DOI: 10.1111/cgf.14438

COMPUTER GRAPHICS *forum*
Volume 41 (2022), number 1 pp. 335–349

Evaluating Data-type Heterogeneity in Interactive Visual Analyses with Parallel Axes

José Matute and Lars Linsen

University of Münster, Münster, Germany
{matutefl, linsen}@uni-muenster.de

Abstract

The application of parallel axes for the interactive visual analysis of multidimensional data is a widely used concept. While multidimensional data sets are commonly heterogeneous in nature, i.e. data items contain both numerical and categorical (including ordinal) attribute values, the use of parallel axes often assumes either numerical or categorical attributes. While Parallel Coordinates and their large variety of extensions focus on numerical data, Parallel Sets and related methods focus on categorical attributes. While both concepts allow for displaying heterogeneous data, no clear strategies have been defined for representing categories in Parallel Coordinates or discretization of continuous ranges in Parallel Sets. In practice, type conversion as a pre-processing step can be used as well as coordinated views of numerical and categorical data visualizations. We evaluate traditional and state-of-the-art approaches with respect to the interplay of categorical and numerical dimensions for querying probability-based events. We also compare against a heterogeneous Parallel Coordinates/Parallel Set approach with a novel interface between categorical and numerical axes . We show that approaches for mapping categorical data to numerical axis representations can lead to lower accuracy in answering probability-based questions and higher response times than hybrid approaches in multiple-event scenarios.

Keywords: information visualization, visual analytics, visualization, user studies, interaction

CCS Concepts: • Human-centred computing → Information visualization; Empirical studies in visualization

1. Introduction

Parallel Coordinates Plot (PCP) [Ins85] is a popular technique for multidimensional data visualization that has successfully been applied to a broad range of fields [Ins09]. The number of extensions during the last decades has increased at a large pace [HW13]. Approaches that aim to increase the understanding of density [HW09], reduce visual clutter [ZYQ*08] or improve the understanding via illustrative rendering [MM08] are among the extensions that can be found in literature. Despite the large variety of extensions, these approaches have mainly focused on the exploration of a single data type, mostly numerical, where different data types are regarded as an afterthought. Recently, Heinrich and Weiskopf [HW13] expressed that one of the main challenges for future PCP-based approaches is how to approach the mapping of categorical data to a metric scale in order to be visualized in the existing approaches. Multiple correspondence analysis (MCA) [RRB*04] is one approach to define a mapping of categorical attributes to PCP axes by generating an order and a spacing, i.e. distances between the locations of the categorical values.

If we focus on categorical (including ordinal) data, its application using the parallel axes metaphor has been mainly relegated to *Parallel Sets* (PS) [KBH06]. When considering numerical data in the context of PS, one may generate a discrete representation of the continuous numerical attribute through the application of binning. However, the discrete representation of continuous values may create interpretation limitations. Kosara *et al*. [KBH06] already stated in their original PS paper that 'showing continuous axes as true parallel coordinate dimensions would of course be the most useful display of this data'. Up to now, there is not yet a clear strategy to deal with the discretization of numerical data for use in PS.

Heterogeneous (or mixed) datasets, where both numerical and categorical dimensions coexist, are ubiquitous in a large number of fields. Coordinated views, where the numerical and categorical components are juxtaposed, may facilitate a joint analysis.

However, it may also present to the user a higher cognitive load when defining queries, where the context switching (i.e. simultaneously exploring numerical and categorical dimensions) is required.

In this paper, we evaluate the interplay of categorical and numerical dimensions for querying probability-based events for mixed datasets. Detection of high- or low-probability events fall into the realm of *value retrieval* and *pattern recognition*. Probability events are ubiquitous in different areas in research such as epidemiology or social science, just to name a few. In Epidemiology [AP14], they describe the aetiologic relationship for the development of diseases. The data collected to represent such events are heterogeneous by nature, e.g. recording the severity of a disease and levels of cholesterol. In social sciences, they [vBGO11] allow for the analysis of survey data without a complete case scenario and whether such values are feasible. PCP and extensions thereof allow for the analysis of such multidimensional data. Understanding the effect that the interplay of data types has is therefore paramount for being able to use these approaches to their full extent. A total of five representations are applied to nine tasks representing differing levels of objectivity.

Besides the already mentioned approaches (PCP, MCA-based PCP, PS, and coordinated views of PCP and PS), we also propose to use heterogeneous PCP (HPCP) as a fifth (novel) representation for our evaluation. In HPCP we display axes of numerical attributes in PCP mode and axes of categorical attributes in PS mode, where we propose a novel interface of adjacent PCP and PS axes.

Our main contributions can be summarized as:

1. First multi-task study for evaluating heterogeneity in interactive parallel-axes approaches.
2. Guidelines for usage of parallel-axes approaches with mixed data.
3. Extension of PCPs for mixed data.

2. Related Work

PCP has had a large number of extensions since its introduction, but few of those approaches have tackled mixing categorical and numerical data in a single view in their analysis. Traditionally, categorical attributes when used within PCPs are mapped to equidistant points without regard for their order [JF16]. One of the first attempts to include categorical data was proposed by Rosario *et al*. [RRB*04], where MCA, a technique akin to PCA devised for categorical data, is used to generate an order and spacing within categorical axes. The numerical attributes were binned and also treated as categorical data causing the inherent order and spacing within the numerical attributes to be lost.

Kosara *et al*. [KBH06, BKH05] described PS, where a frequency-based representation is used to represent categories in the dimensional axes. The frequencies are mapped to blocks representing the percentage of elements with a categorical value within a categorical dimension. Numerical data are interactively binned by selecting value ranges to define the PS blocks.

Few approaches have attempted to mix categorical and numerical data and preserve the continuous behaviour of numerical data. Johannson *et al*. [JJJ08, Joh09], followed up on Rosario *et al*.'s work [RRB*04] by creating an interactive tool for quantification of categorical variables in mixed datasets based on MCA computation. In contrast to Rosario *et al*.'s approaches, it allows for an interactive definition of the categories to define a categorization. The default configuration (without interactions) was defined by a *k*-means approach. Binned numerical attributes used for the MCA computation were kept as numerical attributes for the visual analysis. Despite the improvements made, the approach makes it difficult to analyse how samples behave within a category, as all samples with the same categorical value are mapped to a single point. A more recent approach, named Parallel Bubbles [TEL16], follows the traditional approach by equidistantly placing the categories in the parallel axis. However, at the mapping locations, circles with an area corresponding to the relative frequency is shown. The area representation may have difficulties expressing the true frequency [Mac63]. Thus, only modest gains were achieved when compared to the traditional PCP and no advantage was obtained against PS [TEL18]. The single-task evaluation was focused on a fixed visual representation, where frequency was estimated, i.e. no interaction was allowed. Such a constraint limits the ability to generate insight from parallel axes-based approaches.

Johansson and Forsell recently presented a survey of evaluations of parallel coordinates [JF16]. Focusing on applicability, three of the four presented evaluations [TPM05, AR11, BHGK14] made no explicit reference to their handling of categorical dimensions. The categories were mapped to equidistant locations on the parallel axes. The fourth evaluation was comprised of numerical dimensions only [CMR05]. Evaluations, where parallel axes-based approaches were compared against other approaches, used a similar data handling for categorical dimensions [LMP05, PVF05, Sii03]. To the authors' knowledge, the presented paper is the first evaluation that explicitly explores the interplay of categorical and numerical dimensions in mixed datasets for approaches based on parallel axes that goes beyond frequency-estimation tasks in realistic scenarios based on demographic data.

3. Study

The main goal of the study is to explore the understanding of conditional probability of events on mixed datasets with parallel axes-based approaches. The exploration of conditional probability lies within the task of value retrieval [KARC15] and object comparison [SG17]: If a sample or a set of samples has a certain property or a set of properties, which event is more likely to occur? A total of nine tasks representing events with conditional probability were generated. In probability theory, an event is an outcome of an experiment that has a probability. Conditional probability is the probability of an event occurring given that another event has happened [Jay03]. All interactive filtering on parallel-axes approaches can then be defined as a conditional probability-based query.

We control our stimuli with three main variables: (i) objectivity, (ii) mapping type and (iii) visual representation. Table 1 summarizes the structures of the tasks.

Objectivity: We refer to objectivity as whether the task the user has to perform has an objective directive for actions **Q** or whether

Table 1: *Task definition by variable representation: **A** represents a subjective query, while **Q** represents an objective query. The subscripts represent the data type:* C *for categorical dimension and* N *for numerical dimension.*

Tasks	Query	Objectivity
T1	$P(X = x_i \mid \mathbf{Q}_N)$	Objective
T2	$P(X = x_i \mid \mathbf{Q}_C)$	
T3	$P(X = x_i \mid \mathbf{Q}_C, \mathbf{Q}_N)$	
T4	$P(X = x_i \mid \mathbf{Q}_C, \mathbf{A}_N)$	Mixed—subjective and objective
T5	$P(X = x_i \mid \mathbf{Q}_C, \mathbf{A}_N)$	
T6	$P(X = x_i \mid \mathbf{Q}_C, \mathbf{A}_N)$	
T7	$P(X = x_i \mid \mathbf{Q}_C, \mathbf{A}_N)$	
T8	$P(X = x_i \mid \mathbf{A}_N, \mathbf{A}_C, \mathbf{A}_C)$	Subjective queries
T9	$P(X = x_i \mid \mathbf{A}_N, \mathbf{A}_C)$	

questions are referred to in subjective terms **A**. In objective queries, the task is phrased by defining specific values, e.g. 's elect all samples that have a value below 50'. In subjective queries, instead, the task is phrased with no pre-defined values, e.g. using descriptions such as 'high' or 'low'. Three levels of objectivity were tested: (i) objective, (ii) mixed objectivity and (iii) subjective.

Mapping types: We define three types of mappings, where data are pre-processed into a different data type leading to: (i) continuous mapping, i.e. treatment as a numerical dimension, (ii) discrete mapping, i.e. treatment as a categorical dimension or (iii) hybrid, i.e. data conserve their own initial properties.

Visual representations: A total of five visual representations based on the previous types are explored. PCP, *MCA-based Parallel Coordinates Plot* (MCA), PS, *Coordinated Parallel Coordinates and Sets* (Coor-PCP), and HPCP. A detail description of the representations follows in Section 3.1. An example of each individual representation can be found in Figure 7.

Tasks: The tasks are subdivided into the three objectivity types. There are several considerations in our task generation: An initial assumption is that subjective queries create a higher cognitive load than objective queries. The tasks were developed such that during the study, a relative increase in difficulty, a difficulty plateau and a relative decrease in difficulty is achieved. For objective queries, three tasks were created (one for numerical, one for categorical and one for mixed data). The tasks get progressively 'harder', as familiarity with the approach is gained through previous tasks. The first three tasks, where higher cognitive thought was not necessary, developed familiarity with the approach. Given the now assumed familiarity, four tasks with a similar format were chosen. Given our constraints defined in our study design for cognitive load and time, we limited ourselves to four questions and reduced the number of tasks by half for the last part of our study.

3.1. Visual representations

Given a dataset with multidimensional samples $X = \{\mathbf{x}_i\}$, where each sample $\mathbf{x}_i$ contains a number of mixed numerical and categorical attributes. Then, we use the following parallel axes-based visual representations.

Table 2: *Co-occurrence table of a synthetic dataset. Points in B are evenly distributed while A has an uneven distribution.*

	A_1	A_2	
B_1	50	50	100
B_2	75	25	100
	125	75	200

Figure 1: *Parallel Sets (PS) for synthetic dataset described in Table 2. PS take advantage of the discrete nature of the categorical attributes, where co-occurrence between categorical attributes may be explored. The uneven distribution in attribute A and the 1:1 ratio in attribute B are easily distinguishable.*

PCP: Traditional PCPs are defined for multidimensional numerical datasets. PCPs are constructed by placing axes in parallel with respect to a 2D layout. Each axis represents an attribute in the multidimensional space, where the values of the attributes are mapped using the scaling

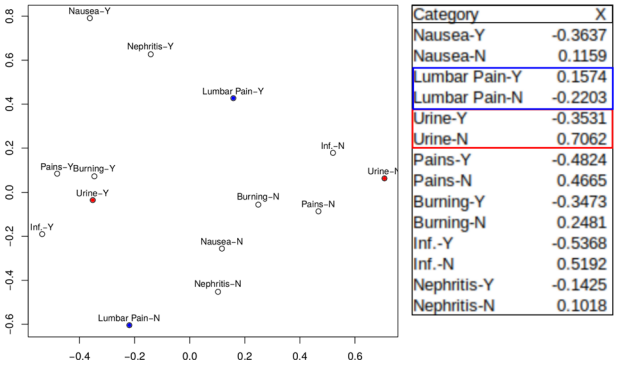
$$g_j(\mathbf{x}_i) = \frac{x_{ij} - min_j}{max_j - min_j} \tag{1}$$

where $[min_j, max_j]$ denotes the value range of attribute j. The mapped values are then connected in order to generate a polyline representing a sample. The interaction mechanisms allow for hovering over values to have them displayed, selection of elements in order to generate filters and swapping of axis. Selected elements are highlighted. Figure 4 displays a typical PCP view for numerical dimensions when only considering the first two axes. In the case of categorical attributes, the values are mapped equidistantly to the axis based on the dataset metadata, i.e. according to order of appearance in the samples.

PS: PS are an extension of PCP to categorical data. PS share the layout with PCP, but the point intersections are replaced with sets of parallelograms that represent the categories. These parallelograms are scaled according to the frequencies of the corresponding categories. PS take advantage of the discrete nature of the categorical attributes, where co-occurrence between categorical attributes may be easily explored. Given the co-occurrence table of two categorical dimensions shown in Table 2, Figure 1 displays the corresponding PS representation. These distributions are observable by examining the relative sizes of the frequency blocks. Numerical dimensions are typically divided into blocks representing numerical ranges. The dimensional axes have therefore two mixed representations when dealing with mixed datasets: a frequency representation for categorical and a range representation for numerical data types. Interactions for categorical dimensions allow for switching the

Category	Nausea-Yes	Nausea-No	Lumbar Pain-Yes	Lumbar Pain-No	Urine-Pushing-Yes	Urine-Pushing-No
Nausea-Yes	29	0	29	0	19	10
Nausea-No	0	91	41	50	61	30
Lumbar Pain-Yes	29	41	70	0	40	30
Lumbar Pain-No	0	50	0	50	40	10
Urine-Pushing-Yes	19	61	40	40	80	0
Urine-Pushing-No	10	30	30	10	0	40
Pains-Yes	29	30	29	30	49	10
...						
...						

Figure 2: *Burt Table of the Acute Inflammation dataset.*



Category	X
Nausea-Y	-0.3637
Nausea-N	0.1159
Lumbar Pain-Y	0.1574
Lumbar Pain-N	-0.2203
Urine-Y	-0.3531
Urine-N	0.7062
Pains-Y	-0.4824
Pains-N	0.4665
Burning-Y	-0.3473
Burning-N	0.2481
Inf.-Y	-0.5368
Inf.-N	0.5192
Nephritis-Y	-0.1425
Nephritis-N	0.1018

Figure 3: *Multiple Correspondence Analysis (MCA) projection of the Acute Inflammation dataset.*

order of categories, selection of categories (cf. Figure 1), and hovering over the parallelograms to display the categories' name. In the case of numerical data, by default four bins are generated, each bin representing 25% of the data. The number and size of the numerical bins can be interactively modified. Swapping of axes is also available. Figure 4 displays a typical PS view for categorical dimensions when only considering the last two axes.

Coordinated Parallel Coordinates and Sets: As the name implies, the coordinated approach of Coor-PCP displays juxtaposed views from traditional PCP and PS. Each approach allows for its own interaction mechanisms as described above.

MCA-based PCP: Correspondence Analysis (CA) [Gre84] is a projection method for categorical attributes based on the same principle as PCA. Instead of the eigenvector analysis of the co-variance matrix, CA makes use of the co-occurrence matrix such as the one shown in Table 2. This approach is limited to pair-wise analysis. MCA [GB06] is its extension to a multidimensional setting where, instead of a co-occurrance matrix, a *Burt Table* is constructed. The Burt Table is built as a symmetric matrix of the product set between the categories. Figure 2 displays a subset of the matrix for the categorical attributes of the Acute Inflammation dataset [CZ03]. It should be noted that the size of the Burt Table is dependent on the total number of categories in the dataset. The interaction mechanisms are equal to the PCP interactions described above.

Figure 3 (left) displays the resulting two-dimensional projection using the first and second principal axis when MCA is performed on the table shown in Figure 2. Figure 3 (right) shows a list of the categories and the value of their first principal axis. The normalized values of the first principal axis are used as numeric representations of the row categories. For mixed data, PCP of the numerical attributes are displayed using Equation (1).

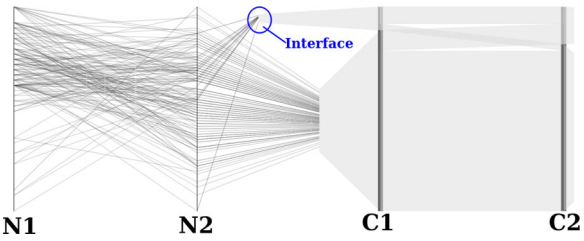


Figure 4: *Numerical–numerical pairs are displayed using traditional Parallel Coordinates Plot (PCP), categorical–categorical pairs are displayed using Parallel Sets (PS) and mixed pairs generate interfaces (such as the one encircled in blue) in the area between the axes.*

HPCP: HPCP can be seen as a natural extension to PCP and PS. The core concept is the handling of the data types according to their best-suited representation: a continuous representation for numerical dimensions and a frequency-based representation for categorical dimensions. Recall the pair-wise locality definition from the previous section: If the dimension pair (A', B') is formed by numerical dimensions, we may apply a traditional PCP approach (or variations thereof). For categorical pairs, PS may be applied. For mixed-type pairs, we propose the generation of a numerical–categorical interface as can be seen in Figure 4 when considering the middle two axes.

The area found between the categorical and numerical dimensions is used as an interface area: For each categorical value one interface is defined and placed in that area. In Figure 4, categorical attribute C1 has two categories, i.e. two interfaces are created that are placed between the axes N2 and C1. The polygon defined in PS between axes is now defined between the categorical axis and the interface, cf. blue circle in Figure 4. Similarly, the lines defined in PCP now go from the numerical axis to the interface. More precisely, the curve defined in the traditional PCP is defined from the numerical dimensions to the edge in the respective interface area using Equation (1). This allows the user to observe the distribution of values that lie within a single categorical value.

The location and the height of the interface leave some design options. In order to distinguish the interfaces, they can be horizontally separated. One suitable option is to place the interfaces equidistantly in the available area. Intuitively, the height of the interface shall be proportional to the height of its respective frequency block. The heights of the interface may be scaled with a proportion α. By selecting a value $\alpha \in (0, 1)$, the heights of the interfaces are diminished, which increases the readability of the plot by reducing the visual complexity due to overplotting. We empirically decided to select $\alpha = 1/3$ for the purpose of this study, which represents a decent trade-off between readability of the interfaces' heights and visual complexity. With our strategy for choosing location and height, we are able to spatially separate the interfaces both horizontally and vertically, which reduces possibly occurring occlusion and respective clutter. Further strategies for interface placement and selecting proportion α are discussed in Section 7 and are available within the code provided as supplementary material. However, an evaluation

of all possible choices of the interfaces' location and scaling is beyond the scope of this paper.

The interaction mechanisms are defined per dimension type. For numerical dimensions PCP, interactions (as described above) are available, while, for categorical dimensions, PS interactions (also described above) are available. Swapping of axes (no matter whether numerical or categorical) is also available. The described interaction mechanisms are selected due to their maturity in well-known software libraries for interactive parallel-axes visualizations [Zhu13, Sie20]. More sophisticated interaction mechanisms have been proposed [HLD02, RSM*16], which can be applied to the approaches presented here and may improve the analysis. However, the evaluation of such more sophisticated interaction concepts is beyond the scope of this paper.

3.2. HPCP design

As described above, the newly introduced HPCP concept elaborates on two main aspects: (i) the usage of an interface between numerical and categorical dimensions in order to (ii) maintain the data type per dimension when using heterogeneous datasets.

Figure 5 documents the benefit of introducing an interface between numerical and categorical dimensions (i), where equidistant interfaces are used for the Iris flower dataset [Fis36]. Creating the frequency blocks (middle and bottom), we can easily observe that the samples are equally distributed into the three categories, which is not so obvious when using PCP (top). However, when directly mapping from the numerical axis to the frequency block without introducing an interface, some occlusions are introduced (middle), which can be resolved by a smarter placement of the interfaces (bottom).

Figure 6 displays a synthetic dataset where the advantage of maintaining the data type (ii) can be observed. While the selection in PCP and HPCP is the same, only the HPCP approach allows for observing that the lowest category is underrepresented in the selection (when compared to the other two categories). Furthermore, by selecting a different placement of the interface based on the median, similar median values can be observed for the upper two categories, while the lower category has a much lower median value.

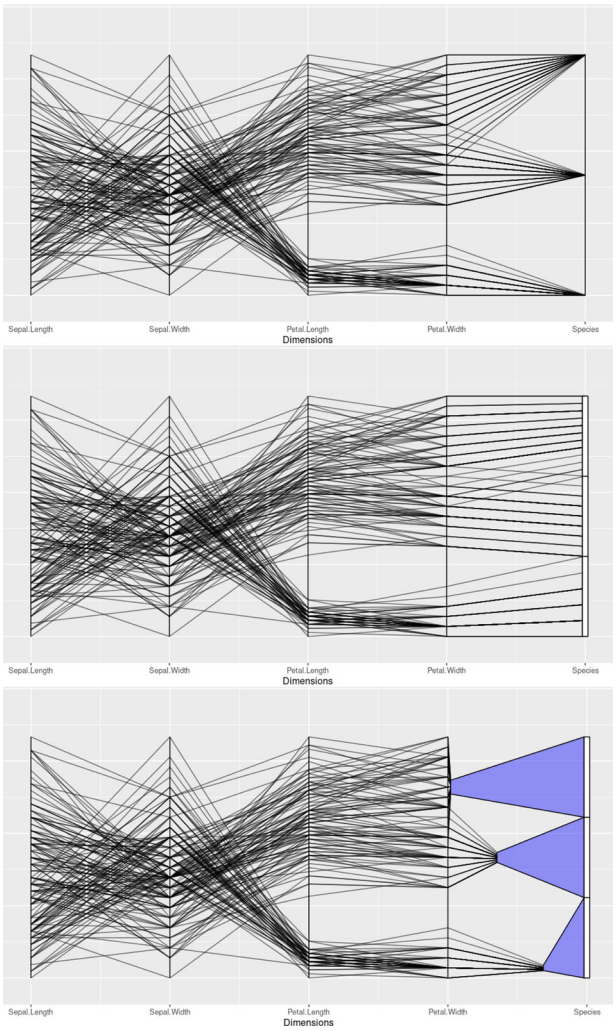


Figure 5: *Possible reduction of occlusion using equidistant interfaces (top) when applied to the Iris flower dataset [Fis36]. Mapping directly to the frequency block (middle) creates some occlusion, which can be resolved by a smarter placement of the interfaces (bottom).*

3.3. Hypotheses

Given the previously defined visual representations, some observations may be derived. PCP and MCA are the approaches with the minimum number of interaction mechanisms, i.e. hovering, filtering and swapping. Thus, in terms of possible interaction mechanism alone, it may provide lower response time. MCA, however, may define categories within a categorical dimension quite close to each other needing thus a finer level of selection, which penalizes interaction time. At the other end of the spectrum, PS has the largest amount of possible interaction mechanisms, rendering it the approach with highest interaction complexity. Coor-PCP and HPCP lie somewhere in between. However, coordinated views may increase the cognitive load for the user, which may increase the interaction time. These considerations would generate, in term of interaction time the following expected relationship: PCP $\succ$ MCA $\succ$ HPCP $\succeq$ Coor-PCP $\succ$ PS. Here $\succ$ denotes that the data representation on the left would result in better performance than that on the right, and $\succeq$ denotes an equal or better performance. Yet, approaches based on numerical approaches may experience difficulties when categorical attributes are present. Therefore, the interaction time for PCP and MCA may be hindered when dealing with categorical dimensions.

In terms of accuracy, PCP and MCA due to their numerical nature may be ill-suited for probability-based queries in categorical dimensions. Here, we define accuracy as the ability to answer the given questions correctly, i.e. to select the correct multiple choice answers. Approaches based on numerical mappings may provide, thus, the worst performance in terms of accuracy. For objective queries, PS may outperform all other approaches. The numerical ranges are concrete and, therefore, not subject to the users' interpretations. However, when the queried range is subjective, PS may find itself at a

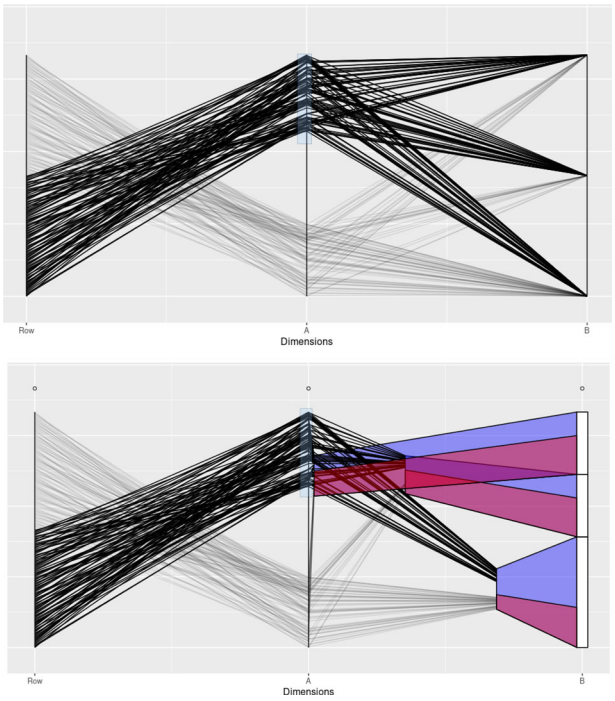


Figure 6: *Comparison of Parallel Coordinates Plot (PCP) (top) and Heterogeneous Parallel Coordinates Plot (HPCP) (bottom) when brushing on high values in the second (numerical) dimension. Only in HPCP, we can observe that the lowest category of the third (categorical) axis is underrepresented in the selection in comparison to the other two categories. Moreover, by selecting a placement of the interface based on the median, the upper two categories exhibit a similar value for their median, while the lower category exhibits a much lower median.*

disadvantage for dealing with numerical dimensions as the number and ranges of bins may affect the interpretability of the data. HPCP and Coor-PCP would, in theory, allow for more flexibility for user queries, and the ability to swap axes where numerical and categorical dimensions are placed side-by-side may provide a boost in accuracy to HPCP.

Therefore prior to the study, we formulated the following hypotheses that were then tested in the study.

For *Objective*, i.e. when well-defined concrete interactions are tasked for querying probabilistic events, we hypothesized that the following relationship in terms of accuracy between types of mappings exist:

(H1) Discrete Mapping $\succeq$ Hybrid $\succ$ Numerical Mapping.

Between visual representations, in terms of accuracy, we assumed:

(H2) HPCP $\succeq$ Coor-PCP $\succeq$ PS $\succ$ PCP $\succeq$ MCA.

In terms of completion time, we assumed the following relationship for mapping types:

(H3) Hybrid $\succ$ Numerical Mapping $\succeq$ Discrete Mapping.

Between visual representations, for completion times we assumed:

(H4) HPCP $\succeq$ Coor-PCP $\succ$ PCP $\succeq$ MCA $\succeq$ PS.

For *mixed objectivity* queries, in terms of accuracy between types of mappings, we assumed:

(H5) Discrete Mapping $\succeq$ Hybrid $\succ$ Numerical Mapping.

Between visual representations, in terms of accuracy we assumed:

(H6) HPCP $\succeq$ Coor-PCP $\succeq$ PS $\succ$ PCP $\succeq$ MCA.

In terms of completion time, for mapping types we assumed:

(H7) Hybrid $\succ$ Numerical Mapping $\succeq$ Discrete Mapping.

Between visual representations, for completion times we assumed:

(H8) HPCP $\succeq$ Coor-PCP $\succ$ PCP $\succeq$ MCA $\succeq$ PS.

For *subjective queries*, we hypothesized relationships similar to the ones above for accuracy in terms of mapping types **(H9)** and visual representation **(H10)** as well as for completion time for mapping types **(H11)** and visual representation **(H12)**.

3.4. Dataset construction

For the purpose of the study, three synthethic datasets were generated. The datasets were modelled after demographic studies of fake populations in science fiction and fantasy universes. Each dataset comprises of four numerical and four categorical dimensions. One dataset was used for tutorial purposes, whose description is found in the supplementary material. The two datasets used and the statistical properties of their dimensions is described below. A table dump with the synthetically generated data is also available in the supplementary material.

Firefly: A synthetic dataset where people in hospitals from four different planets were surveyed and their wage (N), age (N), weight (N), daily exercise minutes (N), surveyed planet (C), political association (C), job (C) and disease (C) were recorded. Value ranges of the attributes are provided in Table 3.

Hitchhiker's Guide to the Galaxy (HHGTTG): A synthetic dataset where the general state of being and eating habits from aliens in three different planets were surveyed and their age (N), time since last meal (N), running speed (N), distance to *Zaphod* (a character from the series) (N), surveyed planet (C), species (C), type of day (C) and whether the participant had drank a *Pan Galactic Gargle Blaster* (drink from the series) (C) were recorded. Value ranges for the data generation of the attributes are provided in Table 4. A single sample was selected from the Betelgeusian species to contain a distance to Zaphod of 0. An explicit reference to *Zaphod* was provided in the introduction of the dataset to the participants.

Table 3: *Description of the value ranges for dimensions in the Firefly dataset. The specific generation rules and dataset can be found in the supplementary material.*

Dimension	Range
Age	[20…70]
Weight	[43…107]
ExerciseMins	[0…60]
Wage	[−34…290]
Planet	*Ariel, Bellerophon, Regina,Miranda*
Association	*Alliance, Browncoat*
Job	*Miner, Physician, Banker, Sales*
Disease	*Pneumoconiosis, Silent Ischaemia, Anxiety Disorder*

Table 4: *Description of the value ranges for dimensions in the Hitchhiker's Guide to the Galaxy (HHGTTG) dataset. The specific generation rules can be found in the supplementary material. *A single sample was selected from the Betelgeusian species to contain a distance to Zaphod of 0.*

Dimension	Values
Age	[0…70]
Time since last meal	[0…10]
Running speed	[1.2…12.6]
Distance to *Zaphod*	[3…50]*
Planet	*Earth, Ursa Minor Beta, Stravromula Beta*
Species	*BogHog, Betelgeusian, Human*
Type of day	*Good, ok, bad*
Drank PGGB	*Yes, No*

3.5. Visual analysis tasks

Given the *Firefly* and *HHGTTG* synthethic datasets, a total of nine tasks (**T1**–**T9**) were developed given the structure defined in Table 1. The answers were given in a multiple choice format. No time limit was set per task. An 'U nable to answer' response was added to the set of possible answers per task, i.e. the participant was able to skip the task at hand.

The participants should select the answer with the highest conditional probability. A small introduction to each of the datasets was given. We include the possible answer set along with its probabilities given the queried condition per possible answer:

1. **T1**: $P(X = x_i|\mathbf{Q}_N)$: (Firefly) *'If you are over 50 in the Firefly universe, which disease are you most likely to develop?'*

   $\arg\max_{x \in S} P(Disease = x_i | age \geq 50)$

   where S = {Pneumoconiosis: 19.9%, Anxiety Disorder: 15.5%, Silent Ischaemia: 64.6%}
2. **T2**: $P(X = x_i|\mathbf{Q}_C)$: (Firefly) *'If you are alliance, which job(s) are you most likely to have?'.*

   $\arg\max_{x,y \in S, x \neq y} P(Job = x_i | assoc = alli.) + P(Job = y_i | assoc = alli.)$

   where S = {Banker–Physician: 68%. , Banker–Sales: 55%, Physician–Miner: 45%, Miner–Sales: 32%}
3. **T3**: $P(X = x_i|\mathbf{Q}_C, \mathbf{Q}_N)$: (Firefly) *'Which disease are you most likely to have, if you are from Regina and under 50 years old?'*

   $\arg\max_{x \in S} P(Disease = x_i | Planet = Regina, age \leq 50)$

   where S = {Pneumoconiosis: 65.3%, Anxiety Disorder: 21.3%, Silent Ischaemia: 13.4 %}
4. **T4**: $P(X = x_i|\mathbf{Q}_C, \mathbf{A}_N)$: (HHGTTG) *'If you are a teenager in Ursa minor beta, what kind of day are you most likely to be having?'*

   $\arg\max_{x \in S} P(Day = x_i | Planet = Ursa, 13 < age < 20)$

   where S = {Good: 66.6%, Ok: 0%, Bad: 33.3%}.
5. **T5**: $P(X = x_i|\mathbf{Q}_C, \mathbf{A}_N)$: (HHGTTG) *'After drinking a pan galactic gargle blaster, what ensures that your day is going to be terrible?'*:

   $\arg\max_{x \in A} P(Day = Bad | DrankPGGB = Yes, \mathbf{A})$

   where A could be defined as (i) being a teenager 52% (ii), being close to Zaphod (distance $\leq$ 15) 100%, (iii) being from Earth 60% (iv) and not eating anything recently (where time $\leq$ 2) 55%.
6. **T6**: $P(X = x_i|\mathbf{Q}_C, \mathbf{A}_N)$: (Firefly) *'Sedentary people from Miranda are most likely to develop what disease?'* where sedentary was defined as people having under 10 min of daily exercise.

   $\arg\max_{x \in S} P(Disease = x_i | Planet = Miranda, Exercise < 20)$

   where S = {Pneumoconiosis: 7.6%, Anxiety Disorder: 69.3 %, Silent Ischaemia: 23.1 %}
7. **T7**: $P(X = x_i|\mathbf{A}_N, \mathbf{Q}_C)$ (HHGTTG) *'We have a reason to believe that Zaphod Beeblebrox was questioned during our survey. Has the Betelgeusian president of the galaxy drank a Pan Galactic Gargle blaster the day of the survey?'* where a single sample was selected from Betelgeusian species to contain a distance to Zaphod of 0.

   $\arg\max_{x \in S} P(DrankPGGB = x_i | distance\ to\ Zaphod = 0)$

   where S = {Yes: 0%, No: 100%}
8. **T8**: $P(X = x_i|\mathbf{A}_N, \mathbf{A}_C, \mathbf{A}_C)$ (HHGTTG) *'Agrajag has been reincarnated, once more, as an alien version of a species that are as fast as Usain Bolt. He is having what could be considered the opposite day than the majority of his species. In which planet was he surveyed?'* BogHogs, whose base speed is higher than other species, have an 80% probability of having a bad day. All randomly generated BogHogs with a good day where placed in the same planet.

   $\arg\max_{x \in S} P(X = x_i | Speed > 7.5, Spec. = BogHog, Day = Good)$

   where S = {Earth: 100%, Ursa Minor Beta 0%, Stravromula Beta: 0%}
9. **T9**: $P(X = x_i|\mathbf{A}_N, \mathbf{A}_C)$: (Firefly) *'We have a reason to believe that within the surveyed people, a gang of traffickers were interviewed. This gang was foolish enough to give their honest earnings, conflicting with their "actual" jobs. In which planet are they hiding?'*. Twenty more samples were added to Regina, where a total of 15 miners from Regina were given a base salary

distributed in the same manner as Bankers and Physicians. High salary was defined as wages over 175. Probability of wages over that amount by profession: Sales 0%, Physician 41.1%, Banker 52.5%, Miner 6.4%. Probability of a miner having a wage <100 equaled 85% describing thus the conflicting job scenario.

$$\arg\max_{x \in S} P(X = x_i | Wage > 175, Job = Miner)$$

where S = {Regina: 100.0%, Miranda: 0 %, Bellerophon: 0 %, Ariel: 0%}

3.6. Study design

We performed a between-subject design such that each participant was only confronted with one visual representation. A between-subject design was selected instead of a within-subject design mainly due to the following reasons:

1. *Minimizing the learning effect and transfer across conditions*: The presented visualization approaches have a strong overlap in interaction mechanisms and visual design. In order to properly reduce the learning effect, the approaches need to be permuted in a within-subject design. The number of permutations would be larger than the application of a between-subject design.
2. *Reduction of cognitive load, time pressure and fatigue*: Previous studies [LMP05, Sii03], where parallel axes-based approaches were included, have shown that in a multi-task environment, the average completion time per approach is ≈ 30 min, which for five approaches leads to an estimated overall duration time of 2.5 h per participant. Long studies result in mental fatigue [VDLFS03], which may further confound the results. The recruitment of participants is also subject to their time availability, another compounding interaction for large studies [TVH05].

Sample size: The evaluation in literature that is closest to our approach compared *PS*, *PCP* and their proposed *Parallel Bubbles* [TEL18]. A large effect size (> 0.80) was found when comparing frequencies between the discrete representation *PS* and the numerical one *PCP*. The reported error rates were 0.2 ± 0.06 (*PS*, $N = 87$) and 0.54 ± 0.05 (*PCP*, $N = 78$), leading to an effect size using Cohen's index of $d = 6.15$. [Coh13]. However, their evaluation allowed no interaction, which may generally reduce the error rate on all tested visual representations. We assume instead a lower effect size of 1.0. Given the nature of our presented hypotheses, i.e. exploring $\succeq$ relationships, a one-tailed directionality is assumed. For a statistical power of 80%, a total of 11 participants per group is desired [BA12].

Subjects: A total of 75 participants were recruited via the usage mailing lists (e.g. an institute-wide e-mail), special interest groups and word of mouth. The study could be performed online via a provided URL. A total of 49 male (age: 27.8 ± 4.4) and 26 female (age: 28.6 ± 7.4) participants were randomly allocated to each of the visual representations. Having 75 participants for five groups, we managed to recruit 15 participants per group, which is more than the necessary 11 participants computed above. Table 5 summarizes the participant allocation to each of the visualization approaches by providing the number of participants per approach that expressed familiarity with multidimensional data analysis and/or visual analysis tools.

Table 5: *Distribution of participants per visualization representation. The number of participants per approach that expressed familiarity with multidimensional data analysis and/or visual analysis tools.*

Method	#Part.	Exp. Multi	Exp. Visual
Coor-PCP	15	8	5
PCP	15	10	7
PS	15	12	10
MCA	15	8	7
HPCP	15	8	6

Coor-PCP: Coordinated Parallel Coordinates Plot; HPCP: Heterogeneous Parallel Coordinates Plot; MCA: Multiple Correspondence Analysis; PCP: Parallel Coordinates Plot; PS: Parallel Sets

Protocol: Initially, a tutorial session was provided where the participants were familiarized with each visual representation and the interaction mechanisms explained. The order of the designated visual representations was selected at random. During the tutorial, each interaction mechanism was tested with the participants. Therefore, only when participants were capable of finishing successfully the complete tutorial, they were allowed to perform the main corpus of the study. A total of nine tasks detailed in Section 3.5 with different levels of objectivity were then tested per participant.

3.7. Statistical analysis

For all tasks, accuracy was determined, and completion time per task was recorded. Following the convention in cognitive science [KARC15], the response time is calculated based only on correct answers. For the timings, skewness and kurtosis were computed in order to test for normality: If the index ranges were $+/-2$, normality was assumed [TD01], a one-way ANOVA ($\alpha = 0.1$) test was applied and iff $p < 0.1$, *post hoc* tests using Tukey's HSD were computed to find pairwise differences. In the case of non-normality, the Kruskal–Wallis test was applied and the Dunn's test for pairwise differences. The p-values are reported accordingly.

In terms of accuracy, we apply the two-proportions z-test. In order to test for $A \succeq B$, we first applied a two-tailed test to observe differences between P_A and P_B. Iff a statistical significant difference is found, we apply the one-tailed test for the $A \succ B$ relationship.

We compute as well the effect sizes for timing and accuracy. Given its continuous nature, we compute Cohen's d index for the analysis of timings [Coh13]. For accuracy, given its binary nature, we compute odds ratio where exposure is defined by the left-side representation of the tested relationships [SF12].

4. Results

Each participant was allowed to interact with his/her allotted visual representation through all interactions inherent to it. Figure 7 displays filters and axis order per visual representation as selected by

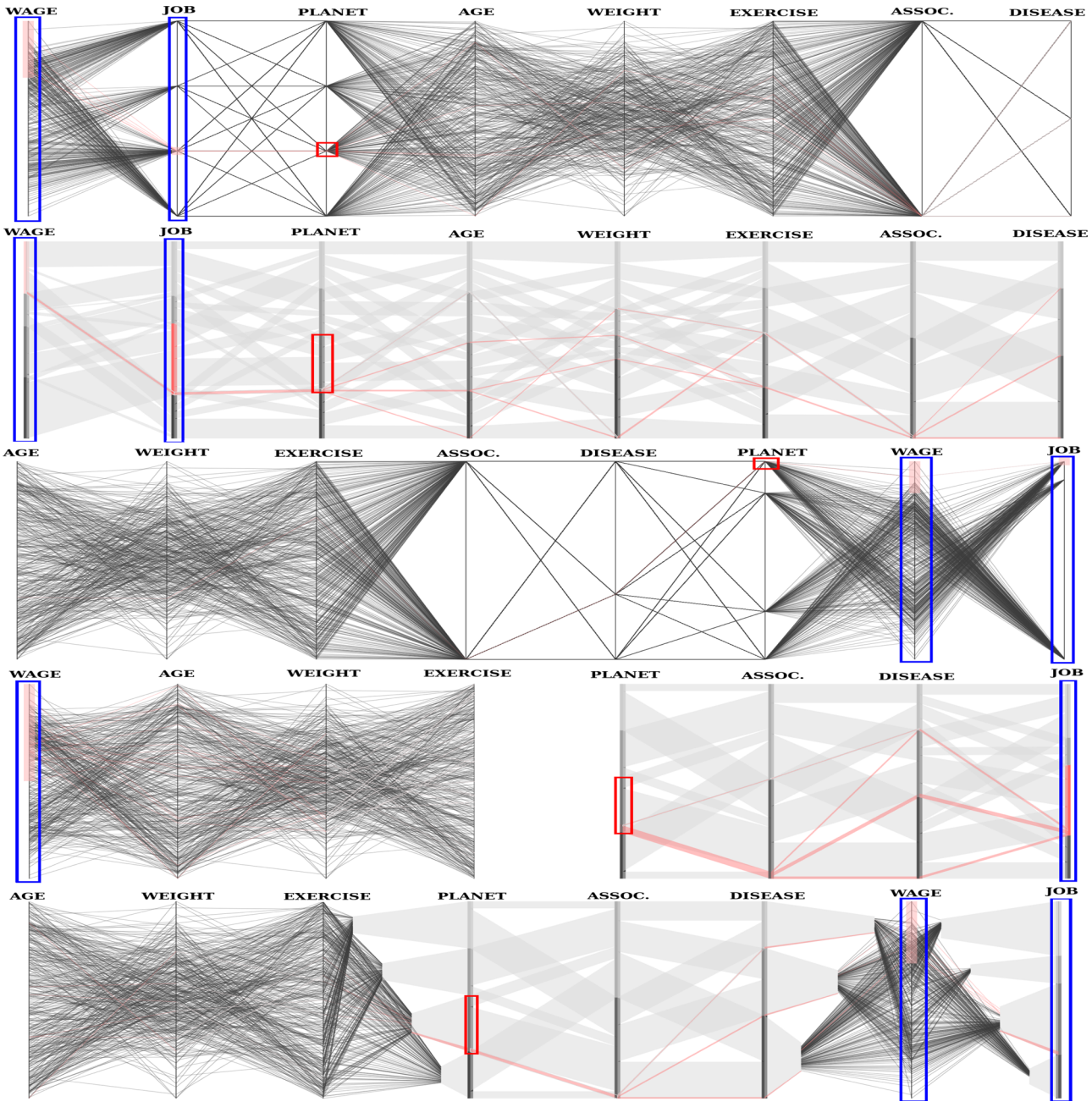


Figure 7: *Filters and axis order as selected by participants when solving Task 9. Visual representation (top to bottom): Parallel Coordinates Plot (PCP), Parallel Sets (PS), Multiple Correspondence Analysis (MCA), Coordinated Parallel Coordinates Plot (Coor-PCP) and Heterogeneous Parallel Coordinates Plot (HPCP). Wages (numerical axis with blue frame) and job as a Miner (categorical axis with blue frame) were selected. All highlighted samples in the planet dimension (red frame) are from Regina. 'High' wages selection range from 122 to 272 as lower interval boundary. A higher quality image for the individual approaches may be found in the supplementary material.*

participants when solving Task 9. Participants generated filters on the axes of the dimensions mentioned in the task, i.e. job and wages. However, when a subjective query was defined, e.g. 'high' wages here in Task 9, the queried values differ. High wages' were defined starting from 122 to 272 (average being $\mu = 197$) up to the maximum wage value. Such observations are typical for all tasks.

4.1. Accuracy

Figure 8 displays the accuracy per task for the tested visual representations. Table 6 summarizes the relationships between the visual representations and the mapping types, respectively, for the tested hypotheses. For p-values highlighting in green denotes that the

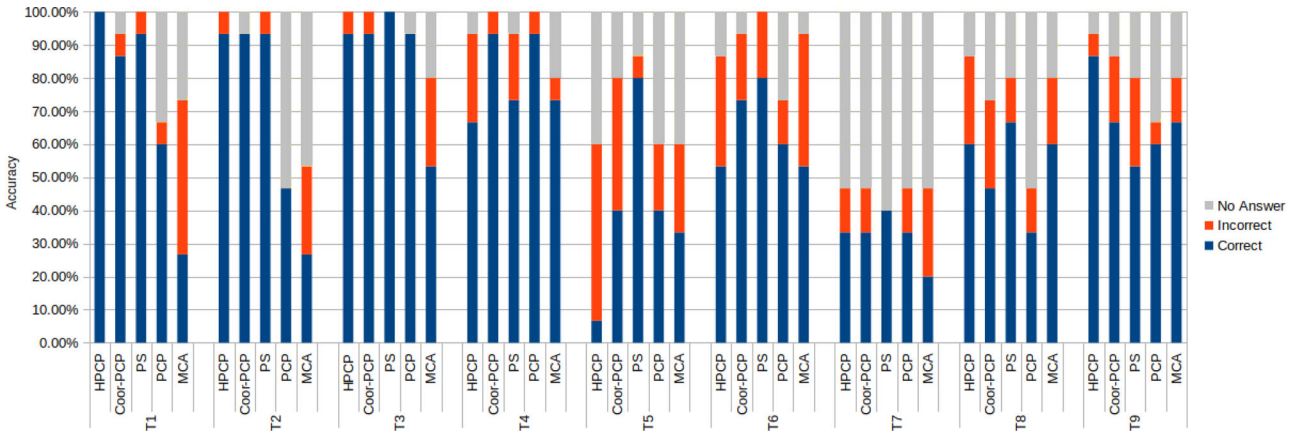


Figure 8: *Accuracy for all tasks per visual representation. Percentages of correct (blue), incorrect (red) and unable to answer (yellow) are displayed. We can observe that concrete directives in T1–T3 result in higher accuracy.*

Table 6: P-*values and effect size for hypotheses related to accuracy. For* p-*values, green denotes that the hypothesized relationship is maintained, while for effect size, green denotes a large effect size (≥3) using odds ratio as effect size index [SF12]. *A relationship Coor-PCP ≻ HPCP was observed.*

Objectivity	Hypotheses	Relationships			
Objective	**H1**	Discrete ≽ Hybrid	Hybrid ≻ Numerical		
	p-value	0.8,1.0,0.60.9075	0.8,1.0,0.63.7e−10		
	effect size	0.7321	0.8,1.0,0.6 9.8043		
Objective	**H2**	HPCP ≽ Coor-PCP	Coor-PCP ≽ PS	PS ≻ PCP	PCP ≽ MCA
	p-value	0.8,1.0,0.6 0.6726	0.8,1.0,0.6 1	0.8,1.0,0.60.009796	0.8,1.0,0.60.006119
	effect size	2.0975	1	0.8,1.0,0.65.1250	0.8,1.0,0.63.625
Mixed objectivity	**H5**	Discrete ≽ Hybrid	Hybrid ≻ Numerical		
	p-value	0.8,1.0,0.60.1249	0.5		
	effect size	1.7272	1.670		
Mixed objectivity	**H6**	HPCP ≽ Coor-PCP	Coor-PCP ≽ PS	PS ≻ PCP	PCP ≽ MCA
	p-value	0.0446/0.9777*	0.8,1.0,0.6 0.8511	0.2881	0.8,1.0,0.60.2733
	effect size	0.444	0.8681	1.3208	1.598
Subjective queries	**H9**	Discrete ≽ Hybrid	Hybrid ≻ Numerical		
	p-value	0.8,1.0,0.6 0.5905	1		
	effect size	1.4201	0.6538		
Subjective queries	**H10**	HPCP ≽ Coor-PCP	Coor-PCP ≽ PS	PS ≻ PCP	PCP ≽ MCA
	p-value	0.8,1.0,0.60.279	0.8,1.0,0.6 1	0.4216	0.8,1.0,0.6 0.7842
	effect size	2.1029	1	0.5604	1.3508

hypothesized relationship is maintained, e.g. in case of ≽ either ≻ or = is maintained, while for effect size, green denotes a large effect size (≥ 3) using odds ratio as effect size index [SF12]. We can observe that **H1** and **H2** were confirmed, e.g. all hypothesized relationships were maintained. A large effect size of 9.8 was observed between discrete and numerical mapping approaches for well-defined concrete interactions. If we restrict our analysis to PS and PCP, i.e. PS ≻ PCP, we can observe similarly to Tour *et al.* [TEL18], a large effect size between the visual representations. The ability to interact with the axes reduced the reported error rate for both approaches, i.e. a reduction of 0.2 to 0.08 for PS and 0.54 to 0.33 for PCP. This is indicative that interaction is capable to reduce error rate as described in Section 3.5. We were also able to observe a large effect size in the relationship PCP ≻ MCA. It may be indicative of the importance of the mapping of values onto the parallel axes. As query subjectivity increased, a lower effect size was observed between PCP and MCA. Medium effect sizes (> 1.5) were also observed for objective and subjective queries for HPCP and Coor-PCP, however the relationship reversed for queries with mixed objectivity.

4.2. Timing

Completion time was recorded per task per participant. Table 7 summarizes the relationships between visual representations and mapping types, respectively, for the tested hypotheses. For *p*-values highlighting in green denotes that the hypothesized relationship is maintained, while for effect size green denotes a large effect size (≥0.8) using Cohen's *d* index for effect sizes [SF12, Coh13].

Figure 9 displays the average completion time for all tasks. For Tasks T1–T3, we can observe that typically the hybrid approaches HPCP and Coor-PCP have a lower completion time for

Table 7: P-*values and effect size for hypotheses related to timing. For* p-*values green denotes that the hypothesized relationship is maintained, while for effect size, green denotes a large effect size (*$\geq$ 0.8*) using Cohen's* d *index for computing effect size [SF12, Coh13].*

Objectivity	Hypotheses	Relationships			
Objective	**H3**	Hybrid $\succ$ Numerical	Numerical $\succeq$ Discrete		
	p-value	0.8,1.0,0.6 7.43e−5	0.8,1.0,0.60.3338		
	effect size	0.6675	0.2486		
Objective	**H4**	HPCP $\succeq$ Coor-PCP	Coor-PCP $\succ$ PCP	PCP $\succeq$ MCA	MCA $\succeq$ PS
	p-value	0.8,1.0,0.6 0.9957	0.8,1.0,0.6 0.0065	0.8,1.0,0.60.9462114	0.8,1.0,0.60.9851
	effect size	0.0877	0.6236	0.339	0.0411
Mixed objectivity	**H7**	Hybrid $\succ$ Numerical	Numerical $\succeq$ Discrete		
	p-value	0.8,1.0,0.6 0.0011	0.8,1.0,0.6 0.0757167		
	effect size	0.5969	0.3992		
Mixed objectivity	**H8**	HPCP $\succeq$ Coor-PCP	Coor-PCP $\succ$ PCP	PCP $\succeq$ MCA	MCA $\succeq$ PS
	p-value	0.8,1.0,0.60.98317	0.8,1.0,0.60.0599	0.8,1.0,0.60.9947	0.8,1.0,0.6 0.2631
	effect size	0.1876	0.7094	0.0981	0.4628
Subjective queries	**H11**	Hybrid $\succ$ Numerical	Numerical $\succeq$ Discrete		
	p-value	0.8,1.0,0.6 0.0205	0.8,1.0,0.60.8557		
	effect size	0.5388	0.1271		
Subjective queries	**H12**	HPCP $\succeq$ Coor-PCP	Coor-PCP $\succ$ PCP	PCP $\succeq$ MCA	MCA $\succeq$ PS
	p-value	0.8,1.0,0.6 0.9919	0.1208	0.8,1.0,0.60.4598	0.8,1.0,0.60.6959
	effect size	0.2698	0.6872	0.4686	0.3978

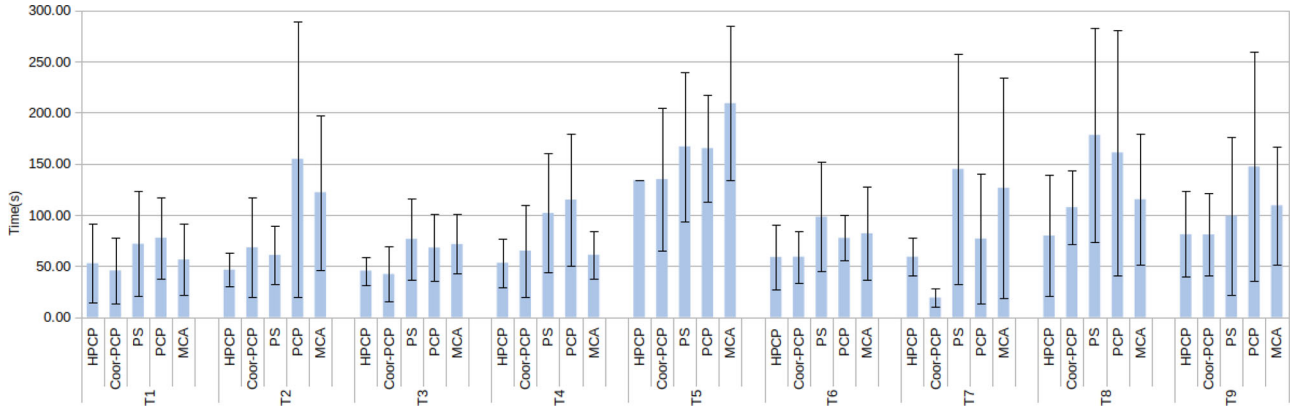


Figure 9: *Task completion time for Tasks T1–T9. We can observe that Heterogeneous Parallel Coordinates Plot (HPCP) and Coordinated Parallel Coordinates Plot (Coor-PCP) have a lower completion time for all tasks. Parallel Sets (PS), Multiple Correspondence Analysis (MCA) and Parallel Coordinates Plot (PCP) have a large completion time variation for Tasks T7–T9.*

these tasks when compared to discrete and numerical mapping approaches. For hybrid approaches, the average completion time was 51.70 s $\pm$ 32.82. Numerical approaches took an average of 87.46 $\pm$ 62.20 s, while the discrete approach took 73.34 $\pm$ 41.23 s. The hypothesized relationships were maintained, thus confirming **H3** and **H4**.

Focusing on mixed objectivity queries, i.e. T4–T7, we can observe that hybrid approaches have a lower completion time than numerical or discrete approaches, see Figure 9. For hybrid approaches, the average completion time was 69.88 s $\pm$ 51.63. Numerical approaches took an average of 105.50 $\pm$ 66.96 s, while the discrete approach took 133.34 $\pm$ 73.96 s. The hypothesized relationships were maintained, thus confirming **H7** and **H8**.

In the case of completely subjective queries, i.e. Tasks T8 and T9, we can observe that hybrid approaches have a lower completion time than numerical or discrete approaches. For hybrid approaches, the average completion time was 91.95 s $\pm$ 59.31. Numerical approaches took an average of 133.6 $\pm$ 92.5 s, while the discrete approach took 145.64$\pm$ 99.92 s. The mapping hypothesized relationships were maintained, thus confirming **H11**.

5. Discussion

If multiple events were to occur with different probabilities as in Tasks T1–T3, mapping categories to numerical locations had a negative effect. Hybrid and Discrete representations displayed higher performance when the conditional probability of the queried

event was ≠100%. The effect size between Hybrid and Numerical approaches is especially large for accuracy comparison with an odds ratio of 13.39. This shows evidence towards avoiding numerical representations for categorical attributes when the conditional probability of the event is not 100%. For individual visual representations, HPCP, Coor-PCP and PS showed an effect size of 10.75, 5.125 and 5.125, respectively, over traditional PCP. The effect size was even larger when comparing against MCA usage. The respective effect sizes were 38.96 (HPCP), 18.57 (Coor-PCP) and 18.57 (PS). However, the differences in terms of accuracy diminish when a low-probability event with a high conditional probability event is queried (T7–T9).

In the case of a discretization of continuous dimensions, no statistically significant difference in accuracy was observed. However, interacting with numerical bins adds complexity to the interactions. Hybrid and Numerical methods were shown to outperform in completion time for all tasks the discrete representation, even taking into account that PS had a larger expertise bias, i.e. a larger percentage of participants for PS had experience in multidimensional analysis and/or visual analysis.

The study focused on the exploration of probabilistic events, which are useful, e.g. for detecting outliers and value retrieval. However, PCP and extensions thereof are also useful for the exploration of correlations between numerical dimensions. Discrete representations such as PS are not suitable for these type of exploration tasks.

Swapping the location of the parallel axes may have a positive effect in the accuracy for queries. However, users not familiar with interactive visualizations may forget the range of possible interactions. We observed that for PCP 20% and MCA 26% of the participants did not apply axis-swapping operations for any of the tasks. For HPCP and PS, the percentage of non-swappers was 30% and 20%, respectively. A possible extension would be to show movement 'hint' to the novice user when hovering over the axes, which could be disabled at any time.

Although the expertise in visualization varied between the groups, few participants had more than a cursory knowledge of PCP or PS. Given the tutorial, each participant should have had the requisite knowledge to fulfil the tasks. However, having a previous encounter to multidimensional data analysis may have given an advantage to PS over all methods. Even so, no large differences between HPCP and PS in terms of accuracy were found.

Initial binning may have strong anchoring effects on the user. Indeed, in the present study for all tasks where queries with numerical abstraction, i.e. a subjective selection of values, were to be performed (T4–T9), eight participants did not modify the number of initial bins, three participants changed the number of bins in one task, two participants in two tasks and two participants in more than two tasks.

Numerical mapping of categorical dimensions was already expressed as a challenge for PCP extensions [HW13], yet special consideration needs to be taken when considering the application of a metric scale to categorical dimensions. It may hide the true frequency of the category or the mapping may result in overlapping events. The middle row of Figure 7 in its last axis displays the MCA mapping of jobs in the Firefly dataset. Only three out of the four jobs are easily observable, as physician and banker are mapped to relatively equal locations. Another example can be exemplified by T1 shown Figure 10 for PCP. Samples over the age of 50 could have any of the provided categories, albeit with different probabilities, e.g. Silent Ischaemia is 3× as likely to occur as Pneumoconiosis and 4× as Anxiety Disorder.

6. Guidelines

Given previous studies [LMP05, PVF05, Sii03, TEL16] and the present study, several observations and guidelines can be provided.

Data: First, in cases where data types are severely unbalanced, i.e. a large majority of the dimensions are either numerical or categorical, it is preferred to use the approach with respect to the large majority: PS or extensions thereof in the case of categorical or ordinal data and PCP or extensions thereof in the case of numerical data. If neither data type has a large majority, a coordinated or hybrid approach should be preferred. If the samples are approximately uniformly distributed for the categorical values in multiple dimensions, i.e. all categorical values have a similar probability, then a coordinated approach does not hinder the evaluation. However, where a non-balanced categorical behaviour occurs, i.e. skewed distributions, then a hybrid approach, where axes of attributes with different data types can be analysed side by side, may provide a better comprehension of the underlying sample distribution. The data guidelines are mainly based on previous studies [TEL16, Sii03, PVF05].

Tasks: Parallel-axes approaches are well-suited for tasks such as value retrieval, outlier detection and trend identification [UHHS96, FJ07, TEL16]. In terms of value retrieval and trend identification, the aforementioned guidelines with respect to the data types are applicable. The approach selection should be suited to the data to be explored. However, in terms of outlier detection in numerical attributes, PCP and extensions thereof may, generally, prove more suitable given that interaction with numerical axes allow for a more fine-grained selection of samples. PS, if the transformation is not defined correctly, might cause tail-end samples to be grouped despite large differences [TEL16]. For heterogeneous data types, this might not be the case as exemplified by tasks T7–T9, where the selected values were outliers in the categorical dimensions.

Interaction: In order to fully utilize the potential of interaction within parallel-axes approaches of mixed datasets, we recommend against mapping to a single data-type representation. When creating a numerical-to-categorical mapping, the strategies for interacting with discretized dimensions are not well-defined, as unintended effects may appear when using either frequency- or range-based representations (cf. Section 3.1). A range-based representation may mislead about the true number of samples within each block, while a frequency-based representation may cause confusion regarding the numerical range of the samples within each block. However, strategies where categorical values are mapped to numerical values, such as MCA, may be used in conjunction with discrete approaches, e.g. the initial placement of the categories within an axis may be based on the order provided by the MCA analysis. The interaction guidelines are mostly derived from our study results.

Users not familiar with interactive visualizations or already used to a different visualization representation may forget the range of

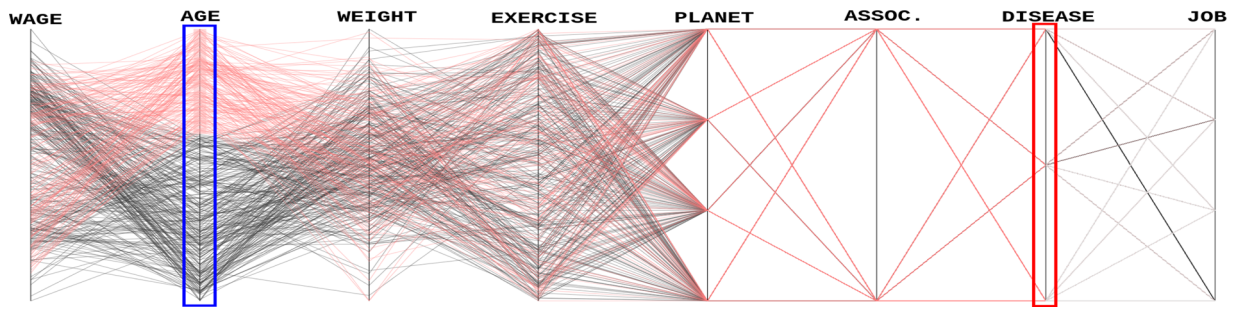


Figure 10: *Selection of T1 by participant unable to answer. It may be difficult to separate multiple events that may concurrently occur with differing probabilities when categories are mapped into as single point in the parallel axis. In blue frame, the selection of samples over the age of 50. In red frame, the disease axis where Silent Ischaemia is* 3× *as likely to occur than Pneumoconiosis and* 4× *as Anxiety Disorder.*

possible interactions. Reminders, in term of *visual cues* to actions, may lead to better performance when performing exploratory data analysis and avoid any strong anchoring effects on the user from the default view.

7. Conclusions and Future Work

To the best of our knowledge, this is the first multi-task study for evaluating heterogeneity in interactive parallel-axes approaches. The study focused on the differences between type of mappings for mixed datasets, i.e. for datasets with categorical and numerical dimensions, and their proposed visual representations with different levels of query objectivity for probabilistic events.

Coordinated views or HPCP approaches are capable of preserving the nature of diverse data types. We performed a first investigation to evaluate their gain. However, further investigation is needed for understanding the advantages and disadvantages of these visual representations. As described in Section 3.1, HPCP introduces a novel interface between numerical and categorical axes that allows for several possibilities for plot configuration. In terms of placements, the horizontal placement may be used to denote similarity within the categorical values, while vertical placement may denote statistical properties of the numerical samples such as mean or median, cf. Figure 6. The height of the interface may also be used in order to encode further statistical properties such as standard deviation or skewness of the samples that fall within a category. The proportion α for the interface may be used to equally scale these values and reduce possible occlusion. The actual selection of the interface depends on the exploratory task, e.g. the distribution of the samples within a category may be explored by selecting the median as placement location and the standard deviation as interface height. The horizontal placement of the interface can also be placed to describe similarity. Such investigations were beyond the scope of this paper and are subject to future work. We provide, however, an implementation of HPCP and the tested visual representations that can be used for further research.

Acknowledgement

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Grant 310876543 (LI 1530/23-1).

References

[AP14] Ahrens W., Pigeot I.: *Handbook of Epidemiology* (vol. *451*). Springer, 2014. New York, NY: Springer.

[AR11] Azhar S. B., Rissanen M. J.: Evaluation of parallel coordinates for interactive alarm filtering. In *Proceedings of the 2011 15th International Conference on Information Visualisation* (London, UK, 2011), IEEE, pp. 102–109. https://doi.org/10.1109/IV.2011.30.

[BA12] Burmeister E., Aitken L. M.: Sample size: How many is enough? *Australian Critical Care 25*, 4 (2012), 271–274.

[BHGK14] Beham M., Herzner W., Gröller M. E., Kehrer J.: Cupid: Cluster-based exploration of geometry generators with parallel coordinates and radial trees. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (2014), 1693–1702.

[BKH05] Bendix F., Kosara R., Hauser H.: Parallel sets: Visual analysis of categorical data. In *INFOVIS: Proceedings of the IEEE Symposium on Information Visualization*. (Minneapolis, MN, USA, 2005), pp. 133–140. https://doi.org/10.1109/INFVIS.2005.1532139.

[CMR05] Caat M., Maurits N., Roerdink J.: *Tiled parallel coordinates for the visualization of time-varying multichannel EEG data*. University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science, 2005. http://www.rug.nl/informatica/organisatie/overorganisatie/iwi Rights: University of Groningen. Research Institute for Mathematics and Computing Science (IWI).

[Coh13] Cohen J.: *Statistical Power Analysis for the Behavioral Sciences*. New York, USA: Academic Press, 2013.

[CZ03] Czerniak J., Zarzycki H.: Application of rough sets in the presumptive diagnosis of urinary system diseases. In *Artificial Intelligence and Security in Computing Systems*. Boston, MA: Springer, (2003), pp. 41–51.

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

[FJ07] Forsell C., Johansson J.: Task-based evaluation of multirelational 3D and standard 2D parallel coordinates. In *Visualization and Data Analysis 2007* (vol. *6495*). R. F. Erbacher, J. C. Roberts, M. T. Gröhnand K. Börner (Eds.). Washington, USA: International Society for Optics and Photonics, SPIE, pp. 111–122. https://doi.org/10.1117/12.697548.

[GB06] Greenacre M., Blasius J.: *Multiple Correspondence Analysis and Related Methods*. London, UK: Chapman and Hall/CRC, 2006.

[Gre84] Greenacre M. J.: *Correspondence Analysis*. Academic Press, London, 1984.

[HLD02] Hauser H., Ledermann F., Doleisch H.: Angular brushing of extended parallel coordinates. In *INFOVIS: Proceedings of the IEEE Symposium on Information Visualization* (Boston, MA, USA, 2002), IEEE, pp. 127–130. https://doi.org/10.1109/INFVIS.2002.1173157.

[HW09] Heinrich J., Weiskopf D.: Continuous parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 1531–1538.

[HW13] Heinrich J., Weiskopf D.: State of the art of parallel coordinates. In *Proceedings of the Eurographics (STARs)* (Girona, Spain, 2013), pp. 95–116.

[Ins85] Inselberg A.: The plane with parallel coordinates. *The Visual Computer 1*, 2 (1985), 69–91.

[Ins09] Inselberg A.: *Parallel Coordinates*. New York, NY: Springer, 2009.

[Jay03] Jaynes E. T.: *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press, 2003.

[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.

[JJJ08] Johansson S., Jern M., Johansson J.: Interactive quantification of categorical variables in mixed data sets. In *Proceedings of the 2008 12th International Conference Information Visualisation* (London, UK, 2008). IEEE, pp. 3–10. https://doi.org/10.1109/IV.2008.33.

[Joh09] Johansson S.: Visual exploration of categorical and mixed data sets. In *VAKD'09: Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration. Paris, France* (New York, NY, USA, 2009), Association for Computing Machinery, pP. 21–29. https://doi.org/10.1145/1562849.1562852.

[KARC15] Kanjanabose R., Abdul-Rahman A., Chen M.: A multi-task comparative study on scatter plots and parallel coordinates plots. *Computer Graphics Forum 34*, 3 (2015), 261–270. https://doi.org/10.1111/cgf.12638.

[KBH06] Kosara R., Bendix F., Hauser H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics 12*, 4 (2006), 558–568.

[LMP05] Lanzenberger M., Miksch S., Pohl M.: Exploring highly structured data: A comparative study of stardinates and parallel coordinates. In *IV'05: Proceedings of the Ninth International Conference on Information Visualisation* (London, UK, 2005), IEEE, pp. 312–320. https://doi.org/10.1109/IV.2005.49.

[Mac63] MacKay D. M.: Psychophysics of perceived intensity: A theoretical basis for Fechner's and Stevens' laws. *Science 139*, 3560 (1963), 1213–1216.

[MM08] McDonnell K. T., Mueller K.: Illustrative parallel coordinates. *Computer Graphics Forum 27*, 3 (2008), 1031–1038.

[PVF05] Pillat R. M., Valiati E. R. A., Freitas C. M. D. S.: Experimental study on evaluation of multidimensional information visualization techniques. In *CLIHC'05: Proceedings of the 2005 Latin American Conference on Human-Computer Interaction, Cuernavaca, Mexico* (New York, NY, USA, 2005), Association for Computing Machinery, pp. 20–30. https://doi.org/10.1145/1111360.1111363.

[RRB*04] Rosario G. E., Rundensteiner E. A., Brown D. C., Ward M. O., Huang S.: Mapping nominal values to numbers for effective visualization. *Information Visualization 3*, 2 (2004), 80–95.

[RSM*16] Rados S., Splechtna R., Matković K., Đuras M., Gröller E., Hauser H.: Towards quantitative visual analytics with structured brushing and linked statistics. *Computer Graphics Forum 35*, 3 (2016), 251–260.

[SF12] Sullivan G. M., Feinn R.: Using effect size–or why the p value is not enough. *Journal of Graduate Medical Education 4*, 3 (2012), 279–282.

[SG17] Sarikaya A., Gleicher M.: Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 402–412.

[Sie20] Sievert C.: *Interactive Web-based Data Visualization with R, Plotly, and Shiny*. London, UK: CRC Press, 2020.

[Sii03] Siirtola H.: Combining parallel coordinates with the reorderable matrix. In *Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization—CMV 2003* (London, UK, 2003), IEEE, pp. 63–74. https://doi.org/10.1109/CMV.2003.1215004.

[TD01] Trochim W. M., Donnelly J. P.: *Research Methods Knowledge Base* (vol. *2*). Atomic Dog Publishing, Cincinnati, OH, 2001.

[TEL16] Tuor R., Evéquoz F., Lalanne D.: Parallel bubbles: Categorical data visualization in parallel coordinates. In *IHM'16: Proceedings of the Actes de La 28ième Conference*

Francophone Sur l'Interaction Homme-Machine, Fribourg, Switzerland (New York, NY, USA, 2016), Association for Computing Machinery, pp. 299–306. https://doi.org/10.1145/3004107.3004142.

[TEL18] Tuor R., Evéquoz F., Lalanne D.: Parallel bubbles—evaluation of three techniques for representing mixed categorical and continuous data in parallel coordinates. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)* (Funchal, Madeira, Portugal, 2018), vol. 3, pp. 252–263.

[TPM05] Tory M., Potts S., Moller T.: A parallel coordinates style interface for exploratory volume visualization. *IEEE Transactions on Visualization and Computer Graphics 11*, 1 (2005), 71–80.

[TVH05] Topi H., Valacich J. S., Hoffer J. A.: The effects of task complexity and time availability limitations on human performance in database query tasks. *International Journal of Human-Computer Studies 62*, 3 (2005), 349–379.

[UHHS96] Unwin A., Hawkins G., Hofmann H., Siegl B.: Interactive graphics for data sets with missing values-manet. *Journal of Computational and Graphical Statistics 5*, 2 (1996), 113–122.

[vBGO11] van Buuren S., Groothuis-Oudshoorn K.: mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software 45*, 3 (2011), 1–67.

[VDLFS03] Van Der Linden D., Frese M., Sonnentag S.: The impact of mental fatigue on exploration in a complex computer task: Rigidity and loss of systematic strategies. *Human Factors 45*, 3 (2003), 483–494.

[Zhu13] Zhu N. Q.: *Data Visualization with D3.js Cookbook*. Birmingham, UK: Packt Publishing Ltd, 2013.

[ZYQ*08] Zhou H., Yuan X., Qu H., Cui W., Chen B.: Visual clustering in parallel coordinates. *Computer Graphics Forum 27*, 3 (2008), 1047–1054.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information