International Symposium on Visual Analytics Science and Technology (2010)
J. Kohlhammer and D. Keim (Editors)

Combining Details of the Chi-Square Goodness-of-Fit Test with Multivariate Data Visualization

T. May [1], J. Davey [1] J. Kohlhammer [1]

[1]Fraunhofer Institute for Computer Graphics Research, Darmstadt, Germany

Abstract

In this work, we combine KVMaps, a visualization technique presented in [May07] for the visualization of statistical aggregations in multivariate contingency tables, with the measures used for the statistical Chi-Square goodness-of-fit test. Goodness-of-fit tests are used to check whether a given distribution of values matches an expected distribution. A single test statistic is calculated to represent the deviation of the complete dataset. By visualizing the deviations for all entries in the contingency table, it is possible to identify the patterns in the distribution of data items, which contribute most to the overall deviation of the dataset. We present two use cases to illustrate how the information about the patterns can be used.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Visualization

1. Introduction

Finding dependencies between a large number of attributes is an important challenge in exploratory data analysis. We combine KVMaps, a visualization technique for the display of multivariate data of up to twenty dimensions with the measures used for the statistical Chi-Square goodness-of-fit test. In the statistical test, the actual distribution of values in a sample data set is compared to an expected distribution, which is specified beforehand. The test computes a single, aggregate value which represents the behavior of the complete dataset. By visualizing the details of this measure with the KVMaps visualization technique, it is possible to explore the deviation from the expected distribution.

It is possible that only a subset of the data is responsible for the deviation. If this subset can be characterized by a specific combination of attribute values, it will emerge as a pattern in the visualization. Thus, the KVMaps technique is extended with the statistical measure to identify and describe the deviation.

In our example, we consider a distribution, which can be expected if all combinations of attributes are mutually independent. Hence, patterns emerging in the visualization relate to dependencies between the associated attributes. We discuss two use cases where this visualization can be applied.

Our method can be used with any other distribution, which can be derived from a hypothesis about the data. The combination of the hypothesis test with the visualization does not only allow the test of the hypothesis itself. In addition, it conveys information on the distribution of data items that cause the test to discard the hypothesis.

Keim et al. [KAF*08] define Visual Analytics as the *combination of automated methods with interactive visualization for effective understanding, reasoning and decision making with large and complex data*. In our case the automated method is Pearons Chi-Square goodness-of-fit test, which is one of the best known tests for categorical data [Sir06, LW92]. We apply the test statistic to the KVMaps visualization technique proposed by May [May07].

From a visualization point of view, the technique is related to pixel and matrix-based visualization techniques using a recursive layout. Keim presented a general view of this class of visualization [Kei00]. Langton et al. [LPWH06] proposed a specific recursive layout, which we currently use

DOI: 10.2312/PE/EuroVAST/EuroVAST10/045-050

in our technique. In pixel-based visualization techniques, every pixel represents one attribute value of a data-item. Instead, our visualization displays a statistical aggregate for a multivariate category of data items. In this work, we propose the aggregate derived from the Chi-Square test.
From the viewpoint of a statistician, our technique is related to the analysis of contingency tables (see Sirkin [Sir06, pp149]). Unlike quantitative measures employed in statistics, the identification of patterns in the visualization of the table is a qualitative result.

2. Preliminaries

In the following section, we will briefly introduce the two basic components of our work. Technically, the statistical test and the visualization technique can easily be combined because of two properties: Firstly, both methods are based on a discretization of all values of data attributes into categories. Secondly, both methods are based on computing statistical aggregate information about the distribution of values for each category and each multivariate combination of categories. The most important difference between the techniques is that the Chi-Square test specifies the aggregate value to be computed for every category. With the KVMap technique, different statistics may be selected for computation depending on the user task.
The Chi-Square goodness-of-fit test combines all information in the dataset, by summing up aggregate information computed for every single category. This sum is the testing criterion for the independency hypothesis. The KVMaps techniques introduces a visual mapping, which exposes the aggregate information separately for every category. Features in the multivariate distribution can be perceived as visual patterns.

2.1. The KVMap Visualization

The *KVMap* is a technique for the visualization of multivariate aggregate data. We will briefly sketch the properties, which are relevant for the topic discussed here. The input for the visualization is a data table containing N data items and C columns. Every data item is defined as a vector $\left(d_{i,1}, d_{i,2}, \ldots, d_{i,C}\right)$ of values. All values of a column $\left(d_{i,j}\right)_{i:1..N}$ are elements of an attribute set A_j. The attributes can be of nominal, ordinal or numerical type.
The second input for the visualization technique is a partitioning of every attribute set. We represent the partitioning as a mapping $\Phi_j : A_j \rightarrow \mathbb{N}$, which may be defined interactively or automatically in a preprocessing step.
The visualization algorithm processes the input data in the following steps (see Figure 1):

1. Compute the vector $\left(\Phi_1(d_{i,1}), \Phi_2(d_{i,2}), \ldots, \Phi_C(d_{i,C})\right)$ defining the partitions each data item belongs to. We call this vector a *profile*. In turn, a profile represents a set of data items. For convenience, we write $k_j = \Phi_j(d_{i,j})$ in the following formulas.
2. Count the number if items in each profile. We will denote the result as $Count\,(k_1, k_2, \ldots, k_C)$
3. Compute the layout for all profiles. The layout defines the position where the information related to a profile will be displayed. The KVMap associates all profiles with the cells in a regular two-dimensional grid (see Figure 2).
4. Compute the statistical aggregate for each profile.
5. Map the aggregate to the color of the associated cell.

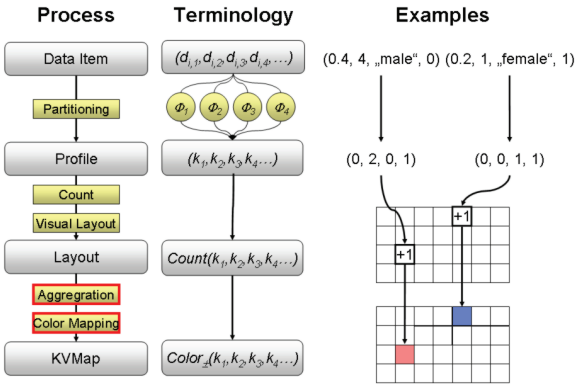

Figure 1: *The visualization process for the KVMap consists of five steps. While the layout scheme is specific for the visualization technique and has been presented in previous works, different statistical aggregates may be chosen for the display. The focus of this work is the calculation of a statistical aggregate derived from the Chi-Square test and its modification for the color mapping (red).*

The focus of the work is the steps 4. and 5. In the following subsection, we will describe how the statistical aggregate is derived from the Chi-Square goodness-of-fit test. In Section 3 we will describe the color mapping.
The first three steps are canonical for the KVMap technique. The partitions and the arrangement of the attributes in the layout can be changed interactively. For details, refer to [May07]. An example of the arrangement of four dimensional profiles into the two dimensional layout is given in Figure 2.

2.2. Chi-Square Goodness-of-fit and Independence test

In statistics, the occurrence of random events $(E_i)_{i:1..n}$ is said to be *independent* from one another, if the following equation holds for the probabilities $p(E_i)$ of every event.

$$\prod_{n}^{i=1} p(E_i) = p(\bigcap_{n}^{i=1} E_i) \qquad (1)$$

The equation implies, that the knowledge about one event does not yield any information on other events. In statistical

(0,0,0,0)	(0,0,0,1)	(0,0,0,2)	(0,0,1,0)	(0,0,1,1)	...		(0,0,5,2)
(0,1,0,0)	(0,1,0,1)	(0,1,0,2)	(0,1,1,0)	(0,1,1,1)	...		(0,1,5,2)
(0,2,0,0)	(0,2,0,1)	(0,2,0,2)	(0,2,1,0)	(0,2,1,1)	...		(0,2,5,2)
(0,3,0,0)	(0,3,0,1)	(0,3,0,2)	(0,3,1,0)	(0,3,1,1)	...		(0,0,5,2)
(1,0,0,0)	(1,0,0,1)	(1,0,0,2)	(1,0,1,0)	(1,0,1,1)	...		(1,0,5,2)
(1,1,0,0)	(1,1,0,1)	(1,1,0,2)	(1,1,1,0)	(1,1,1,1)	...		(1,1,5,2)
...	...	...	...	...	...	...	...
					...		
(4,3,0,0)	(4,3,0,1)	(4,3,0,2)	(4,3,1,0)	(4,3,1,1)	...		(4,3,5,2)

Figure 2: *This table shows a layout example for the KVMap with four attributes. Every attribute is aligned either horizontally or vertically. The layout is recursive, ensuring that every profile is unambigously associated with a specific position in the map. It also ensures that every attribute is connected to a unique spatial frequency, which is exploited in the perception of the patterns. This schema can be extended to an arbitrary number of dimensions up to the resolution of the display.*

data analyis, the data is represented as a series of random events. In this case, it must be assumed that the data representing different data items (i.e. rows) is mutually independent. Instead, we are interested in the dependencies between different attributes (i.e. columns) of the data table. In our case we consider the fact that "attribute j of an item falls into a specific category $k_j \in \Phi_j(A_j)$" as a primitive random event. If a data item is selected randomly from the data, the relative frequency serves as an estimator for the probability of these events:

$$p\left(k_j\right) = \frac{\left|\left\{i : \Phi_j(d_{i,j}) = k_j\right\}\right|}{N} \tag{2}$$

With the assumption, that the different attributes are independent, we can define an estimator for the probability that an item belongs to a profile $(k_1, k_2, \ldots, k_C)$ by virtue of equation 1:

$$p\left(k_1, k_2, \ldots, k_C\right) = \prod_{c}^{j=1} p(k_j) \tag{3}$$

With the estimator it is possible to compare the *expected* number of items for every profile with the *actual* number of items which are counted in step two of the KVMaps visualization. In short, if the expected number and the actual number of items are similar for every profile, this is a strong indication that the attributes are mutually independent. When this estimator is used for two attributes, the goodness-of-fit test is known as the *Chi-Square test for independence*.

The difference measure must satisfy the property that separate profiles must be comparable with regard to the fact, that the number of samples in each profile may be different. A large relative difference with small samples is not as significant as the same relative difference with a large sample size. Because this property has to be satisfied in statistical tests as well, we choose the distance function applied in the Chi-Square goodness-of-fit test. We do not directly apply this test in the visualization technique. We give a brief introduction here, because in the following section we will discuss how we modify this test to be used as a statistical aggregate in the KVMap.

The Chi-Square goodness-of-fit test is applied to support or discard the hypothesis that a sampling distribution is equal to an expected distribution if the samples are drawn randomly from a set. Because there is never an exact match, it is necessary to define a threshold for a deviation, which is still accepted to support the hypothesis. The threshold is compared against a test statistic. We chose the test statistic for the Chi-Square goodness-of-fit test [Sir06], because it can be directly applied to the measures calculated for every profile. In the terms introduced above, the difference measure for a single profile is defined as:

$$diff\left(k_1, \ldots, k_C\right) = \frac{\left(Count\left(k_1, \ldots, k_C\right) - N \cdot p\left(k_1, \ldots, k_C\right)\right)^2}{N \cdot p\left(k_1, \ldots, k_C\right)} \tag{4}$$

The Chi-Square test statistic sums up the differences of all profiles:

$$X^2 = \sum diff\left(k_1, \ldots, k_C\right) \tag{5}$$

The sum yields a single, positive value for the complete data set. If the actual distribution exactly matches the expected distribution the sum will be zero. In this case, the hypothesis that the distributions match is supported at all levels of significance. If the difference exceeds the threshold, the hypothesis is discarded.

3. Using the Chi-Square distance function in KVMaps

Instead of summarizing the dataset in a single testing value, we retain the difference measures for every profile to visualize them individually in the KVMap. By using the visualization technique, we are able to identify patterns in the profiles which do match the expected distribution. For instance, we are interested in subsets of attributes which correlate. This visualization may convey more information on dependencies than the test statistic.

In this section, we describe how the difference measure $diff\left(k_1, \ldots, k_C\right)$ of a profile is mapped to the color of its associated cell in three steps. As a first step, we discard all profiles where the expected number of items is "too small", because the Chi-Square test statistic requires a minimum number of items to yield a valid result. A minimum number of five is cited in various sources (see, for example, Lehn and Wegmann [LW92, p148]). Thus, profiles with an expected number of items smaller than five are displayed with the background color (white).

The difference has no predefined maximum. Hence, in a second step the difference is mapped from the interval $[0..\infty)$ to the interval $[0..1]$. This mapping is not straightforward. Scaling by the actual maximum or minimum value is not an

option. This strategy could introduce visual differences in the distribution even if the numerical differences are negligible.

We suggest a mapping with a cut-off value τ as defined in the following equation:

$$color(k_1,\ldots,k_C) = \begin{cases} 1 & : diff(k_1,\ldots,k_C) > \tau \\ \frac{diff(k_1,\ldots,k_C)}{\tau} & \text{otherwise.} \end{cases} \quad (6)$$

When setting the cut-off value, one has to consider that the expression $\sqrt{diff(k_1,\ldots,k_C)}$ is approximately normally distributed with mean 0 and variance 1, if the hypothesis is actually true [LW92, p86]. Hence, the difference is measured in units of the variance of the normal distribution. Currently we use a conservative value of τ representing ten units of variance as an initial setting. We assume that every deviation greater than this value can be considered significant. In addition, the cut-off value can be adjusted interactively.

As the last step, we compute a signed difference to distinguish the profiles whose frequencies lie above and below the expected distribution. The sign of the deviation can be defined by

$$signum = sgn(Count(k_1,\ldots,k_C) - N \cdot p(k_1,\ldots,k_C)) \quad (7)$$

The signed difference for a profile can be calculated as

$$color_{\pm}(k_1,\ldots,k_C) = signum \cdot color(k_1,\ldots,k_C) \quad (8)$$

The signed difference is mapped to the interval $[-1..1]$. We use a diverging colormap as suggested by Wijffelaars et al. [WVvWvdL08] (see Figure 3). Profiles matching the expected distribution should be unobtrusive and are displayed in grey color. Profiles below and above the expected distribution are shown in different saturation levels of blue and red respectively. This colormap emphasizes the difference between these two groups.

The patterns in the KVMap visualization are sets of cells of similar coloring, which repeat in the horizontal or vertical direction. The layout of the visualization ensures that every attribute has its unique frequency. Because a pattern may include multiple frequencies, it is possible to discern multivariate dependencies in the display.

Figure 3: *This color-map is used with the KVMap visualization to emphasize the difference between profiles matching the expected distribution, and profiles whose actual frequencies lie above or below the expected distribution with the independency assumption.*

4. Using the KVMap

The regular layout of the KVMap shows all possible profiles. Statistical information for every profile is encoded in color. Patterns can be identified as repeating variations of color. The layout must guarantee that every attribute is associated with a specific horizontal or vertical spatial frequency. Because visual attributes like form and size do interfere with the perception of frequency, the KVMap always uses a regular uniform grid. Dependencies involving a specific attribute will create a pattern which includes the characteristic frequency of the attribute. The visualization exploits the fact that multiple superimposed frequencies still can be perceived as a pattern. Hence, dependencies between more than two attributes can be identified.

The arrangement of attributes affects the appearance of the patterns. We observed that patterns forming continguous blocks seem to draw more attention than sparse patterns (see Figure 4). In order to mitigate this effect the attributes can be rearranged interactively.

The identification of a pattern does not mean that it can be read and understood immediately. In fact, a dependency which still can perceived as a pattern may well include a number of attributes that exceeds the working memory of the user.

We shortly sketch the approach to circumvent this cognitive bottleneck. For further details refer to [MK08]. The approach couples visual perception and automated analysis for the interpretation of the patterns.

The user interactively selects the cells which belong to the same pattern. Any subset of cells defines a subset of the original dataset. After selection an automatic classification is done. The resulting classifier represents a formula which separates the selected and the unselected cells. Among other possible candidate methods for classification, a decision tree is used because of two reasons: Firstly, the tree can directly be read by the user. Secondly, it allows the estimation of the relevance of the attributes for the selected pattern.

Furthermore the classifier gives a visual feedback by suggesting a completion of a pattern which has not been fully selected yet. The visual match between the selected and the completed part of the pattern supports an evaluation of the classifier model and a visual hint for the faster identification of complex patterns.

In the following, we present two use cases where the mutual independency is a prerequisite for further analysis. The KVMap can be used as a preliminary step to identify a latent bias which might distort the results of the analysis.

4.1. Use case 1 - Questionnaire calibration.

Consider a questionnaire for a test with a number of questions for which a right or wrong answer can be given. In many cases, the score of the test is the number of correct answers. In the perfect case the questions are mutually independent: The answer to one question does not convey information, as to whether other questions will be answered correctly or not. If this is not the case, either a maximum subset of independent questions must be found or the scoring scheme must be adapted to the bias.

We applied our technique to questionnaire data consisting of eleven questions. Every test result is a data item. Every attribute of the data item indicates whether the respective question has been answered correctly or not. Because all attributes are binary, all partitions Φ_j are trivial. Hence the KVMap displays all 2048 possible combinations for the test results (see Figure 4). A typical result involving three attributes reads as follows: "Questions A, B and C have been answered correctly *together* ten times more often than expected". The analyst may proceed by checking the questions directly for their content. But it is also possible to interactively remove questions (i.e. attributes) from the visualization to test which attribute contributes most to this deviation.

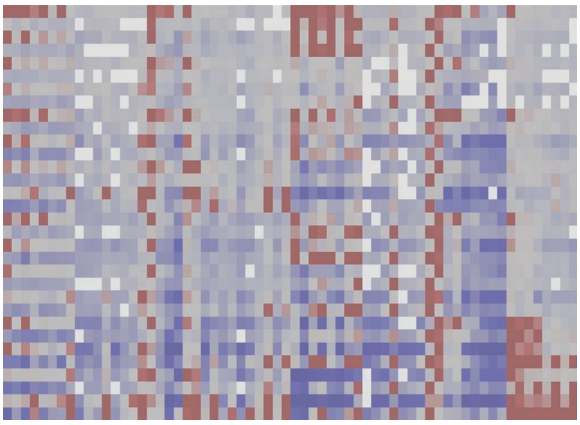

Figure 4: *This figure shows the results for a questionnaire calibration. With eleven questions, there are 2048 possible combinations of wrong and right answers, which are displayed in the map. The map shows multiple dependencies between groups of four variables. The contiguous pattern in the lower right corner defines a result combination including four questions.*

4.2. Use case 2 - Preparing Classification.

The requirements for the calibration of the questionnaire also apply to classification tasks. A classification model represents a function mapping *independent* attributes of a data-item to a *dependent* attribute. In preparation for a classification, the visualization can support the testing of the selected attributes for independency, which is required for some methods. John et al. [JKP94] showed the effect of a latent bias between selected independent attributes for decision tree induction, resulting in a tree identifying trivial relationships. This visualization can be used to identify a latent bias.

In addition, it is possible to visualize the independent and the dependent attributes together. If no patterns can be found, the classification is likely to produce weak results. If patterns can be found, it is possible to exclude independent attributes, which do not appear in a pattern together with the dependent attribute. In this case the visualization can be used as a tool for the attribute selection.

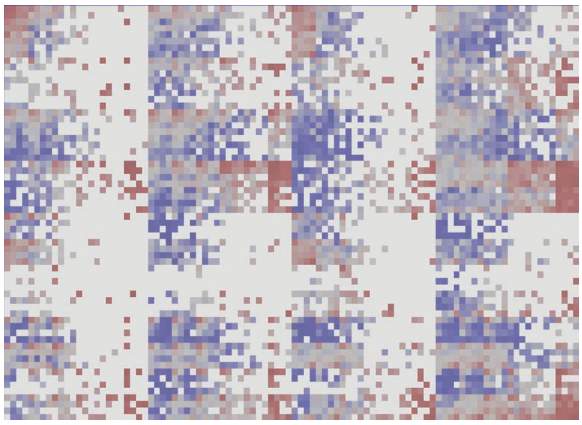

Figure 5: *This map shows the latent dependencies between eight attributes of the US-census dataset. The effect of discarding profiles with a low expected frequency is also shown. The corresponding cells are drawn in the background color (white).*

5. Discussion

In contrast to the approach presented, the dependencies between a large number of attributes also can be computed by applying the Chi-Square test to all pairs of attributes. The results of such a test is a correlation matrix, which shows the pairwise dependencies in a compact way (see, for example, MacEachren et al. [MDH*03]). However, dependencies which manifest in three or more attributes of the dataset do not necessarily appear when only pairs of attributes are considered.

The KVMap is specifically designed to expose potential interdependencies between all attributes which are actually shown. That way, the deviation from the expected distribution is not only described by quantitative statistical measures, but also by qualitative information. The user is able to discern if the deviation follows a regular pattern or if it may be the result of random noise. With this information, the user can assess if and how the model of the dataset can be refined.

A limitation of our approach is imposed by the number of data items available. The statistical test requires a minimum number of items per profile to guarantee significant results. An increasing number of attributes and/or profiles often results in "'white space"' spreading on the map (see also Figure 5). This effect is in fact a visual manifestation of the *curse of dimensionality*. It is not necessarily a drawback

of the visualization technique: One could argue, that if the empirical base simply is too small for sound analysis, a visualization should expose the underlying uncertainty and protect the user from false deductions.

A strategy to cope with the curse of dimensionality is the interactive replacement of attributes which contribute least information about the data. These attributes can be identified by reading the decision tree. The goal of this strategy is to find the minimum number of profiles which still represent most information about the dataset.

In our examples the expected distribution has been derived from the independency hypothesis for the attributes. By using a visualization technique for the display of multivariate contingency tables, latent dependencies between attributes can be revealed as patterns in the display. The visualization could be applied in scenarios where the dependency or independency between attributes is an important prerequisite.

An issue, which is yet to be solved is the choice of the colormap and the cut-off parameter τ. It has been selected because it emphasizes the differences between frequencies which are "too low" or "too high" compared to the expected frequencies. However, it is possible that the most notable patterns are distinguishable only by the amount of their deviation. Interactively shifting the difference measures before applying them to the colormap could be one solution, because this could emphasizes other differences in the colormap.

Furthermore, it is not mandatory to use the *Chi-Square* goodness-of-fit test. It has been chosen, because it is well known and the computation and interpretation of its statistic is quite straightforward. Technically, any other goodness of fit test operating on contingency tables could possibly be used.

6. Conclusion & Future Work

We presented a method to visualize the summands of the Chi-Square goodness-of-fit test with the KVMaps technique. The technique reveals the details of the set of profiles, which are responsible for the deviation from an expected distribution. The visualization clearly separates the cases where this set is random and the cases where the set constitutes a pattern in the display. In addition to the general information on the match between the actual and expected distribution, the patterns can be used to refine the assumptions about the distribution of items in the dataset.

We showed two use cases for the variant, testing the independence hypothesis between attributes. However, the technique can be applied to other hypotheses and the corresponding expected distributions of data items. Future work on this technique will include its application to other distributions.

Another direction of future work will involve the management of patterns found with different configurations of the visualization. After our first tests, we believe that much additional information about an attribute can be derived by comparing the visualizations which include and the visualizations which exclude this attribute. While in this technique the detection of patterns remains the domain of human perception, methods which automatically consolidate the information is the next step in coping with multidimensional dependencies on a greater scale.

References

[JKP94] JOHN G., KOHAVI R., PFLEGER K.: Irrelevant features and the subset selection problem. In *Proceedings of the 11th International Conference on Machine Learning, New Brunswick, NJ* (1994), Morgan Kaufmann Publishers, San Francisco, CA, pp. 121–129. 5

[KAF*08] KEIM D., ANDRIENKO G., FEKETE J.-D., GÖRG C., KOHLHAMMER J., MELANCON G.: Visual analytics: Definition, process, and challenges. In *Information Visualization*, vol. 4950 of *Lecture Notes in Computer Science*. Springer-Verlag, 2008, ch. 7, pp. 154–175. 1

[Kei00] KEIM D. A.: Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics 6* (2000), 59–78. 1

[LPWH06] LANGTON J. T., PRINZ A. A., WITTENBERG D. K., HICKEY T. J.: Leveraging layout with dimensional stacking and pixelization to facilitate feature discovery and directed queries. In *Proceedings of the 1st Visual Information Expert Workshop (VIEW)* (2006), vol. 4370 of *Lecture Notes in Computer Science*, Springer, pp. 77–91. 1

[LW92] LEHN J., WEGMANN H.: *Einführung in die Statistik*. Vieweg + Teubner, Wiesbaden, DE, 1992. 1, 3, 4

[May07] MAY T.: Working with patterns in large multivariate datasets - karnaugh-veitch-maps revisited. In *IV '07: Proceedings of the 11th International Conference Information Visualization* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 277–285. 1, 2

[MDH*03] MACEACHREN A. M., DAI X., HARDISTY F., GUO D., LENGERICH E.: Exploring high-d spaces with multiform matrices and small multiples. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)* (2003), IEEE Computer Society. 5

[MK08] MAY T., KOHLHAMMER J.: Towards closing the analysis gap : Visual generation of decision supporting schemes from raw data. In *EuroVis '08: Computer Graphics Forum (Special Issue on Eurographics Symposium on Visualization)* (Washington, DC, USA, 2008), IEEE Computer Society, pp. 277–285. 4

[Sir06] SIRKIN M.: *Statistics for the Social Sciences, 3. Edition*. Sage Publication Inc., Thousand Oaks, CA, 2006. 1, 2, 3

[WVvWvdL08] WIJFFELAARS M., VLIEGEN R., VAN WIJK J. J., VAN DER LINDEN E.-J.: Generating color palettes using intuitive parameters. In *EuroVis '08: Computer Graphics Forum (Special Issue on Eurographics Symposium on Visualization)* (Washington, DC, USA, 2008), IEEE Computer Society. 4