.*, vol. 43, no. 4, pp. 423–444, Aug. 2003.
- [18] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1–3, pp. 1–6, 1998.
- [19] A. Unwin, *Graphical data analysis with R*. CRC Press, 2015, vol. 27.
- [20] N. Tierney, D. Cook, M. McBain, C. Fay, M. O'Hara-Wild, J. Hester, and L. Smith, "Naniar: Data structures, summaries, and visualizations for missing data," *R Package*, 2019.
- [21] A. Lex, N. Gehlenborg, H. Strobel, R. Vuillemot, and H. Pfister, "Upset: visualization of intersecting sets," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1983–1992, 2014.
- [22] A. Pihlhofer, "Extracat: Categorical data analysis and visualization," *R package version*, pp. 1–7, 2014.
- [23] 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.
- [24] P. Brix, "miP: Multiple imputation plots," <https://CRAN.R-project.org/package=miP>, 2012.
- [25] X. Cheng, D. Cook, H. Hofmann *et al.*, "Visually exploring missing values in multivariable data using a graphical user interface," *Journal of Statistical Software*, vol. 68, no. 6, pp. 1–23, 2015.
- [26] J. Honaker, G. King, and M. Blackwell, "Amelia ii: A program for missing data," *Journal of Statistical Software, Articles*, vol. 45, no. 7, pp. 1–47, 2011.
- [27] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, "Profiler: Integrated statistical analysis and visualization for data quality assessment," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 2012, pp. 547–554.
- [28] T. Gschwandtner and O. Erhart, "Know your enemy: Identifying quality problems of time series data," in *2018 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2018, pp. 205–214.
- [29] J. A. Triana, D. Zeckzer, H. Hagen, and J. T. Hernandez, "Vafusq: A methodology to build visual analysis applications with data quality features," *Information Visualization*, vol. 18, no. 4, pp. 384–404, 2019.
- [30] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann, "A systematic view on data descriptors for the visual analysis of tabular data," *Information Visualization*, vol. 16, no. 3, pp. 232–256, 2017.
- [31] A. Cedilnik and P. Rheingans, "Procedural annotation of uncertain information," in *Visualization 2000. Proceedings*. IEEE, 2000, pp. 77–84.
- [32] Z. Xie, S. Huang, M. O. Ward, and E. A. Rundensteiner, "Exploratory visualization of multivariate data with variable quality," in *In Proceedings of the IEEE Symposium on Visual Analytics Science and Technology*, 2006, pp. 183–190.
- [33] C. Arbesser, F. Spechtenhauser, T. Mühlbacher, and H. Piringer, "Visplause: Visual data quality assessment of many time series using plausibility checks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 641–650, 2017.
- [34] B. L. W. Wong and M. Varga, "Black holes, keyholes and brown worms: Challenges in sense making," in *Proceedings of the Human Factors and Ergonomics Society 56th Annual Meeting*, 2012, pp. 287–291.
- [35] A. Kirk, "Visualizing zero: How to show something with nothing," <http://blogs.hbr.org/2014/05/visualizing-zero-how-to-show-something-with-nothing/>, May 2014.
- [36] C. Eaton, C. Plaisant, and T. Drizd, "Visualizing missing data: graph interpretation user study," in *Human-Computer Interaction-INTERACT 2005*. Springer, 2005, pp. 861–872.
- [37] R. Andreasson and M. Riveiro, "Effects of visualizing missing data: an empirical evaluation," in *2014 18th International Conference on Information Visualisation*. IEEE, 2014, pp. 132–138.
- [38] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay, "Uncertainty displays using quantile dotplots or cdfs improve transit decision-making," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.
- [39] H. Song and D. A. Szafrir, "Where's my data? evaluating visualizations with missing data," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 914–924, 2018.
- [40] P. T. Quinlan and G. W. Humphreys, "Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on feature and conjunction searches," *Perception & psychophysics*, vol. 41, no. 5, pp. 455–472, 1987.
- [41] J. Johansson, M. Cooper, and M. Jern, "3-dimensional display for clustered multi-relational parallel coordinates," in *Proceedings IEEE International Conference on Information Visualization, IV05*, 2005, pp. 188–193.
- [42] B. S. Dayal, "Application of feedforward neural networks and partial least squares for modelling kappa number in a continuous kamyr digester," *Pulp and Paper Canada*, vol. 95, pp. 26–32, 1994.
- [43] K. Bache and M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [44] A. M. Graziano and M. L. Raulin, *Research methods: A process of inquiry*. HarperCollins College Publishers, 1993.
- [45] H. T. Kahraman, S. Sagirolu, and I. Colak, "Developing intuitive knowledge classifier and modeling of users' domain dependent data in web," *Knowledge Based Systems*, vol. 37, pp. 283–295, 2013.
- [46] I.-C. Yeh, "Modeling of strength of high performance concrete using artificial neural networks," *Cement and Concrete Research*, vol. 28, no. 12, pp. 1797–1808, 1998.
- [47] M. Little, P. McSharry, S. Roberts, D. Costello, and I. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMedical Engineering OnLine*, vol. 6, no. 23, 2007.
- [48] G. Cumming and S. Finch, "Inference by eye: confidence intervals and how to read pictures of data." *American psychologist*, vol. 60, no. 2, p. 170, 2005.
- [49] A. L. Anwyl-Irvine, J. Massonnié, A. Flitton, N. Kirkham, and J. K. Evershed, "Gorilla in our midst: An online behavioral experiment builder," *Behavior Research Methods*, pp. 1–20, apr 2019.
- [50] A. Asuncion and D. J. Newman, "UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences," 2007, <http://www.ics.uci.edu/~mllearn/MLRepository.html>.