Eurographics Conference on Visualization (EuroVis) 2022
R. Borgo, G. E. Marai, and T. Schreck
(Guest Editors)

Volume 41 (2022), Number 3

Where did my Lines go? Visualizing Missing Data in Parallel Coordinates

A. Bäuerle[1†], C. van Onzenoodt[1†], S. der Kinderen[1], J. Johansson Westberg[2], D. Jönsson[1,2], and T. Ropinski[1],

[1]Ulm University, Germany
[2]Linköping University, Sweden

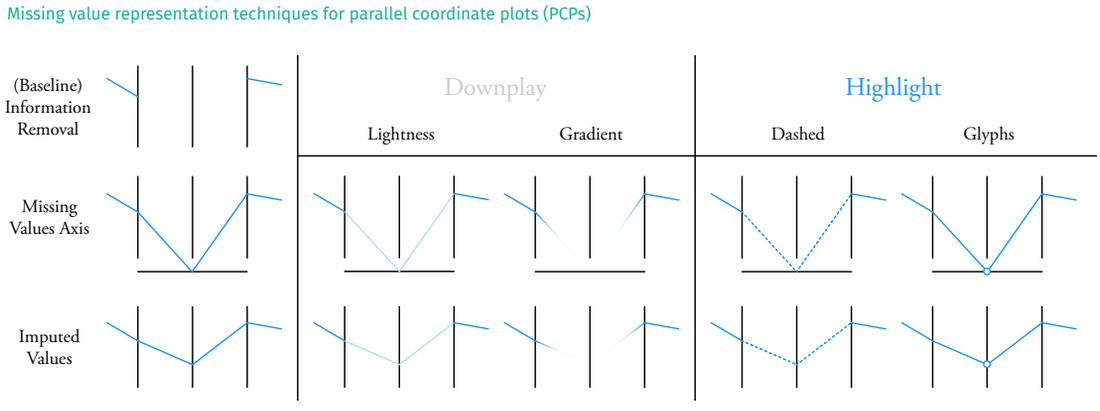

Figure 1: *Illustrations of the evaluated concepts (left) and variations (middle, right) for representing missing values in parallel coordinates.* Information removal *leaves out the missing values,* missing values axis *introduces a separate axis onto which missing values are projected, while* imputed values *calculates replacements for the missing values and shows these estimates on their corresponding axis. For both of the latter techniques, we propose downplay and highlight variations.*

Abstract

We evaluate visualization concepts to represent missing values in parallel coordinates. We focus on the trade-off between the ability to perceive missing values and the concept's impact on common tasks. For this purpose, we identified three missing value representation concepts: removing line segments where values are missing, adding a separate, horizontal axis onto which missing values are projected, and using imputed values as a replacement for missing values. For the missing values axis *and* imputed values *concepts, we additionally add* downplay *and* highlight *variations. We performed a crowd-sourced, quantitative user study with 732 participants comparing the concepts and their variations using five real-world datasets. Based on our findings, we provide suggestions regarding which visual encoding to employ depending on the task at focus.*

CCS Concepts

• ***Computing methodologies*** → ***Perception;*** • ***Human-centered computing*** → *User studies;* ***Visualization techniques;***

1. Introduction

With the widespread use of sensor technology and ambitious data collection strategies, high-dimensional datasets constantly gain importance, while their size and complexity increase at the same time. Such datasets often require visual analysis to understand and detect trends, outliers, or other associations between dimensions. One of the most prominent techniques for visualizing high-dimensional datasets is parallel coordinates [Ins85]. As illustrated in Figure 1, parallel coordinates represent the dimension axes by parallel, often vertical, lines. The datapoints connect these axes with lines passing through their corresponding axes' coordinates.

It is not uncommon for attributes of high-dimensional datapoints to be missing in real-world datasets, as exemplified in recent works [SS18, JBF*19, MGU*21]. The reasons for missing values can be myriad, e.g., broken sensors, incomplete forms, or physical measurements that could not be collected during a clinical study. Even in the most widely used high-dimensional example datasets,

† Both authors contributed equally to this research.

such as the Palmer Archipelago penguins [HHG20] and the Auto MPG dataset [Qui93], missing values can be found. At the same time, confident decisions can only be made if one is aware of missing values [AR14]. Visualizing missing values can be important even when they are evenly distributed, e.g., when investigating prediction errors of a deep learning model, visualizing missing values of the input data can be important when reasoning about those instances. Thus, the omnipresence of missing data and the importance of visualizing missing values call for measures to represent missing values in techniques such as parallel coordinates.

Despite the prevalence of missing values, they are often disregarded by removing the datapoint [JBF*19]. However, removing an entire line obfuscates that there are missing values in the data. As this clearly is undesirable, there have been several works that represent missing values in parallel coordinates, e.g., by introducing a missing values axis [MGU*21] or using information removal [JBF*19]. However, little is known about the value of these techniques, as their performance with respect to missing value discovery, or interference with common parallel coordinates analysis tasks, has not yet been evaluated.

In this paper, we investigate missing value representation techniques designed for parallel coordinates. Based on related work on missing data [SS18], we initially separate these techniques into three distinct concepts, as illustrated on the left of Figure 1. The first technique, *information removal*, removes the part of the line connected to the axis that contains the missing value. The second technique, *missing values axis*, introduces a separate axis for missing values onto which they are projected and connected to. This *missing values axis* is oriented horizontally and positioned below the vertical axes. The third technique, *imputed values*, uses reconstructed values based on existing data in place of the missing values. Imputation obtains an estimated point at which the line can be connected to. Because these techniques may interfere with common parallel coordinates analysis tasks, we introduce novel variations designed to *downplay* and *highlight* missing values [SS18]. The *downplay* variations are designed to de-emphasize missing values by decreasing the opacity of the lines, while the *highlight* variations emphasize the missing values using dotted lines or glyphs.

Summary of the user study/ evaluation

To assess the effectiveness of missing value representation techniques, we evaluate the eleven varieties described above within a crowd-sourced, quantitative user study with 732 participants. We investigate the ability to estimate the number of missing values, as well as the performance on different parallel coordinates tasks, namely, value retrieval, cluster interpretation, and outlier detection. Our user study is based on real-world datasets, where data points are removed randomly to an increasing degree, while the entire dataset is used to compute a baseline for each task. Thus, within this paper we make the following contributions:

- We conducted an **evaluation of visualization concepts** for representing missing values in parallel coordinates.
- We **developed visual variations** to control how missing values are perceived.
- We provide **recommendations for visual encodings** depending on whether the visualization focuses on communicating missing values or general patterns.

2. Related Work

One of the earliest works on parallel coordinates was by Inselberg [Ins85, ID90] who proposed the technique for visual analysis of high-dimensional geometry. Later, Wegman [Weg90] introduced the technique for visualization of statistical data, which was the basis for most research on parallel coordinates. While there are myriad extensions to parallel coordinates that address everything from visual clutter [HW13, JF16] to axis ordering [KHG03, Weg90], we focus on missing values. Therefore, our evaluation does not consider any of these advancements. Research most closely related to this work deals with evaluating parallel coordinates encodings and the visualization of missing data.

2.1. Evaluating Parallel Coordinates

Can novice users participate in parallel coordinates studies? There was the longstanding assumption that parallel coordinates would be an expert-only technique. However, Siirtola et al. found in their study that, on the contrary, users learn the visual encoding used in parallel coordinates quickly [SLHR09]. Furthermore, to evaluate parallel coordinates with novice users, Kwon and Lee investigated different online education methods [KL16]. In their study, they found that video tutorials and interactive guides can get novices up to speed in short time-frames. Following these findings, showing that performing parallel coordinates studies with novice users are feasible, we also use interactive movies and interactive guides to train the novice participants in our study.
Which analysis tasks have been evaluated for parallel coordinates? Quadri and Rosen [QR21] summarized evaluations of different visualization approaches and found that parallel coordinates have been investigated in terms of value retrieval, filtering, sorting, characterizing distributions, and clustering tasks. Kanjanabose et al. found that parallel coordinates can outperform scatter plots for clustering, outlier detection, and change detection [KARC15]. Raidu et al. added visual enhancements to parallel coordinates to support trend estimation [REB*15]. As tasks, they included outlier detection, pattern discernability, and discovering obstructed data. For cluster detection, Holten and van Wijk evaluated nine different parallel coordinates visualization techniques [HVW10], while Blumenschein et al. [BZP*20] evaluated the axis order itself. Based on these findings, we distilled three general parallel coordinates tasks encapsulating analysis on a per-line basis, between axes tracing and line aggregation (clustering), c.f. Section 5.2.

2.2. Visualizing Missing Data

While this work focus on parallel coordinates, there have been many works dealing with missing data for other common visualization techniques [TAKP11, ANI*17]. Eaton et al. [EPD05] compared missing value encodings for connected scatterplots and, similar to our work, investigated participants' ability of trend perception and value comparison with missing data. Song and Szafir [SS18] investigated visualization of incomplete datasets for bar and line charts using *highlight*, *downplay*, *annotation* and *information removal* visual encoding variations of imputed values. We adopted the *highlight* and *downplay* variations for parallel coordinates in this work to visually differentiate missing values. Andreasson and

Riveiro [AR14] evaluated the effects of missing value encodings for line charts on decision-making processes and show that the visual encoding can have effects on confidence and risk-friendliness when making decisions based on such visualizations. While not the focus of this work, we also ask the participants how confident they are and provide these results to the community.

In terms of missing value techniques for parallel coordinates, Sjöbergh et al. [ST17] introduced a horizontal axis below parallel coordinates and connected the missing value lines to this axis. This technique was also adopted by Muller et al. [MGU*21] and is referred to as *missing values axis* in this work, see Figure 1. Jönsson et al. [JBF*19] applied what we will refer to as *information removal* for dealing with missing data in brain cohort study data. Johansson Fernstad and Johansson Westberg [JFJW21] introduced the *Missingness Glyph*, providing information about several missing data patterns from earlier work [Fer19]. This *Missing Glyph* is intended to be used either as a standalone visualization or as an enhancement for multivariate visualizations such as parallel coordinates. The missingness glyph encodes missing values using several other visual representations, rendering it out of scope for this work. Instead, we focus on evaluating missing value encodings solely by augmenting the lines of parallel coordinates, which require no special introduction or training to be understood.

While there exist techniques for visualizing missing values in parallel coordinates, no evaluation to compare their performance has been performed so far. Thus, guidance on how to visually encode missing data for other visualization techniques is available, this is still largely unknown for parallel coordinates.

3. Parallel Coordinates Missing Value Encoding

The core idea of parallel coordinates is to facilitate lines that represent rows of a tabular dataset going through parallel axes for each column in the dataset. While parallel coordinates have been a common visual encoding for a long time, there exists no clear guidance on how to represent missing values while not interfering with tasks not related to missing values. In this section, we explain the visual encodings we investigated to communicate missing values.

The encodings in this work visualize missing values by changing the lines in parallel coordinates. Another option to signify missing data would be to modify the axes of a parallel coordinates visualization. However, we not only want to show how much data is missing but also enable visualization users to perform more detailed analyses, including which lines contain missing values and how missing values are distributed. Thus, we only investigate techniques that encode missing data directly on the corresponding lines in the parallel coordinates visualization. We formally divide these encodings into *concepts*, which are different methods to project missing values, and *variations*, which add further diversification and can be applied to different concepts in the same way.

3.1. Missing Value Encoding Concepts

We investigate three concepts to visualize missing data in parallel coordinates (cf. Section 2.2). These concepts are *information removal* [JBF*19], adding a *missing values axis* [ST17], and *imputed values* [TAF12, CCH15]. The concepts will be detailed in the following and are illustrated in Figure 2, left.

Information Removal. An intuitive notion of the data not being present can be obtained by simply *not* drawing lines to axes where data is missing, see Figure 1, top left.

Missing Values Axis. Missing values are inherently *out of distribution* as they cannot be mapped to a position on the axis containing these missing values. Thus, the *missing values axis* encoding (Figure 1, left center) displays missing values on an axis separately from coordinates with complete data. A horizontal axis, which specifically encodes missing data, is placed below the conventional parallel coordinates axes. Whenever an axis contains missing values, those coordinates containing missing values are routed to this horizontal axis, right below the axis for which values are missing.

Imputed Values. To preserve the core property of parallel coordinates, connecting all lines to all axes, this technique replaces missing values with *imputed values*. Since parallel coordinates typically include a high number of dimensions and data samples, we cannot use the same imputation method as Song and Szafir [SS18]. For our parallel coordinates imputation method, we aim at preserving general data patterns. Thus, the imputation method used must reflect the trend of the original data across all dimensions, i.e., imputation needs to take into account all axes for reconstructing the missing value independent of axis ordering. Although there are many advanced imputation methods [SS07], they are computationally expensive, which is undesirable for large numbers of dimensions and values to be imputed, as for parallel coordinates [BE10, BFG*15]. We use nearest neighbor imputation [BS16] to estimate missing values. Interpolation is performed on the basis of lines that are next to the line that contains the missing value in each dimension except for where the line is missing values. Internal tests were performed to verify that this imputation method works well in our setting.

3.2. Missing Value Encoding Variations

To further diversify the tested visualization designs, we propose variations on the basis of the work by Song and Szafir [SS18], by adapting the *downplay* and *highlight* attention manipulators to the *missing values axis* and *imputed values* concepts. Examples for these conditions can be seen in Figure 2, right.

Parallel coordinates have specific traits that restrict the types of suitable visual encodings. Datasets typically contain many entries, e.g., table rows, resulting in many lines drawn close to and on top of each other. This line proximity trait limits the possibility of adding additional visual elements around lines, as it would further clutter the visualization. Preliminary experiments with adding elements as annotations [SS18] in addition to the variants we propose, e.g., error bars, were performed. However, we found that it is both hard to trace which line the added elements are associated with and that they additionally clutter the view to such a high degree that it significantly obstructs other analysis tasks. Due to the line proximity constraint causing overdraw, we chose to minimize the number of added elements and exclude annotations that add additional uncertainty. The following *downplay* and *highlight* variations take into account that overdraw can be a problem for parallel coordinates by either modifying the lines themselves in terms of drawing style or opacity, or adding minimally occlusive glyphs.

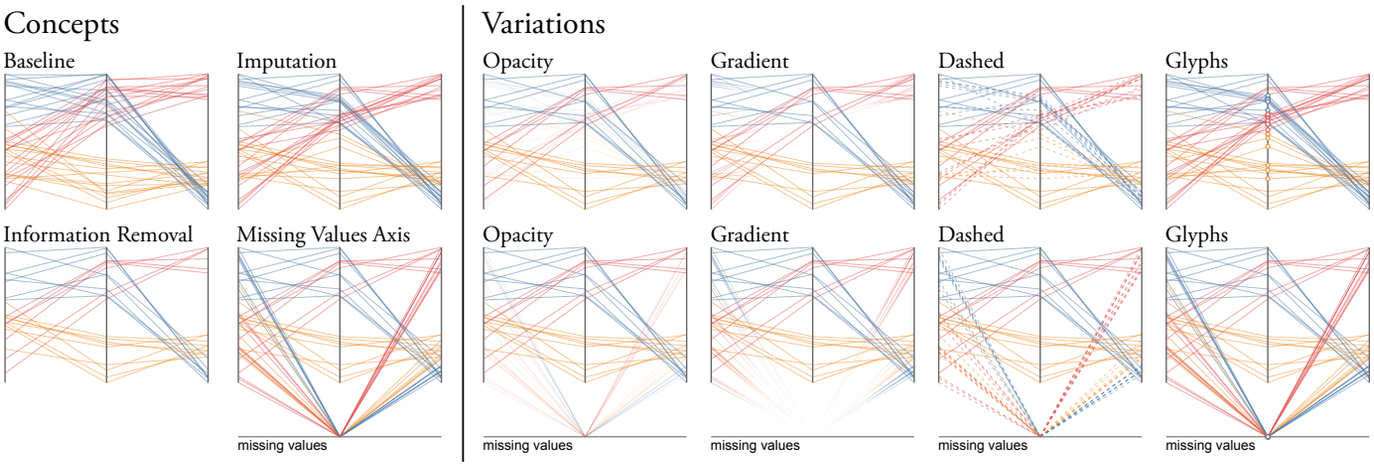

Figure 2: *Example illustrations of all evaluated visual encodings. On the left, one can see the different concepts we used to encode data. Baseline is the full dataset, information removal means that lines where data is missing are not drawn, imputation replaces missing values with imputed data, and missing values axis adds an axis onto which missing data is projected. The imputation and missing values axis conditions additionally have variations that downplay (opacity, gradient) or highlight (dashed, glyphs) missing values.*

3.2.1. Downplay

Downplay is a technique to make individual values less salient compared to the rest of the data [SS18]. This can be used to indicate the presence of a missing value, while at the same time not drawing visual attention. The focus of this work is on modifying the opacity to create a downplay effect as it is associated with the line itself and is commonly not used to visually encode other dimensions. We consider two distinct downplay techniques.

Opacity. We decrease the opacity of a line connecting a missing value (Figure 2, third column) to lower its visual prominence. This technique preserves the entire line as an element of the plot. The opacity needs to be set such that it is visually distinct from the other lines, while still ensuring that they are clearly visible. It was experimentally decided that an opacity of 40% compared the other lines fulfilled these criteria for the datasets used in the study.

Gradient. Lines are gradually faded away as they approach missing values (Figure 2, fourth column). The amount of gradient effect was experimentally decided to start with the same opacity as other lines, linearly decreasing to 75% opacity until reaching the midpoint and finally linearly decreasing to zero. The exact axis position is thus not visible anymore; creating the impression that the coordinate passes behind the axis whenever a value is missing.

3.2.2. Highlight

Highlight makes missing values more salient compared to non-missing data. Thus, missing values are emphasized using a perceptually dominant visual encoding, intentionally drawing visual attention to missing values. To this end, we propose two visually distinct methods; a line modification technique and a glyph technique that adds minimally occluding elements to the visualization.

Dashed. Dashed lines for all missing values (Figure 2, fifth column) draws attention to this irregular pattern. This way, connections to missing data are still clearly visible and even emphasized compared to filled lines. The dashed line effect is created using a regular pattern with a repetition of five opaque and five transparent pixels, resulting in a 50/50 split. While for this variation, less pixels are colored overall, we still argue that dashed lines are a form of *highlight*, as they can serve as a preattentive visual attribute which visually separates from the other lines in this context [Tre85].

Glyphs. Are filled circles with a border of the same color as the line at the location of the missing value. Glyphs indicate exactly where a value is missing, while keeping the visual appearance of all lines the same. Because only missing values are represented by such glyphs, they draw visual attention, highlighting the presence and position of missing values. However, glyphs are more likely to overlap than lines and might introduce overdraw. Overdraw is also one reason why we placed the glyphs where values are missing instead of where lines end or start, as this would double the number of glyphs drawn. Additionally, placing glyphs at line starts and ends does not show where a value is missing and leads to ambiguities when multiple subsequent axes contain missing values.

Apart from these, there is a potentially endless set of variations. Color is a natural candidate for missing value encoding [SS18], but is often used to encode one data dimension of parallel coordinates, making it less suitable for encoding missing data. We experimented with different variations, such as glyph types and placement, curved lines, different gradients and opacity levels, color, lightness, etc. but found the above-mentioned to be the most promising. Although we would have liked to include more variations, this would have exceeded the limits of a crowdsourced user study. Therefore, we selected these variations for this evaluation.

4. Hypotheses

We believe that each concept and variation has characteristics that make it perform better or worse for a given task. Thus, we formulated and investigated the following hypotheses to determine if these characteristics impact the visual analysis performance.

H1 Using the *information removal* concept, estimating the number of missing values will be significantly harder.
As missing values are represented by lines that end at one axis and continue a few axes later, they are hard to see. Thus, we assume that performance will be significantly worse for this condition.
H2 With the *imputed values* concept, highlight and downplay enable spotting missing values.
As the *imputed values* condition in its default representation does not display missing data and complete data differently, spotting missing data will be impossible with this technique.
H3 We suspect *imputed values* to perform worse than *information removal* for the common parallel coordinates tasks, but better than *missing values axis*. The *imputed values* concept will help preserve general patterns within the data while not completely obstructing missing values.
While *information removal* makes it hard to spot missing values as they are simply not drawn, the *missing values axis* concept might interfere with general patterns in the data. We expect the *imputed values* condition to lie somewhere in between, not performing as well as the *information removal* concept in preserving patterns as well as the *missing values axis* concept with respect to missing data. Although its upsides are not as strong, we expect that *imputed values* will serve as a solid trade-off between showing missing data and preserving the original data pattern.
H4 *Imputed values* leads to lower accuracy in missing value estimation compared to the *missing values axis* concept, but outperforms *information removal*.
We suspect that *imputed values* hinders the ability to spot missing values, as imputed lines are mixed with all other lines, resulting in more visual clutter compared to *missing values axis*. However, we still suspect that *imputed values* leads to higher accuracy when estimating missing values compared to the *information removal* concept, as *information removal* simply removes missing lines.
H5 The *missing values axis* concept will perform best in terms of missing value estimation, but will harm common parallel coordinates tasks.
Explicitly rerouting missing values to a separate axis visually highlights these values. Thus, we expect them to be easier to spot with this technique. On the other hand, this interferes with the general pattern common to parallel coordinates. In turn, we expect that this improvement in missing value estimation will come at the cost of lower accuracy in the other tasks.
H6 The *highlight* variations will perform best in terms of missing value estimation but will harm common parallel coordinates tasks.
Highlighting missing values helps in spotting those values. However, it also draws attention needed to solve other tasks. Therefore, we assume that while *highlight* will improve missing value estimation, other task will be made harder through this variation.
H7 The *downplay* variations will help to preserve patterns within the data, but make missing value estimation harder.
Contrary to the *highlight* variations, *downplay* de-emphasizes missing values. In turn, we expect missing value estimation to be harmed by downplay while other tasks can be performed with higher accuracy.

5. User Study Design

We designed a quantitative user study to evaluate the 11 visualization techniques shown in Figure 1. Performance in four tasks (missing value estimation, trend estimation, outlier detection, and line tracing) are evaluated with different amounts of missing values, and five datasets.

The study is focused on two main aspects. First, we evaluate how well missing values can be perceived with each visual encoding. Second, we test how much each encoding interferes with common parallel coordinate tasks, i.e., whether the visual encoding aids or harms these tasks. For both aspects, participants were asked how confident they were in their answers on a five-point Likert scale to determine if the technique gave a false sense of confidence.

5.1. Datasets

Five real-world datasets with diverse properties have been selected to reflect data distributions in realistic scenarios. Each dataset has a different number of datapoints, dimensions, and the datasets contain a variety of numerical and categorical data. An overview of the datasets can be seen in Table 1.

Table 1: *Datasets used in our evaluation. ◬○ represents the number of categorical axes while ∼ represents the number of continuous axes. Brackets indicate dataset sizes after cleanup.*

Name	Rows	Dimensions	Source
Airline Safety	56	8 (1 ◬○, 7 ∼)	[Fiv21]
Bad Drivers	51	8 (1 ◬○, 7 ∼)	[Fiv21]
Iris	150	5 (1 ◬○, 4 ∼)	[Fis36, And36]
MPG	398 (392)	9 (3 ◬○, 6 ∼)	[Qui93]
Penguins	344 (333)	7 (3 ◬○, 4 ∼)	[HHG20]

To evaluate the influence of missing data, it is necessary to precisely control which and how much data is missing. However, real-world data can already contain missing values, which is for instance the case for the Penguins and MPG datasets. Thus, the rows containing missing values have been removed to obtain a valid ground truth for our user study. Thus, we removed 9 rows from the Penguins dataset and 6 rows from the MPG dataset in total.

Categorical axes were placed on the first or last axes because users focus more on the center of parallel coordinates [NVE*17]. For all other axes, we ordered them so that trends and outliers could be seen between a pair of axis for every dataset. However, as we ask for trends and outliers in the context of an axis pair, enabling us to perform these tasks with crowd-workers, and value retrieval is not axis-order dependant, we argue that the axis order is not a critical factor for our findings. Furthermore, we have omitted all semantics, by using capital letters for axis annotations, so that we could avoid bias resulting from prior knowledge about the data.

5.2. Tasks

The following tasks were designed to test user performance of both missing value estimation and common parallel coordinates tasks:

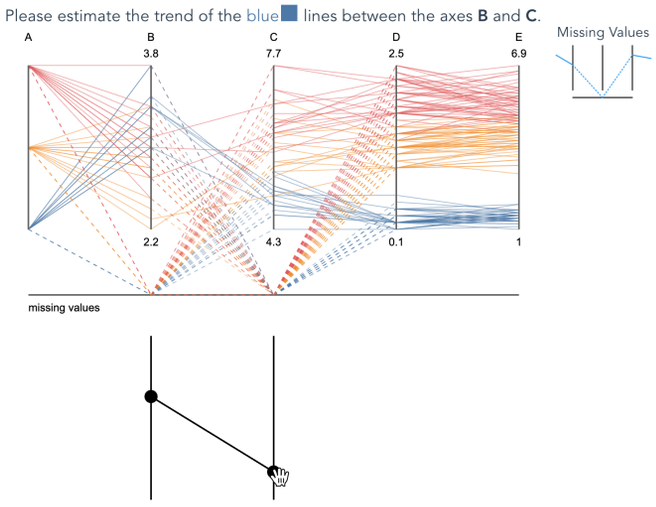

Figure 3: *Trend estimation task example with the Iris dataset, missing values axis concept, and dashed-line highlighting. Participants manipulated the knobs on the axes below the main plot to indicate their estimate of where most lines of a given color start and end.*

Trend Estimation. Relates to perceiving patterns that involve groups or clusters between two axes, e.g.,, lines starting and ending close to each other or lines with the same color. Participants were asked to, e.g., *"estimate the trend of the red lines between the axes X and Y"* by defining a trend vector, using an interactive element displayed below the axes of interest. The color we selected in the question was varied for different stimuli. An example stimulus for the trend estimation task can be seen in Figure 3. The ground truth was computed as the average value of the subset of lines we asked for. This way, we obtained a ground truth and a prediction for each of the two axes in question. To calculate the correctness of an answer, we used the euclidean distance between the ground truth values and the answer given by a participant. Data was removed randomly from either one or both axes between which the trend was to be estimated. We selected the axes between which the trend was to be selected once per dataset to make sure that trends were visible in all stimuli.

Outlier Detection. Relates to finding values that are significantly different from others. Participants were asked to *"select the line between axes X and Y that is most different compared to all other lines (disregarding the missing values)"* by clicking on the line in the visualization, see Figure 4. To measure participants' accuracy, outliers between the two axes were calculated using the Local Outlier Factor [BKNS00] with recommended settings [sld21]. A participant could be either correct if they selected an outlier or wrong if they selected a line that was not deemed an outlier, resulting in a binary answer. Data was randomly removed from either one axis which the outlier line was connected to or both, while ensuring that the outlier remained. We selected the axes between which the outlier was to be spotted once per dataset to make sure that one or more outliers were present in all stimuli.

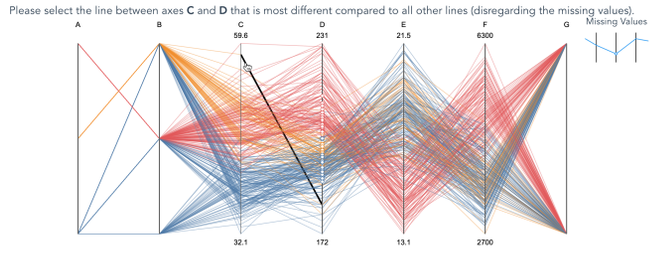

Figure 4: *Outlier detection task example with the Penguins dataset, imputation concept, and glyph highlight. Participants were asked to select the line between two axes, in this case C and D, that was most different. The selected outlier is black and thicker.*

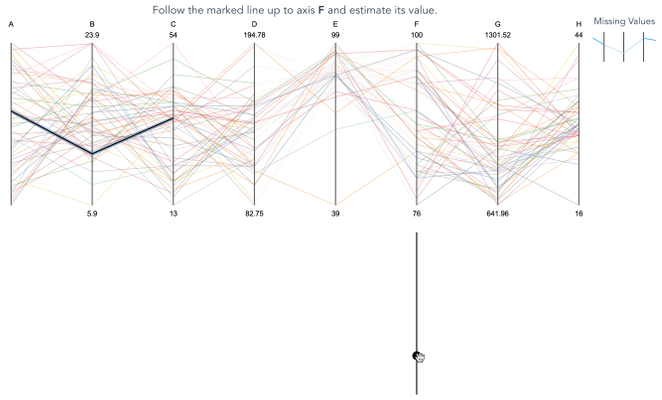

Figure 5: *Value retrieval task example with the Bad Drivers dataset, imputation concept, and opacity-based downplay. Participants were asked to follow the highlighted line (black with glow in original color) and estimate its axis-value three axes away using an interactive slider placed below the axis in question.*

Value Retrieval. Relates to tracing a line to retrieve its value for another axis. We asked participants to trace a marked line and retrieve its value for the third following axis to the right of where the marked line ended, see Figure 5. For example, given a marked line ending at axis *B* participants needed to retrieve the value from axis *E*. Participants had to adjust the knob of a vertical slider below the corresponding axis to answer the question: *"follow the marked line up to axis X and estimate its value"*. We calculated the accuracy for this task as the absolute distance from the target value. We randomly removed data from one or both axes between the highlight and axis for which the value was to be estimated. We selected the start and retrieval axes once per dataset and made sure that neither axis (start, retrieval, and between) was categorical.

Missing Value Estimation. After answering to the stimulus, participants were asked to *"estimate the number of missing values for axis X"*. The participant could adjust a slider going from one to the total number of items, or tick a checkbox stating *"No missing values"* to make this an explicit choice. The error was determined by the difference between the participant's answer and the correct number of missing values. As for the removal settings described for the other tasks, values could be missing from one or two axes, but we always asked for an estimate for only one axis.

Cluster detection was intentionally left out due to its similarity to the trend estimation task, the difficulties in identifying ground

truth clusters in real-world datasets, and the relatively poor performance in cluster identification of real-world datasets compared to synthetic ones [BZP*20].

5.3. Stimuli

Each graph was generated using a custom JavaScript library (available on GitHub) and stored in vector format (SVG) to guarantee pixel-perfect quality. Missing data was simulated by randomly removing values from dimensions related to the tasks in steps of 5 values, 10%, 50%, and 70%. A fixed number was used to represent a small number of missing values independent of dataset size while the percentage values were included to simulate different missing/non-missing ratios. The maximum was set to 70% because imputation does not make sense for higher percentages of missing values [MGU*21]. As a baseline for the techniques, we asked the same questions for a standard parallel coordinates visualization, where we showed the original, unmodified dataset. The Tableau 10 color palette [Sto16] was applied to one categorical dimension of each dataset to reflect more realistic scenarios compared to applying a constant color, and to allow for questions relating to a group of lines (i.e., same category). The opacity was set to 0.5 for all datasets with back-to-front blending as otherwise overdraw would have made some tasks impossible to solve. A random drawing order was used to draw each row of a dataset to ensure that the order did not have a major impact on the visual appearance. Not every visualization platform supports interaction, e.g., papers and newspapers. To provide a uniform guideline and isolate the effect of the visual encoding from interaction techniques, filtering, hovering, and other graph interactions not related to the tasks were disabled.

5.4. Independent Variables

Of the aforementioned concepts, the *missing values axis* and the *imputed values* concept were drawn in their default setting and with variations (4), which amounts to 5 encodings per variation concept. For the *information removal* concept, variations were not applicable as they could not be drawn without a line to be drawn in the first place. Thus, we investigated 11 different visual encodings in total. Each of these encodings was drawn with different amounts of data removal (5). As explained in Section 5.3, we removed either 5 datapoints, 10%, 50%, or 70% of the data in an axis. Data removal was performed on different axis configurations (3). For each task, we selected either one or both of two candidate axes. Thus, this amounted to 132 stimuli. Additionally, a baseline stimulus was generated where no data was removed. Altogether, we thus produced 133 different stimuli per task and dataset.

These visualizations were used for the tasks (3) we evaluated as described in Section 5.2. For the trend task, the Bad Drivers dataset could not be used as it does not contain categorical data we could color by (the one categorical axis contains 51 individual values). In turn, the trend task was evaluated with the remaining datasets (4) introduced in Section 5.1. Thus, for this task, we had a total of 532 stimuli. For the other tasks (2), we used the full set of datasets (5). In turn, each of these tasks included 665 stimuli. Altogether, we produced 1,862 stimuli. We ensured that each stimulus was viewed exactly three times, requiring 5,586 responses.

5.5. Procedure

The study consisted of four phases. In the first phase, we asked each participant to confirm that they had no color deficiency, since we used colors in our visualizations and asked for colors in the trend task. Note that additional color checks were performed during the study, as detailed below. In the second phase, we introduced the concept of parallel coordinates. Third, we outlined the tasks that participants had to solve. Then, we explained the concept of missing data. Following, we introduced our visual encoding techniques. To present the study workflow, we then showed an animated GIF explaining how to solve the task. Then a series of example questions explaining the task was presented. Finally, participants had to answer three training tasks for which we used hand-crafted data to make them easily solvable, where the performance was not measured to get comfortable with the interface. Lastly, the formal study was performed.

To simplify the introduction and prevent learning effects, each participant was only presented with either the trend, outlier, or value retrieval task. We deceided to limit the number of stimuli per participant to further prevent learning and fatigue. For the trend task, we only show twelve stimuli per participant (three per dataset), for the value task ten (two per dataset), and five for the outlier task (one per dataset). These stimuli have been counterbalanced using Latin square for the outlier task, and using random order for the other tasks, whereby we ensured to not present the same dataset twice in a row.

Each stimulus for each task was followed up by the missing value estimation task. For both, the task response and the missing values estimation task, before showing the visualization, we first displayed only the question for each task. When the user pressed the space bar, we revealed the visualization and started a timer to measure the response time. We additionally asked participants to indicate their confidence for each answer on a five point Likert scale from *Not Confident* (1) and *Completely Confident* (5). Attention checks, i.e., asking the participants to click a box of a specific color from a set of four colors, were presented in-between every fourth stimulus. Failure to click the correct box either meant that the participant is color deficient, or that they did not pay attention. One failed attention check could be well attributed to a simple misclick, so we required two failed attention checks to be excluded. None of the participants failed more than one of these checks, which meant that we did not have to exclude any data from the study. Afterwards, we conducted a short demographic questionnaire including age, location, gender, and experience with parallel coordinates.

5.6. Participants

Our study was conducted via the crowdwork platform Prolific. In total, we had 732 participants (379 female, 349 male, eight other, three did not respond, $M_{age} = 26.15$, $SD = 14.92$, 108 out of these participant reported having seen plots like the ones in the study before. As discussed in Section 5.4 and Section 5.5, our tasks required different numbers of participants. Therefore, 133 participants were assigned to the trend task, 399 to the outlier detection task, and 200 to the value retrieval task.

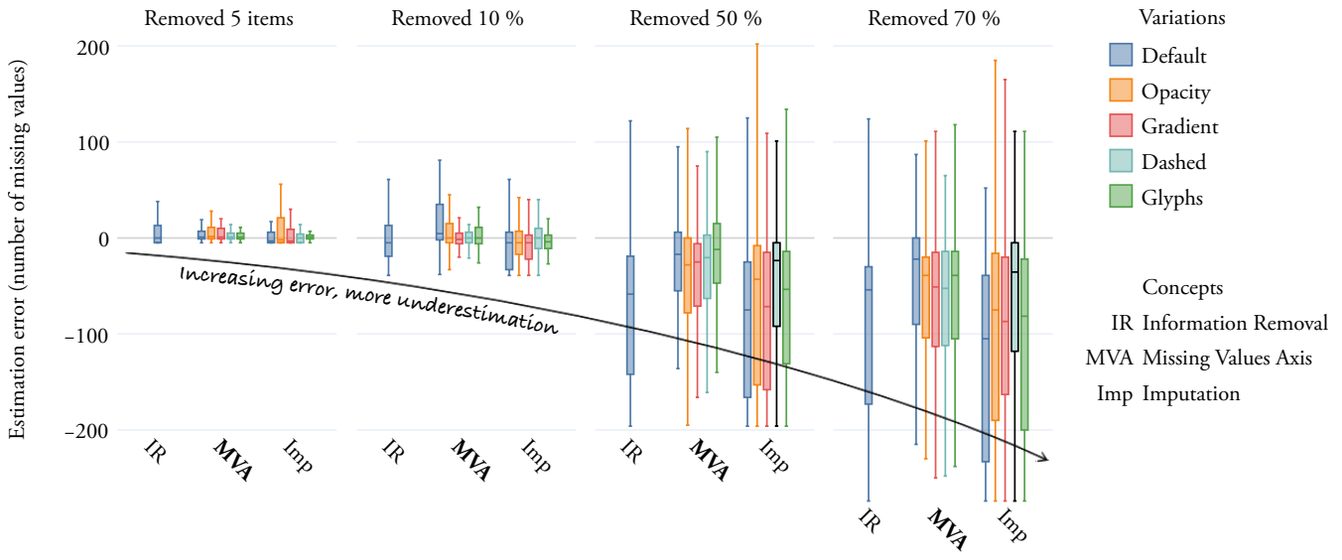

Figure 6: *Missing value estimation error increases with number of removed values. This decrease in performance can be observed for all concepts and variations, indicating that the degraded accuracy indeed stems from the missing values themselves, not from our encodings. The default variation for* imputed values *acts as a random guess baseline as it is impossible to spot missing values. The concept with best average along the horizontal axis is highlighted in bold. The* missing values axis *concept performs better than the other concepts. However, the* imputed values *concept combined with the dashed lines variation (highlighted with black outlines) almost performs on par.*

Participants in the trend task had to provide answers to 12 stimuli. To make sure that they did not see the same stimulus twice, we showed all four datasets used in this task and varied the color for which we requested a trend estimate. For the outlier detection task, participants had to answer 5 stimuli. This number was chosen to prevent a learning effect, as we did not want participants to see the same dataset with the same outliers twice. The value retrieval task consisted of 10 stimuli. We showed each dataset twice, while a learning was mitigated through constant variation of the line that had to be traced for value retrieval. Since these three tasks required different amounts of time to complete, we made sure that on average, each worker was paid 7.5£ per hour.

6. Results

Below, we present the results of our experiments. We conducted Kruskal-Wallis-Tests and used the Mann-Whitney rank test for pairwise post hoc anlysis, with Bonferroni correction. For binary results (outlier detection), we conducted the chi-square test with pairwise post hoc comparisons, also using Bonferroni correction. A summary of our findings can be found in Table 2.

6.1. Missing Value Estimation

We first analyzed the ability to estimate missing values, as illustrated in Figure 6. Here, *Imputed values* with default variation acts as a baseline since imputed values can not be differentiated from non-missing values.

Concepts. Comparing our concepts, we found a significant effect on accuracy (*Missing values axis* ($Mdn = -3.0$, $IQR = 36.25$), *Imputed values* ($Mdn = -10.0$, $IQR = 60.0$), *Information removal* ($Mdn = -5.0$, $IQR = 39.0$), $H(3) = 114.47$, $p < .001$, *Missing values axis* $\leftrightarrow$ *Imputed values* ($p < .001$), *Missing values axis* $\leftrightarrow$ *Information removal* ($p < .001$), *Imputed values* $\leftrightarrow$ *Information removal* ($p = .66$)), showing that the *missing values axis* concept provides the highest accuracy when trying to estimate missing values partially supporting H4.1 and H5.1. We also found a significantly worse accuracy for the *information removal* concept, sup-

Table 2: *Findings for each of our hypotheses. Marks can be traced to detailed statistics in Section 6.*

Summary of findings
Information removal makes missing value estimation significantly less accurate H1, thus **supporting** H1.
Imputed values can support missing value detection **if** variations are used H2, thus **supporting** H2.
For common parallel coordinates tasks, *Imputed values* performs better than *information removal* H3.1 but not significantly worse than *missing values axis* H3.2, thus **rejecting** H3.
Missing values axis makes missing value estimation significantly more accurate H4.1 whereas *information removal* makes it significantly less accurate H4.2, thus **supporting** H4.
Missing values axis makes missing value estimation significantly more accurate H5.1, but harms common parallel coordinates tasks H5.2, thus **supporting** H5.
The highlight variations did perform best in terms of missing value estimation H6.1 but did not harm common parallel coordinates tasks H6.2, thus **rejecting** H6.
The downplay variations made missing value estimation harder H7.2, but it did not help to preserve patterns within the data H7.1, thus **rejecting** H7.

porting the second part of H4.2 and H1. We found a similar effect for the confidence in missing value estimation between our concepts (*Missing values axis* ($M = 2.95$, $SD = 1.11$), *Imputed values* ($M = 2.84$, $SD = 1.13$), *Information removal* ($M = 2.82$, $SD = 1.14$), $H(3) = 16.27$, $p < .001$, *Missing values axis* $\leftrightarrow$ *Imputed values* $p < .001$, *Missing values axis* $\leftrightarrow$ *Information removal* $p < .05$, *Imputed values* $\leftrightarrow$ *Information removal* $p = 1.0$)). However, when analyzing response times, we found that the *missing values axis* concept requires most time for this task (*Missing values axis* ($Mdn = 17.3s$, $IQR = 17.8s$), *Imputed values* ($Mdn = 14.4s$, $IQR = 16.0s$), *Information removal* ($Mdn = 15.0s$, $IQR = 18.3s$), $H(3) = 52.76$, $p < .001$, *Missing values axis* $\leftrightarrow$ *Imputed values* $p < .001$, *Missing values axis* $\leftrightarrow$ *Information removal* $p < .01$, *Imputed values* $\leftrightarrow$ *Information removal* $p = .85$)). This indicates that *missing values axis*, while taking the longest time to estimate the number of missing values provides the most accurate results, together with high confidence.

Variations. Comparing individual variations, we found significant effects (Dashed ($Mdn = -5.0$, $IQR = 36.0$), Glyph ($Mdn = -5.0$, $IQR = 40.0$), Gradient ($Mdn = -6.5$, $IQR = 50.0$), Opacity ($Mdn = -5.0$, $IQR = 44.75$), Default ($Mdn = -5.0$, $IQR = 42.0$), $H(5) = 17.36$, $p < .01$ (Gradient $\leftrightarrow$ Dashed $p < .001$, Dashed $\leftrightarrow$ Default $p < .05$, others not significant.)), suggesting the dashed variation for missing value estimation. Although not significant, we see generally lower variances for missing value estimation for the *highlight* variations, supporting H6.1 and H7.2.

***Imputed values* - Variations.** When using *imputed values*, we found a significant effect on the missing value estimation error between our variations (Dashed ($Mdn = -5.0$, $IQR = 36.0$), Glyphs ($Mdn = -11.0$, $IQR = 61.25$), Gradient ($Mdn = -13.0$, $IQR = 71.5$), Opacity ($Mdn = -7.0$, $IQR = 63.25$), Default ($Mdn = -25.0$, $IQR = 70.0$), $H(5) = 44.12$, $p < .001$, (Gradient $\leftrightarrow$ Dashed ($p < .0001$), Dashed $\leftrightarrow$ Glyphs ($p < .001$), Dashed $\leftrightarrow$ Default ($p < .0001$), Glyphs $\leftrightarrow$ Default ($p < .05$), Opacity $\leftrightarrow$ Default ($p < .001$), others not significant). This suggests that the dashed variation seems to be a good choice in combination with the *imputed values* concept. With the exception of Gradient, all variations also perform significantly better than Default, which does not reveal missing values, supporting H2.

***Missing values axis* - Variations.** The same pattern could be found when investigating variations for the *missing values axis* concept (Dashed ($Mdn = -3.0$, $IQR = 36.5$), Glyphs ($Mdn = -2.0$, $IQR = 32.0$), Gradient ($Mdn = -5.0$, $IQR = 38.25$), Opacity ($Mdn = -4.0$, $IQR = 38.0$), Default ($Mdn = -1.0$, $IQR = 34.5$), $H(5) = 18.94$, $p < .001$, Gradient $\leftrightarrow$ Default ($p < .0001$), Dashed $\leftrightarrow$ Default ($p < .05$), Opacity $\leftrightarrow$ Default ($p < .05$)) however, in this concept the default variation significantly outperformed all others except the glyphs. This is to be expected, because the only difference between the Default and the Glyphs variation is a single dot on the *missing values axis* axis, as seen in Figure 2.

6.2. Common Parallel Coordinates Tasks

Next, we present the results of our study regarding common parallel coordinates tasks.

Value Retrieval. For the value retrieval task, we did not find a significant effect between our conditions.

Trend Estimation. For the trend estimation task, we found a significant difference between the *missing values axis* concept and the other conditions (*Missing values axis* ($Mdn = 0.19$, $IQR = 0.22$), *Imputed values* ($Mdn = 0.16$, $IQR = 0.2$), *Information removal* ($Mdn = 0.15$, $IQR = 0.17$), $H(3) = 7.67$, $p < .05$, (*Missing values axis* $\leftrightarrow$ *Imputed values* $p < .14$, *Missing values axis* $\leftrightarrow$ *Information removal* $p < .05$, *Imputed values* $\leftrightarrow$ *Information removal* $p = .64$)). This indicates that for the trend estimation task, *imputed values* and *information removal* perform approximately on par, while the *missing values axis* concept performs significantly worse.

Outlier Detection. For the outlier detection task, we also found significantly worse performance using the *missing values axis* condition (*missing values axis* (77.22%), *imputed values*(85.22%), *information removal* (92.31%), ($H(3) = 34.62$, $p < .001$), (*missing values axis* $\leftrightarrow$ *imputed values* $p < .001$, *missing values axis* $\leftrightarrow$ *information removal* $p < .001$, *imputed values* $\leftrightarrow$ *information removal* $p < .05$)). Altogether, these results indicate that the *missing values axis* concept harms both, the trend and outlier detection performance, further supporting H5.2. Only for the value retrieval task could we not find a significant difference. These findings support H3.1 in that *imputed values* performs better than *missing values axis*. However, H3.2 could not be supported since *information removal* did not outperform *imputed values*.

Variations. We could not find a significant effect for either task (Trend estimation ($p = .95$), Value retrieval $p = .26$, Outlier detection ($p = .38$)) between our variations *highlight*, *downplay*, and default. We also could not find a effect between the individual variations (Trend estimation ($p = 0.94$), Value retrieval ($p = 0.12$), Outlier detection ($p = 0.39$)). In turn, these findings reject H6.2 and H7.1. For the trend estimation task with the *missing values axis* concept, the *downplay* variation showed descriptively higher values, but no significant effect could be measured. The results of our experiments regarding our tasks can be seen in Figure 7.

7. Discussion

In the following, we draw conclusions and provide guidelines with respect to using the evaluated concepts and variations.

Comparing the different concepts. For the missing value estimation task, *missing values axis* performed best followed by *imputed values* and *information removal*. However, *missing values axis* also resulted in the longest task completion time. Nevertheless, if the main focus is on communicating missing values, we recommend the *missing values axis* concept. For value retrieval, none of the concepts outperformed the other. Regarding trend estimation, we found that *missing values axis* performs significantly worse, while the other two concepts are approximately on par. In turn, we suggest not to use *missing values axis* for trend-related tasks. Finally, for the outlier detection task, we found that *information removal* performs best, followed by *imputed values* and *missing values axis*. Consequently, we propose to use the *information removal* concept when the objective is to detect outliers. All tasks combined, while *imputed values* is not the single best concept in any of them, it often performs on par. Therefore, we suggest to use *imputed values* for visualization where a balanced task performance is desired.

Comparing the different variations. We did not find significant effects for any of the tasks. While overall, the impact of different variations on the different tasks was low, this is different for the

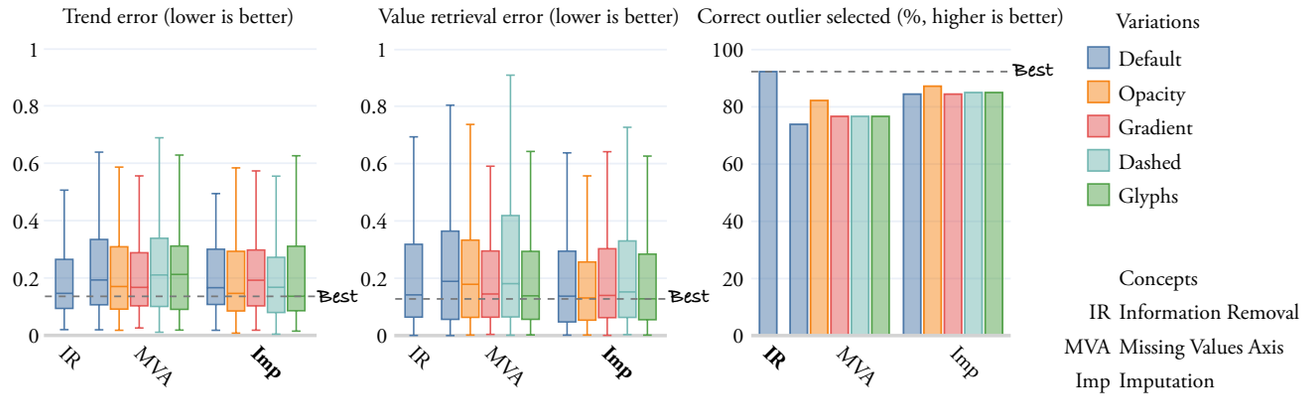

Figure 7: *Depictions of task accuracy errors to understand the methods impact on parallel coordinates tasks (outliers removed for presentation purposes). Information removal represents a base level as it has no elements interfering with, or aiding, the tasks. Most methods are on par with information removal. However, the* missing values axis *concept using dashed lines stands out as interfering most with the tasks.*

missing value estimation task. Here, we found that gradient and default are significantly worse than dashed. However, all other combinations were not significantly different which is why we can only recommend not to use gradient or default. Especially with the *imputed values* concept, variations have to be included, as the default variation does not reveal missing values. When using *imputed values*, we found that using the dashed variation leads to the highest accuracy when trying to identify missing values.

Comparing all concepts and variations. For missing value estimation, *missing values axis* with the default encoding performed better than each of the *imputed values* conditions and *information removal*. Additionally, it significantly outperformed the gradient variation of *missing values axis*. This was the best combination of concept and variation for the missing value estimation task. For the trend estimation and value retrieval tasks, we could not find significant effects between the individual concept/variation combinations. For the outlier task, the *information removal* condition outperformed all concept/variation combinations except for *imputed values* with opacity. This is in line with our recommendation to use *information removal* when outlier detection is at focus. However, since *information removal* was the worst in terms of missing value estimation, we do not recommend this encoding in every case.

8. Limitations and Future Work

Since our evaluation provides high-level guidance visualizing missing data in parallel coordinates it naturally come with limitations.

Interaction Techniques and Encoding Variants. We did not include interaction techniques or test different configurations of our visual encodings, e.g., imputation algorithm, opacity, gradient falloff, dash configuration, and glyph designs. Such parameter explorations could help to further improve the perception of parallel coordinates with missing values.

Distribution of Missing Data. While this work provides general guidance on how to best preserve patterns and visualize data incompleteness, one might also need to take the data collection process into account. For example, if the missing value distribution is biased, imputation might introduce non-existent patterns.

Uncertainty Visualization. We did not experiment with uncertainty visualization techniques. Future work could build on our insights and add uncertainty visualization techniques such as roughly sketched lines, manipulating gradients, or other techniques.

Scalability. Our study includes datasets with up to 398 datapoints. We found that already at this level, overdraw is a big issue for parallel coordinates and in turn did not include even larger datasets. Integrating missing value encodings with techniques to reduce overdraw, such as bundling techniques, would be another interesting research question to address.

More encodings can be evaluated using the same procedures and data. As no additional baseline is required, new results can be directly compared to our findings, thus further sampling the space of possible missing value encodings.

9. Conclusion

In this paper, we evaluated different visualization concepts to show missing values for parallel coordinates. Additionally, these conditions were augmented with variations to further diversify the set of visualization options. Our quantitative user study, which included 732 participants, indicates that the best concept for the respective tasks were *missing values axis* for missing value estimation, *information removal* for outlier detection, *information removal* and *imputed values* for trend estimation, and none of the concepts could significantly outperform any other in value retrieval. On this basis, our discussion provides first indications on which encoding to select depending on what task is at focus, namely, the aforementioned if one of the tasks is especially important, and *imputed values* if a fair tradeoff of the tasks is the target of the visual encoding. If *imputed values* is used, we suggest to combine it with the dashed variation so that missing values estimation performance remains high. Although this evaluation serves as a first source of guidance on how missing data can be represented in parallel coordinates, we hope that further research can expand these guidelines to more concept, variations, and parallel coordinates techniques.

Acknowledgements This project was supported by the Wallenberg Autonomous Systems and Software Program (WASP) as well as Ynnerman KAW Scholar.

References

[And36] ANDERSON E.: The species problem in iris. *Annals of the Missouri Botanical Garden 23*, 3 (1936), 457–509. 5

[ANI*17] ALEMZADEH S., NIEMANN U., ITTERMANN T., VÖLZKE H., SCHNEIDER D., SPILIOPOULOU M., PREIM B.: Visual analytics of missing data in epidemiological cohort studies. In *VCBM* (2017), pp. 43–51. 2

[AR14] ANDREASSON R., RIVEIRO M.: Effects of visualizing missing data: an empirical evaluation. In *2014 18th International Conference on Information Visualisation* (2014), IEEE, pp. 132–138. 2, 3

[BE10] BARALDI A. N., ENDERS C. K.: An introduction to modern missing data analyses. *Journal of school psychology 48*, 1 (2010), 5–37. 3

[BFG*15] BÖGL M., FILZMOSER P., GSCHWANDTNER T., MIKSCH S., AIGNER W., RIND A., LAMMARSCH T.: Visually and statistically guided imputation of missing values in univariate seasonal time series. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2015), IEEE, pp. 189–190. 3

[BKNS00] BREUNIG M. M., KRIEGEL H.-P., NG R. T., SANDER J.: Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (2000), pp. 93–104. 6

[BS16] BERETTA L., SANTANIELLO A.: Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making 16*, 3 (2016), 197–208. 3

[BZP*20] BLUMENSCHEIN M., ZHANG X., POMERENKE D., KEIM D. A., FUCHS J.: Evaluating reordering strategies for cluster identification in parallel coordinates. In *Computer Graphics Forum* (2020), vol. 39, Wiley Online Library, pp. 537–549. 2, 7

[CCH15] CHENG X., COOK D., HOFMANN H.: Visually exploring missing values in multivariable data using a graphical user interface. *Journal of statistical software 68*, 1 (2015), 1–23. 3

[EPD05] EATON C., PLAISANT C., DRIZD T.: Visualizing missing data: Graph interpretation user study. In *IFIP Conference on Human-Computer Interaction* (2005), Springer, pp. 861–872. 2

[Fer19] FERNSTAD S. J.: To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization. *Information Visualization 18*, 2 (2019), 230–250. 3

[Fis36] FISHER R. A.: The use of multiple measurements in taxonomic problems. *Annals of eugenics 7*, 2 (1936), 179–188. 5

[Fiv21] FIVETHIRTYEIGHT: Fivethirtyeight datasets, 2021. URL: https://github.com/fivethirtyeight/data/tree/master/bad-drivers. 5

[HHG20] HORST A. M., HILL A. P., GORMAN K. B.: *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*, 2020. R package version 0.1.0. URL: https://allisonhorst.github.io/palmerpenguins/, doi:10.5281/zenodo.3960218. 2, 5

[HVW10] HOLTEN D., VAN WIJK J. J.: Evaluation of cluster identification performance for different pcp variants. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 793–802. 2

[HW13] HEINRICH J., WEISKOPF D.: State of the Art of Parallel Coordinates. In *Eurographics 2013 - State of the Art Reports* (2013). 2

[ID90] INSELBERG A., DIMSDALE B.: Parallel coordinates: a tool for visualizing multi-dimensional geometry. In *Proceedings of the First IEEE Conference on Visualization: Visualization90* (1990), IEEE, pp. 361–378. 2

[Ins85] INSELBERG A.: The plane with parallel coordinates. *The visual computer 1*, 2 (1985), 69–91. 1, 2

[JBF*19] JÖNSSON D., BERGSTRÖM A., FORSELL C., SIMON R., ENGSTRÖM M., YNNERMAN A., HOTZ I.: A Visual Environment for Hypothesis Formation and Reasoning in Studies with fMRI and Multivariate Clinical Data. In *Eurographics Workshop on Visual Computing for Biology and Medicine* (2019), The Eurographics Association. 1, 2, 3

[JF16] JOHANSSON J., FORSELL C.: Evaluation of parallel coordinates: Overview, categorization and guidelines for future research. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2016), 579–588. doi:10.1109/TVCG.2015.2466992. 2

[JFJW21] JOHANSSON FERNSTAD S., JOHANSSON WESTBERG J.: To explore what isn't there—glyph-based visualization for analysis of missing values. *IEEE Transactions on Visualization and Computer Graphics* (2021). 3

[KARC15] KANJANABOSE R., ABDUL-RAHMAN A., CHEN M.: A multi-task comparative study on scatter plots and parallel coordinates plots. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 261–270. 2

[KHG03] KOSARA R., HAUSER H., GRESH D. L.: An Interaction View on Information Visualization. In *Eurographics 2003 - STARs* (2003), Eurographics Association. doi:10.2312/egst.20031092. 2

[KL16] KWON B. C., LEE B.: A comparative evaluation on online learning approaches using parallel coordinate visualization. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), pp. 993–997. 2

[MGU*21] MULLER J., GARRISON L., ULBRICH P., SCHREIBER S., BRUCKNER S., HAUSER H., OELTZE-JAFRA S.: Integrated dual analysis of quantitative and qualitative high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics 27*, 6 (2021), 2953–2966. doi:10.1109/TVCG.2021.3056424. 1, 2, 3, 7

[NVE*17] NETZEL R., VUONG J., ENGELKE U., O'DONOGHUE S., WEISKOPF D., HEINRICH J.: Comparative eye-tracking evaluation of scatterplots and parallel coordinates. *Visual Informatics 1*, 2 (2017), 118–131. 5

[QR21] QUADRI G. J., ROSEN P.: A survey of perception-based visualization studies by task. *IEEE Transactions on Visualization and Computer Graphics* (2021). 2

[Qui93] QUINLAN J. R.: Combining instance-based and model-based learning. In *Proceedings of the tenth international conference on machine learning* (1993), pp. 236–243. 2, 5

[REB*15] RAIDOU R. G., EISEMANN M., BREEUWER M., EISEMANN E., VILANOVA A.: Orientation-enhanced parallel coordinate plots. *IEEE Transactions on Visualization and Computer Graphics 22*, 1 (2015), 589–598. 2

[sld21] SCIKIT-LEARN DEVELOPERSE: Outlier detection with local outlier factor (lof), 2021. URL: https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html. 6

[SLHR09] SIIRTOLA H., LAIVO T., HEIMONEN T., RÄIHÄ K.-J.: Visual perception of parallel coordinate visualizations. In *2009 13th International Conference Information Visualisation* (2009), IEEE, pp. 3–9. 2

[SS07] SONG Q., SHEPPERD M.: Missing data imputation techniques. *International journal of business intelligence and data mining 2*, 3 (2007), 261–291. 3

[SS18] SONG H., SZAFIR D. A.: Where's my data? evaluating visualizations with missing data. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 914–924. 1, 2, 3, 4

[ST17] SJÖBERGH J., TANAKA Y.: Visualizing missing values. In *2017 21st International Conference Information Visualisation (IV)* (2017), IEEE, pp. 242–249. 3

[Sto16] STONE M.: How we designed the new color palettes in tableau 10, 2016. URL: https://www.tableau.com/about/blog/2016/7/colors-upgrade-tableau-10-56782. 7

[TAF12] TEMPL M., ALFONS A., FILZMOSER P.: Exploring incomplete data using visualization techniques. *Advances in Data Analysis and Classification 6*, 1 (2012), 29–47. 3

[TAKP11] TEMPL M., ALFONS A., KOWARIK A., PRANTNER B.: Vim: visualization and imputation of missing values. *R package version 2*, 3 (2011). 2

[Tre85] TREISMAN A.: Preattentive processing in vision. *Computer vision, graphics, and image processing 31*, 2 (1985), 156–177. 4

[Weg90] WEGMAN E. J.: Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association 85*, 411 (1990), 664–675. 2