A Study on Quality Metrics vs. Human Perception: Can Visual Measures Help us to Filter Visualizations of Interest?

Dirk J. Lehmann, Sebastian Hundt, Holger Theisel

Abstract: The number of visualizations being required for a complete view on data non-linearly grows with the number of data dimensions. Thus, relevant visualizations need to be filtered to guide the user during the visual search. A popular filter approach is the usage of quality metrics, which map a visual pattern to a real number. This way, visualizations that contain interesting patterns are automatically detected. Quality metrics are a useful tool in visual analysis, if they resemble the human perception. In this work we present a broad study to examine the relation between filtering relevant visualizations based on human perception versus quality metrics. For this, seven widely-used quality metrics were tested on five high-dimensional datasets, covering scatterplots, parallel coordinates, and radial visualizations. In total, 102 participants were available. The results of our studies show that quality metrics often work similar to the human perception. Interestingly, a subset of so-called Scagnostic measures does the best job.

ACM CCS: Human-centered computing → Visualization → Empirical studies in visualization

Keywords: Quality Metric, Perception Study, Visualization

1 Introduction

Nowadays, the visual analysis of high-dimensional data has become a major task within the field of data analysis. Generally, dealing with high-dimensional data means to be faced with a scalability issue: the number of visualizations grows non-linearly with the number of data dimensions. *Quality metrics* measure a certain quality of a visualization – e.g., correlations, clusters, and trends – and map it onto a real number (see [Lea12b] for details). They are a tool to filter relevant visualizations. If they resemble the human perception system, such quality metrics are useful. To figure this out it is important to continuously evaluate if the available quality metrics are related to the human perception system. In this work, we therefore evaluate a popular set of quality metrics for different visual tasks.

2 Related Work & Background

In our study, we consider bivariate *scatterplots* (SP), *parallel coordinate plots* (PCP) [Ins85] and multivariate *radial visualizations* (RadViz) [Hea97]. A scatterplot is a visualization of pairwise dimensions generated by an orthographic projection of the data records onto a two-dimensional plane. Interpreting two data dimensions in a plane as vertical axes yields a PCP: therein, a data record is represented as a line where the start/end vertex is placed on the axes with an intercept that relates to the component value of the record w.r.t. the related dimension. For radial visualization, the data dimensions are represented as anchor points within a radial layout. A spring force is assumed as being the component of the related dimension w.r.t. a certain record. The position of a record is where the spring forces vanish. In general, an a priori classification of the data is visually stressed by labeling the visualization, denoted as a *classified visualization*. For this, different color schemes, icons, or glyphs might be used to visually emphasize different classes. Consequently, an unlabeled visualization is denoted as *unclassified visualization*. Figure 1 illustrates the three considered visualization approaches.

A broad set of quality metrics for visualizations are known. Sips et al. [Sea09] and Tatu et al. [Tea09, Tea11] proposed a quality metric collection for both scatterplots and parallel coordinates. The *Rotating Variance Measure* detects correlations for unclassified scatterplots. Classified scatterplots can be handled with the *Class Density Measure* and the *Histogram Density Measure*, which measure the sepa-

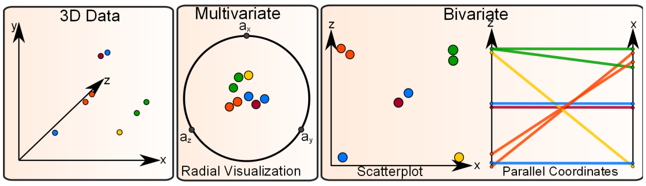


Figure 1: Classified visualization approaches for high-dimensional data. Unclassified visualizations look similar but they would be b/w instead. The visualized data are implied by a 3D dataset (left).

ration of classes. The *Hough Space Measure* detects clusters in unclassified parallel coordinates, whereas the *Similarity Measure* and the *Overlap Measure* detect clusters in classified parallel coordinates. For unclassified RadViz, the *Cluster Density Measure* is proposed to measure cluster separation. Wilkinson et al. [Wea05] presented a graph-theoretic scagnostics – based on the scagnostics from Tukey et al. [TT85] – with nine graph-based metrics for scatterplots. They consist of measures for outliers, shape, trend, density, and coherence. Later on, Bertini et al. [Ber11] systematized quality metrics. Such a classification supports the comparison of different quality metrics.

A set of studies regarding perceptiveness and quality metrics are known: In [WW08], criteria are evaluated that should be met in order to use scagnostic indices for visual data analysis. Sips et al. [Sea09] presented a quality metrics-based study for scatterplots with ten study participants and six high-dimensional datasets. They investigated human perception vs. two quality metrics and they got promising results. Tatu et al. [Tea10] presented a study for quality metrics for scatterplots with 18 study participants based on the *Wine* dataset w.r.t. four quality metrics. A rather general validation was done in [Bea13]. The authors evaluate parameter settings for well memorable visualization techniques. In [Sea12] and [Lew12] studies are presented where Cluster Separation Metrics for 2D scatterplots are compared with human judgments. The former evaluates the quality metrics that were presented by Sips et al. [Sea09] on 800 scatterplots of 75 real and synthetic datasets which are judged by the first two authors of the article. The result of this qualitative data study is rather disappointing, since in about 50 % of the cases the metrics fail to match with the human judgments. The latter investigated seven cluster quality measures on 19 2D datasets, each with nine different cluster versions that were judged by 12 non-expert participants and 5 expert participants. They received partly promising but also partly disappointing results. Both studies conclude that the success of a quality metrics depend on the underlying dataset. However, in our study we also consider quality metrics for correlations and trends, the popular scagnostics indices, further visualization approaches, and our study has been conducted with a larger group of participants. Sedlmair et al. [Sea13] presented an empirical overview of aptitude for 2D/3D scatterplots for cluster separation based on dimension reduction techniques. Different scenarios of evaluation in information visualization have been analyzed by Lam et al. [Lea12a] in order to provide suggestions for well-designed evaluations. Our study design is inspired by this work and it will be explained subsequently.

3 Study Design

This section explains and illustrates our basic study design.

3.1 Considered Data & Visualizations

We use five high-dimensional datasets to conduct the study: *Iris* [Fis36], *Yeast* [HN96], *Wine*, *Wdbc*, and *Cars* [AN07], which have different properties, as can be seen in Table 1. We studied the visualization techniques scatterplots, parallel coordinates, and RadViz; classified and unclassified. Scatterplots and parallel coordinates are the most frequently used bivariate visualization approaches. Hence, it was logical to include them in our study. The RadViz is a multivariate (projective) projection approach. In the past, we used them often to address, e.g., multi-class issues for our collaboration partners from fluid dynamics. Thus, we were interested in to disclose whether quality metrics might help to guide us during a visual search in RadViz. Therefore, we also included RadViz. In total, 7549 different visualizations resulted on which the quality metrics of Section 3.2 were applied.

Dataset	Dimensions	Records	Classes
Iris	5	150	3
Yeast	10	1484	10
Wine	14	178	3
Wdbc	32	569	2
Cars	33	7755	52

Table 1: The datasets we considered in our study.

3.2 Considered Quality Metrics

In our study, we consider seven quality metrics which we already know regarding their properties by some of our previous works, such as [Lea12b, Aea10]:

Class Density Measure (CDM) [Tea09]: measures the separation between several classes. It was applied to classified scatterplots and RadViz. The $CDM(\mathbf{v})$ of the points $\mathbf{p} \in \mathbf{v}$ of either a classified scatterplot or a RadViz $\mathbf{v}$ with M different classes is given by

$$CDM(\mathbf{v}) = \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \sum_{i=1}^{P} ||\mathbf{p}_k^i - \mathbf{p}_l^i||.$$

Cluster Density Measure (C$_l$DM) [Aea10]: measures the quality of separated clusters within unclassified RadViz. The $C_lDM(\mathbf{v})$ of an unclassified RadViz $\mathbf{v}$ is given by

$$C_lDM(\mathbf{v}) = \frac{1}{K} \sum_{k=1}^{K} \sum_{l=k+1}^{K} \frac{d_{k,l}^2}{r_k r_l},$$

with K being the number of clusters, $d_{k,l}^2$ being the p_2-norm between two cluster centroids $d_{k,l}^2 = ||\mathbf{c}_k, \mathbf{c}_l||$ and r_i being a cluster diameter. For details on how this parameter set is chosen please see [Aea10].

Hough Space Measure (HSM) [Tea09]: measures the separation of clusters. It was applied to unclassified parallel coordinates. The HSM is based on the observation that clusters of separated lines form sharp density points within a Hough Space. For unclassified parallel coordinates $\mathbf{v}$, the $HSM(\mathbf{v})$ is given by

$$HSM(\mathbf{v}) = 1 - \frac{p_{\mathbf{v}}}{w \cdot h},$$

with $p_{\mathbf{v}}$ being the number of pixels in the Hough Space Image $H_{\mathbf{v}}$ (resolution $w \times h$) that fulfills $H_{\mathbf{v}} > median(H_{\mathbf{v}})$.

Overlap Measure (OM) [Tea09]: A measure that penalizes the overlap between different classes in classified parallel coordinates $\mathbf{v}$. The $OM(\mathbf{v})$ for M different classes and a number of pixels P in the class-dependent Hough Space Image H_v is given by

$$OM(\mathbf{v}) = \sum_{k=1}^{M-1} \sum_{l=k+1}^{M} \sum_{i=1}^{P} ||\mathbf{H}_{\mathbf{v}}{}_k^i - \mathbf{H}_{\mathbf{v}}{}_l^i||.$$

Scagnostics: The Scagnostics indicies [Wea05] are graph attributes (except for the measure *Monotonic*). They are designed for unclassified scatterplots. For this, the visualized points are interpreted as nodes in a graph. In this study, we consider the indicies *Monotonic*, *Striated*, and *Stringy*.

Monotonic (Mono): measures the trend of the scattered points, based on the Pearson correlation coefficient by

$$Monotonic(\mathbf{v}) = \left(\frac{cov(x_i, y_i)}{\sigma(x_i)\sigma(y_i)}\right)^2,$$

with $cov(x_i, y_i)$ being the covariance of the components and $\sigma(a_i)$ being the standard deviation.

Striated (Striat): measures if scattered points form straight line patterns being space-coherent. The $Striated(\mathbf{v})$ of the scatterplot $\mathbf{v}$ is given by

$$Striated(\mathbf{v}) = \frac{1}{|\mathbf{V}^{(2)}|} \sum_{v \in \mathbf{V}^{(2)}} |\cos(\angle(e(v,a), e(v,b)))|,$$

with $e(v,a)$ being an edge in the Minimum Spanning Tree (of the points) and $\mathbf{V}^{(2)}$ being the set of nodes of degree 2.

Stringy (String): measures if thin shapes occur in scattered points. The $Stringy(\mathbf{v})$ of a scatterplot $\mathbf{v}$ is given by

$$Stringy(\mathbf{v}) = \frac{\Phi(MST(\mathbf{v}))}{l(MST(\mathbf{v}))},$$

with Φ being the diameter and l being the shortest path of the Minimum Spanning Tree.

3.3 Types of Conducted Studies

A high-quality study should fulfill two criteria: the result should be free of missed values and the number of participants should be large enough to get significant results.

Thus, a study should be conducted under supervised conditions, since it guarantees that multiple participations, biased study data, or inappropriate participants are avoided. Unfortunately, such lab studies are expensive, since staff, rooms, and equipment are required. Therefore, lab studies are commonly conducted with a small number of participants. The web enables an alternative to increase the number of participants by conducting an online study. Of course, it cannot be supervised in any situation. Thus, such a study is unsupervised and bears the risk that, e.g., missed values occur, but it also offers the chance to gain numerous participants.

In order to use the advantages of both worlds, we conduct a lab study as well as an online study. Informally speaking, our idea is to reveal study results from the lab study and to confirm their statistical significance with the online study, based on the assumption that a match of both study results confirms the reliability of the lab study. Note that the inverse implication is not necessarily true. Formally speaking, the reliability of the lab study can be falsified with the aid of the online study.

Our lab study was conduced within a controlled environment with 22 participants, composed of undergraduate and graduate students with knowledge of visualization. For our online study, we informed potential participants via social networks, such as Facebook. We gained 80 participants for our online study.

3.4 Sequence of Tasks for a Participant

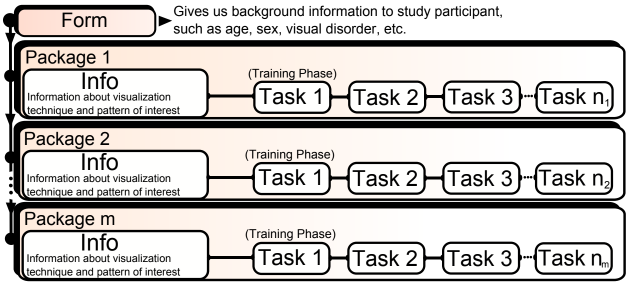

Form
Gives us background information to study participant, such as age, sex, visual disorder, etc.
Package 1
Info
Information about visualization technique and pattern of interest
(Training Phase)
Task 1
Task 2
Task 3
Task n1
Package 2
Info
Information about visualization technique and pattern of interest
(Training Phase)
Task 1
Task 2
Task 3
Task n2
Package m
Info
Information about visualization technique and pattern of interest
(Training Phase)
Task 1
Task 2
Task 3
Task nm

Figure 2: Study procedure for each study participant.

Figure 2 illustrates the guideline for a study participant.

Form: At the beginning, a study participant had to fill out a form with reference data, such as *sex*, *age*, etc.

Package: Then, several work packages had to be completed. Within a package, we collect data to analyze the accordance between a user-based ranking and a quality metric-based ranking of the same visualizations. Each package has the same design – explained below – merely the visualization technique and quality metric are exchanged.

Package - Info: At first, a text informs the user about the visualization technique that is considered and about the pattern of interest that should be detected by the user. For instance, one goal was: "Please find such visualizations which show the best cluster separation". Even though the values of the underlying quality metric are known, the values will not shown to the participant in order to get bias-free selection results.

Package - Task: Our package is designed in a way that a participant had to conduct the same task several times for different datasets. As Figure 3 (top) shows, our task consists of three steps:

- 1) **Set of Visualizations**: a set of k visualizations is presented to a study participant. We will not go into too much detail here, but an appropriate number *k* of visualizations depends on the histogram, given over the quality metric values w.r.t. a visualization technique. This *k* is chosen in a way to get enough votes per visualization in order to guarantee that the study results can be statistically analyzed.

- 2) **Selection**: the participant is asked to select a number of the 25% (plus one) best visualizations from the presented ones w.r.t. the selection goal.
- 3) **Sorting**: the user sorts the selected visualizations in order of perceived quality.

By using a random order of different tasks, they barely influence the respective results. We used an integrated training phase in order to familiarize the participants: the first task in a package is considered as training to become familiar with the study tool. Thus, the results of all first tasks (per package) are discarded.

Figure 3 shows four task scenarios that were captured from different study participants. These scenarios are randomly selected. *Task Scenario 1* shows a task for selecting classified scatterplots that visualize best the class separation of the *Wine* dataset being compared with the Class Density Measure. The participant was able to select and rank two of the three best plots compared to the measure. Nevertheless, the best plot have been overlooked. *Task Scenario 2* shows a task for selecting unclassified scatterplots that visualize best the correlation of the *Wdbc* dataset being compared with the Monotonic measure. Here, the participant was able to select and rank the three best plots. *Task Scenario 3* shows a task for selecting classified parallel coordinates that visualize best the class separation of the *Wdbc* dataset being compared with the Overlap measure. Our participant was able to select and rank the four best plots. *Task Scenario 4* shows a task for selecting classified parallel coordinates that visualize best the class separation of the *Cars* dataset being compared with the Overlap measure. Our participant was not able to select and rank the four best plots compared to the Overlap measure.

In total, different quality metrics and visualization techniques are considered by our packages. Five datasets are considered within the different tasks per package. The packages and tasks were presented in random order to minimize bias caused by familiarization. Our study framework enables the investigation of a large spectrum of visualization techniques, quality metrics, and datasets.

3.5 Conducting our Study in a Nutshell

Each participant completed 7 packages presented in random order. A package is a unit of a visualization technique and a quality metric. Based on Section 3.2 and 3.1 the following configurations were part of the study:

- Package 1: unclassified RadViz, Cluster Density
- Package 2: classified Scatterplot, Class Density
- Package 3: unclassified Scatterplots, Striated
- Package 4: unclassified Scatterplots, Monotonic
- Package 5: unclassified Scatterplots, Stringy
- Package 6: classified Parallel Coord., Overlap
- Package 7: unclassified Parallel Coord., Hough Space

Per package, 5 datasets (cf. Sec. 3.1) were considered, i.e., 5 tasks had to be completed. In total, each participant conducted 35 tasks shared over 7 packages either in the lab or online study (cf. Sec. 3.3). This scheme allows to study the human perception for 7 quality metrics and 3 visualization techniques. In the following, we discuss decisions we made during our study design.

3.6 Discussion of Our Study Design

We designed our study to examine whether quality metrics resemble the human perception. Our design is based on the hypothesis: *the larger the accordance between the selected visualizations of the quality metrics vs. our participants, the more conform are the metrics with human perception.*

In this respect, quality metrics are designed for the simplest version and parameter set of a visualization technique. We also used solely visualization techniques in the simplest version in order to make the comparison fair. Moreover, a visual search is an iterative process in practice where the analyst deals with different visualization techniques, large amounts of visualizations, and different interaction techniques at the same time. Our presentation scheme mimics such a real life situation. To minimize the bias and influence of further psychological phenomena, we asked our participants to select the visualizations as quickly as possible.

What our study finally measures is the accordance of selected visualizations by both the quality metric and our participants. A good match does not imply that a causal relation between the properties of a quality metric and the perception of a human exists. It remains unclear whether such a correlation is caused by further hidden variables, such as cultural background, healthiness, form of the day, education, etc. However, this way we can measure which quality metrics already tend to resemble the human perception.

4 Results

In this section, we evaluate the results of our study. Section 4.1 illustrates our study characteristics, Section 4.2 provides an analysis on how good the lab study matches with the online study. Section 4.3 describes the accordance of the visualization selection based on the quality metrics and our participants. This is the main result of our study. In Section 4.4, we evaluate how stable the quality metrics' behavior is according to their perception properties and in Section 4.5, we evaluate the correlation between the accordance and the stability. We close with an interpretation of our study results in Section 4.6. For the interested reader, we provide additional material of our study under: `http://QM.dirk-lehmann.de/AddMaterialPDF.pdf`.

4.1 Evaluation: Characteristics of our Study

Table 2 shows that 102 participants were involved.

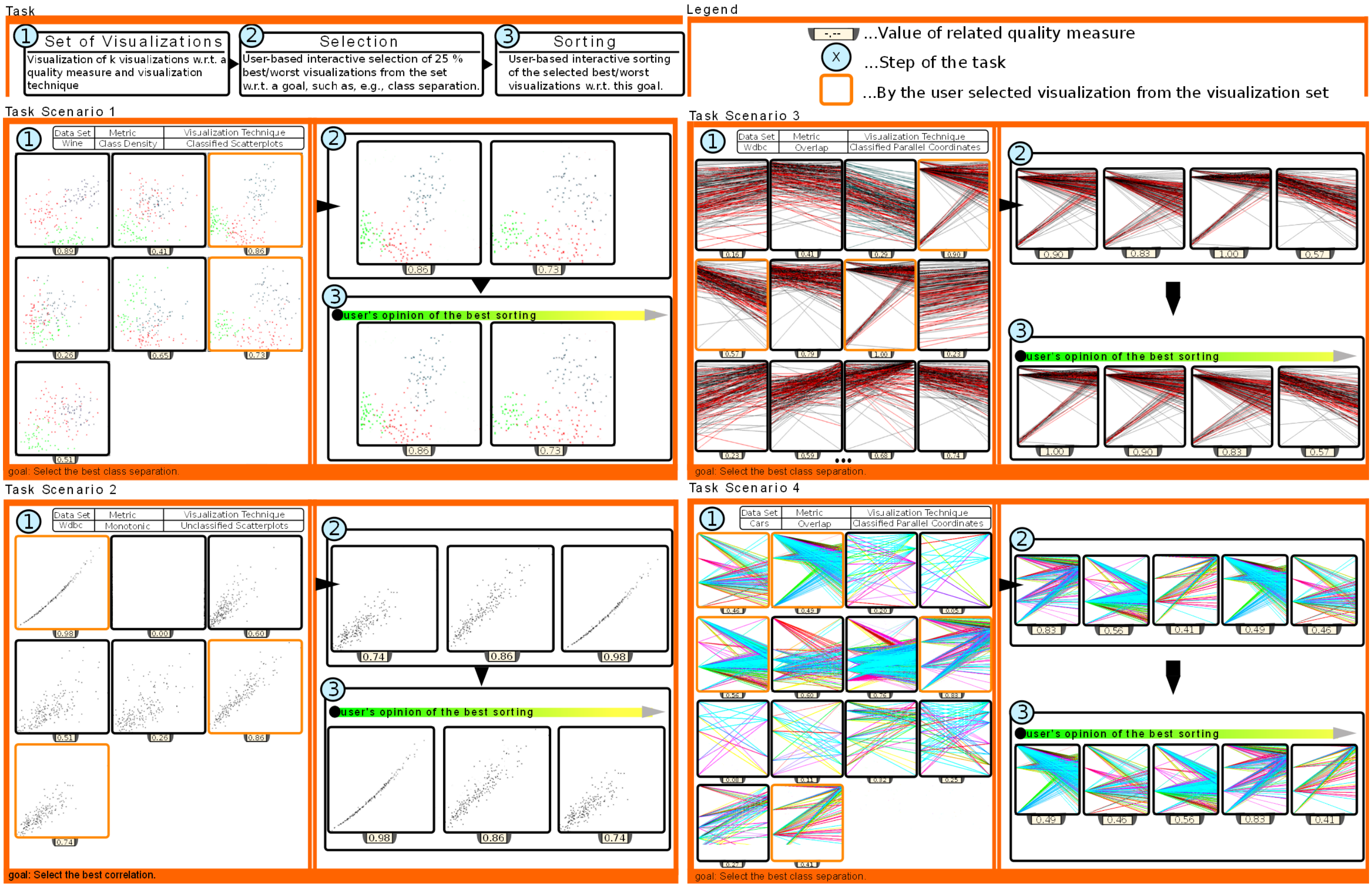
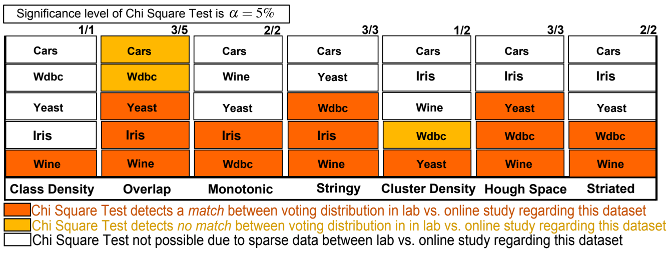

Task
1 Set of Visualizations
Visualization of k visualizations w.r.t. a quality measure and visualization technique
2 Selection
User-based interactive selection of 25 % best/worst visualizations from the set w.r.t. a goal, such as, e.g., class separation.
3 Sorting
User-based interactive sorting of the selected best/worst visualizations w.r.t. this goal.
Legend
-.-- ...Value of related quality measure
X ...Step of the task
...By the user selected visualization from the visualization set
Task Scenario 1
Data Set Wine
Metric Class Density
Visualization Technique Classified Scatterplots
0.89
0.41
0.86
0.26
0.65
0.73
0.51
0.86
0.73
user's opinion of the best sorting
0.86
0.73
goal: Select the best class separation.
Task Scenario 2
Data Set Wdbc
Metric Monotonic
Visualization Technique Unclassified Scatterplots
0.98
0.00
0.60
0.51
0.26
0.86
0.74
0.74
0.86
0.98
user's opinion of the best sorting
0.98
0.86
0.74
goal: Select the best correlation.
Task Scenario 3
Data Set Wdbc
Metric Overlap
Visualization Technique Classified Parallel Coordinates
0.90
0.83
1.00
0.57
user's opinion of the best sorting
1.00
0.90
0.83
0.57
goal: Select the best class separation.
Task Scenario 4
Data Set Cars
Metric Overlap
Visualization Technique Classified Parallel Coordinates
0.83
0.56
0.41
0.49
0.46
user's opinion of the best sorting
0.49
0.46
0.56
0.83
0.41
goal: Select the best class separation.

Figure 3: Task scenarios captured from our participants.

Study	Males	Females	Total	μ_{age}	σ_{age}	Votes
Lab	12	10	22	27.25	3.7	3432
Online	25	55	80	27	5.59	6880

Table 2: Composition of our participants for both studies.



Significance level of Chi Square Test is $\alpha = 5\%$

1/1	3/5	2/2	3/3	1/2	3/3	2/2
□ Cars	🟨 Cars	□ Cars	□ Cars	□ Cars	□ Cars	□ Cars
□ Wdbc	🟨 Wdbc	□ Wine	□ Yeast	□ Iris	□ Iris	□ Iris
□ Yeast	🟧 Yeast	□ Yeast	🟧 Wdbc	□ Wine	🟧 Yeast	□ Yeast
□ Iris	🟧 Iris	🟧 Iris	🟧 Iris	🟨 Wdbc	🟧 Wdbc	🟧 Wdbc
🟧 Wine	🟧 Wine	🟧 Wdbc	🟧 Wine	🟧 Yeast	🟧 Wine	🟧 Wine
Class Density	Overlap	Monotonic	Stringy	Cluster Density	Hough Space	Striated

🟧 Chi Square Test detects a *match* between voting distribution in lab vs. online study regarding this dataset
🟨 Chi Square Test detects *no match* between voting distribution in in lab vs. online study regarding this dataset
□ Chi Square Test not possible due to sparse data between lab vs. online study regarding this dataset

Figure 4: Match between lab and online study.

Note that we extended related studies, such as [Tea09, Sea09]. For this, a visualization data base with a total of 7549 visualizations has been used, which is the sum of visualizations obtained from the datasets *Iris*, *Yeast*, *Wine*, *Wdbc*, and *Cars*. Even though the number of participants was large, in some cases there were not enough votes for analysis purposes. Additionally, about 15 % of the participants stopped the online study before its end. Since the packages were randomly presented, this behavior also leads to a lack of votes in some cases. Thus, we only considered study results for which we got enough votes in order to allow a statistical analysis.

4.2 Match: Lab vs. Online Study

The *Chi-square test for similarity* [Ken70] enables to measure how good two (or more) random variables fit to the same distribution behavior. We use it to investigate whether our lab study results match with the results of the online study:

How good both studies match, is stated in Figure 4, separated for each dataset and provided by the Chi test. The orange boxes are related to a match of the distribution of votings between the lab and the online study and regarding a certain quality measures. It is illustrated by a light orange box if the test failed, i.e., there is no match between both studies in this case. The probability that these results are wrong is given by the significance level $\alpha = 5\%$. Since a sparse data situation, the preconditions to conduct such a test was not for each dataset given, which is emphasized by a white colored box. For instance, for five datasets regarding the OM a test was able. For three datasets the test was positive regarding a match between both studies, and thus a ratio 3 out of 5 (3/5) follows.

From this it follows that the quality metrics CDM, Mono, String, HSM, and Striat match for both studies. The selection results of the measures OM and C_lDM do not completely match. In addition, only one dataset was available to check the quality of the match between the studies for the CDM. Therefore, the match for the CDM cannot be confirmed due to the sparse data situation. For the remaining quality metrics, a appropriate number of results from $2,3,4$ or even 5 datasets were available in order to check the match between the studies. In total, our analysis reveals that the subsequent evaluation-based statements apply rather for the quality metrics Mono, String, Striat, and HSM. For these measures, the statements are more reliable than for the remaining measures.

4.3 Accordance between the Visualization Selection of the Quality Metrics and our Participants

The *accordance* $A(QM,DS)$ of a quality metric QM and a dataset DS is a statistical measure, which is given by the number n_{match} of the event "vote of a study participant and a QM is equal" and the total number n_{total} of votes regarding the QM and the DS: $A(QM,DS) = n_{match}/n_{total}$. The results for the accordance are given in Figure 5: The orange bar charts illustrate the accordance per dataset for each of the considered quality metrics and separated for the lab study and the online study. A *total accordance* $tA(QM)$ of a quality metric QM is given by the mean value

$$tA(QM) = \frac{1}{n}\sum_{i=1}^{n} A(QM,DS_i),$$

with n being the number of datasets. Figure 5 (top left) shows the values of the total accordance per quality metric. Please note that the total accordance gives a condensed accordance view, which gives us an entry point for the study interpretation: A large (total) accordance value (close to one) means that the votes of a participant perfectly match with the votes of a quality metric, i.e., the related quality metric seems to be strongly perceptive. If the quality metric would just randomly vote, the related accordance value would be 0.25 (or below). Hence, the larger the abs. difference between the accordance value and 0.25, the more perceptual a quality metric is.

It follows that none of the investigated measures are "strongly perceptive", since the total accordance of them is much smaller than one. But trends can be seen: the metrics Mono, String, Striat, and HSM often work similar as the participant. They can be classified as "perceptive" with a total accordance larger than 0.50. The metrics OM and the C$_l$DM can be classified as "weakly perceptive" with a total accordance larger than 0.40. In contrast, the CDM appears as "not perceptive".

4.4 Volatility of Visualization Selection for Quality Metrics vs. our Participants

A *Pearson correlation* [Ken70] is a statistical linear regression measure to measure the binding of variables w.r.t. a certain model, here a linear function. A large value relates to a strong linear binding between the visualization rankings from the quality metric and the participants. The Pearson correlation is related to a certain model: If the value is large, the result fit to a linear model. Otherwise, it possibly fits to another model. Thus, if this value varies over different datasets, the results likely fits to various models. A good quality metric should fit to the same model, even if this model is hidden. Hence, it is logical that a good quality metric ought to have a low variance w.r.t. regression values over different data. We denote the standard deviation of the regression value over different dataset as *volatility*.

Figure 6 provides the volatility results of our study: (top left) shows in orange bar charts the values of volatility for each quality metric. The remaining bar charts illustrate the Pearson correlation per dataset for each of the considered quality metrics, separated for the lab and the online study. Note that the volatility (top left) gives us an entry point for the study interpretation: If this value is larger than 0.3, we label the measure as "volatile". It can be seen that the C$_l$DM and the CDM are volatile. For instance, the CDM turns out as being volatile if it is applied to different datasets, e.g., it shows promising results for the *Wine* dataset, but disapointing for the *Yeast* dataset. Interestingly, we get similar regression results for CDM in the *Wine* dataset as Tatu et al. [Tea09], which confirm the results of their study. The metrics OM, Mono, String, HSM, and Striat depict better results. Especially the measures String and HSM show a stable behavior with a volatility smaller than 0.24. The stablest behavior is given by OM with a volatility smaller than 0.2.

4.5 Correlation between Accordance and Volatility

Figure 7 shows a comparison between the *total accordance* (cf. Fig. 5) and the *volatility* (cf. Fig. 6) as a scatterplot. On the left, the comparison is based on the results for both the lab and the online study. There is only a weak correlation verifiably, with a correlation coefficient $\rho = -0.201$. However, we already know from the abovementioned evaluations that the results for the measures CDM and C$_l$DM are inconsistent, mainly due to a sparse data situation. In addition, both measures can be qualitatively considered as to be outliers in Figure 7 (left). By excluding these two measures, we get a very high positive correlation between the accordance and the volatility for the remaining measures with $\rho = 0.903$. This is a bit of surprising result, since the accordance grows with the volatility for this subset of measures. On the right in Figure 7, the comparison is provided separated into the lab study (top) and the online study (down). The variables barely correlate, with $\rho = -0.08$ (lab) and $\rho = 0.165$ (online). By disregarding again the CDM and C$_l$DM, there is only a weak positive correlation $\rho = 0.330$ (lab) and $\rho = 0.367$ (online). In total, the accordance of the considered quality metrics does not generally relate to the volatility, but there are correlating subsets of measures possible.

4.6 Discussion

Here, we discuss our study results and the helpfulness of using quality metrics in practice. By considering the men-

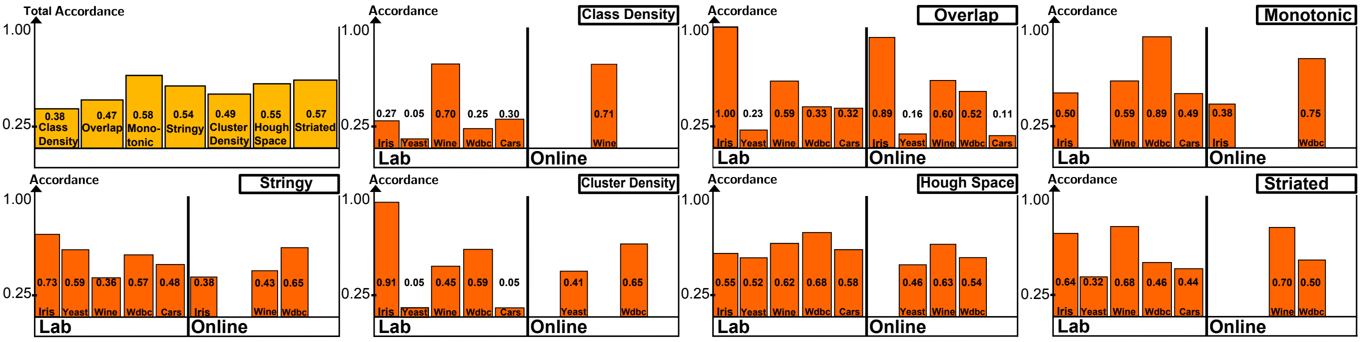


Figure 5: Accordance between the votes of the quality metrics and our participants.

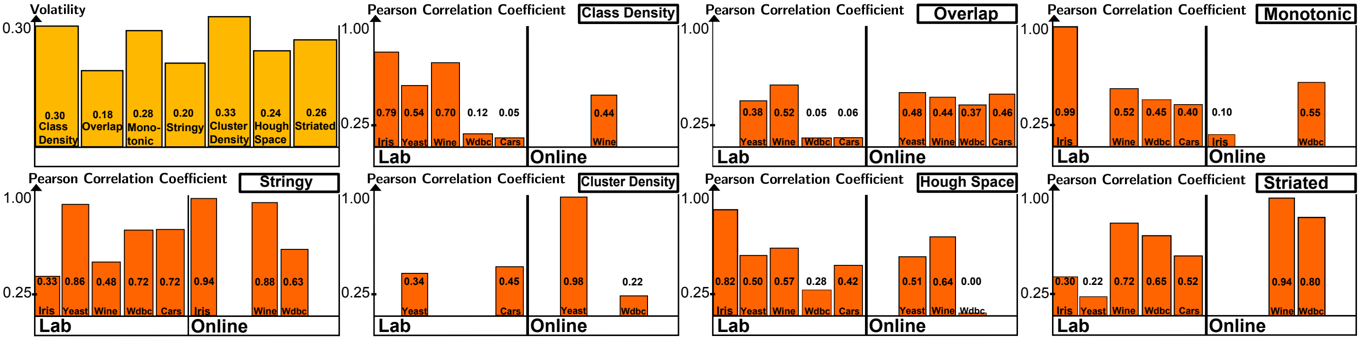


Figure 6: Volatility and regression analysis regarding the votes of the quality metrics and our participants.

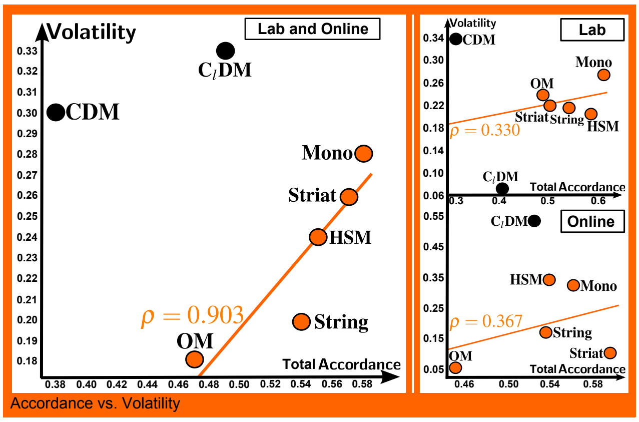


Figure 7: Accordance vs. Volatility

tioned criteria *match of both studies*, *accordance*, and *volatility*, we are able to identify such quality metrics that are best w.r.t. all criteria and have formally the best relation to human perception. These metrics are: **Monotonic**, **Stringy**, **Striated**, and **Hough Space Measure**. These measures might be used whenever human perception needs to be mimicked. Since the inverse statement is not true, it does not mean that the remaining measures are useless, but for them our study is not able to statistically prove their relation to the human perception.

For the interested reader, a detailed analysis of the seven quality metrics follows: In order to detect correlations in unclassified plots or RadViz, the selection results regarding the Mono between participants and quality metrics substantially match. The detection of thin shapes and scattered lines in unclassified plots also matches well between metrics and participants. Thus, we recommend Mono, String, and Striat for practical use. The C_lDM for detecting separated clusters in plots or RadViz seems to be volatile and cannot be recommended for practical use. The Hough Space measure for detecting cluster separation in unclassified parallel coordinates delivered promising results. Participants and metric made the same selection decisions w.r.t. best views. Thus, this measure is recommended to support the user in practice. The accordance of the CDM is indeed larger than for the OM, but the volatility is larger for the CDM and the situation of available data is worse for it. Thus, the OM is to be preferred for use in practice in order to detect a class separation in classified parallel coordinates.

An interesting observation is that the more classes are presented in a view the less the participants were able to recognize any reasonable patterns therein, and the less accordance could be observed for the selected views between metric and participant. Especially the selection of class separation in parallel coordinates of the *Yeast* and *Cars* data-set (by CDM and OM) with more than 10 classes was nearly impossible for our participants. However, for few classes (*Iris*, *Wine*) the accordance between metrics and participants was convincing. Apparently, the participants are perceptively (and maybe cognitively) overwhelmed if the number of classes is larger than nine. On the other hand, quality metrics are still able to select views showing a class separation – not in a perceptive sense, but in an analytical sense. Thus, our study illustrates that quality metrics might support the selection and filtering of views for those cases where human perception fails.

Can visual measures help us to filter visualizations of interest? The answer is: yes, for two reasons:

Early Reject: The quality metrics Mono, String, Striat, and HSM can be applied to early reject at least 75 % of bad unclassified plots or parallel coordinates related to human perception, if patterns such as correlations, skinny shapes, straight lines, or separated clusters are of interest. Since it was necessary to reveal a sufficient number of votes, we asked our participants to select the best 25 % w.r.t. a metric. Thus, we can only be sure that 75 % of the worst views can be rejected. Since our study framework scales well, it can be used as template for further studies.

Overcome Perception Limitations: Participants reach the limit of their perception if a visualization becomes complex. Our study illustrates this for classified visualization techniques. Quality metrics might help to overcome such a limitation, since they allow the selection of views that show relevant information, even though they might be too complex to be found by a participant.

Perceptivity of Quality Measure vs. the Data: Beyond the answer which quality metric is perceptive, another observation of our study is relevant. Figure 5 shows an uneven pattern regarding the rate of accordance w.r.t. the different data. Similar observations were already given in [Lew12] and [Sea12]. Apparently, the perceptivity of a quality metric depends on the chosen dataset. This is an unexpected and curious behavior. Its reason is unknown. Our impression is, that there is a relation of certain data structures to the users' perceptivity. This is an interesting observation, however, we leave this issue for further studies in the future.

Further Studies and Research Options: We learned from our study that further studies require to treat and to investigate why the perceptivity depends on the underlying data. For this, a vast amount of data should be treated within a study to clarify which sort of structures relate to the user's perception. Based on such a study, novel perception-based quality metrics might be developed. We also discovered a lack of studies regarding multivariate projection techniques. Are there already quality metrics available that detect nonlinear embeddings and trends similar to the user's perception? Also the stability of quality metrics w.r.t. noise has not been systematically investigated yet. How sensitive is a quality metric if the data is slightly disturbed? How strong is the impact regarding the perceptivity of this quality metric? In addition, our study reveals that a set of quality metrics for unclassified visualization techniques is available being related to human perception. For classified visualization techniques, quality metrics are missed being assuredly related to human perception. There is a gap of such metrics. Thus, in the future it is advisable to focus more on the design of quality metrics for classified and complex visualization techniques.

However, from experience we recommend to conduct online studies. In comparison to the traditional laboratory study set up, online studies lead to a large number of participants, are budget-friendly, and they can run over long time periods.

5 Conclusion

We presented a study to depict relations between quality metrics and human perception. There are two aspects that we learned about quality metrics within our study. First, the investigated scagnostics indices were the best regarding their perceptivity properties. Second, there is a strong relation between perceptivity and the underlying data from which a couple of questions arise, which we will treat in the future.

Literature

[Aea10] ALBUQUERQUE G., ET AL.: Improving the visual analysis of high-dimensional datasets using quality measures. In *IEEE VAST* (2010).

[AN07] ASUNCION A., NEWMAN D.: UCI machine learning repository, 2007.

[Bea13] BORKIN M. A., ET AL.: What makes a visualization memorable? *IEEE TVCG (Proc. of InfoVis)* (2013).

[Ber11] BERTINI E.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE TVCG 17* (2011).

[Fis36] FISHER R. A.: The use of multiple measurements in taxonomic problems. *Annals of Eugenics* (1936).

[Hea97] HOFFMAN P., ET AL.: Dna visual and analytic data mining. In *Proc. of the 8th conference on Visualization* (1997).

[HN96] HORTON P., NAKAI K.: A probabilistic classification system for predicting the cellular localization sites of proteins. In *International Conference on Intelligent Systems for Molecular Biology* (1996).

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

[Ken70] KENDALL M. G.: *Rank Correlation Methods*. London, England, 1970.

[Lea12a] LAM H., ET AL.: Empirical studies in information visualization: Seven scenarios. *IEEE TVCG 18*, 9 (2012).

[Lea12b] LEHMANN D. J., ET AL.: Selecting coherent and relevant plots in large scatterplot matrices. *Computer Graphics Forum* (2012).

[Lew12] LEWIS J.M. ACKERMAN M. D. S. V.: Human cluster evaluation and formal quality measures: A comparative study. *Proc. 34th Conf. of the Cognitive Science Society* (2012), 1870 – 1875.

[Sea09] SIPS M., ET AL.: Selecting good views of high-dimensional data using class consistency. *EuroVis 28*, 3 (2009).

[Sea12] SEDLMAIR M., ET AL.: A taxonomy of visual cluster separation factors. *Computer Graphics Forum 31* (2012), 1335–1344.

[Sea13] SEDLMAIR M., ET AL.: Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE TVCG 19*, 12 (2013).

[Tea09] TATU A., ET AL.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *IEEE VAST* (2009).

[Tea10] TATU A., ET AL.: Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In *Proc. of the International Conference on Advanced Visual Interfaces* (2010), AVI '10.

[Tea11] TATU A., ET AL.: Automated analytical methods to support visual exploration of high-dimensional data. *IEEE TVCG 17* (2011).

[TT85] TUKEY J., TUKEY P.: Computing graphics and exploratory data analysis: An Introduction. In *Proc. of the Sixth Annual Conference and Exposition: Computer Graphics 85. Nat.* (1985).

[Wea05] WILKINSON L., ET AL.: Graph-theoretic scagnostics. *IEEE InfoVis* (2005).

[WW08] WILKINSON L., WILLS G.: Scagnostics Distributions. *Journal of Computational and Graphical Statistics 17*, 2 (June 2008), 473–491.

Dr.-Ing. Dirk J. Lehmann is a Postdoctoral Research Fellow at the Computer Science Department at the University of Magdeburg, Germany. In 2008, he received the M.Sc. in Computational Visualistics and in 2012 a Ph.D. in Computer Science from the University of Magdeburg. His research interests focus on flow visualization, information visualization, and visual analytics.

Address: Otto-von-Guericke-Universität, Fakultät für Informatik, Universitätsplatz 2, 39106 Magdeburg E-Mail: dirk@isg.cs.uni-magdeburg.de

Dipl.-Ing. Sebastian Hundt successfully studied Comptational Visualistics and received a M.Sc. from the University of Magdeburg in 2012. His research interests focus on information visualization, computer graphics, and visual analytics. Recently, he left the university for industry.

Address: Otto-von-Guericke-Universität, Fakultät für Informatik, Universitätsplatz 2, 39106 Magdeburg

Prof. Dr.-Ing. Holger Theisel is professor for Visual Computing at the Computer Science Department at the University of Magdeburg, Germany. In 1994, he received the diploma in Computer Science, in 1996 a Ph.D. in Computer Science, and a habilitation (venia legendi) in 2001 from the University of Rostock. His research interests focus on flow and volume visualization as well as on CAGD, geometry processing and information visualization.

Address: Otto-von-Guericke-Universität, Fakultät für Informatik, Universitätsplatz 2, 39106 Magdeburg E-Mail: theisel@ovgu.de