Conceptual and Methodological Issues in Evaluating Multidimensional Visualizations for Decision Support

Evanthia Dimara, Anastasia Bezerianos, and Pierre Dragicevic

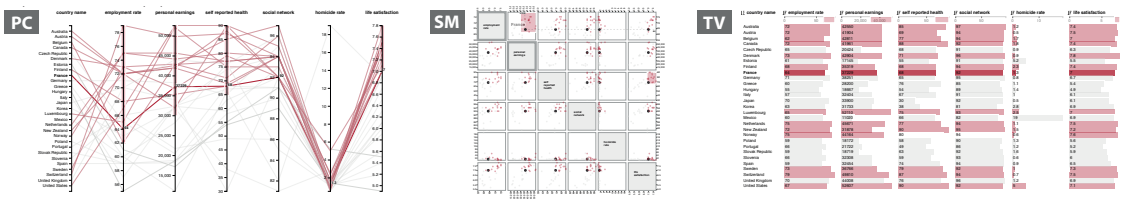

Fig. 1: The visualizations we evaluated: Parallel Coordinates (PC), Scatterplot Matrix (SM) and Tabular Visualization (TV)

Abstract—We explore how to rigorously evaluate multidimensional visualizations for their ability to support decision making. We first define *multi-attribute choice tasks*, a type of decision task commonly performed with such visualizations. We then identify which of the existing multidimensional visualizations are compatible with such tasks, and set out to evaluate three elementary visualizations: parallel coordinates, scatterplot matrices and tabular visualizations. Our method consists in first giving participants low-level analytic tasks, in order to ensure that they properly understood the visualizations and their interactions. Participants are then given multi-attribute choice tasks consisting of choosing holiday packages. We assess decision support through multiple objective and subjective metrics, including a *decision accuracy* metric based on the consistency between the choice made and self-reported preferences for attributes. We found the three visualizations to be comparable on most metrics, with a slight advantage for tabular visualizations. In particular, tabular visualizations allow participants to reach decisions faster. Thus, although decision time is typically not central in assessing decision support, it can be used as a tie-breaker when visualizations achieve similar decision accuracy. Our results also suggest that indirect methods for assessing choice confidence may allow to better distinguish between visualizations than direct ones. We finally discuss the limitations of our methods and directions for future work, such as the need for more sensitive metrics of decision support.

Index Terms—decision making, multidimensional visualization, parallel coordinates, scatterplot matrix, tabular visualization, evaluation

1 Introduction

Suppose Yannis needs to book a hotel for his honeymoon in Paris. A range of websites exist that provide advanced filtering and searching tools to find hotels. But having heard of the power of data visualization, Yannis seeks instead a dataset that he can visualize, so as to fully understand and compare all the options available to him. He downloads an up-to-date dataset with about two hundred hotels with a dozen attributes such as price, room size, bed size, or user ratings. The dataset is not a particularly challenging one to visualize: manageable size, no missing or uncertain data, and all values conveniently encoded in quantitative or ordinal format. Many systems and techniques are available that can visualize such a dataset. Which system should we as visualization experts recommend to Yannis? Which visualization technique is the most likely to help him choose the best hotel?

Surprisingly, there is very little empirical data to help us decide which visualization best supports making such decisions. Yannis could either choose to use a general-purpose multidimensional visualization tool based on scatterplot matrices or parallel coordinates [52, 60], or use a visualization system specifically designed for decision support [15, 37, 61]. However, to our knowledge, no previous work has evaluated such tools for their ability to support decision-making tasks. Most existing studies are either qualitative studies without a comparison baseline [6–8, 37, 61, 64, 89], or use elementary analytic tasks such as value retreival [55, 84] or correlation estimation [57, 89] instead of decision-making tasks. Although supporting elementary analytic tasks is likely an important precondition for supporting informed decisions, data-driven decision making differs from data analysis and data exploration. For example, recent visualization studies have suggested that decision tasks are more error-prone than equivalent analytic tasks [28], and that people can make irrational decisions even when they properly understood the data [27]. Thus, good performance with elementary analytic tasks does not guarantee good performance in decision making.

Since many decision tasks have no clear ground truth, evaluating visualizations for their ability to support decisions is difficult, and there is a lack of methodological guidance in the information visualization literature on how to do so. This article attempts to bridge this gap by exploring conceptual and methodological issues in evaluating visualizations for their ability to support decisions. We first define our target task, called a *multi-attribute choice task*, and conduct a systematic analysis of existing multidimensional visualization techniques and the extent to which they are appropriate for such tasks. Based on our analysis, we chose to focus on evaluating three generic and commonly-used elementary visualization techniques: parallel coordinates, scatterplot matrices, and tabular visualizations. Each technique supported basic interactions. Participants first received extensive training with each technique, and performed elementary analytic tasks identified as possible components of higher-level decision making tasks. They then used each visualization to choose an ideal holiday package for themselves. We measured the quality of participant's decisions using a range of metrics. In particular, we introduced a novel decision accuracy metric based on the consistency between the choice made and self-reported preferences for attributes. We discuss which of these metrics were the most able to capture meaningful differences between the three techniques.

2 Background

We first define the type of task we want to support, i.e., *multi-attribute choice tasks*, and then articulate the link between these tasks and multidimensional data visualization. We next review the different types of

- *Evanthia Dimara is with Inria, France. E-mail: evanthia.dimara@gmail.com.*
- *Anastasia Bezerianos is with Univ. Paris-Sud, CNRS, Inria, Université Paris-Saclay, France. E-mail: anastasia.bezerianos@lri.fr .*
- *Pierre Dragicevic is with Inria, France. E-mail: pierre.dragicevic@inria.fr .*

Manuscript received 31 Mar. 2017; accepted 1 Aug. 2017.
Date of publication 28 Aug. 2017; date of current version 1 Oct. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2017.2745138

multidimensional visualizations and discuss to what extent they may be appropriate for multi-attribute choice tasks. Finally, we review interactive visualization systems specifically meant for decision support and the methods used to evaluate them.

2.1 Multi-Attribute Choice Tasks

In this article we focus on supporting *multi-attribute choice tasks*, a common type of decision making task. We refer to a multi-attribute choice as a task that consists of *finding the best alternative among a fixed set of alternatives, where alternatives are defined across several attributes*. One example is buying a camera at an online store, where each available camera is defined by its price and a number of technical features such as size, weight, or resolution.

There is no unique way of defining a "good" alternative, and the best definition depends on the context. "Goodness" can be defined in objective terms (e.g., Pareto dominance [27]) or in subjective terms (e.g., personal satisfaction with the choice). Possible metrics of goodness will be further discussed in Sections 4.9 and 4.10. For now, we note that multi-attribute choice tasks generally involve users' personal preferences and rarely have an obvious "right" answer.

Our notion of multi-attribute choice task is similar to the *preferential choice* previously introduced in HCI by Bautista and Carenini [11], and defined as "the process of selecting the best option out of a set of alternatives". Our term is more closely related to the terminology of Multi-Criteria Decision Making (MCDM), a discipline that studies procedures to aid decision making in areas like business intelligence and finance [76,91]. MCDM problems however, refer to a broader class of decision-making tasks. Some MCDM problems involve an infinite number of alternatives [76], and even when the alternatives are finite (referred to as Multi-Attribute Decision Making or MADM [76]), they are not necessarily known in advance [91]. Some MADM problems also involve ordering or classifying alternatives rather than identifying the best [76]. In this article we focus solely on the task of identifying the best among a finite number of alternatives known ahead of time.

A key difference between our work and MCDM as a discipline is our focus on supporting spontaneous decision-making aided by visualizations, *without any imposed procedure or strategy*. While MCDM methods are extremely useful for critical team decisions such as choosing a long-range business investment scheme [91], we are interested in how common visualizations can benefit a broad range of users without prior training in decision analysis. Thus we treat visualizations not as tools to guide users in their decisions, but rather as tools to help them better understand the information on which they base their decisions.

Since in a multi-attribute choice task all alternatives are *i)* known in advance, and *ii)* defined across a set of attributes, all information can be provided as a data table [60] where rows are alternatives and columns are attributes [27]. Rows are also commonly called "data cases", while columns are often called "dimensions". In order to help users understand this type of dataset, information visualization has contributed a range of multidimensional visualization tools which we discuss next. Though many of these tools are used to analyze big datasets, most of them are also adapted to the small datasets typical of common multi-attribute choice tasks (e.g., booking a hotel).

2.2 Multidimensional Visualization Approaches

Many approaches exist to visualize multidimensional datasets. Here we provide a systematic analysis of existing approaches and discuss their relation to multi-attribute choice tasks. We group them into three major families: techniques based on dimensionality reduction, non-geometric approaches, and what we call "lossless" geometric visualizations.

2.2.1 Dimensionality Reduction

Some multidimensional visualizations rely on *dimensionality reduction* to collapse multiple dimensions into a smaller number of dimensions, typically two [45,62]. Two common approaches are principal component analysis [62] and multidimensional scaling [45]. Although dimensionality reduction can reveal hidden structures in complex datasets and can show similar and dissimilar data cases, the resulting dimensions are often hard to interpret [70]. Furthermore, raw values are lost during the reduction process, whereas multi-attribute choice generally requires users to be able to read attribute values directly. A related family of techniques is *dimension filtering*, which automatically removes dimensions that are either redundant or unimportant according to some criteria [88]. However, in a context of multi-attribute choice, the importance of dimensions (attributes) can rarely be deduced from the data itself as it requires personal judgment and varies across decision makers [78]. Thus, in the absence of prior information, it seems safer to use visualizations that initially give all dimensions equal importance.

2.2.2 Non-Geometric Visualization Techniques

Keim and Kriegel [52] (also [60]) classified multidimensional visualizations into six categories, the first being *geometric projection*. Geometric projection is a broad class of techniques that encompasses both dimensionality reduction (Sect. 2.2.1) and simpler forms of projections discussed in Sect. 2.2.3. We discuss non-geometric approaches here.

Typical non-geometric approaches are *icon-based* techniques, where data cases are visualized side-by-side as icons or glyphs [34]. Examples include Chernoff faces [19] and star glyphs [54]. Although icons presumably tap into our ability to visually process shapes, they can make comparisons across dimensions challenging [60], as some dimensions may be perceived as more salient than others [20,54,60].

In *pixel-oriented* techniques, each data case is encoded as a single colored pixel [52]. These techniques are very space-efficient and mostly useful when the number of data cases is very high. However, for common multi-attribute choice tasks, the number of data cases is rarely that high. Furthermore, color is not the most effective visual variable [22] and can impede decision making [13].

Two other categories are *hierarchical* and *graph-based* techniques [52]. These techniques assume the existence of structural relationships between attributes that may not be available in multi-attribute choice situations. Finally, *hybrid* techniques combine multiple visualizations either in-place or side-by-side [60]. Although combining different approaches can be powerful, the strengths and weaknesses of elementary visualization techniques need to be better understood before we know how to combine them effectively.

2.2.3 Lossless Geometric Projection

Keim and Kriegel's taxonomy [52] can be refined by splitting *geometric projection* techniques into *lossy* and *lossless*. As we discussed, visualizations employing dimensionality reduction are lossy because raw values are lost and cannot be retrieved by looking at the visualization. For example, an MDS projection can lay out cameras on a 2D space so that similar cameras are close to each other [45], but users cannot read the price or resolution of cameras unless separate detail-on-demand techniques are provided. In contrast, in a lossless projection, *any attribute value from any data case can be visually retrieved without interactions beyond basic scrolling and panning operations*. Thus, although in practice lossless projections may require interaction if the dataset is too large to fit the screen, in principle no interaction is required if the display is sufficiently large.

Lossless geometric projection approaches employ simple visual encodings and encompass some of the most commonly used multidimensional visualization techniques [59,86].

A table dataset with two dimensions can be visualized losslessly with a *2D scatterplot*. 2D scatterplots can be extended to more dimensions by employing either higher-dimensional scatterplots (e.g., 3D scatterplots) or star coordinates [50]. However, since the location of each data point on the display encodes a vector sum, both techniques are lossy. A lossless alternative involves creating 2D scatterplots for every pair of dimensions and arranging them in a *scatterplot matrix* [31]. Many variations of scatterplot matrices have been proposed, including versions that use color encodings [35,77], or extensions that support categorical data [31,44] or continuous multidimensional functions [82].

Another classic lossless geometric projection technique is the *parallel coordinates plot*, where dimensions are parallel axes, and data cases are polylines that intersect the axes at their corresponding values [46]. Variations of parallel coordinates exist that are circular [42], hierarchical [33], bundled [90], curved [5] or use 2D-3D layouts [48,81].

However, according to a recent survey [47], there is not enough empirical support to suggest that the alternative configurations outperform the original representation. A hybrid technique has also been proposed that combines parallel coordinates with scatterplot matrices [79].

A third lossless technique is the *tabular visualization*, i.e., a numerical table whose cell values are encoded visually [63, 83]. Common encodings include length (bars) [14, 63, 65], luminosity or hue (i.e., shaded cells) [63, 83], and area (e.g., circles) [14, 63, 71]. Tabular visualizations are supported to some extent by most modern spreadsheet software through a "conditional formatting" feature, where numerical values are generally displayed on top of the encodings [63].

Stacked bar charts and *grouped (or clustered) bar charts* are analogous to tabular visualizations that use bars to encode values, except bars are stacked or displayed next to each other instead of being aligned. Although stacked and grouped layouts are commonly used in statistical charts, studies have suggested that the aligned layout of tabular visualizations has perceptual benefits [22, 39, 75, 87]. Furthermore, stacked and grouped bar charts need to encode bars of the same category with color, which limits their scalability as multidimensional visualizations due to human limitations in color discrimination [80].

2.2.4 Evaluations

Previous studies suggest that 2D scatterplots outperform bivariate parallel coordinates for correlation tasks [57], and that scatterplots embedded within parallel coordinates outperform parallel coordinates alone for cluster detection [43]. A more recent study [55] compared parallel coordinates with three simplified forms of scatterplot matrices (where only a subset of the plots is shown) for basic value retrieval tasks, and found that one of the simplified forms outperformed parallel coordinates. However, it remains unclear whether complete scatterplot matrices (i.e., that include all $n(n-1)$ pairs of dimensions) would also outperform parallel coordinates in value retrieval if screen real-estate is controlled for. At the same time, simplified scatterplot matrices hide most bivariate relationships, and thus, may not be as suitable for overview tasks such as identifying highly correlated dimensions.

Evaluation of multidimensional visualization techniques is still in its infancy. We know little about how elementary multidimensional visualizations compare in terms of elementary analytic tasks, and even less so in terms of how they support decision tasks.

2.3 Visualizations Used in Decision Support

After reviewing these general approaches for visualizing multidimensional data, we now move to visualization tools specifically meant for decision support. We start by briefly reviewing domain-specific tools.

2.3.1 Domain-Specific Tools

A major application area for multi-attribute choice tasks is product comparison. The vast majority of product comparison charts produced for magazines and for the Web are tables[1], with various combinations of text and visual encodings (e.g., colors, checkmarks). Similarly, a number of interactive product comparison tools present products in a tabular visualization [58, 74]. Exceptions include ProductExplorer [67] which uses parallel coordinates. SmartClient [64] shows a subset of product alternatives in a scatterplot display, with a table for the remaining criteria and parallel coordinates if the users wish to apply constraints to many criteria. EZChooser shows products as an image collection and encodes criteria as bargrams (i.e., histograms whose bars have been tipped over and lined up end-to-end) [85].

Other domain-specific visualization tools for decision support exist, e.g., in areas such as financial investment [25, 69], software release planning [9], health [1] or lighting design [73]. In contrast to product comparison tools, these tools are not intended for a general audience and their design is hard to generalize beyond their specific application domain. As we focus on assessing the effectiveness of visualizations that can be used by a large audience and in a range of multi-attribute choice tasks, we devote the rest of this section to domain-agnostic tools.

[1] As of 24 Feb 2017, the twenty top results of the search query "product comparison" on Google Images are all tables.

2.3.2 General-Purpose Tools

Some visualization tools designed for exploratory data analysis are thought to aid multi-attribute choice due to their support for interactive querying and filtering. ScatterDice is a multidimensional data exploration tool based on scatterplot matrices [30]. In one scenario, a user browses cameras to buy, by creating a lasso query and refining it while transitioning between scatterplots [30]. HomeFinder represents data cases as dots on a map, while other attributes are represented as dynamic query widgets that can be used to progressively refine a query [84]. FilmFinder generalizes HomeFinder by changing the map into a scatterplot display [2]. Both HomeFinder and FilmFinder focus on specific domains (houses and films), but their widget-based query approach can be used with arbitrary datasets [32].

Some visualization tools support multi-attribute choice more explicitly, by allowing users to combine multiple attributes into a single aggregate score. ValueCharts [15] and LineUp [37] initially show the choice dataset as a tabular visualization where columns can be resized to express attribute importance. The entire visualization can then be collapsed into a stacked bar chart and sorted. This approach is effectively an interactive implementation of the weighted sum method in Multi-Criteria Decision Making [76]. WeightLifter [61] extends the approach by adding analytic and visualization tools such as parallel coordinates. CommonGIS supports decision tasks with geographical components (e.g., ranking counties by their need for funding, or choosing a skiing resort) [6, 7]. It also supports interactive weighted sums and implements a range of visualizations such as scatterplot matrices, parallel coordinates and tabular visualizations, all linked to a map.

As we saw, the majority of visualization tools meant to support multi-attribute choice employ lossless geometric projections. One exception is Dust & Magnet [89], where queries are embodied by magnets that are displayed in the same 2D space as data cases. The more a data case satisfies a query, the more it is attracted to the magnet. A scenario illustrates how a user can select cereals based on their dietary composition, by placing and moving magnets. Similarly, the Data Context Map [17], which features a scenario involving choosing a university, displays alternatives, attributes, and query results in the same unified 2D space. Since these representations are lossy projections, detail-on-demand techniques are provided to let users retrieve individual values.

2.3.3 Evaluations

Many visualization tools meant to support decisions have not been evaluated. Some of their design features have been assessed through qualitative studies, but without comparison to other techniques [6–8, 37, 61, 64, 89]. The few exceptions are controlled experiments comparing either variations of the same visualization [11, 24], or comparing a visualization with non-visualization base cases, such as web forms [67], static numerical tables [85], or Q&A systems and textual formats [84].

Despite the lack of comparative evaluations, the tasks and metrics from previous studies provide insights on how visualizations tools for decision support can be evaluated. Some studies have examined subjective user ratings [8, 61, 67]. More formal evaluations have employed tasks such as value retrieval [11, 24, 84], range tasks [67, 84], finding extrema [11, 24, 89], finding outliers [84], and identification of patterns [84], correlations [89], and clusters [89]. Other studies involved more complex analytic tasks combining multiple low-level tasks [8, 67]. In other words, a number of evaluations have employed analytic tasks.

Analytic tasks are informative when evaluating visualization tools for decision support, because good decisions require a good understanding of the relevant data. However, understanding the relevant data does not necessarily yield good decisions due to limits in human reasoning [49]. For example, in a recent visualization study, participants were almost 100% accurate in selecting good (non-dominated) alternatives using a scatterplot, but their decision appeared irrationally influenced by the presence of irrelevant (dominated) alternatives [27]. Therefore, it seems important to also include actual decision-making tasks when evaluating visualization tools meant to support decision making.

A few studies have indeed evaluated visualization tools using multi-attribute choice tasks. In the evaluation of EZChooser [85], participants were asked to choose among cameras and mutual funds, and

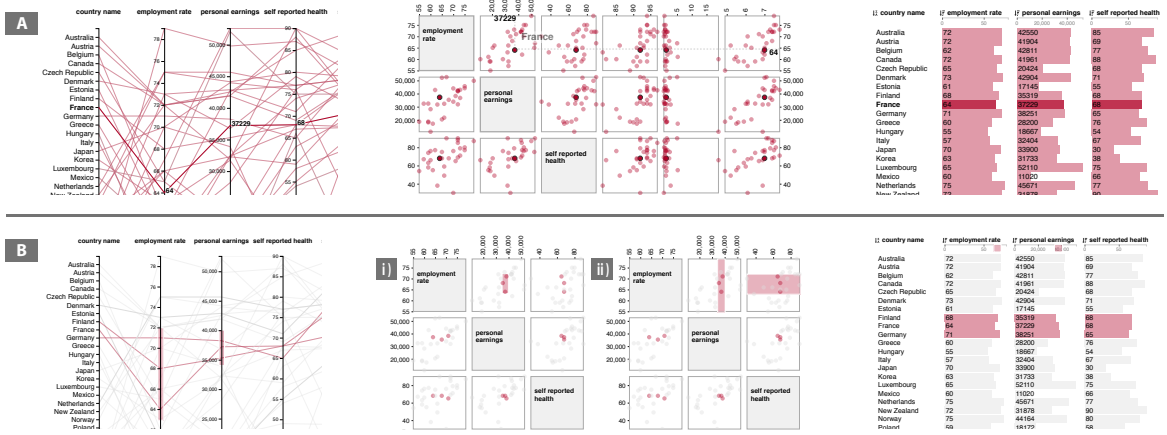

Fig. 2: A) Single data case "France" highlighted on hovering. B) Range selection on two dimensions. The three data cases at the intersection of the selected ranges (63-72 for "Employment rate", and 340000-40000 for "Personal earnings") are "France", "Germany" and "Finland".

the independent measures included *decision time*, *subjective ratings* of technique preference, as well as *satisfaction* and *confidence* in one's choice. Value Charts were evaluated twice [11, 24]. In the first study, participants were asked to choose among houses, cell phones and tourist attractions, and the study examined the *number of insights* acquired during the decision process [11]. In the second study, participants chose among universities and restaurants, and the study examined *time*, *choice satisfaction* and *confidence* [24]. In the evaluation of Dust & Magnet [89], participants indicated which attributes of a cereal brand they consider important, and were then asked to choose a cereal brand. However, the study focused on observing user interactions rather than reporting metrics of success. While all these studies involved actual decision making tasks, none of them used objective metrics of decision quality, and none of them used alternative visualizations as a basis of comparison — EZChooser was compared with numerical tables, Value Charts were compared with variations of the same tool, while Dust & Magnet did not have a control condition [11, 85, 89].

Although generic visualization tools for decision support are likely extremely useful, there still remain important limitations in their evaluation methodology: a lack of good baselines of comparison, a limited use of actual decision tasks and a lack of metrics for decision quality. We provide a first attempt at filling this gap by comparing three elementary multidimensional visualizations for their ability to support decisions. We decided to focus on elementary visualization techniques instead of complete tools such as Value Charts or LineUp [11, 37], because very little data exists on elementary visualizations and we believe that basic visual encodings and basic interactions need to be better understood before we can examine how they work in combination.

3 Technique Design

We focus our evaluation on three, commonly used, elementary lossless geometric projection visualizations: the parallel coordinates (PC), the scatterplot matrix (SM), and the tabular visualization (TV). Our evaluation methodology relies on two major principles: *(i)* include all features that are considered standard for each visualization, *(ii)* keep the visualizations as comparable as possible through a consistent visual design, a consistent interaction design, and by having all interactions present the same amount of information across visualizations.

3.1 Visual Encodings

Our implementation employs the most commonly used visual marks to represent data cases: polylines for PC, dots for SM, and bars for TV. We keep the visual design as consistent as possible across the techniques, to facilitate comparison. For example, visual marks share the same color across all techniques (translucent red by default, or translucent gray when outside a range selection), while decorations (e.g. axes, fonts) are consistently displayed in gray or black. The three techniques occupy similar vertical screen space, although the total area of SM is smaller due to its square aspect ratio that is not adapted to the typical landscape orientation of computer displays. More details are given next.

Parallel Coordinates (PC). We use the original representation introduced by Inselberg in 1960 [46]: a polylines diagram where the dimensions are represented as parallel axes and the data cases as polylines that intersect the axes at their corresponding values [47] (see PC in Fig. 1). This representation is considered standard in several infovis textbooks and surveys [59, 86]. As we saw in the Background section, many variations exist, but there is not enough empirical evidence that they outperform the original layout [47].

Scatterplot Matrix (SM). We use the full matrix, defined by Emerson et al. [31] as *"a grid of scatterplots showing the bivariate relationships between all pairs of variables in a multivariate data set"* (see SM in Fig. 1). As we have seen, simplified forms of scatterplot matrices exist that only show a subset of plots [55], but the complete scatterplot matrix (either square or triangular) has the advantage of showing all attribute pairs and is by far the most widely used [10, 16, 21, 30, 31, 56, 57, 59, 86].

Tabular Visulization (TV). We encode cell values by length (bars) [14, 63] (see TV in Fig. 1). Although other encodings exist (see Background section), we followed the Table Lens example [65] of choosing length, because it is more accurately perceived than other visual encodings [21], and because it is commonly used in tabular visualizations for decision support [6, 15, 37, 61]. We also display the numerical values on top of the bars, as is usually done in current spreadsheet software through the "conditional formatting" feature [63].

3.2 Interaction Techniques

Interaction is essential for analytic and visual exploration tasks, and likely also for decision making tasks. We chose to support three types of interactions which are either considered standard for at least two of the techniques, or have proven useful in decision making tools:

Highlighting of individual data cases with *linking* and *detail-on-demand* to support value retrieval across all criteria [15, 37, 61, 67] (see Fig. 2 A). Single data cases can be highlighted by hovering over a data case, which changes the opacity of the entire data case (line, dot or bar depending on the visualization) from the default 40% to 100%. Hovering over a data case or a dimension axis displays the precise values of the data case with tool-tips. Brushing and linking is commonly used in all techniques to highlight one or several data cases so as to help users relate their values across dimensions [72] or of the same row. In SM the data case highlighted in all plots (linking) assists users to relate the different views [12, 30, 59, 86].

Range selection on one or more dimensions to support dynamic filtering and queries [15, 37, 58, 64, 67] (see Fig. 2 B). This results in graying out all data cases outside the selection. If range selection is performed across multiple dimensions their intersection is shown, i.e., data cases that simultaneously fulfill the selected ranges for each dimension. Range selection in PC and TV is performed by brushing an axis, which in TV is located below the column titles [36]. Range selection is slightly different in SM given the bivariate nature of scatterplots (Fig. 2 i and ii). Instead of brushing individual axes, users draw selection rectangles inside the scatterplots. This effectively selects two ranges

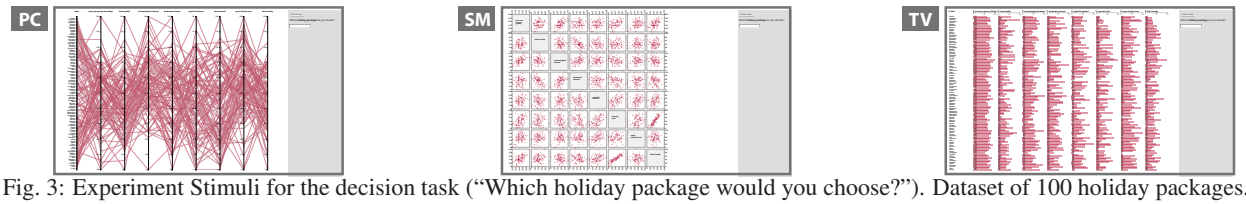

Fig. 3: Experiment Stimuli for the decision task ("Which holiday package would you choose?"). Dataset of 100 holiday packages.

at the same time (one for each dimension of the scatteplot). All range selections are re-sizable and drag-able through handles that appear on hover. While drawing or adjusting a range, the value of the range limits is displayed on the corresponding axes (not visible in the figures).

Dimension reordering to allow users to sort attributes by preference [30,37,67]. Rearrangement of dimensions brings together the ones relevant to the task. In our implementation, it is performed by dragging axis titles for all techniques. Reordering is fairly common in PC [72], and SM occasionally includes methods for manually or automatically reordering dimensions [30]. Unlike PC, though, in SM all possible pairs of dimensions are shown and thus reordering is not essential to perform side-by-side comparisons of dimensions. Reordering is also considered essential in TV, and research prototypes typically support manual reordering not only of columns, but also of rows [26,41,63,71]. Thus we also allow manual reordering of rows (data cases) in TV. Reordering data cases is impossible in PC and SM since the position of visual marks is determined by the data. Most research prototypes of tabular visualizations also support automatic reordering of rows or columns based on similarity [26,41,63,71], but we considered these features as too advanced for a comparison of elementary visualization techniques. Nevertheless, column sorting (a simple form of reordering) is a central feature of all commercial spreadsheet software tools, so we decided to include it in TV as well (both ascending and descending).

4 Experiment

Our goal is to explore how to evaluate elementary multidimensional visualizations for their ability to support decisions. To this end, we compare PC, SM and TV (see Fig. 1) according to how well they can support *i) basic data exploration*, by giving participants analytic tasks; and *ii) decision making*, by giving participants multi-attribute choice tasks. The reason behind this dual evaluation is that elementary analytic tasks can be thought as necessary (but not necessarily sufficient) components of multi-attribute choice tasks. By starting with basic tasks, we can train participants in reading and interacting with the visualizations before they proceed with the decision task. Doing so also allows us to ensure that they properly understood the techniques when they performed the choice task, thus eliminating potential confounds (e.g., a technique yielding poor decisions because participants did not know how to use it). Finally, a dual evaluation may uncover potentially interesting interactions between a technique's ability to support analytic tasks and its ability to support decision tasks. Again, there is currently little empirical data we can draw from on how the three techniques compare even for basic analytic tasks.

4.1 Research questions

Prior to data collection we framed the following research questions:

Q1 [ACCURACY] *How do the three techniques compare in terms of accuracy in **a**) analytic tasks and in **b**) decision tasks?*

Q2 [TIME-ON-TASK] *How do the three techniques compare in terms of speed in analytic tasks?*

Q3 [SUBJECTIVE PREFERENCE] *Which technique people prefer overall for **a**) analytic tasks **b**) decision tasks?*

Q4 [SUBJECTIVE CHOICE ASSESSMENT] *How do the three techniques compare in decision tasks in terms of choice **a**) satisfaction, **b**) confidence, **c**) easiness, and **d**) attachment?*

We did not consider time for decision tasks as part of our initial research questions, as we wanted to focus on accuracy and subjective satisfaction. All metrics are described in sections 4.9 and 4.10.

4.2 Tasks

We used three *analytic* tasks inspired from standard visualization taxonomies [3,68] and one *decision* task:

Value Retrieval. The task consisted of identifying the alternative having a certain attribute value and finding the value of another of its attributes [55]. Reading individual attribute values is likely very common in multi-attribute choice tasks. Value retrieval is also often considered as a building block of tasks like "find extrema" or "sorting" [4,55], that are both common in decision making [37,61].

Range. The task consisted of finding how many alternatives have their attribute X within a given range, and their attribute Y within another given range. This task is analogous to the "compute derived value" task [4]. It is likely involved in multi-attribute choice tasks when filtering alternatives, including when discarding unattractive options that do not match the decision makers' preferences and constraints.

Correlation. The task consisted of finding the pair of attributes that have the strongest correlation. This is an overview task, in contrast to correlation tasks that require to estimate the correlation of a single pair of dimensions [4,68] or to compare the correlation between two pairs of dimensions [66]. Identifying strong correlations can be involved in decision tasks where relations and trade-off comparisons between pairs of attributes are important [61]. For example, detecting a high correlation between two attributes such as quality and price may lead to a search for outliers which are particularly "good deals".

Decision. The task was a multi-attribute choice task as defined in Sect. 2.1. It consisted of finding the best alternative (in terms of subjective desirability) among a fixed set of alternatives (see Fig. 3).

4.3 Datasets and Task Generation

We used three different datasets:

Training. For the training, we used a dataset of country indicators from www.oecdbetterlifeindex.org, from which we selected 29 countries and 6 dimensions (e.g., life satisfaction, homicide rate, etc.).

Analytic tasks. For the analytic tasks, we used 18 synthetically generated datasets of student grades, containing 100 data cases each (students) and 8 dimensions (grades for different subjects such as English, math, biology, etc.). Grades were between 0 and 100.

Decision task. For the decision task, we used 3 synthetically generated datasets of holiday packages, containing 100 data cases each (holiday packages) and 8 dimensions: price per person (euro/day), hotel quality, archaeological interest, landscape interest, night life interest, security level, sport activity level, and kids friendly. Prices were between 100€ and 200€. All other dimensions were ratings from 0 to 100. Package names were generated using the City & Town Name Generator (www.mithrilandmages.com/utilities/CityNames.php).

For both the analytic and the decision datasets, correlated data was generated by sampling from random positive definite covariance matrices using the R packages `clusterGeneration` and `MASS`. Datasets were regenerated until the difference between the highest and the second highest absolute correlation was at least 0.3. For the analytic dataset, the highest correlation additionally had to be positive, and its two attributes had to be separated by at least a column. For the holiday dataset, price had to be positively correlated with all other dimensions.

Each analytic dataset yielded a *correlation task*. In addition, we generated a *value retrieval* task by randomly choosing a data case and two attributes (one to locate the data case, one to read the value), such that *i)* the attributes are separated by at least a column, and *ii)* the value of the attribute used to locate the data case is separated from the closest value by at least 0.02 (axes normalized between 0 and 1). We also

generated one *range task* per dataset by choosing two random attributes and value ranges such that *i)* the two attributes are separated by at least a column, *ii)* each endpoint of each range is separated from the closest value by at least 0.02, *iii)* range widths are between 0.1 and 0.8, *iv)* each range contains 1 to 5 data cases, and *v)* the intersection between the two ranges contains fewer data cases than either range alone.

4.4 Apparatus

We used a 1920x1080 resolution screen, with mouse and keyboard as input. The visualization software was implemented in D3, and questionnaires were shown on Google web forms.

4.5 Techniques

The three techniques (PC, SM and TV) are illustrated in Fig. 3 and explained in detail in Sect. 3. Each visualization entirely filled the vertical display space, and for PC and TV, the horizontal display space. Each visualization could accommodate the seven attributes without the need for scrolling, and with legible fonts. TV could display the 100 data cases simultaneously without the need for vertical scrolling.

4.6 Participants

We recruited 21 participants (6 female) among students, engineers and researchers working in computer science, with a mean self-reported experience in data visualization of 6.0/10 (range 2–9, σ : 1.66).

4.7 Experiment Design

We used a within-subjects design with independent factor the *visualization technique* (PC, SM and TV). The experiment was divided into two sessions. In the *analytic* session, participants performed the three analytic tasks in a fixed order: *four* trials of the value retrieval task, then *four* trials of the range task, then *two* trials of the correlation task, using the "student grades" dataset described in Sect. 4.3. During pilot testing the correlation task took much longer, so we decided to only include two trials to keep the experiment time manageable. Two training trials were performed before each new task. The presentation order for visualization technique was counterbalanced using a latin square.

In the *decision* session participants performed one decision task per technique, using the "holiday packages" dataset described in Sect. 4.3. This dataset was generated in a similar manner as the analytic dataset, but used different random values as well as different names for attributes and data cases in order to prevent the analytic session from influencing decisions and strategies used in the decision session. The order of the decision tasks was fixed while the technique presentation order followed that of the analytic session, effectively counterbalancing the dataset/technique pairing.

4.8 Procedure

We conducted a pilot study to ensure the clarity of the instructions and estimate task time. Our final experiment lasted on average 1.4 hours (ranging from 1.1 to 1.7 hours) and consisted of the following steps.

Technique Training: At the beginning of the experiment and before each change of technique, participants were shown, in paper, a table representation of a minimalistic dataset (5 data cases) next to the introduced technique. The experimenter then explained how to read the visualization. Participants were next shown the interactive version with the training dataset described in Sect. 4.3 (see Figs 2). For each interaction (highlighting, range selection and reordering), the experimenter explained the interaction and invited the participants to try on their own. A summary of all instructions was provided on a cheat-sheet paper that was visible by participants during the experiment.

Task Training: After technique training, participants moved into performing the analytic tasks as described previously. Each type of task was preceded by two training trials, one performed by the experimenter to illustrate the task, and one by the participant. When participants indicated they had understood the task, they moved on to performing the experimental trials without assistance. Participants typed their answer (value in the retrieval task, number of items in the range task, and pair of dimensions in the correlation) in a text field provided to the right of the screen (see Fig. 3). At the end, participants filled in a technique preference questionnaire described in 4.10.1.

Decision Task: After performing all analytic tasks with all techniques, participants were told they would now use the techniques to make a personal choice. They were asked to imagine planning their vacations and looking for the ideal holiday package. The meaning of each of the attributes of the holiday dataset was explained, and they were informed that they would see a different set of holiday packages each time. Participants conveyed their choice by copying the package's name in a text field provided to the right side of the screen.

As we will explain in Sect. 4.9.1, before the first, and after each decision task (4× total) participants filled-in a questionnaire to indicate which attributes they consider important. After each task, they also filled in a questionnaire to assess their satisfaction with their choice. At the very end of all decision tasks, participants filled in a questionnaire on their overall technique preference for decision tasks.

4.9 Objective Performance Metrics

We collected accuracy and time-on-task measures for both tasks. Accuracy in particular is a challenging measure to define in decision making, an inherently subjective task. Details are provided next.

4.9.1 Accuracy

For all tasks, we used a normalized measure of *accuracy* ranging from 0 to 1, with 1 being a fully correct answer. We used continuous measures whenever possible to maximize statistical power.

Analytic tasks: In the *value retrieval* task, where participants needed to find the value of an attribute, we gave a binary score (1 = correct, 0 = incorrect). A partially correct answer was difficult to define as values close to the correct value were often shared with other items. Thus there was no way to determine if an incorrect response was due to an incorrectly identified data case or due to a misread value. In the *range* task, where participants needed to count data cases, accuracy was defined $a = 1 - ¼|correct - response|$. All range tasks involved from 1 to 5 items, thus the normalizing term $5 - 1 = 4$. In the *correlation* task, accuracy was defined as $a = 1 - |correct - response|$, where *correct* was the highest correlation in the dataset, and *response* was the correlation between the two attributes given as a response.

Decision tasks: There is no simple way to define the accuracy of a multi-attribute choice task, given its subjective nature. Although Pareto dominance is one such measure [27], the selection of a dominated alternative is unlikely in our dataset given the number of alternatives and attributes. We thus decided to use as an indicative measure of accuracy the *consistency between the choice made by a participant and her self-reported preferences*. As mentioned before, participants rated the *importance* of each of the 8 holiday package attributes between 0 and 10. They also indicated the *direction* of their preference, i.e., whether they prefer the attribute to be high or low. For example, a holiday package with lots of physical activity can be perceived as desirable by an athletic person but undesirable by someone with reduced mobility. As preferences may evolve during the session, the questionnaire was administered before and after each decision task (4× total).

Based on this data, we can roughly estimate how desirable each alternative should be using a weighted sum approach [76]. For each participant and decision dataset, we compute a *desirability score* per alternative as follows: for each attribute, *i)* divide its value by the maximum allowed value, *ii)* if the user's preference is toward small values, replace the value with $1 - value$, *iii)* multiply the value by the attribute's importance obtained from the questionnaire (0–10). Once done, sum up all attribute values to obtain a desirability score d for that particular alternative. We repeat the process for all alternatives, then normalize all d scores between 0 and 1. Thus, the "optimal" alternative in the dataset has a d of 1 while the worst one has a d of 0.

Desirability scores can be computed using the preferences elicited either before the decision task (d_{pre}), or after the task (d_{post}). Since preferences can evolve while exploring options, d_{post} may seem more indicative of the "true" desirability. However, a participant may also update their preferences *after* the choice was made, e.g., as a way of rationalizing their choice. Thus, we consider both d_{pre} and d_{post} and

define the accuracy of a decision task as $a = \max(d_{pre}, d_{post})$, with d_{pre} and d_{post} being the desirability scores of the chosen alternative. This score is an approximation and is not meant to capture decision quality perfectly. The elicited preferences may not be completely reliable, and cannot fully capture the complexity of choice criteria (i.e., someone may want an attribute to be neither too high nor too low). However, if a visualization happens to be misleading or particularly hard to use, we can expect participants to make choices that are clearly inconsistent with their preferences, thus yielding an abnormally low precision score.

4.9.2 Time-on-task

Analytic tasks: We consider the time participants took to complete each analytic task, from the moment the task page is displayed to pressing the ENTER key after giving the answer.

Decision tasks: We did not consider completion times for decision tasks in our planned analysis, but considered including them in post-hoc analyses. Time was measured from the moment the decision dataset was shown, to when participants pressed ENTER to confirm their choice.

4.10 Subjective Metrics

We considered two types of subjective metrics: technique preference (for analytic and decision tasks) and choice assessment (for decision tasks). All responses were reported on an 11-point scale, from 0 to 10.

4.10.1 Technique Preference

We asked participants to rate the techniques based on overall preference.

Analytic tasks: After completing all analytic tasks (value, range and correlation) with all techniques, participants were asked to rate how easy and helpful they found found each technique. They were orally instructed not to focus on a specific analytic task but on their overall experience. They were also given the option to justify their ratings.

Decision tasks: Similarly, after completing all decision tasks with all techniques, participants were asked to rate how easy and helpful they found each technique for choosing a holiday package.

4.10.2 Choice Assessment

After completing each decision task (one per technique) and before the next preference elicitation questionnaire, participants evaluated the choice they just made according to the following criteria:

- *satisfaction*: Participants were asked to what extent they are satisfied with their choice ranging from "not satisfied at all" to "very satisfied";
- *confidence*: They were asked to what extent they are confident about their choice ranging from "not confident at all" to "very confident";
- *easiness*: They were asked to what extent they consider this choice as easy to make ranging from "very difficult" to "very easy";
- *attachment*: Participants were asked to imagine that an automatic recommender system could suggest another choice from the dataset taking into account their preferences, and were asked whether they would switch to this choice ranging from "I would definitely stick to my initial choice" to "I would definitely switch".

The first three subjective metrics are often used in decision support tool evaluations [11, 37, 89]. They are meant to complement the objective accuracy metrics described previously, by explicitly asking the participants to evaluate their choice. The fourth metric (attachment) is based on Chernev [18], and provides a more indirect way of asking participants to evaluate their choice. Chernev used this metric as the primary dependent variable in a decision making study involving the evaluation of consumer choices, assuming that participants who are confident in their choice will have less propensity to switch [18].

5 Results

We analyze, report and interpret all our inferential statistics using interval estimation [29]. Experimental stimuli, data and analyses are available at http://www.aviz.fr/dm.

5.1 Planned Analyses

In this section, we focus on the analyses planned before data was collected. Each subsection corresponds to one of our research questions stated in Sect. 4.1, with the same notation **Qx**. All differences between techniques are computed within-subjects (paired samples).

5.1.1 Q1 – Accuracy

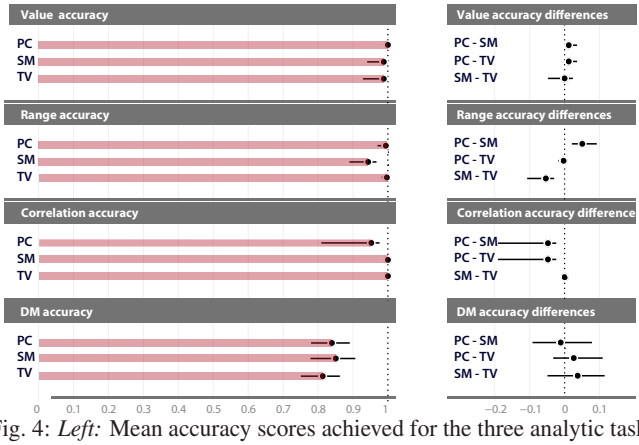

Fig. 4: *Left:* Mean accuracy scores achieved for the three analytic tasks and the decision making (DM) task, using the parallel coordinates plot (PC), the scatterplot matrix (SM), and the tabular visualization (TV). *Right:* Mean differences in accuracy scores between each pair of techniques — a positive value indicates that the left technique outperforms the technique on the right. All error bars are 95% CIs (n=21).

Results for accuracy are reported in Fig. 4. Each of the four horizontal panels shows the results for one type of task. The top three panels report accuracies for the analytic tasks (value retrieval, range and correlation), while the bottom panel reports accuracies for the decision task. The bar charts on the left show the *mean accuracy* score for each technique, while the dot plots on the right show the *mean differences in accuracy* between techniques. A positive value (to the right of the zero axis) indicates that the left technique outperforms the right one. For each mean, a point estimate is reported together with a 95% confidence interval (CI) indicating the range of plausible values for the population mean [29]. All confidence intervals are 95% BCa bootstrap CIs [53].

Q1a. We can see that participants achieved a perfect or close-to-perfect accuracy score in almost all analytic tasks. The two exceptions are the range task carried out with SM, and the correlation task performed with PC. In both cases, participants were reliably less accurate than with the other techniques, but the differences are relatively small. This means that participants followed the instruction to be as accurate as possible, and completion times (analyzed thereafter) should give a good indication of overall performance with analytic tasks.

Q1b. For the multi-attribute choice tasks (DM), participants were on average fairly accurate with all techniques in terms of how consistent their choices were with their self-reported preferences. That said, no technique yielded a perfect or close-to-perfect average accuracy score, meaning that participants rarely made an "optimal" choice regardless of which technique they were using. Interestingly, there is no sign of a clear difference in accuracy between the three techniques.

5.1.2 Q2 – Time-on-Task

Fig. 5 presents the average amount of time spent by participants in each analytic task. As before, mean observations (in seconds) are reported to the left. This time, raw measurements were log-transformed to correct for positive skewness and reduce the influence of extreme observations, and were then antilogged at the end of the analysis [29]. As a consequence, all reported means are geometric means, and differences between techniques are expressed as ratios of mean completion times (reported to the right). A value to the left of the x=1 axis means that the numerator technique is faster than the denominator technique. All confidence intervals are exact confidence intervals for the normal distribution, computed on the logged observations.

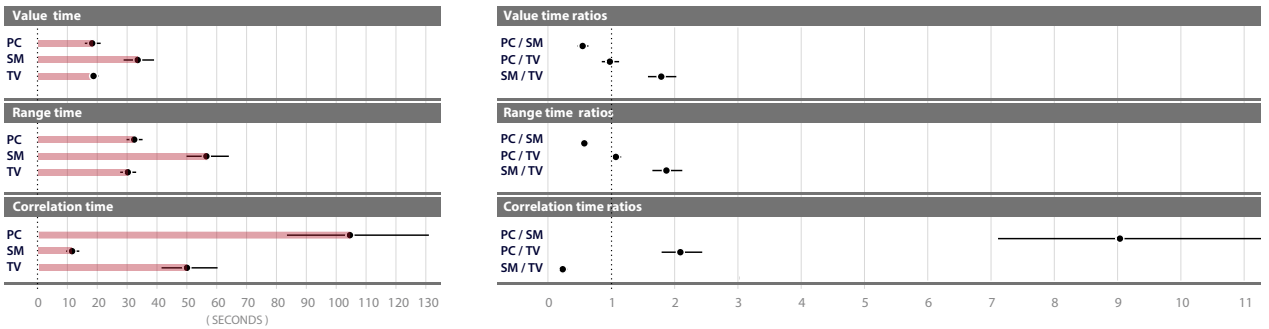

Fig. 5: *Left:* Average time (in seconds) spent on each analytic task for techniques PC, SM and TV. *Right:* Average time ratios between each pair of techniques — a value less than one indicates that the left technique is faster than the technique on the right. All error bars are 95% CIs (n=21).

Q2. The top horizontal panel in Fig. 5 reports completion times for the *value retrieval* task. The figure provides strong evidence that participants were much slower on average (almost twice as slow) with SM than the other two techniques, which are comparable in speed. The results are similar for the *range* task (second panel), with SM being again almost twice as slow as the other two techniques. TV is possibly slightly faster than PC (ratio PC/TV of 1.1, 95% CI [0.99,1.2]). The results are very different for the *correlation* task. SM is remarkably fast: about 9 times faster than PC and 4 times faster than TV. Here PC is clearly the slowest, with TV being about twice as fast as PC.

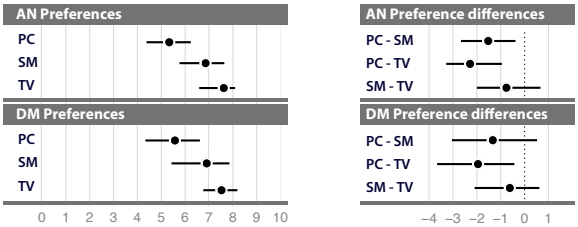

Fig. 6: *Left:* Mean rating for each technique, for the analytic (AN) and for the decision (DM) tasks. *Right:* Mean differences in ratings between each pair of techniques — a positive value indicates a preference for the technique on the left. Error bars are 95% CIs (n=21).

5.1.3 Q3 – Subjective Preference

Fig. 6 presents mean participant ratings, in terms of how easy and helpful they felt the techniques was when carrying out analytic tasks (top panel) and decision making tasks (bottom panel). On the difference plots, a positive value indicates that the left technique is on average preferred to the technique on the right (conversely, a negative value indicates a preference for the right technique). All CIs are 95% BCa bootstrap confidence intervals.

Q3a. For analytic tasks, PC appears as the least preferred. Our data does not provide enough support for a difference between SM and TV.

Q3b. For decision making, results suggest that participants prefer TV over PC. We do no have enough evidence to draw other conclusions.

5.1.4 Q4 – Subjective Choice Assessment

Fig. 7 reports how participants evaluated the choice they made in the decision-making task, depending on the technique used. Each horizontal panel presents the results according to a different choice assessment metric (see Section 4.10.2). On the difference plots, a positive value indicates a higher average rating for the technique on the left. All CIs are again 95% BCa bootstrap confidence intervals.

Q4a. There is no evidence of a major difference between techniques in terms of average participants' satisfaction with their choice. We cannot conclude as to the direction of the effects, but the differences are likely no more than ±1 point on an 11-point Likert scale.

Q4b. The data is also inconclusive regarding participants' confidence in their choice, except we know that large effects are implausible (likely not above ±1.5 points).

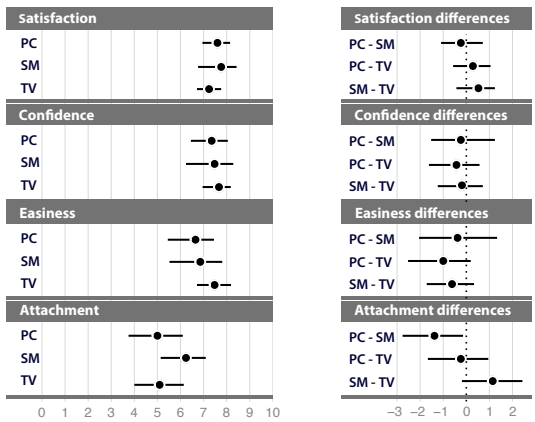

Fig. 7: *Left:* Mean rating for each choice assessment metric and each technique. *Right:* Mean differences between each pair of techniques — a positive value indicates a benefit (e.g., higher choice satisfaction) for the technique on the left. All error bars are 95% CIs (n=21).

Q4c. Concerning perceived easiness, the precision of our estimates is again low, but it is not implausible that decisions made with TV are perceived as easier to make on average than with PC and SM. However, the evidence is rather weak.

Q4d. The data suggests that on average, participants may be slightly more attached to their choice if they made it using SM than if they used either PC or TV. There is no evidence for a major difference between PC and TV in terms of attachment.

5.2 Additional Analyses

We now report additional (unplanned) analyses, to better understand in what respects the three techniques differ.

5.2.1 Time-on-Task for Decision Making

When framing our research questions, we reasoned that time-on-task was of secondary concern for decision making, as the answers themselves seemed more important than the time it took to reach them. Time-on-task is also difficult to interpret for open-ended tasks, as increased time can be a sign of both increased difficulty and increased engagement [28].

However, the three techniques turned out to be hard to distinguish in terms of decision accuracy and subjective choice assessment. The effects there are likely small (i.e., likely not more than a ±10% difference in accuracy and ±15% for subjective metrics, see Figures 4 and 7), requiring a large statistical power to be estimated reliably. Therefore, the time metric can be a useful differentiating factor. Time can also be of particular interest when decisions have to be made rapidly.

Fig. 8 shows the average amount of time it took participants to make their choice with each technique. The analysis method was the same as for the analytic tasks (Sect. 5.1.2). As we can see, there is some evidence that decisions were made more rapidly with TV than with the other two techniques: both SM and PC took on average 1.3 times longer, with 95% CI [1.1, 1.6] for SM and 95% CI [0.96, 1.7] for PC.

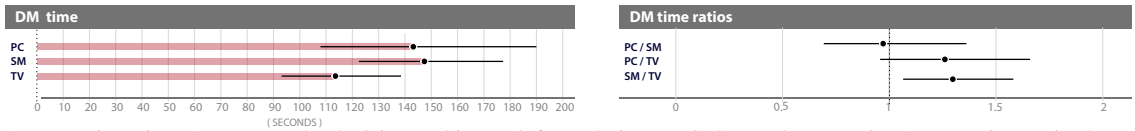

Fig. 8: *Left:* Average time (in sec) spent on the decision making task for techniques PC, SM and TV. *Right:* Average time ratios between each pair of techniques — a value less than one indicates that the left technique is faster than the technique on the right. All error bars are 95% CIs (n=21).

5.2.2 Qualitative Feedback

A text field allowed participants to justify their technique ratings on decision tasks, and 13 out of the 21 participants did so. Two raters (co-authors) independently broke down text responses into positive and negative comments (Cohen's kappa = 0.90 for segmenting+classification). The 13 respondents produced a total of 50 comments.

PC received *9 negative* comments, characterizing PC as hard to use for comparing alternatives, as well as for searching, isolating and selecting a single alternative. PC received *4 positive* comments, on how polylines allow performing a quick evaluation of individual alternatives.

SM received *5 negative* comments, mainly on its visual complexity, and on the amount of information presented that can be overwhelming. Meanwhile, *11 positive* comments referred to SM's advantages in overview tasks making it possible to see patterns and trade-offs between attributes, as well as its ability to filter multiple attributes at once.

TV received *4 negative* comments, mainly because alternatives can only be sorted by one attribute, making it hard to perform overview comparisons that take into account all attributes. TV received *17 positive* comments, stating that it was easy and straightforward in a range of elementary tasks, e.g., comparing and identifying alternatives, or isolating and selecting them. One participant also found TV's support for manual reordering of alternatives very useful for making decisions.

Overall, PC received the largest number of negative comments, mostly because it did not allow to easily compare alternatives. SM and TV received the largest number of positive comments, mostly because they supported well overview (for SM) and elementary tasks (for TV). Meanwhile, some comments about SM appeared strongly negative (e.g., "extremely difficult" to understand at first), while none of the negative comments on TV seemed to have reported strong drawbacks. TV has also received the largest number of positive comments (17 vs. 11 and 4). Although no strong conclusion can be derived from this observation alone, it appears consistent with the preference ratings (Fig. 6).

5.3 Summary and Discussion

In order to verify that participants were able to understand the visualizations and use them effectively, we first evaluated the visualizations on analytic tasks. All three techniques yielded close-to-perfect accuracy. There were however large differences in completion times: SM was slowest for value retrieval and range tasks, but by far the fastest in correlation tasks. The lower performance of SM in the two low-level analytic tasks can be explained by the lower resolution of SM's axes compared to PC (see Fig. 3), and by the difficulty of dealing with two axes concurrently. As one participant noted "*I felt I had to pay more attention to which axis corresponded with which variable, and my eyes where on the axes while dragging on the dots*". On the other hand, the efficiency of SM for correlation tasks is not surprising, as scatterplots are known to convey correlation effectively [51, 57]. Furthermore, SM shows all pairwise correlations simultaneously, while both PC and TV required manual attribute reordering to examine them in sequence. Though PC is often considered a good choice for conveying correlations [38,40,57], it was outperformed by TV both on time and accuracy.

The second part of our evaluation involved actual decision-making tasks. Overall, we found our techniques to be comparable across metrics, with a slight speed advantage for TV. Participants also preferred TV over PC overall. Participants reported being more attached to choices they made with SM on average, a result that needs to be confirmed by further studies. The reasons for this are currently unclear, although one explanation could be that SM's better support for overview tasks (confirmed by our results with the correlation task) made participants more confident that they did not miss a particularly interesting alternative. However, this difference was not captured by the confidence metric.

Our metrics for decision support overall showed a large variability in responses compared to the analytic tasks. This is likely due to the fact that our multi-attribute choice tasks were that involve personal preferences are inherently subjective. In addition, participants may not be able to perfectly express (or be aware of) their criteria preferences, which likely adds further noise to our accuracy metrics. As a result, many of our metrics were not sensitive enough to capture differences across conditions that likely exist [23]. Additional work is needed on establishing more sensitive metrics of choice quality, considering also non preference-based choices (e.g., data-driven medical decisions). It seems though that the time metric can become a useful tie breaker when participants achieve sufficient decision accuracy across techniques.

6 Conclusions

There has been little work on how to evaluate visualizations for decision support. In this work we explored conceptual and methodological issues in assessing decision support in multidimensional visualizations. We first defined the notion of multi-attribute choice task and provided a systematic analysis of multi-dimensional visualizations that may support this task. We then identified various objective (accuracy, time) and subjective metrics (satisfaction, easiness, attachment, preference) that can be used to assess decision support. We illustrated how we can use these metrics by empirically comparing three common general-purpose multidimensional visualizations. Overall, tabular visualizations seem to be a compelling choice, despite the low attention they have received in the literature on multidimensional visualization.

Although our evaluation focused on elementary and generic visualization techniques, it can inform the design and evaluation of more complex visualization tools targeted at decision support. For example, our findings on TV provide an empirical justification for the use of a tabular layout in decision support visualization systems such as ValueCharts [15], LineUp [37], and WeightLifter [61]. Besides, the support for interactive weighted-sum ranking provided by these tools is likely to be a very useful addition to the basic TV technique. Our finding on choice attachment with SM also provides some support for the addition of scatterplots to provide overviews [61]. Complete systems however remain to be evaluated more formally in the future, possibly using a similar method as the one we proposed here.

In our evaluation we limited our choice sets to 100 items, which is a reasonable size for an everyday choice task. It would nonetheless be interesting to expand such evaluations to larger choice sets, in which case some form of data aggregation (e.g., through attribute clustering) may need to be added, an aspect that could complicate the evaluation. It would also be interesting to generalize our methodology to other types of datasets and decision-making tasks.

Evaluating visualizations for their ability to support decision making is challenging. The quality of a decision is hard to capture with objective measures, as decisions often involve personal preferences which are themselves hard to capture reliably. Self-reported measures of confidence or satisfaction are informative but also inherently noisy. As a workaround, previous work has often evaluated visualizations using analytic tasks with a well-defined ground truth. Testing a representative set of such tasks can indeed help determine whether decision makers can understand the information on which to base their decision. Nevertheless, testing real decision tasks can provide more insights. More work is however needed on identifying objective and subjective metrics of decision quality that are sensitive enough to detect small (but possibly meaningful) differences between techniques.

Acknowledgments

We would like to thank Jean-Daniel Fekete, Sriram Karthik and Frédéric Vernier for their help. Work partly supported by Digicosme grant.

REFERENCES

[1] S. Afzal, R. Maciejewski, and D. S. Ebert. Visual analytics decision support environment for epidemic modeling and response evaluation. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 191–200. IEEE, 2011.

[2] C. Ahlberg and B. Shneiderman. Visual information seeking: Tight coupling of dynamic query filters with starfield displays. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 313–317. ACM, 1994.

[3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pp. 111–117. IEEE, 2005.

[4] R. Amar and J. Stasko. Best paper: A knowledge task-based framework for design and evaluation of information visualizations. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pp. 143–150. IEEE, 2004.

[5] D. F. Andrews. Plots of high-dimensional data. *Biometrics*, pp. 125–136, 1972.

[6] G. Andrienko, N. Andrienko, and P. Jankowski. Building spatial decision support tools for individuals and groups. *Journal of Decision Systems*, 12(2):193–208, 2003.

[7] N. Andrienko and G. Andrienko. Informed spatial decisions through coordinated views. *Information Visualization*, 2(4):270–285, 2003.

[8] T. Asahi, D. Turo, and B. Shneiderman. Using treemaps to visualize the analytic hierarchy process. *Information Systems Research*, 6(4):357–375, 1995.

[9] B. A. Aseniero, T. Wun, D. Ledo, G. Ruhe, A. Tang, and S. Carpendale. Stratos: Using visualization to support decisions in strategic software release planning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1479–1488. ACM, 2015.

[10] K. E. Basford and J. W. Tukey. *Graphical Analysis of Multi-Response Data*, vol. 6. CRC Press, 1998.

[11] J. Bautista and G. Carenini. An empirical evaluation of interactive visualizations for preferential choice. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 207–214. ACM, 2008.

[12] R. A. Becker and W. S. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[13] P. Bera. How colors in business dashboards affect users' decision making. *Communications of the ACM*, 59(4):50–57, 2016.

[14] J. Bertin. *La graphique et le traitement graphique de l'information*. Number 91 (084.21) BER. 1977.

[15] G. Carenini and J. Loyd. Valuecharts: analyzing linear models expressing preferences and evaluations. In *Proceedings of the working conference on Advanced visual interfaces*, pp. 150–157. ACM, 2004.

[16] D. B. Carr, R. J. Littlefield, W. Nicholson, and J. Littlefield. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82(398):424–436, 1987.

[17] S. Cheng and K. Mueller. The data context map: Fusing data and attributes into a unified display. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):121–130, 2016.

[18] A. Chernev. When more is less and less is more: The role of ideal point availability and assortment in consumer choice. *Journal of Consumer Research*, 30(2):170–183, 2003.

[19] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368, 1973.

[20] H. Chernoff and M. H. Rizvi. Effect on classification error of random permutations of features in representing multivariate data by faces. *Journal of the American Statistical Association*, 70(351a):548–554, 1975.

[21] W. S. Cleveland. *Visualizing data*. Hobart Press, 1993.

[22] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[23] J. Cohen. The Earth is round ($p<.05$). *American psychologist*, 49(12):997, 1994.

[24] C. Conati, G. Carenini, E. Hoque, B. Steichen, and D. Toker. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. In *Computer Graphics Forum*, vol. 33, pp. 371–380. Wiley Online Library, 2014.

[25] M. Daradkeh, C. Churcher, and A. McKinnon. Supporting informed decision-making under uncertainty and risk through interactive visualisation. In *Proceedings of the Fourteenth Australasian User Interface Conference-Volume 139*, pp. 23–32. Australian Computer Society, Inc., 2013.

[26] A. de Falguerolles, F. Friedrich, and G. Sawitzki. A tribute to J. Bertin's graphical data analysis. In W. Bandilla and F. Faulbaum, eds., *SoftStat '97 (Advances in Statistical Software 6)*, pp. 11–20. Lucius & Lucius, 1997.

[27] E. Dimara, A. Bezerianos, and P. Dragicevic. The attraction effect in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):471–480, 2017.

[28] E. Dimara, A. Bezerianos, and P. Dragicevic. Narratives in Crowdsourced Evaluation of Visualizations: A Double-Edged Sword? In *ACM Conference on Human Factors in Computing Systems (CHI)* . Denver, United States, May 2017. doi: 10.1145/3025453.3025870

[29] P. Dragicevic. Fair statistical communication in hci. In *Modern Statistical Methods for HCI*, pp. 291–330. Springer, 2016.

[30] N. Elmqvist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1539–1148, 2008.

[31] J. W. Emerson, W. A. Green, B. Schloerke, J. Crowley, D. Cook, H. Hofmann, and H. Wickham. The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1):79–91, 2013.

[32] J.-D. Fekete. The infovis toolkit. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pp. 167–174. IEEE, 2004.

[33] Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of the conference on Visualization'99: celebrating ten years*, pp. 43–50. IEEE Computer Society Press, 1999.

[34] J. Fuchs, P. Isenberg, A. Bezerianos, and D. Keim. A systematic review of experimental studies on data glyphs. *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[35] G. W. Furnas and A. Buja. Prosection views: Dimensional inference through sections and projections. *Journal of Computational and Graphical Statistics*, 3(4):323–353, 1994.

[36] N. Golmie and B. Kules. Highlight and selection control for dynamic table visualization. CMSC 838S Term Paper, 1999.

[37] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, 2013.

[38] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014.

[39] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 203–212. ACM, 2010.

[40] J. Heinrich, Y. Luo, A. E. Kirkpatrick, and D. Weiskopf. Evaluation of a bundling technique for parallel coordinates. *Proceedings of International Conference on Information Visualization Theory and Applications*, pp. 594–602, 2012.

[41] N. Henry and J.-D. Fekete. Matrixexplorer: a dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):677–684, 2006.

[42] P. E. Hoffman. *Table visualizations: a formal model and its applications*. PhD thesis, University of Massachusetts Lowell, 1977.

[43] D. Holten and J. J. Van Wijk. Evaluation of cluster identification performance for different pcp variants. In *Computer Graphics Forum*, vol. 29, pp. 793–802. Wiley Online Library, 2010.

[44] J.-F. Im, M. J. McGuffin, and R. Leung. Gplom: the generalized plot matrix for visualizing multidimensional multivariate data. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2606–2614, 2013.

[45] S. Ingram, T. Munzner, and M. Olano. Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics*, 15(2):249–261, 2009.

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

[47] J. Johansson and C. Forsell. Evaluation of parallel coordinates: Overview, categorization and guidelines for future research. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):579–588, 2016.

[48] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing structure in visualizations of dense 2d and 3d parallel coordinates. *Information Visualization*, 5(2):125–136, 2006.

[49] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[50] E. Kandogan. Star coordinates: A multi-dimensional visualization tech-

nique with uniform treatment of dimensions. In *Proceedings of the IEEE Information Visualization Symposium*, vol. 650, p. 22. Citeseer, 2000.

[51] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE transactions on visualization and computer graphics*, 22(1):469–478, 2016.

[52] D. A. Keim and H.-P. Kriegel. Visualization techniques for mining large databases: A comparison. *IEEE Transactions on knowledge and data engineering*, 8(6):923–938, 1996.

[53] K. N. Kirby and D. Gerlanc. Bootes: An r package for bootstrap confidence intervals on effect sizes. *Behavior research methods*, 45(4):905–927, 2013.

[54] A. Klippel, F. Hardisty, and C. Weaver. Star plots: How shape characteristics influence classification tasks. *Cartography and Geographic Information Science*, 36(2):149–163, 2009.

[55] X. Kuang, H. Zhang, S. Zhao, and M. J. McGuffin. Tracing tuples across dimensions: A comparison of scatterplots and parallel coordinate plots. In *Computer Graphics Forum*, vol. 31, pp. 1365–1374. Wiley Online Library, 2012.

[56] D. J. Lehmann, G. Albuquerque, M. Eisemann, M. Magnor, and H. Theisel. Selecting coherent and relevant plots in large scatterplot matrices. In *Computer Graphics Forum*, vol. 31, pp. 1895–1908. Wiley Online Library, 2012.

[57] J. Li, J.-B. Martens, and J. J. Van Wijk. Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization*, 9(1):13–30, 2010.

[58] R. Liu, T. Chao, C. Plaisant, and B. Shneiderman. Manylists: product comparison tool using spatial layouts with animated transitions. *University of Maryland Technical Report*, 2012.

[59] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[60] M. C. F. d. Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, 2003.

[61] S. Pajer, M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Muller, and H. Piringer. Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):611, 2017.

[62] K. Peason. On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2(11):559–572, 1901.

[63] C. Perin, P. Dragicevic, and J.-D. Fekete. Revisiting bertin matrices: New interactions for crafting tabular visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2082–2091, 2014.

[64] P. Pu and B. Faltings. Enriching buyers' experiences: the smartclient approach. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 289–296. ACM, 2000.

[65] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 318–322. ACM, 1994.

[66] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. In *Computer Graphics Forum*, vol. 29, pp. 1203–1210. Wiley Online Library, 2010.

[67] P. Riehmann, J. Opolka, and B. Froehlich. The product explorer: decision making with ease. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pp. 423–432. ACM, 2012.

[68] S. F. Roth and J. Mattis. Data characterization for intelligent graphics presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 193–200. ACM, 1990.

[69] S. Rudolph, A. Savikhin, and D. S. Ebert. Finvis: Applied visual analytics for personal financial planning. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 195–202. IEEE, 2009.

[70] M. Sedlmair, M. Brehmer, S. Ingram, and T. Munzner. Dimensionality reduction in the wild: Gaps and guidance. *Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03*, 2012.

[71] H. Siirtola. Interaction with the reorderable matrix. In *Information Visualization, 1999. Proceedings. 1999 IEEE International Conference on*, pp. 272–277. IEEE, 1999.

[72] H. Siirtola and K.-J. Räihä. Interacting with parallel coordinates. *Interacting with Computers*, 18(6):1278–1309, 2006.

[73] J. Sorger, T. Ortner, C. Luksch, M. Schwarzler, E. Groller, and H. Piringer. Litevis: Integrated visualization for simulation-based decision support in lighting design. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):290–299, 2016.

[74] M. Spenke, C. Beilken, and T. Berlage. Focus: the interactive table for product comparison and selection. In *Proceedings of the 9th annual ACM symposium on User interface software and technology*, pp. 41–50. ACM, 1996.

[75] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2152–2160, 2014.

[76] E. Triantaphyllou. *Multi-criteria decision making methods: a comparative study*, vol. 44. Springer Science & Business Media, 2013.

[77] L. Tweedie and R. Spence. The prosection matrix: a tool to support the interactive exploration of statistical models and data. *Computational Statistics*, 13(1):65–76, 1998.

[78] M. Velasquez and P. T. Hester. An analysis of multi-criteria decision making methods. *International Journal of Operations Research*, 10(2):56–66, 2013.

[79] C. Viau, M. J. McGuffin, Y. Chiricota, and I. Jurisica. The flowvizmenu and parallel scatterplot matrix: Hybrid multidimensional visualizations for network exploration. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1100–1108.

[80] C. Ware. *Information visualization: perception for design*. Elsevier, 2012.

[81] R. Wegenkittl, H. Loffelmann, and E. Groller. Visualizing the behaviour of higher dimensional dynamical systems. In *Visualization'97., Proceedings*, pp. 119–125. IEEE, 1997.

[82] J. J. Wijk and R. Liere. Hyperslice visualization of scalar functions of many variables. 1994.

[83] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 2012.

[84] C. Williamson and B. Shneiderman. The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '92, pp. 338–346. ACM, New York, NY, USA, 1992. doi: 10.1145/133160.133216

[85] K. Wittenburg, T. Lanning, M. Heinrichs, and M. Stanton. Parallel bargrams for consumer-based information exploration and choice. In *Proceedings of the 14th annual ACM symposium on User interface software and technology*, pp. 51–60. ACM, 2001.

[86] P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization*, pp. 3–33, 1994.

[87] M. A. Yalçına, N. Elmqvistb, and B. B. Bedersona. Evaluating multicolumn bar charts and treemaps for dense visualization of sorted numeric data. 2015.

[88] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pp. 105–112. IEEE, 2003.

[89] J. S. Yi, R. Melton, J. Stasko, and J. A. Jacko. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4(4):239–256, 2005.

[90] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual clustering in parallel coordinates. In *Computer Graphics Forum*, vol. 27, pp. 1047–1054. Wiley Online Library, 2008.

[91] S. Zionts. Mcdm—if not a roman numeral, then what? *Interfaces*, 9(4):94–101, 1979.