DOI: 10.1111/cgf.12385
Eurographics Conference on Visualization (EuroVis) 2014
H. Carr, P. Rheingans, and H. Schumann
(Guest Editors)

Volume 33 (2014), Number 3

Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets

Jürgen Bernard[1], Martin Steiger[1], Sven Widmer[2], Hendrik Lücke-Tieke[1], Thorsten May[1], Jörn Kohlhammer[1]

[1]Fraunhofer IGD, Germany
[2]Graduate School of Computational Engineering, TU Darmstadt, Germany

Abstract

The analysis of research data plays a key role in data-driven areas of science. Varieties of mixed research data sets exist and scientists aim to derive or validate hypotheses to find undiscovered knowledge. Many analysis techniques identify relations of an entire dataset only. This may level the characteristic behavior of different subgroups in the data. Like automatic subspace clustering, we aim at identifying interesting subgroups and attribute sets. We present a visual-interactive system that supports scientists to explore interesting relations between aggregated bins of multivariate attributes in mixed data sets. The abstraction of data to bins enables the application of statistical dependency tests as the measure of interestingness. An overview matrix view shows all attributes, ranked with respect to the interestingness of bins. Complementary, a node-link view reveals multivariate bin relations by positioning dependent bins close to each other. The system supports information drill-down based on both expert knowledge and algorithmic support. Finally, visual-interactive subset clustering assigns multivariate bin relations to groups. A list-based cluster result representation enables the scientist to communicate multivariate findings at a glance. We demonstrate the applicability of the system with two case studies from the earth observation domain and the prostate cancer research domain. In both cases, the system enabled us to identify the most interesting multivariate bin relations, to validate already published results, and, moreover, to discover unexpected relations.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques I.5.2 [Computer Graphics]: Design Methodology—Pattern analysis

1. Introduction

The value of research data and the (potential) benefit to society is widely accepted [CMZW13]. We classify research data as data that is gathered in scientific domains with the goal to gain or verify knowledge. Research data is a basis for data-driven science [HTT09]. It can be assumed that the data contains undiscovered knowledge, which makes it attractive for exploratory data analysis. In addition, various initiatives for open access of research data exist. Digital object identifiers (DOI) are increasingly attached to research data making the data citable. The developments contribute to an increased popularity of research data for collaborative research.

In many cases, the aim to derive and validate hypotheses shifts the analysis of research data towards an exploratory process [WR09]. For specific domain problems, visual analytics approaches are already successfully developed, e.g., in Biology [MWS*10]. However, state-of-the-art visual data representation techniques from the information visualization and visual analytics domains are still rarely applied in many data-centered research domains [TDN11]. In contrast, many scientists actually perform at least parts of their analyses using general purpose tools - most notably, Excel. This mismatch leaves the question whether visual analytics techniques can be used as a generic baseline technique. In this work, we want to support heterogeneous, multivariate data consisting of numerical, ordinal, and categorical attributes. Almost every research domain generates mixed research data sets. We aim to support scientists to validate hypotheses and provide exploratory means to generate new hypotheses. This raises the following problems to be solved:

*C*1 **Mixed Data Problem** The analysis of mixed data is generally considered difficult [JJJ08]. To make different attribute types comparable, unification strategies have to be applied. Yet the nature of different data types should still be reflected by the functional capability and the interactions of the analytical system.

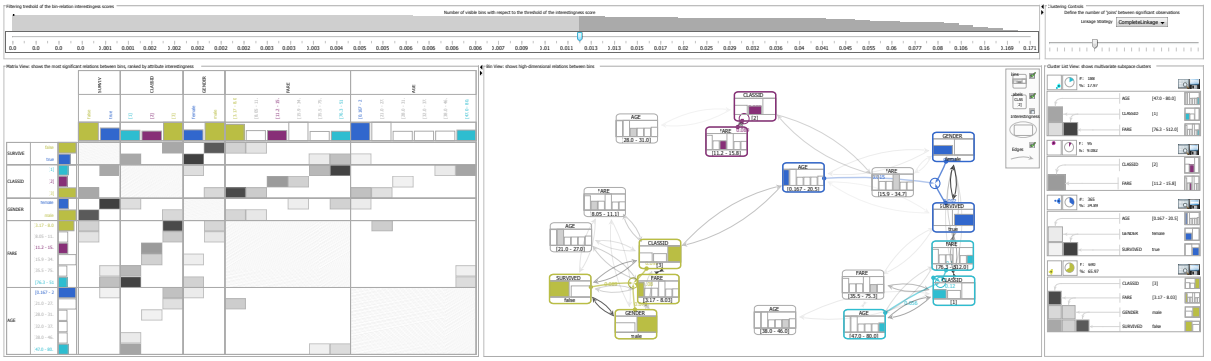

Figure 1: *The Titanic data set as a proof of concept. Left: Attribute View with ranked attributes and all bin relations above the interestingness filter. Center: the Bin View aligns related bins close to each other. Right: the Cluster List View provides compact representations of clustered subspaces. The blue cluster relates to the Birkenhead Drill: 'women and children first!'.*

*C*2 **Multiple Granularity Problem** Many analysis techniques search for relations defined for the entire dataset. This may mask the specific behavior of different subgroups (also called cohorts) in the data. We call this analysis granularity the *bin-level*, where observations are made for subsets of the data, defined by one or more attributes. We argue that an effective approach for research data should provide analytical capabilities to identify relations on the attribute-level (the entire data) and the bin-level.

*C*3 **Measuring Interestingness** Users need guidance that leads to interesting relations. An interesting relation might be something entirely new to the expert. Alternatively, it might be a relation which satisfies or disproves expectations of the expert. In any case, it is challenging to define interestingness in the context of the current domain knowledge. Furthermore, this definition must be adequate for mixed data. While many analytical approaches use similarity measures to identify interesting relations, we argue that the measure must be based on a generic dependency measure. Of course, specific measures like linear correlations allow for more specific conclusions. However, we want to ensure that more general patterns are not overlooked in the first place.

*C*4 **Foraging Multiple Hypotheses** Hunting for interesting relations in complex data sets may be time-consuming, especially if single relations have to be validated one by one in a 'batch-process'. In contrast, the system could just as well provide a broad range of available relations at startup as an automated service. In other words, due to the exploratory information seeking behaviors, systems should prefer recall over precision. This change in the analytical workflow would shift the *batch process* problem towards a *relation-space overview* problem. A careful choice of data abstractions is needed to provide a) an overview of the data set, b) an overview of available relations, and c) tight coupling of both. Thus, visual representations must be easy to translate to the domain knowledge and vice versa, while remaining generic enough to cover different types of findings.

*C*5 **Keeping Track of the Relation Space** Challenge *C*4 describes the breadth-first search to support the hypotheses generation process. Moreover, systems should also provide depth-first search to support efficient hypotheses validation. Scientists may themselves want to determine the number of shown relations. Focusing on subsets of the provided information to match specific domain knowledge or to prove expectations to the relations may be supported by drill-down capability. A solution to this requires scalability in terms of dimensionality and complexity. Again in order to solve this task, a careful choice of filters for both is needed to adapt the shown proportion of the data set, and the relation-space.

*C*6 **Structuring Subspaces** Subspaces may consist of arbitrary intersections of attributes (columns) and instances (rows). Both the internal structure of subspaces and the interrelation between different subspaces may be unknown. However, for an overview of the data set these subspace structures are crucial. We identify the need of clustering techniques that support scientists in the exploration of these structures.

We contribute a system that enables scientists to identify direct and indirect relations between multiple attributes in mixed data subsets. The system enables the analysis of relations on the attribute-level and on the bin-level. The interestingness of bin relations is based on statistical dependency measures. The system provides complementary linked views, showing bin relations in a matrix, a node-link, and a list metaphor. The number of shown bin relations is interactively steerable. Interactive drill-down based on expert knowledge and algorithmic support is provided in the matrix view where scientists can apply filters to focus on most interesting bins and attributes. The node-link view enables scientists to analyze multivariate bin relations. To this end, highlighting of multivariate bin relations is supported by interactive subspace clustering. The revealed high-level abstractions can be analyzed with respect to internal and external relations. To support different information seeking behaviors of scientists, three clustering techniques are provided.

We present a proof of concept (see Figure 1) based on the titanic data set and two case studies from earth observation and prostate cancer research to demonstrate the applicability of our techniques. In both case studies, we compare identified relations to the ground truth provided by domain experts and, moreover, contribute unexpected findings.

2. Related Work

In this section, we first outline related work in visual analytics on two prerequisites for our methods. Firstly, we give an overview of the related work on measuring the strength of relations. Secondly, we focus on categorical data only to present binning techniques. Afterwards, we present competing approaches for the related tasks of feature subset selection and subspace clustering.

Measuring Interestingness The interestingness of a relation depends on domain knowledge as well as on its type or strength. From statistics, a plethora of measures are known and we present some of them coarsely ordered from more specific to more generic: similarity measures may be applied to attributes of multivariate data to identify redundancies, to cluster or arrange attributes in a visualization (see for example Yang's *Value and Relation display* [YHW*07]). While many types of similarity measures exist, we consider similarity as one of the most specific forms of a relation. For linear relations in numerical data, one of the most well-known approaches is *Pearson's Correlation-Coefficient*, which, for example, has been applied in *DimStiller* [IMI*10], or in combinations with scatterplot and scatterplot matrices [CM84]. This approach is a specific form of regression analysis, where data is fit to a predefined model and the fitness translates to the strength of the relation in terms of this model. Statistical (in)dependency tests like Pearson's χ^2-test [RS81] or the *G-Test* [MS99] are based on entropy measures. For our technique, we modified these because they allow for a broad search for *all* types of relation as a baseline approach. Interestingly, statistical tests essentially compare something *assumed* about the data with something actually *measured* in the data. The parallel coordinates metaphor has been used for the analysis of mixed data, as presented in ParallelSets [KBH06], CComViz [ZKG09], and VisBricks [LSS*11]. These techniques support the identification of relations between clustering results and attributes. A visual analytics system for large scale categorical data was presented by Alsallakh et al. [AAMG12]. Based on the provided aggregation concept, their Contingency Wheel++ is capable for millions of data records as shown for movie ratings. Similar to our approach, the bin-level is chosen as the targeted analysis granularity and Pearson's χ^2 test was applied. In contrast, analysis is bound to a single target attribute. Johannson et al. [JJJ08] use MCA to quantify categorical data, while Friendly [Fri99] applies MCA to show multidimensional bin distributions arranged in a scatterplot.

Binning As we unify all data attributes on a categorical level, we have to consider different binning techniques. We distinguish between automated and interactive techniques. Interactive binning is basically identical to the selection of split-points in decision trees, for which Ankherst et al. [AEK00] present a comprehensive analysis of both automated and interactive approaches. Visual-interactive techniques were presented by Hao [HDS*10], as well as Novotny and Hauser [NH06]. Most recently, a visual-interactive regression analysis approach was presented where data aggregation is heavily applied for attribute visualization [MP13]. Clustering is a prominent aggregation technique for multivariate data, automated [Han05] and interactive approaches [SBTK09] have been presented. Visual-interactive aggregation techniques for hierarchies [EF10], time series [BRG*12], and text corpora [LKC*12] exist.

Feature Subset Selection Selecting appropriate candidates out of possibly large sets of features is mandatory to improve the predicting performance, the efficiency, and the transparency. A survey from the machine learning research was presented by Guyon et al. [GE03]. A prominent class of approaches transforms variables by linear or non-linear functions, e.g., dimension reduction. While most of the presented techniques are automated, interactive variants are iPCA [JZF*09] and the DimStiller framework [IMI*10]. However, our work is based on the selection, not on the transformation of existing features. Visual-interactive feature subset selection based on the Mutual information measure is SmartStripes [MBD*11]. Similar to this work, the approach is based on binned multivariate attributes. Guidance-based concepts for visual-interactive selection of interesting features were presented based on descriptor comparison [BvLBS11] and cluster labeling [BRS*12b].

Subspace Clustering Subspace clustering refers to data mining techniques closely related to our idea. Its challenge is the interdependency between the selection of attributes and the selection of subsets to identify clusters. Jain [Jai10] and Sim et al. [SGZC13] provide comprehensive overviews, mentioning subset clustering as a prominent challenge. From the visual analytics domain, subspace clusters have been tackled by Assent et al. [AKMS07] and Tatu et al. [TZB*12]. Like our approach, both approaches present an overview of the attribute and value space. However, these approaches perform clustering on numerical attributes, while our approach tackles the problem for categorical data.

3. Data Abstraction and Algorithmic Capabilities

In this section, we present data structures and algorithmic capabilities of the system, including data aggregation (Section 3.1), interestingness calculation (Section 3.2), and subspace clustering techniques (Section 3.3).

3.1. Creating Bins: Aggregation of Data

We define a *bin* as the atomic data object for the provided techniques. Every bin represents a set of value(s) of a given

attribute. For all attributes, the data is aggregated to bins to provide the following requirements:

- Unification of input variables ($C1$)
- Support for the analysis of bins and attributes ($C2$)
- Basis for interestingness calculation ($C3$)
- Scalability for large data sets ($C4$)

The binning process has great influence on the calculation of interesting relations. As a first step, the system estimates the data type of each loaded attribute. For Excel or WEKA files, we can fall back to the file header information. We provide different binning models depending on the attribute types, each of them able to handle missing values. *Binary* attributes are aggregated to two bins. For other attribute types, a parameter a defines the number of resulting bins produced by the models. If no expert knowledge is available and an automatic definition is necessary, we suggest to choose a value of 5-8 for a in order to avoid overlooking local substructures, but also to have bin populations high enough to guarantee statistical reliability. *Categorical* attributes are binned in a way that the most frequent $a-1$ values are assigned to individual bins, while the remaining categories are grouped together. The categorical model supports splitting (if possible), merging, and re-ordering functionality. For *numerical* attributes, we provide domain-preserving and frequency-preserving binning variants [MP13]. A method based on a goodness-of-fit measure [SS07] allows to estimate an optimal value for a before the frequency-based binning is applied. We set the frequency-based binning strategy as a default since it produces statistically reliable populations and caused less embarrassments in the discussions with the scientists. Interactively defining the number of bins as well as splitting and merging of bins is supported. We store binning results in a config file to accelerate future program starts.

3.2. Identification of Interesting Relations Between Bins

Our interestingness measure must satisfy the following properties: firstly, it can be applied to mixed data ($C1$) and secondly allows to consider bins separately ($C2$). Thirdly, it allows for the search for deviations between assumptions and measurements in the data ($C3$). We consider two methods to represent interestingness. The first is a standard dependency measure - *Mutual Information* [Gra11], which is applied to the bivariate distribution of bins. The second is developed on inquiry of the scientists, who where interested in calculating multiple significance scores at once.

Given binnings $(X_i)_i, (Y_j)_j$ of attributes X and Y we write p_i resp. p_j for relative frequencies of bins. The first measure reads as follows:

$$MI(X,Y) = \sum_i \sum_j p_{ij} \cdot log\left(\frac{p_{ij}}{p_i p_j}\right) \quad (1)$$

The values $p_{ij} = p(X_i \cap Y_j)$ are the observed relative frequencies of bivariate bins. If no prior knowledge exists, the measure models the deviation from the independency assumption $p_{ij} = p_i \cdot p_j$ for the attributes X and Y.

While the mutual information is an aggregation over the entire contingency table, we are also interested in deviations of single bin combinations. For this reason, we modify this test statistic to separate a single bin from the rest of the data.

$$Int_{XY}(i,j) = p_{ij} \cdot log\left(\frac{p_{ij}}{p_i p_j}\right) + (1-p_{ij}) \cdot log\left(\frac{1-p_{ij}}{1-p_i p_j}\right) \quad (2)$$

When derived from small samples (< 10 samples per bivariate bin), the values are ignored due to statistical unreliability. The measure *Int* is stored for each pair of attributes and bins in a matrix and additionally, in a graph-based data structure to simplify the retrieval of bin relations. Both data structures are direct sources for all views of the system ($C4$).

The second method is based on the observation that researchers sometimes do statistical testing in "batch-mode". Multivariate data offers multiple ways to create hypotheses, and foraging through hypotheses is difficult. The significance score (or *p-value*) is a common means to judge whether a hypothesis should be discarded. The p-value is not a measure; it does not allow to compare relations. However, scientists are typically more familiar with the p-value than with dependency scores. Hence, we choose to calculate this measure for all attributes and bins and use it as a filter. The statistic for the calculation of the p-value is *Pearson's Chi-Square-Statistic*, N denotes the size of the data:

$$\chi^2 = N \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(p_{ij} - p_i p_j)^2}{p_j p_j} \quad (3)$$

It has been shown that this statistic is proportional to the mutual information [Mor02]. If the independency assumption holds, this statistic follows a χ^2 distribution with $(m-1)(n-1)$ degrees of freedom. This resulting p-value is evaluated against the predefined score for the actual filtering. Like the mutual information, the p-value can be applied to the deviation of individual bins.

3.3. Clustering Bin Relations

We provide three different clustering techniques to support scientists in the identification of complex subspace structures ($C6$). These multivariate bin groupings are calculated on the basis of interesting bin relations above the filter status. The subspaces revealed by the clustering techniques may overlap since every bin can be assigned to multiple clusters. These bins may reveal indirect relationships between different subspaces (see e.g. Figure 4). We provide clustering techniques for both the detection of expected relations as well as the discovery of new relations, described as follows.

Bin Clustering provides information about interesting relations of single bins and is valuable for scientists to test hypotheses against a specific data subspace. For a targeted bin, the multivariate subspace is revealed based on interesting relations. Furthermore, multiple bins can be clustered at the same time. If two of the targeted bins have a direct relation, however, a single subspace structure is calculated.

Attribute Clustering supports scientists to focus on one or more target attributes. The algorithm identifies all subspaces in the data set which contain at least one bin of every

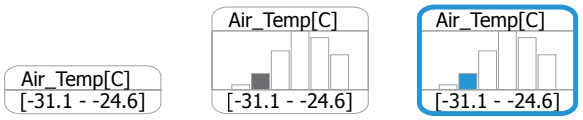

Figure 2: *Glyph layout of bins and attributes. Left, center: levels of detail. Right: cluster color indication.*

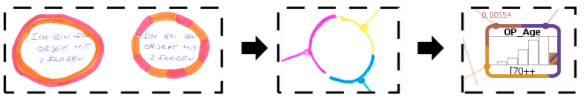

Figure 3: *Question in a formative user study:* 'I am a multi-colored object'. *The shown variants of the bin glyph reflect the design process of the multi-cluster color assignment.*

targeted attribute. Indirect relations between revealed subspaces can be identified by bins with multi-cluster assignment. As shown in the second case study, the selection of multiple attributes is an efficient means to reveal multivariate bin relations of a specific attribute focus (see Figure 9).

Exploratory Clustering emphasizes the most interesting bin relations and thus helps to gain an overview. Instead of focusing on a target variable, exploratory clustering takes all interesting bin relations into account (see Figure 1). Scientists are therefore able to quickly derive most apparent hypotheses at a glance. For this purpose, we use hierarchical, agglomerative clustering of bins with a user-defined aggregation level. Supported merge criteria include single-, median-, average-, and complete-linkage [Han05].

4. Visual Design

Our system is structured in three linked views (see Figures 1,8, and 9). The matrix on the left enables the analysis on the attribute-level (Section 4.2), the node-link diagram at the center focuses on the analysis of bins (Section 4.3), and on the right, we support visual cluster analysis (Section 4.4). A glyph design represents bins and attributes in all views (Section 4.1). All views are sensitive to an *Interestingness Filter* control at the top of the system, which enables to remove bin relations below an interestingness threshold. Thus, scientists can steer the number of shown bin relations (*C*5).

4.1. A Glyph for Attributes and Bins

We carried out a glyph design for the visual representation of bins and associated attributes (*C*2). The chosen barchart metaphor shows the distribution of a binned attribute (see Figure 2). When the glyph is applied to represent a single bin rather than an attribute at a glance, the respective bin is highlighted. The glyph supports the following interactions:

- Filtering single bins, e.g., if a bin is not interesting (*C*5).
- Merging bins. For categorical attributes and neighboring bins of numerical attributes.
- Splitting bins. For bins with more than one distinct value.
- Reordering of bins for categorical attributes. Due to expert knowledge or information gain (cf. [JJJ08]).
- (Multi-)selection of bins to trigger bin clustering (*C*6).

We use the glyph metaphor to link all views whenever attribute and bin information is visually encoded (*C*1, *C*4). The visual encodings and interactions of the glyph are optimized with respect to the results of a formative laboratory design study with 14 non-experts. Most of the participants were conducted several times within the development process to leave iterative feedback. In earlier phases, we identified problems of the participants to make use of the glyph, due to the visual complexity. Thus, the most decisive result of the study was to keep the visual encodings as simple as possible, with the option of interactive adaption. The barchart metaphor was chosen horizontally, accompanied by an attribute and a bin label. Different coloring concepts (one color per attribute, one color per bin) were withdrawn due to the cause of confusion. In the end, we use color explicitly for linking clusters (see Section 4.5). Different glyph outline techniques were compared. The final representation keeps the glyph compact and supports cluster color encoding. Concepts for indicating an additional interestingness score per bin were set to optional. According to the provided clustering techniques, multiple cluster assignments are possible at the same time. The design choice for assigning multiple colors to a single bin was also part of the user study (see Figure 3). At the left, two insufficient variants are shown. The concept of outlines with different sizes caused unintended ranking indication, while the second variant caused more distractions. Moreover, both variants lack color orientation. Finally, we implemented the third design conceptk, which resolves all described drawbacks.

4.2. Attribute View

The Attribute View enables the analysis of two-dimensional bin relations on both the attribute-level and the bin-level (*C*1, *C*2). The supported analysis tasks are as follows:

- Overview: getting an overview of attributes and bins (*C*4).
- Comparison: view and compare bin relations (*C*4).
- Drill-down: reducing the complexity by filtering (*C*5).

A matrix-based visualization represents pairwise bin relations at a glance (*C*4) (see Figure 6). The quadratic layout supports the visualization of asymmetric relations. The matrix is structured in major divisions corresponding to the attributes. Furthermore, minor divisions represent the bins of each attribute, using the glyph as described in Section 4.1, including all interactive capabilities. The ordering of attributes is provided by means of interestingness-based rankings. Depending on the analysis goals and the data characteristics, scientists can switch between different presets (means, median, and maximum scores). Cells of the matrix represent the interestingness value between two particular bins. All pairwise relations above the threshold of the interestingness filter are mapped to the alpha value (dark cells represent highest interestingness). The Attribute View enables to filter attributes as a whole (*C*5). Filtered attributes (and bins) are

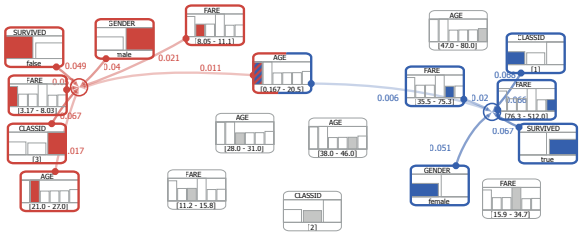

Figure 4: *Bin clustering of the titanic data set. The target bins 'Class=3' (red) and 'Survived=true' (blue) reveal separated clusters, except of the children's bin 'AGE [0-20.5]'.*

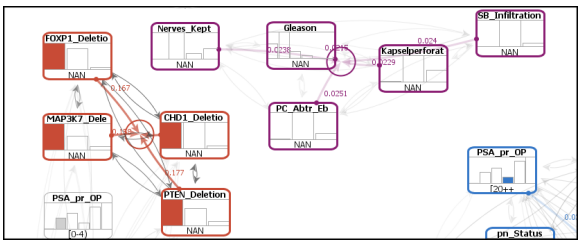

Figure 5: *Some multivariate bin relations are less interesting. We remove two subspaces with jointly occurring missing values of genetical (red) and clinical attributes (purple).*

visually represented in the corresponding filter panels (see Figure 6, bottom) and can be re-included on demand.

4.3. Bin View

The Bin View at the center of the system enables scientists to analyze multivariate bin relations (see e.g. Figure 4). Position information is used for each bin to identify related bins as neighbors in a topology-preserving layout, visually represented as a node-link diagram. Again, we use the glyph presented in Section 4.1 to visually represent single bins. As a result, the Bin View shows bin relations regardless of a definition on the bin-level or on the attribute-level. The presented topology of the bins reveals the structure of subspaces by means of an ordered, abstract visual representation (*C*6). Just as well, indirect bin relations can be uncovered, as shown in Figure 4. The Bin View complements the Matrix View, showing the same data through a different perspective and with respect to different analysis tasks [GFC04]:

- Overview: all available bins are aligned with respect to a generated topology (*C*4).
- Lookup and comparison: entities in different views can be re-identified and cluster structures can be compared (*C*4).
- Visual filter support: using the interestingness filter, bins without interesting relations can be removed (*C*5).
- Bin Clustering: (multi-)selection of bins reveals cluster structures based on bin clustering (*C*6).

By interpreting the interestingness measures as similarities between bins, we are able to use a standard projection technique (such as MDS) to obtain default coordinates for each bin in 2D. A layout based on repulsion forces reduces local overlap of neighboring bin glyphs, respectively. The Bin View provides an interactive legend for the individualization of the bin glyphs and the node-link structure on demand. In the title Figure 1, the visual representation of edges (interestingness of bin relations) is activated. Similar to the Attribute View, the Bin View enables the comparison of different entities (here: bins and bin clusters), linking entities interactively (by hovering), and the reduction of visual complexity by drill-down interaction (right-click for filtering bins). The Bin View is sensitive to the global interestingness filter. The selection of bins triggers the bin clustering algorithm, which is described in the subsequent section in detail.

4.4. Interactive Clustering of Multivariate Bin Relations

Up to this point, the described interactions focus on the *reduction* of the number of displayed attributes, bins and bin relations. However, we also enable scientists to additionally *highlight* subsets of bins using our clustering techniques (*C*6). Depending on the information seeking behavior, three different techniques (see Section 3.3) can be applied:

- by (multi-)selection of bins.
- by (multi-)selection of attributes.
- by operating the exploratory cluster slider in the upper right of the system.

No matter which of the clustering technique is applied, the cluster result is not limited to bivariate bin relations. Instead, the resulting data subspaces can be based on multivariate bin relations. We want to mind the fuzzy nature of the clustering techniques. A single bin may belong to multiple clustered subspaces. A compactified overview of each cluster is displayed in the list-based Cluster View (see Figure 7). In this way, scientists can further analyze the clustering results. Moreover, the view can be applied as a means of result communication, e.g., for a scientific publication. The visualization of a single cluster consists of four elements that describe the multivariate bin relation in detail, a) the label of the attribute, b) the label of the bin, c) the glyph of the bin, and d) a diagonal matrix to represent the internal cluster relations. The line-based linking between the bin labels and the matrix is inspired by the NodeTrix visualization [HFM07]. The bin glyph enables a quick allocation of bins within the domain of each attribute. Hence, scientists are able to read clustering results such as 'the first class relates to the oldest passengers, with expensive fares'. The absolute and the relative size of the subspace are shown. As a default, we calculate the size of a subspace by the join of pairwise bin cut sets. For a further analysis of the subset in detail, a tabular view can be opened. The respective icon is located at the upper right. Finally, to support the export of findings, subspaces can be stored in a separate file by clicking the disc icon.

4.5. Using Color for Linking Clusters

We use the visual variable *color* explicitly for linking cluster results in the three provided views. In this manner, we

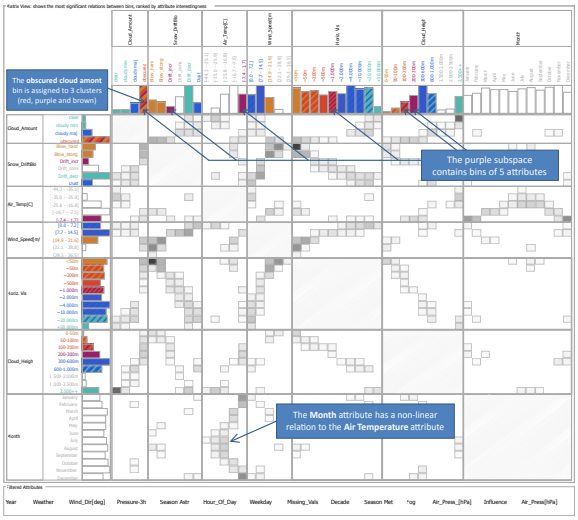

Figure 6: *Matrix View showing 7 weather-related attributes in Antarctica. Different non-linear correlations can be seen. Horizontal Width and Cloud Height are selected, revealed subspace clusters are highlighted with color-coding.*

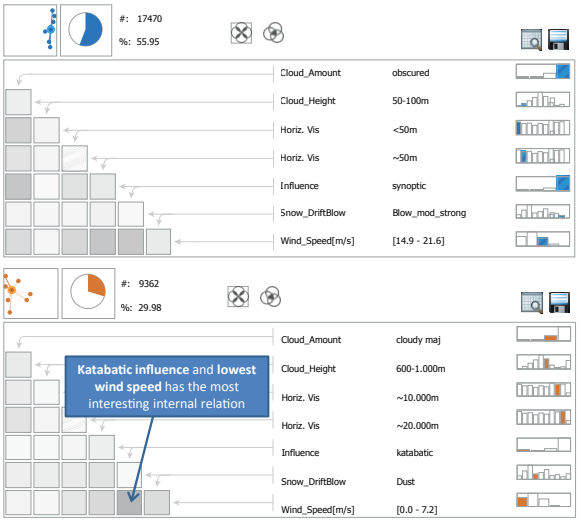

Figure 7: *Cluster List View. Matrices show internal cluster relations. Depending on the cluster mode, not every cell needs to have a relation above the interestingness value. Two weather phenomena in Antarctica are shown.*

combine the individual advantages of each view to facilitate reasoning about the clustered subspace structures (*C6*). The choice of colors needs to discriminate dissimilar clusters, and simultaneously to indicate similar (and possibly intersecting) clusters. For this purpose, we apply the results of the topology-preserving bin projection algorithm (see the Bin View 4.3) and map the 2D coordinates to a 2D colormap. With the chosen RGB 2D colormap, we aim to exploit large parts of the available color space [BRS*12a]. Nonetheless, adapting the applied color map based on user preference or semantic application context is easily possible [BvLBS11].

5. Case Study

We report the results of two real-world scenarios with research data from the meteorological and the medical domain. Scientists confirmed that we were able to detect important expected relations in both scenarios. Moreover, we were able to communicate interesting relations that were unexpected.

5.1. Meteorological Synoptical Observations

In this study, we explore multivariate weather phenomena in Antarctica. Since March 1981, a meteorological observatory program has been carried out at Neumayer Station (NM) (70°37‘S, 8°22‘W), located in Antarctica. NM is an integral part of many international networks, organized e.g., by the World Meteorological Organization (WMO). The data helps to close gaps in the global weather and climate observing networks. Our contacted domain expert is Dr. Gert König-Langlo, scientific leader of the meteorological observatory of Neumayer. The provided mixed data set consists of 26 attributes with measurements every three hours for 30 years (92902 time stamps) [RLKLI12]. We remove columns and lines of the data set with too many missing values. Next, we re-name some of the attributes and bins due to a domain-specific encoding. For the exploration of temporal relations, we integrate binnings for *Year*, *Month*, *Season*, *Weekday*, and *Hour of Day*. We apply frequency-preserving binning for numerical attributes as a default, and population-based binning for *Air Temp[C]* and *Wind Speed*. The mutual information measure is chosen as the interestingness measure.

The number of 150 bins at startup reveals the limits of the visual scalability of the system. Based on the attribute ranking in the Attribute View, we first remove less interesting attributes. Likewise, bins representing missing values are removed. We take advantage of the additional display space and sort the bins of categorical attributes (see *Horiz.Vis* and *Cloud Height* in Figure 6). For further information drill-down, we raise the interestingness filter to the 66% position. We identify a variety of correlations in the Attribute View. Some of these show periodic, non-linear behavior (see *Month* in Figure 6). Based on the gained overview, we next apply the exploratory clustering technique. The most interesting relations in the data set are presented. We obtain multivariate subspaces containing popular weather phenomena, verified by Mr. König-Langlo. For a further investigation, we select the observed attributes *Wind Speed*, *Snow Drift*, *Influence* (wind direction), *Horizontal Visibility*, *Cloud Height*, and *Cloud Amount* as the basis for attribute clustering. Figure 7 illustrates the two resulting subspace clusters in detail. The blue subspace consists of bins with low horizontal visibility, low cloud height, high wind speeds, obscured sky conditions, synoptic weather influence (wind coming from the east), and blowing snow drifts. The domain ex-

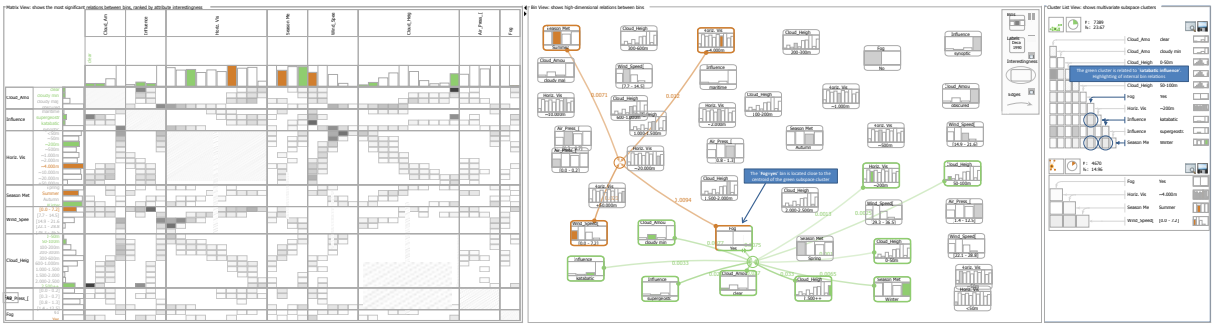

Figure 8: *Exploration of undiscovered relations. The bin Fog=Yes occurs in winter (green) and in summer (orange). While in the winter period katabatic winds are observed, foggy weather in summer is not significantly influenced by maritime winds.*

perts points out that we have just identified one of the most prominent weather situations at NM, a bad weather situation that usually occurs in the winter period. In contrary, the orange cluster contains low wind speeds, vast horizontal sights, and katabatic winds (coming from the south) indicate fine weather, which mainly occurs in summer. The scientist was pleased to recommend page 30 of one of his publications where our explored weather conditions are already published [KLL07]. Even though we are no experts, we were still able to discover unknown relations (from our point of view), and thus to independently validate previously published findings.

For the discovery of something unexpected, Mr. König-Langlo suggests to focus on foggy weather situations in summer and winter in combination with wind influences, which are not yet fully understood in the scientific community. We apply bin clustering and define *Fog=Yes*, *Season=Summer*, and *Season=Winter* as target bins. The results are shown in Figure 8. The *Fog=Yes* bin is connected to both a *Summer* cluster (orange) and a *Winter* cluster (green). It can be seen in the Bin View that the green cluster contains more bins and is more compact in contrast to the orange cluster. Just as well, the position of the *Fog=Yes* bin indicates a strong relation to the green cluster. The scientists explains that the 'green' fog behavior in winter may be influenced by the continental climate of the inner Antarctica, represented by katabatic winds, (so-called radiation fog). Fortunately, the Cluster List View at the right lists katabatic influence in the green cluster. An additionally calculated p-value analysis reveals a probability for a statistical dependency between *Influence=katabatic* and *Fog=Yes* of 99.9%. We can validate that foggy weather in winter is influenced by katabatic winds; a new discovery. However, in summer things look different. The scientist explains that in summer the open sea is very close to NM since most of the sea ice has melted. Thus, wet air coming from the near sea may support foggy weather, which the researchers call advection fog. However, the maritime influence appears to be weak. The bin is not contained in the orange cluster. From a scientific perspective, we suggest to withdraw the hypothesis of the maritime influence to foggy weather.

5.2. Prostate Cancer Research

In prostate cancer research, much effort is spent in the discovery of prognostic features, which predict biological cancer events or medical treatment success. Hence, research in the medical domain performs a lot of multivariate data analysis. The limited number of clinical treatment attributes makes the discovery of new findings difficult, unless e.g., genomic attributes are additionally involved. However, only few scientific labs are able to perform the analysis of both sources at large data scale. The Martiniklinik Hamburg Eppendorf (UKE) has one of the largest (anonymized) data bases with both: attributes of clinical treatment and genomic indicators. Our contacted domain experts Prof. Dr. Thorsten Schlomm and Dr. Pierre Tennstedt are interested in 'associations' of genomic deletions (indicators) to clinical attributes in order to improve patient treatment. On our obtained data set, we first apply prepreocessing steps to re-name attributes and clean categorical data. Moreover, we bin some numerical attributes with respect to definitions given by the scientific community. Based on inquiry, we choose frequency-preserving binning for numerical attributes and χ^2 with the inverted p-value statistics to measure interestingness. We begin with 13571 anonymized patients containing 41 mixed attributes. With 160 bins, we again reach the limits of visual scalability. The Attribute View helps to drill-down the number of attributes. We remove attributes with expected, and thus less interesting relations (from the doctor's point of view). An example is shown in Figure 5, where two clusters of missing values were identified and removed. Apparently, missing values of different attributes seem to occur frequently for identical patients. Next, we apply exploratory clustering in order to reveal the most interesting subspaces. We detect that pathology results of prostate cancer surgery highly indicate patient well-being in the post-surgery phase (see e.g. [WCK*09]). The *biological recurrence (BCR)* is one of the strongest indicators for patient well-being. Among others, we detect expected relations to the size of the tumor, the Gleason score, and the PSA-value. Finally, we aim to discover 'associations' between genomic deletions (mutations based on missing parts of a chromosome or a sequence of DNA) and clinical attributes. The domain experts point

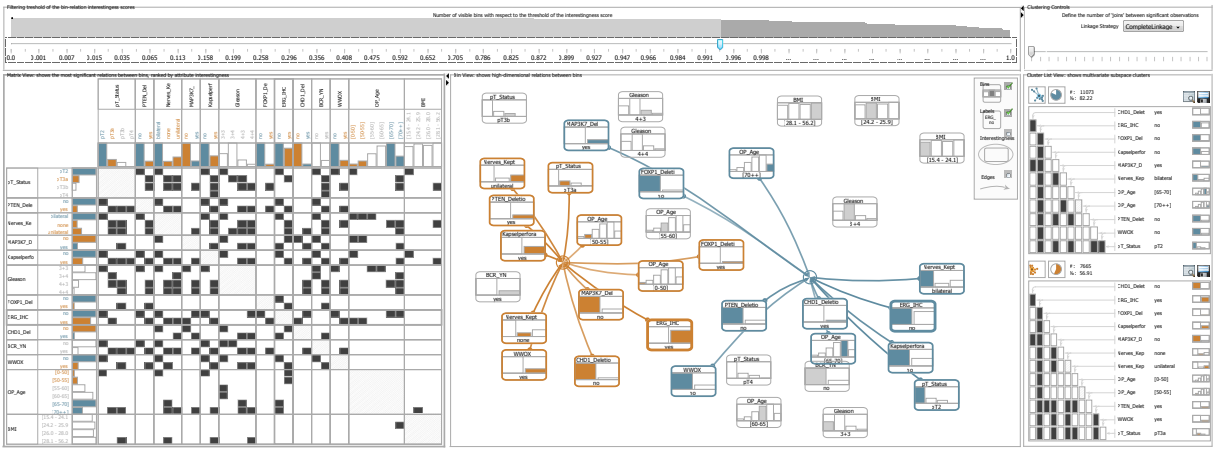

Figure 9: *Attribute clustering of the prostate cancer data set. We discovered an association of the genomic 'deletion' ERG_IHC attribute to bins of the OP_AGE attibute. The finding will be published in the prostate cancer research community soon.*

out that prostate cancer early-treatment can benefit substantially from new associations. One candidate is the *ERG_IHC* transcription factor, which we chose for attribute clustering. It can be seen in Figure 9 that the two expression bins *yes* (orange cluster) and *no* (blue cluster) have a relation to bins of the *OP_Age* attribute (the p-value for the associations is <0.001). This finding was completely new to the domain experts. A second discovery is a multivariate relation of *ERG_IHC* to other deletions (*PTEN*, *MAP3K7*, *FOXP1*, *CHD1*, and *WWOX* in particular). With this finding, we supported the discovery of 'additional effects' among deletions, which will be a vast research area in future. We are happy about the discovery of two findings, which both will be published in the prostate cancer research community soon.

6. Summary and Conclusion

We presented techniques for the identification of interesting bin relations in large mixed research data sets. Three complementary visual components provide different views onto the data, linked together in an interactive system. The interestingness of bin relations calculated with the Pearson's χ^2 test or the Mutual Information measure enables scientists to directly apply discovered findings as the basis for domain-specific research. From an analytical point of view, our system supports both, to derive and to validate hypotheses. Based on the three linked views our system enables scientists to get an overview of large mixed data sets. Moreover, interactive drill-down capability is provided by means of filtering and attribute individualization. Finally, different clustering techniques support scientists in revealing most interesting and potentially multivariate relations hidden in the data. The visual appearance of the system was also targeted towards supporting scientists in the communication of analysis results visually. Two case studies were conducted to prove the applicability of the system for research tasks in practice.

A possible limitation of the approach is the dependency on the pre-calculated binning results. Generally speaking, it would be beneficial to study the dependency of different binning results on the interestingness calculation results. Another issue is the genericity of the interestingness measures. Indeed, we enable the scientist to find relations, but the relations are unspecific and need the interpretation of experts, e.g., to deduce a linear correlation. The latter limitation is associated to a possible extension of the system: the interactive definition of estimators for the χ^2 test for assumptions about the data. Eventually, the visual scalability of the approach is limited to about 30 attributes, or 150 bins, respectively. Based on the findings in the use cases, we aim to prove to which extent the presented topology of bins in the Bin View can be generalized as a means for scientific reasoning. While proven meaningful in this approach, the application of 2D colormap concepts for depicting similarity can further be taken into account. Perceptional linearity is only one of the interesting objectives of a possible study.

References

[AAMG12] ALSALLAKH B., AIGNER W., MIKSCH S., GRÖLLER M. E.: Reinventing the Contingency Wheel: Scalable Visual Analytics of Large Categorical Data. *TVCG, IEEE 18*, 12 (2012), 2849–2858. 3

[AEK00] ANKERST M., ESTER M., KRIEGEL H.-P.: Towards an Effective Cooperation of the User and the Computer for Classification. In *ACM SIGKDD* (New York, NY, USA, 2000), KDD '00, ACM, pp. 179–188. 3

[AKMS07] ASSENT I., KRIEGER R., MÜLLER E., SEIDL T.: VISA: Visual Subspace Clustering Analysis. In *ACM SIGKDD Explorations Special Issue on Visual Analytics, Vol. 9, Issue 2* (New York,NY,USA, 2007), ACM, pp. 5–12. 3

[BRG*12] BERNARD J., RUPPERT T., GOROLL O., MAY T., KOHLHAMMER J.: Visual-Interactive Preprocessing of Time Series Data. In *SIGRAD* (2012), pp. 39–48. 3

[BRS*12a] BERNARD J., RUPPERT T., SCHERER M., KOHLHAMMER J., SCHRECK T.: Content-based layouts

for exploratory metadata search in scientific research data. In *JCDL* (New York, NY, USA, 2012), ACM, pp. 139–148. 7

[BRS*12b] BERNARD J., RUPPERT T., SCHERER M., SCHRECK T., KOHLHAMMER J.: Guided discovery of interesting relationships between time series clusters and metadata properties. In *i-KNOW* (New York, NY, USA, 2012), ACM, pp. 22:1–22:8. 3

[BvLBS11] BREMM S., VON LANDESBERGER T., BERNARD J., SCHRECK T.: Assisted descriptor selection based on visual comparative data analysis. In *VGTC, Eurographics / IEEE* (2011), EuroVis'11, Eurographics Association, pp. 891–900. 3, 7

[CM84] CLEVELAND, W. S., MCGILL, R.: The Many Faces of a Scatterplot. *JASA 79*, 388 (1984), 807–822. 3

[CMZW13] COSTAS R., MEIJER I., ZAHEDI Z., WOUTERS P.: The Value of Research Data-Metrics for datasets from a cultural and technical point of view. *A Knowledge Exchange Report* (2013). URL: www.knowledge-exchange.info/datametrics. 1

[EF10] ELMQVIST N., FEKETE J.-D.: Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines. *IEEE TVCG 16*, 3 (2010), 439–454. 3

[Fri99] FRIENDLY M.: Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Computational and graphical Statistics 8*, 3 (1999), 373–395. 3

[GE03] GUYON I., ELISSEEFF A.: An introduction to variable and feature selection. *JMLR 3* (2003), 1157–1182. 3

[GFC04] GHONIEM M., FEKETE J.-D., CASTAGLIOLA P.: A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations. In *Information Visualization, IEEE* (Washington, DC, USA, 2004), IEEE, pp. 17–24. 6

[Gra11] GRAY R. M.: *Entropy and Information Theory*. Springer-Verlag Inc., New York, NY, USA, 1990, 2011. 4

[Han05] HAN J.: *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., SF, CA, USA, 2005. 3, 5

[HDS*10] HAO M. C., DAYAL U., SHARMA R. K., KEIM D. A., JANETZKO H.: Variable binned scatter plots. *Information Visualization 9*, 3 (2010), 194–203. 3

[HFM07] HENRY N., FEKETE J.-D., MCGUFFIN M.: NodeTrix: a Hybrid Visualization of Social Networks. *Visualization and Computer Graphics, IEEE 13*, 6 (2007), 1302 –1309. 6

[HTT09] HEY A. J. G., TANSLEY S., TOLLE K. M.: *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009. 1

[IMI*10] INGRAM S., MUNZNER T., IRVINE V., TORY M., BERGNER S., MÖLLER T.: DimStiller: Workflows for dimensional analysis and reduction. In *VAST* (2010), pp. 3–10. 3

[Jai10] JAIN A. K.: Data Clustering: 50 Years Beyond K-means. *Pattern Recogn. Lett. 31*, 8 (June 2010), 651–666. 3

[JJJ08] JOHANSSON S., JERN M., JOHANSSON J.: Interactive Quantification of Categorical Variables in Mixed Data Sets. In *Proc. of the 12th Int. Conf. on Information Visualization* (2008), IV '08, IEEE Computer Society, pp. 3–10. 1, 3, 5

[JZF*09] JEONG D. H., ZIEMKIEWICZ C., FISHER B., RIBARSKY W., CHANG R.: iPCA: An Interactive System for PCA-based Visual Analytics. In *Proc. of the 11th Eurographics Conf. on Visualization* (2009), EuroVis'09, pp. 767–774. 3

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data. *IEEE TVCG 12*, 4 (July 2006), 558–568. 3

[KLL07] KÖNIG-LANGLO G., LOOSE B.: The Meteorological Observatory at Neumayer Stations (GvN and NM-II) Antarctica. *Polarforschung2006 76*, 1 (2007), 25–38. 8

[LKC*12] LEE H., KIHM J., CHOO J., STASKO J., PARK H.: iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *CG Forum 31*, 3pt3 (2012), 1155–1164. 3

[LSS*11] LEX A., SCHULZ H.-J., STREIT M., PARTL C., SCHMALSTIEG D.: VisBricks: Multiform Visualization of Large, Inhomogeneous Data. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2291–2300. 3

[MBD*11] MAY T., BANNACH A., DAVEY J., RUPPERT T., KOHLHAMMER J.: Guiding feature subset selection with an interactive visualization. In *VAST* (2011), IEEE, pp. 111–120. 3

[Mor02] MORRIS A.: *An information theoretic measure of sequence recognition performance*. Tech. rep., 2002. 4

[MP13] MÜHLBACHER T., PIRINGER H.: A Partition-Based Framework for Building and Validating Regression Models. *Transactions on Vis. and CG. 19*, 12 (2013), 1962–1971. 3, 4

[MS99] MANNING C., SCHÜTZE H.: *Foundations of statistical natural language processing*. MIT, MA, USA, 1999. 3

[MWS*10] MEYER M. D., WONG B., STYCZYNSKI M. P., MUNZNER T., PFISTER H.: Pathline: A Tool For Comparative Functional Genomics. *CG. Forum 29*, 3 (2010), 1043–1052. 1

[NH06] NOVOTNY M., HAUSER H.: Outlier-Preserving Focus+Context Visualization in Parallel Coordinates. *Visualization and Computer Graphics, IEEE 12*, 5 (2006), 893–900. 3

[RLKLI12] RIMBU N., LOHMANN G., KÖNIG-LANGLO G., IONITA M.: 30 years of synoptic observations from Neumayer Station with links to datasets., 2012. 7

[RS81] RAO J. N. K., SCOTT A. J.: The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Assoc. 76*, 374 (1981), 221–230. 3

[SBTK09] SCHRECK T., BERNARD J., TEKUŠOVÁ T., KOHLHAMMER J.: Visual Cluster Analysis of Trajectory Data With Interactive Kohonen Maps. *Palgrave Macmillan Information Visualization 8* (2009), 14–29. 3

[SGZC13] SIM K., GOPALKRISHNAN V., ZIMEK A., CONG G.: A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery 26*, 2 (2013), 332–397. 3

[SS07] SHIMAZAKI H., SHINOMOTO S.: A Method for Selecting the Bin Size of a Time Histogram. *Neural Comput. 19*, 6 (2007), 1503–1527. doi:10.1162/neco.2007.19.6.1503. 4

[TDN11] TOMINSKI C., DONGES J. F., NOCKE T.: Information Visualization in Climate Research. In *Information Visualization* (Washington, DC, USA, 2011), IEEE, pp. 298–305. 1

[TZB*12] TATU A., ZHANG L., BERTINI E., SCHRECK T., KEIM D., BREMM S., VON LANDESBERGER T.: ClustNails: Visual analysis of subspace clusters. *Tsinghua Science and Technology 17*, 4 (2012), 419–428. 3

[WCK*09] WALZ J., CHUN F. K.-H., KLEIN E. A., REUTHER A., SAAD F., GRAEFEN M., HULAND H., KARAKIEWICZ P. I.: Nomogram predicting the probability of early recurrence after radical prostatectomy for prostate cancer. *The Journal of urology 181*, 2 (2009), 601–608. 8

[WR09] WHITE R., ROTH R.: Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services 1*, 1 (2009), 1–98. 1

[YHW*07] YANG J., HUBBALL D., WARD M. O., RUNDENSTEINER E. A., RIBARSKY W.: Value and Relation Display: Interactive Visual Exploration of Large Data Sets with Hundreds of Dimensions. *TVCG, IEEE 13* (2007), 494–507. 3

[ZKG09] ZHOU J., KONECNI S., GRINSTEIN G.: Visually comparing multiple partitions of data with applications to clustering. *Proc. SPIE 7243* (2009). doi:10.1117/12.810093. 3