Eurographics Conference on Visualization (EuroVis) 2015
H. Carr, K.-L. Ma, and G. Santucci
(Guest Editors)

Volume 34 (2015), Number 3

Data-driven Evaluation of Visual Quality Measures

M. Sedlmair[1] and M. Aupetit[2]

[1]Visualization and Data Analysis Group, University of Vienna, Austria
[2]Qatar Computing Research Institute, Doha, Qatar

Abstract

Visual quality measures seek to algorithmically imitate human judgments of patterns such as class separability, correlation, or outliers. In this paper, we propose a novel data-driven framework for evaluating such measures. The basic idea is to take a large set of visually encoded data, such as scatterplots, with reliable human "ground truth" judgements, and to use this human-labeled data to learn how well a measure would predict human judgements on previously unseen data. Measures can then be evaluated based on predictive performance—an approach that is crucial for generalizing across datasets but has gained little attention so far. To illustrate our framework, we use it to evaluate 15 state-of-the-art class separation measures, using human ground truth data from 828 class separation judgments on color-coded 2D scatterplots.

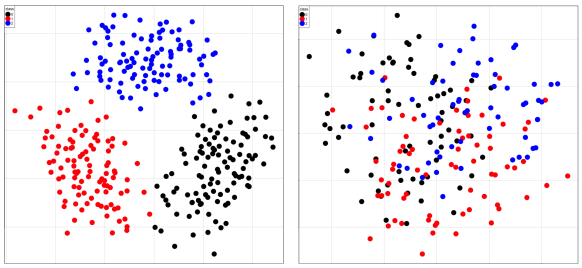
Categories and Subject Descriptors (according to ACM CCS): H.5.0 [Information Interfaces and Presentation]: General

1. Introduction

The general idea behind visual quality measures is to algorithmically mimic the human perception of patterns such as class separability, correlation, or outliers. Previous work has shown that such perception-centered measures have a great potential for improving visualization tools and guiding human analysts, for instance, to find good 2D projections in high-dimensional datasets [SNLH09, TAE*09, WA05].

We focus on how to evaluate such visual quality measures. Given their goal of imitating human perception, the effectiveness of such measures seems to be best evaluated in comparison to human judgments [BTK11, TBB*10]. The most common current form of evaluation is *usage examples*: measures are applied to a small set of example datasets, results are shown in the paper, and the reader is asked to individually compare their own (human) judgment with the measure's results. As an evaluation method, however, usage examples give only a limited and selected view [IIC*13]. To overcome these limitations, researchers have suggested to empirically study quality measures with either *controlled user studies* [SNLH09, TBB*10], or manual *qualitative data studies* [STTM12]. However, there seems to be disagreement on how effective quality measures are, even among this small set of empirical work. Carefully reading these studies suggests that the different findings might likely stem from the different methodological approaches. The largest deficiency of user studies seems to be the small number of datasets used, which considerably limits the generalizability of findings to other datasets. Qualitative data studies, on the other hand, include a large number of different datasets, however, come with a high time cost of manual human judgments.

To overcome these limitations, we propose an alternative evaluation approach for quality measures, which we call *data-driven evaluation*. The basic idea is to take a machine learning perspective and to automatically compare measures based on how they would predict human judgements on unseen new datasets. To illustrate our ideas, we instantiate the framework for the evaluation of *visual separation measures* that seek to algorithmically quantify the degree of how well a class is visually separable in scatterplots as shown in Figure 1. To evaluate such measures, we first take a large set of color-coded scatterplots with human class separation judgments ("ground truth" data). Assuming these scatterplots to be a representative sample of the larger "population" of all scatterplots, we then use bootstrapped classification techniques to predict how well a measure would predict human class judgments on unseen scatterplots. We argue that this data-driven approach has several advantages over current approaches: improving external validity by generalizing over datasets, putting a stronger focus on actual human class perception, and providing a more automatic way of evaluation that is, nevertheless, grounded in human perception.



(a) (b)

Figure 1: *Examples of two scatterplots, each showing synthetic data with three different classes, color-coded in the plots. Scatterplot (a) visually separates the classes nicely as can be seen by a human; an effective visual separation measure should score high. In scatterplot (b) the classes are not visually separable; the measure should score low.*

We use this approach to *evaluate a set of 15 parametric and non-parametric separation measures* from the visualization and machine learning community (Section 4). The study adds to the current small body of empirical evaluations of separation measures in visualization. Finally, we discuss benefits and limitations of our approach, and derive guidelines for quality measure evaluation and future work in this area (Section 5). In summary, our work makes the following contributions:

- a general framework for data-driven evaluation of quality measures,
- a concrete instantiation of the framework for evaluating visual separation measures,
- an empirical study of 15 separation measures using this framework, and
- a set of guidelines for visual quality measure evaluation.

2. Related Work

The question of how to evaluate visualization research has gained much attention, and many have argued for a facetted spectrum of different evaluation methods in visualization research [Car08, IIC*13, LBI*12].

In our work, we focus on the evaluation of *visual quality measures* (or metrics, or indices). Starting with the venerable work of Friedman and Tukey on projection pursuit [FT74], quality measures have been used for various purposes in visualization, such as supporting human analysts in finding interesting projections in high-dimensional scatterplot matrices [SNLH09, TAE*09, WA05], ordering axes in parallel coordinate plots [DK10, TAE*09], or guiding data abstractions [JC08]. Our main focus is on *visual separation measures*, a particularly vibrant area of research in visualization [AEM11,MMdALO15,STTM12,SNLH09,TAE*09]. In our study, we test 15 of these measures.

Quality measures have been evaluated with different goals and different methodological approaches. Many of the original *technique papers* that proposed quality measures used a small set of example datasets and showed visual encodings, such as scatterplots, along with how they were judged by the measure. Isenberg et al. characterized such usage example-based evaluations as one of eight typical evaluation strategies in visualization research [IIC*13]. Sips et al. [SNLH09] additionally conducted a small user study and found a positive correlation between scatterplot rankings from humans and measures. Naturally, evaluations in these original technique papers were focused on confirmatory evidence, underlining the benefits of the proposed measures.

Richer and more objective comparisons are usually done in pure *evaluation papers*. Most closely related to our work, are the studies by Tatu et al. [TBB*10] and Sedlmair et al. [STTM12]. Tatu et al. [TBB*10] conducted a user study with 15 participants comparing four separation measures on one dataset. While two measures seemed to work better than the others, the general conclusion was that all tested measures performed well. They also proposed a framework for conducting user studies for quality measure evaluation derived from their study. Sedlmair et al. [STTM12] pointed out that a major drawback of this framework is that it focuses on a limited number of example datasets. They therefore conducted a qualitative data study to further evaluate the effectiveness of the two separation measures that performed best. In a time-intensive process, two expert coders manually judged the separation effectiveness of these measures on a broad set of scatterplots. They found a high rate of failure cases, contradicting previous work. Other studies on related measures also fall into this dichotomy. Lewis et al. [LAdS12] conducted a user study on clustering measures (no overlap between color-coded classes). Wilkinson and Wills [WW08] conducted a quantitative data study to better understand the empirical distributions of scagnostic measures without focusing on human judgments.

Our approach seeks to combine the benefits of both approaches by using machine learning techniques. In that sense, our idea is similar to Albuquerque et al.'s of using psychophysics studies to learn visual quality measures [AEM11]. However, we use different learning approaches, different data, and overall follow a different goal, namely designing a better evaluation framework, not proposing a novel measure. We are also comparing a broader set of measures than previous studies, including 35 parameterized measure instances (derived from 15 different measures). Previous studies including human judgments tested seven [LAdS12], four [TBB*10], two [STTM12] and one measure respectively [SNLH09].

3. Data-driven Evaluation Approach: Overview

Figure 2 provides a global overview of our data-driven evaluation framework for visual quality measures. The major

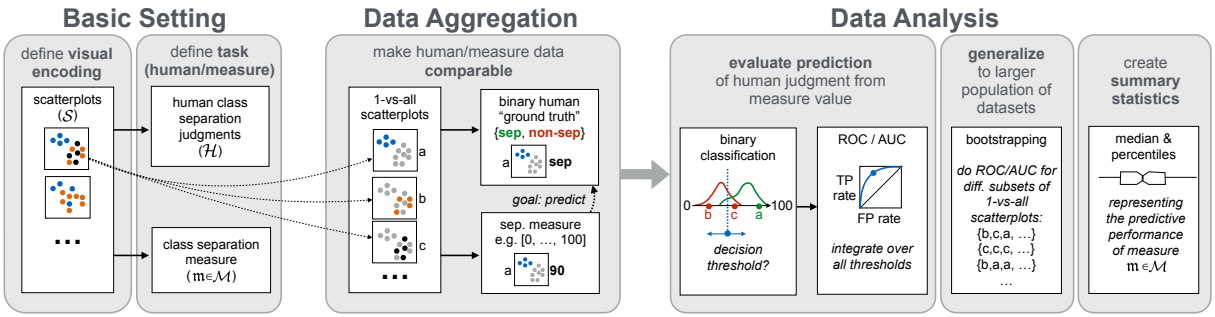


Figure 2: *Overview of our data-driven framework for quality measure evaluation. The overall process is broken down into three major steps. The gray boxes show the generic steps of our framework, while the white boxes within them describe how we instantiated the framework for testing class separation measures in color-coded scatterplots.*

goal of the framework is to provide researchers with a way to evaluate how well visual quality measures predict human judgements.

The general framework consists of three major steps. In the *basic setting* step a researcher first needs to define which visual encoding techniques, human tasks, and which quality measure she wants to test. For our purpose, we instantiate the framework with the specific setting of color-coded 2D scatterplots and the task/measures of class separability (white boxes in Figure 2). The general framework (gray boxes), however, can also be instantiated for other visual encodings, tasks and measures. Before testing human judgement and measure data, further cleaning and aggregation might be necessary (*data aggregation*). In our case, we, for instance, extrapolate multi-class scatterplots into 1-vs-all scatterplots with a "target-class" and all other classes merged into an "other-class". The most crucial step is the *data analysis* step. Here, the aggregated data is used to evaluate how "good" a measure would predict the human judgment on unseen data. For instance, we learn a classifier that predicts a binary human class judgment (separable or not) from a measure's value. Machine learning techniques like bootstrapping or cross-validation help to generalize the results beyond the actual sample data—scatterplots in our case—that have been used in the study. Repeating this process for different quality measures will then allow to compare their effectiveness in terms of how well they predict human judgments.

In the following, we focus on our specific instantiation of this framework for evaluating class separation measures in color-coded scatterplots, and provide details and formal definitions for each of these steps. We use bold fonts when we refer to specific components in Figure 2.

3.1. Basic Setting

Our base setting has three different components:

- The most fundamental component is the set of 2D **scatterplots** $\mathcal{S} = \{\mathfrak{s}_1, \ldots, \mathfrak{s}_S\}$. We focus on classified data, that is, each point in a scatterplot has a color that encodes its unique class membership to a class $\mathfrak{c} \in \mathcal{C}_\mathfrak{s}$, from a set of k classes in this scatterplot $\mathcal{C}_\mathfrak{s} = \{\mathfrak{c}_1, \ldots, \mathfrak{c}_k\}$. For each class within a scatterplot, we need two things.
- First, we need **human judgments** $\mathcal{H} = \{\mathfrak{h}_1, \ldots, \mathfrak{h}_H\}$ on the visual separability of classes. This judgment could, for instance, be provided as a natural number quantifying "how separable" a class is by a human.
- Second, for each class we also have "judgments" from a set of **separation measures** $\mathcal{M} = \{\mathfrak{m}_1, \ldots, \mathfrak{m}_M\}$.

The goal is to find the optimal measure $\mathfrak{m}^\star \in \mathcal{M}$ that most accurately predicts human judgments on unseen scatterplots.

3.2. Data Aggregation

Given the goal of predicting human class judgements from separation measures, we first need to ensure that human and measure judgments are comparable. To do so, we aggregate the data from the basic setting in two ways: (i) extrapolating 1-vs-all scatterplots, and (ii) aggregating human judgments.

i) Extrapolating 1-vs-all scatterplots: One of our design goals is to base our framework on the actual low-level perceptual task of visual class separability. Hence, we need judgments of the separability of actual classes rather than integrated judgments over all classes of a scatterplot, as commonly produced by current state-of-the art separation measures. Many of the current measures operate on entire scatterplots not on separate classes. That is, in multi-class scatterplots these measures cannot reliably tell the separability of actual classes, but only the whole scatterplot.

To avoid internally adapting these measures, we decided to split each multi-class scatterplot $\mathfrak{s} \in \mathcal{S}$ into several **1-vs-all scatterplots**, one for each class $\mathfrak{c}_t \in \mathcal{C}_\mathfrak{s}$. $\mathfrak{c}_t$ is the "target-class", and we want to test how separable $\mathfrak{c}_t$ is from all other classes at once, which thus get combined into the "other-class" $\mathfrak{c}_o = \mathcal{C}_\mathfrak{s} \backslash \mathfrak{c}_t$. Formally, we refer to these 1-vs-all scatterplots as our elementary data items $d_{\mathfrak{s},\mathfrak{c}_t}$, which are uniquely determined by a specific target-class $\mathfrak{c}_t \in \mathcal{C}_\mathfrak{s}$ within

a specific scatterplot $\mathfrak{s} \in \mathcal{S}$. For each of these 1-vs-all scatterplots $d_{\mathfrak{s},\mathfrak{c}_t}$, we have a human judgment and a measure judgment about the separability of class $\mathfrak{c}_t$. The human judgment of the separability of class $\mathfrak{c}_t$ can be considered as the **ground truth** in our setting, which we want to **predict** using a separation measure. The assumption then is that we can use these 1-vs-all judgements as a perceptual surrogate of class separability of the underlying multi-class scatterplots. We will further discuss this assumption in Section 5.

ii) Aggregating human judgments: There might be several human judgments available for the same target-class within the same scatterplot, either from different humans or from the same human under different conditions. In this case, we need to aggregate these judgments in order to come up with a single certain value reflecting the human judgment as a ground truth. Moreover, human judgments might be multi-valued, for instance, based on a Likert scale from 1 to 5 [SMT13]. In our framework, we propose to simply aggregate this multi-valued judgment scale to a binary scale $\{0,1\}$. That is, the human judgement(s) of a target class $\mathfrak{c}_t$ gets aggregated to either 1-separable (or short *sep*), or 0-non-separable (*non-sep*).

We now have a dataset $\mathcal{D}_\mathcal{S}$ that we can use for further analysis. Denoting the aggregated human judgement into sep/non-sep as a function $\underline{\mathfrak{h}}$, we can now formally describe this dataset as:

$$\mathcal{D}_\mathcal{S} = \{(d_{\mathfrak{s},\mathfrak{c}_t}, \underline{\mathfrak{h}}(d_{\mathfrak{s},\mathfrak{c}_t})) | \mathfrak{s} \in \mathcal{S}, \mathfrak{c}_t \in \mathcal{C}_\mathfrak{s}\}$$

In other words, $\mathcal{D}_\mathcal{S}$ is the set of all 1-vs-all-scatterplots $d_{\mathfrak{s},\mathfrak{c}_t}$ labeled with the aggregated human judgement $\underline{\mathfrak{h}}(d_{\mathfrak{s},\mathfrak{c}_t})$ that tells us how visually separable the class $\mathfrak{c}_t$ was judged by one or more humans. $\mathcal{D}^1_\mathcal{S}$ denotes the subset that was judged as 1 (*sep*), $\mathcal{D}^0_\mathcal{S}$ the subset judged 0 (*non-sep*). Other data that comes in this form can directly be plugged into our instantiated framework.

Undoubtedly, this binary differentiation between *sep* and *non-sep* limits the rich human perception into a coarse-grained dichotomy. However, given that our understanding of perceptual class separability is still at a preliminary stage [STTM12], it gives us a way to ensure a certain degree of reliability of human judgment data. We propose this binary aggregation as a first step and leave more fine-grained predictions for future work.

3.3. Data Analysis

Now we focus on how we can use this dataset $\mathcal{D}_\mathcal{S}$ to find the best separation measure.

A binary classification setting: State-of-the-art separation measures usually give a scalar value. This value is supposed to be monotonically increasing with the actual separation of the target-class, imitating human perception. Therefore, a decision threshold has to be set upon this separation measure in order to decide about the label *sep* or *non-sep* of a class in a scatterplot. Classes whose separation measure is greater than the threshold value would be assigned the *sep* label, and *non-sep* label if less than the threshold. A separation measure together with a decision threshold value defines a classifier. In order to maximize the probability that such a classifier will predict the correct label of new unseen data (scatterplots that have not been judged by humans), we need to identify which separation measure to use and which value to set the threshold to. This is a standard binary classification setting where the predicted label can be positive (*sep*) or negative (*non-sep*) and can be true or false regarding the correct human ground-truth label to be predicted. There is a vast literature on this topic (see for instance Bishop [Bis06]).

ROC and AUC analysis: Framing the measure evaluation as a classification problem, we can argue about two possible kinds of errors: the Type I error or False Positive (FP) when the classifier predicts the *sep* label while the human ground-truth data label is *non-sep*; and the Type II error or False Negative (FN) when the outcome from the classifier is *non-sep* while it should be *sep*. The quality of a binary classifier can be evaluated by counting the amount of True Positives (TP) against the amount of FP. A good classifier would have a high number of TPs and a low number of FPs.

The value of the decision threshold, however, greatly impacts the tradeoff between the two types of errors, FPs and FNs. In the extreme case, if the range of the outcome from the separation measure is $[0,1]$, setting the decision threshold to 0 will end up classifying any data as positive. That is, it would lead to no FNs but a large number of FPs equal to the number of truly *non-sep* data all being predicted as *sep*. Setting the decision threshold to 1 will end up with the exact opposite result. Hence, setting the decision threshold depends strongly on the intended use and on how many FNs or FPs are acceptable to the user. The left side of Figure 3 visually illustrates this tradeoff when setting a decision threshold.

To guarantee an objective evaluation, we do not want to make assumptions on any specific intended use. Therefore, we vary the threshold from the minimum to the maximum values of the separation measure. When drawing the TP rate against the FP rate for all threshold variations, we get the Receiver Operating Characteristic curve (ROC curve, see Figure 3, right side) [Faw06]. The closer the curve passes to the upper left corner and the farther from the diagonal of this diagram, the better the classifier is. The Area Under the ROC curve (AUC)—the integral of the ROC curve on the $[0,1]$ domain—can then be used as a summary statistics of a classifier's quality independent of a specific threshold value.

Formally, the $AUC_{\mathfrak{m},\underline{\mathfrak{h}},\mathcal{D}_\mathcal{S}}$ score maps a set of pairs of outcomes from the separation measure $\mathfrak{m}$ and the human judgments $\underline{\mathfrak{h}}$ relative to the same data $\{(\mathfrak{m}(d_{\mathfrak{s},\mathfrak{c}_t}), \underline{\mathfrak{h}}(d_{\mathfrak{s},\mathfrak{c}_t}))\}_{\mathcal{D}_\mathcal{S}}$ to a value within $[0,1]$. The closer the AUC is to 1, the better the classifier is. $AUC_{\mathfrak{m},\underline{\mathfrak{h}},\mathcal{D}_\mathcal{S}}$ is the probability that the separation measure $\mathfrak{m}$ will assign a higher value to a datum randomly

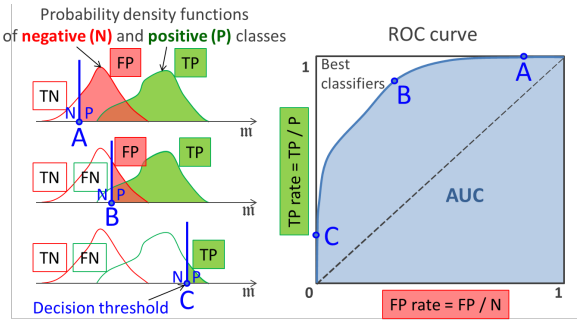


Figure 3: ***(left)*** *The horizontal axis represents a continuum of all possible values a separation measure ($\mathfrak{m}$) can take. Onto this axis, one can now draw all measure values as a probability density function. Measure values associated with separable classes in the human ground truth data are drawn in green. The ones associated with non-separable classes in red. A classifier is now defined by a decision threshold (blue dot) acting on this axis. Any data located over the decision threshold is assigned to the positive class (P, that is, it would predict separable), else to the negative class (N, predicting non-separable). By moving the threshold along the axis the tradeoff between false positives (FP) and false negatives (FN) changes drastically, as indicated by the three different threshold positions A, B, and C.* ***(right)*** *The Receiver Operating Characterictic (ROC) curve results from varying the decision threshold value from the lowest available separation measure value to the highest one. The Area Under the ROC Curve (AUC) quantifies the quality of the classifier over all possible decision threshold values.*

chosen in the set of truly separable data $\mathcal{D}_{\mathcal{S}}^{1}$ than to a datum randomly chosen in the set of truly non-separable data $\mathcal{D}_{\mathcal{S}}^{0}$. ROC and AUC are graphically illustrated in Figure 3.

Bootstrapping: The dataset $\mathcal{D}_{\mathcal{S}}$ is only a sample of the unknown population of all possible data (1-vs-all scatterplots). Computing the AUC for this sample only would provide a *descriptive* statistic of this particular sample. Instead, what we would like to have is an *inferential* statistic that predicts, or generalizes, from the specific representative sample $\mathcal{D}_{\mathcal{S}}$ to the larger population of scatterplots.

To facilitate this generalization, we use bootstrapping [ET93]. Bootstrapping is a well-understood resampling technique that simulates new data by generating same-size random samples with replacement of the observed dataset. In each bootstrapping sample random data points (1-vs-all scatterplots) are missing while others are weighted in different ways. This characteristic allows us to make inferential statements as it imitates random draws from the unknown, underlying data population. We chose bootstrap as it is straightforward to implement with only one parameter to tune—the number of bootstrap samples. An alternative, but equivalent approach would have been to use cross-validation.

A bootstrap sample $\mathcal{D}_{\mathcal{S}_{boot}}$ is a specific multiset based on $\mathcal{D}_{\mathcal{S}}$ with the same total number of elements, but where some of them are missing while others are replicated multiple times. Consider, for instance, a set $\{a,b,c\}$, with a, b, c being labeled 1-vs-all scatterplots in our case. Valid bootstrap samples of this set might look like $\{a,b,b\}$, $\{c,c,c\}$, or $\{a,b,c\}$.

Using this approach, we generate B bootstrap samples of $\mathcal{D}_{\mathcal{S}}$. The AUC is computed for each of these bootstrap samples. We therefore get a set of different AUC values whose distribution approximates the distribution of AUC values we would observe by drawing many random data samples from the population. We can then use the average of the AUC bootstrap distribution to estimate the expectation of the AUC value over the whole population. The greater the average of the AUC bootstrap distribution of a separation measure, the higher the probability that this separation measure will give a high value to a truly separable unseen data, than to a truly non-separable one.

Summary statistics: The AUC bootstrap distribution can then be used to compare the measures $\mathcal{M}$. Specifically, we propose that the best separation measure $\mathfrak{m}^* \in \mathcal{M}$ is the one with the highest average of the AUC bootstrap distribution. The separation measures can be ranked according to their AUC bootstrap average value. Bootstrap percentiles can be displayed as whisker plots to evaluate the variance, as indicated at the very right in the overview picture (Figure 2).

4. Empirical Study

We now use this framework to conduct a study of 15 state-of-the-art separation measures.

4.1. Data

Given our inferential evaluation approach, we sought to base our study on a *large* and *representative* set of *reliable* human judgments. As generating such a human ground-truth dataset is very expensive, we decided to take a large and carefully crafted dataset on human class perception from a previous study by Sedlmair et al. [SMT13]. In their work, Sedlmair et al. had two human expert coders sifting through 816 2D, 3D, and multi-D scatterplot matrices. The coders individually judged the separability of all color-coded classes in these scatterplots on a scale from 1 (not separable at all) to 5 (nicely separable). The scatterplots were generated from a representative sample of 75 different datasets, both real and synthetic, reduced with four different dimension reduction techniques, such as Principal Component Analysis (PCA) [SMT13]. Sedlmair et al. used this data to evaluate scatterplot visual encoding and dimension reduction technique choices. We use it to evaluate visual separation measures in scatterplots.

An important question is how reliable such human judgments are. Sedlmair et al. argued that the diversity in datasets

m	Short description of measure (and parameters if applicable)	Param.	Reference
ABTN	Between-class average distances	-	[LAdS12]
AWTN	Within-class average distances	-	[LAdS12]
ABW	Between-class ABTN over within-class AWTN average distances ratio	-	derived by us
WII	Average between-class over average within-class distances weighted by the respective size of the classes	-	[Str02]
CAL	Centers-of-mass between-class square distances over points-to-centers-of-mass within-class square distances	-	[CH74]
LDA	Centers-of-mass between-class distances over points-to-centers-of-mass within-class distances ratio with optimal linear transformation of the points to maximize this ratio	-	[Fuk90]
DUNN	Maximum within-class distance over the minimum between-class distance ratio	-	[Dun74]
GAM	Normalized comparison of numbers of within-class distances smaller or greater than between-class distances	-	[BH75]
SIL	Difference of between-class and within-class average distances normalized by the maximum of them	-	[Rou87]
HM	Average differences between distances from each point to its other-class and its same-class nearest-neighbor	-	[GBNT04]
CS	Average proportion of same-class neighbors of each point in minimum spanning tree	-	[MMdALO15]
DSC*	Proportion of points x whose the nearest class-center-of-mass belongs to the same class as x	-	[SNLH09]
CDM_K	Pixel-wise class-density differences with class-density estimated at pixel z as the inverse distance to its K^{th} nearest point of this class (here, $K \in \{1,2,3\ldots,10\}$)	K	[TAE*09]
DC_σ	Average of the class entropy for each pixel computed over the classes of its ε-neighbors (here, $\varepsilon = \sigma * \Delta(X_\mathfrak{s})$ where $\sigma \in \{0.1\%, 0.2\%, 0.5\%, 1\%, 2\%, 5\%, 10\%, 20\%\}$ and Δ is the maximal Euclidean distance between points of the evaluated scatterplot)	ε	[SNLH09]
HDM_{Nb}	Entropy measure of the classes in each cell and their adjacent cells in a square-grid partition. N_b is the number of cells in each direction (here, $N_b \in \{5, 10, 20, 40, 80\}$), and β the level of neighboring cells (here, we used a fixed $\beta = 1$, i.e., 8 neighboring cells).	N_b, β	[TAE*09]

Table 1: *List of tested separation measures, ordered by algorithmic similarities. The last three measures are parametric; we tested them with different parameter settings as indicated. The best measure in our experiments is marked with an *.*

outweighs the subtle differences in human judgments of class separability. This argument is supported by a high inter-coder reliability in their study (Krippendorff's alpha was 0.858), but also by other empirical studies from Lewis et al. [LAdS12], and from Tatu et al. [TBB*10]. Both studies indicate little variance in class separability judgments among humans. Following this empirical evidence, we assessed these human judgments as reliable for our purpose [SMT13].

Given the different purpose of our study, we needed to filter and clean the data in order to get it into a suitable form for our intended use. This process included five steps (technical details can be found in the supplemental material):

1. We excluded 3D and multi-D scatterplots.
2. We needed to correct for misalignments of points in the scatterplots caused by the un-normalized scales used in the original study. We used different image processing algorithms for that purpose.
3. To guarantee a fair comparison between human judgments and measures, we removed fully occluded points which could not be seen by the human coders.
4. After that, we excluded scatterplots for which only one class remained visible. We then also removed datasets with more than 14 classes and those with over 1000 (visible) points due to limitations of human perception; reliable judgments cannot be guaranteed over a certain scale.
5. The human judgment in Sedlmair et al.'s study was done by two coders using a 5-point Likert scale. Let (u,v) be the judgments of the two coders, then we aggregate the judgments as follows: $(5,5),(4,5),(5,4)$ were aggregated into *sep*, and $(1,1),(1,2),(2,1)$ into *non-sep*. We excluded all other instances due to lower reliability with respect to our binary classification setup.

After cleaning, we ended up with 828 data items $d_{\mathfrak{s},\mathfrak{c}_t}$, that is, 1-vs-all scatterplots, from overall 56 multidimensional datasets. 408 1-vs-all scatterplots stemmed from synthetic datasets, and 420 from real datasets.

4.2. Separation Measures

Our goal was to include a broad set of different measures into our study. With this goal in mind, we selected a set of 15 state-of-the-art measures discussed either in the visualization or the machine learning community. Table 1 shows the measures we tested and gives a brief summary of how they operate. For measures that needed to be parameterized, we tested between 5 and 10 different parameter settings. Overall, this process led to 35 measure instances that we evaluated. For all measures, the higher the measure value is, the greater the separation of the target class. In the supplemental material, we provide additional details, including mathematical definitions, as well as a high-level algorithmic classification of these measures in terms of locality criteria, notion of discrepancy, and computational complexity.

4.3. Results

Using the cleaned dataset, we generated 10,000 bootstrap samples, a number that generated highly consistent results at 0.1% precision (see supplemental material for more information). We computed the AUC bootstrap distribution for each of the 35 measure instances. Each bootstrap sample had the same number of items as the underlying dataset $d_{\mathfrak{s},\mathfrak{c}_t}$, that is, 828. The experiments were run on a standard desktop computer using Matlab.

Figure 4 shows the different measures ranked from top to bottom in decreasing order of the AUC bootstrap average. We now highlight some interesting findings.

Winner: We found that the *DSC* measure by Sips et al. [SNLH09] gave the best AUC bootstrap average of 82.5%. *DSC* is the proportion of points whose nearest class-center-of-mass (class centroid) belongs to their own class. This result means that *DSC* should give a greater value to a truly human-separable unseen data than to a truly non-separable one in 82.5% of the cases.

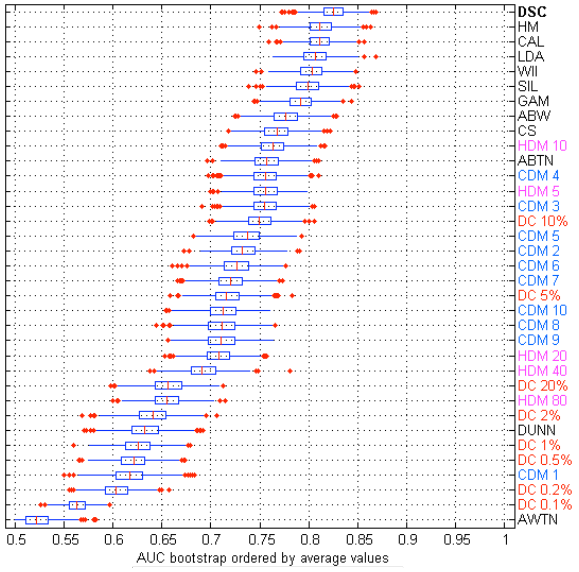


Figure 4: *Results of our study. Each row represents a separation measure with a box plot encoding the AUC bootstrap distribution: median (red center line); interquartile range (IQR), i.e., from 25 to 75 percentile (box);* 25/75 *percentile +/- 1.5 times the IQR, including* 99.3% *of the data (whiskers); and outliers (red points). The measures are ranked in decreasing order of the AUC bootstrap average. A score of* 0.5 *denotes a random guess, while* 1 *would indicate perfect separation prediction on unseen (but similar) data.*

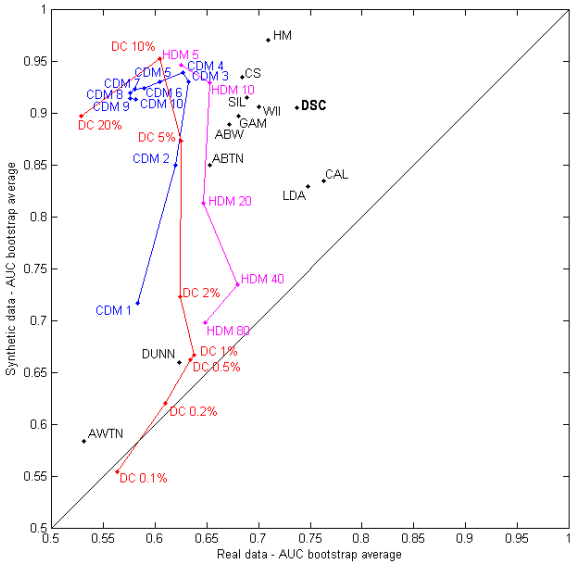


Figure 5: *Results of real and synthetic datasets separately. Each separation measure is represented as a point with coordinates encoding the AUC bootstrap average on real datasets (x-axis) and synthetic datasets (y-axis). Observations: (1) Points fall in the upper left part, that is, measures score better on synthetic than on real datasets. (2) Points of parametric measures are colored and connected in sequence (by parameter value). AUC scores vary greatly, and the parameter dependence is stronger for synthetic datasets.*

In a previous study by Tatu et al. [TBB*10], *DSC*† was also found the best measure, along with another measure, *HDM*. Our study confirms that *DSC* is among the best measures, despite the different methodological approaches in terms of used datasets, users and the evaluation process. The *DSC* pole position in both studies is a sign that *DSC* captures characteristics of the scatterplots that might be relevant to the human visual separation process in general. However, our study gives *HDM* a worse position than *DSC* compared to Tatu et al. with an AUC bootstrap average of 76.4% for *HDM*10. Being a histogram-based entropy measure, *HDM* captures characteristics of the scatterplots different from the ones captured by point distance-based *DSC* measure. The evaluation using our framework suggests that *HDM* is less robust and generalizable to other scatterplot characteristics than *DSC*. In the following, we further study dependencies on dataset characteristics and parameters to better understand the different findings.

Dataset characteristics (synthetic vs. real): As our set of scatterplots was originally obtained from synthetic and real datasets [SMT13], we explored how the measures performed on both synthetic and real sets separately. The results are shown in Figure 5.

We found that almost all of the separation measures have a higher AUC bootstrap average score for synthetic datasets than for real datasets, with differences up to 36.6% in the AUC score for *DC*20% (53.1% on real and 89.7% on synthetic datasets). Note that such a difference indicates a jump from very accurate for synthetic (close to 100%) to almost random for real data (close to 50%). Therefore, using only synthetic data to evaluate separation measures would give an optimistic view of these measures. Our study thus confirms that most of the separation measures are able to quantify correct separation in simple cases but do not perform with high accuracy on more realistic data [STTM12]. In terms of the *HDM* measure that was found best in Tatu et al.'s study [TBB*10] but not ours, Figure 5 shows that *HDM*10 scores equal as *DSC* on synthetic datasets, while *DSC* is clearly better on real datasets. So a reason for the discrepancy of our results is the different sets of scatterplots used to evaluate the measures.

These results underline the importance of having large and representative datasets to evaluate separation measures, which is fostered by our framework. This finding is fur-

† *DSC* was named *CCM* in this study

ther supported by a sanity check experiment with 30 simple, highly synthetic datasets, for which 26 of the 35 measures tested got a perfect 100% AUC bootstrap average. This experiment is further described in the supplemental material.

Parameters: Three measures were parametric in our study: *CDM*, *HDM*, and *DC*. Testing 5-10 different parameterizations for each of these, we found that different parameterizations led to considerable differences of the AUC bootstrap values as displayed in the Figure 4. This finding contradicts comments in the original measure papers that these measures are relatively insensitive to parameter choices [SNLH09, TAE*09], but confirms similar observations by Sedlmair et al. [STTM12]. The best *CDM* we found was for $K = 4$, the best *HDM* for $N_b = 10$, and the best *DC* for $\sigma = 10\%$ (Table 1 gives definitions of these parameters).

When separating real and synthetic datasets as in Figure 5, we see that parametric measures are mostly influenced by synthetic datasets (scores differ mostly along the vertical, synthetic axis). Depending on the particular datasets used in the original papers [SNLH09, TAE*09], this finding could explain that no strong dependency on the parameter had been observed in their case. We also observe that for *DC* and *CDM* both the synthetic and the real scores increase with the parameter value up to some maximum and then decreases again. These optima indicate good parameter settings, which however differ for synthetic and real data.

Toward improving separation measures: The winner *DSC* was closely followed by several other measures *HM*, *CAL*, *LDA*, *WII*, *SIL*, and *GAM*. While all these measures score around 80%, remember that a 50% score is a random guess. So, there is still more than a third of the way to go in order to get to a theoretically perfect measure. This finding echoes Tatu et al.'s message suggesting room for further improvements [TBB*10].

Analyzing the characteristics of the best ranked measure *DSC*, but also the second best *HM*, reveals that both involve a non-parametric Nearest Neighbor approach and give a distinct role to between and within class distances. These characteristics might therefore play an important role in the human visual separation process and could give hints to develop even better separation measures. Regarding the specific case of parametric separation measures, we used the same parameter value for all the scatterplots as it is standard in the literature. However, it is possible that the human visual separation process adapts the scale parameter to each scatterplot, and possibly to each part of a scatterplot. Taking such characteristics into account might provide additional room for improvements on the parameterization scheme of parametric separation measures.

5. Discussion

While the above analysis focuses on visual separation measures, we believe that the general ideas behind our framework are also applicable to other visual encodings and quality measures. Similarly, the actual analysis steps within the framework can be easily replaced by alternatives: instead of a classifier one might, for instance, use regression models if the goal is to learn a continuous mapping, say for correlation [HYFC14, RB10]; or, instead of bootstrapping one can use other ways to facilitate generalizability such as cross-validation, just to name a few alternatives. To use a simple metaphor, we see the white boxes that describe these characteristics in Figure 2 as a set of building blocks that can easily be replaced by others.

For our study, we used a specific scatterplot dataset from previous work [SMT13]. While this dataset stems from dimensionally reduced (DR) data, our approach would, of course, similarly work with other types of scatterplot data. In fact, we believe that the visual cluster separation characteristics of DR data, specifically from linear techniques such as PCA, are similar to other scatterplots, such as scatterplots stemming from axis-aligned projections (as in "normal" scatterplots). However, to the best of our knowledge an explicit comparison of characteristics has not been done yet.

We now revisit the major aspects of our framework and summarize them as a set of four guidelines for visual quality measure evaluation.

Predict how measures would perform on previously unseen data: The major goal of our evaluation framework is to predict how visual separation measures would perform on unseen data. That is, we want to generalize findings over datasets. In inferential statistics this ability to generalize is called *external validity*. In visualization, human-computer interaction (HCI), and psychology, generalizing over human subjects is very common (and important). In contrast, an inferential lens to generalize over datasets has not gained much attention so far. Given the very nature of visualization research being tied to users and data [PVW09], however, we believe that such a lens is important when evaluating our research in general, and quality measures in particular.

A natural question is how well we can generate representative subsets of the larger "population" of all (relevant) datasets. In that respect, we echo the recommendations by Sedlmair et al. [STTM12]: use large samples of datasets to study measures; gradually integrate more and more new datasets into evaluations; and, take care of *ecological validity* by integrating real datasets as far as possible. Our 828 1-vs-all scatterplots cover many different separation characteristics [STTM12] and, thus, we are confident that it is a good first step. Nevertheless, we hope that others will extend upon this work, adding new data with new characteristics.

Separate human judgment studies from measure studies: The most time consuming part of testing visual quality measures is to get reliable human judgments for a broad selection of different datasets. One way to overcome such limitations is to use crowdsourcing services such as Amazon's Mechanical Turk that allow outsourcing tasks to a large population of

users [HB10]. Such studies work particularly well with low-level perceptual tasks that are existent in the general public. Current research indicates that the perception of class separability, as focused on in our study, might fulfill this criterium [LAdS12]. For other, more complex tasks, however, care needs to be taken in terms of the reliability, and expert judgments might be necessary [LvdMdS12].

Orthogonal to that, we want to advocate a separation of (a) running perceptual studies with human subjects, from (b) evaluating quality measures. Running perceptual studies with human subjects can help us to gradually build up a more and more reliable "human ground truth" dataset. Using this ground truth data as an input, we can then evaluate separation measures in a more automatic fashion, without directly involving human subjects. While such calls have been made before [TBB*10], we are the first to offer an evaluation framework that fosters this separation.

Be objective with respect to the intended use: When looking into previous work, we found that the interpretation of measure evaluation results seemed to be strongly related to the subjective expectations of researchers. When testing measures intended for pre-selecting views in a scatterplot matrix, Sips et al., for instance, noted that *"even for a large number of dimensions the [tested measure] correlates to over 50% with people's judgment of good views"* [SNLH09]. With the intention to use measures for guiding crucial visual encoding and abstraction choices, Sedlmair et al. wrote: *"The two studied measures failed to provide a robust and reliable judgment in nearly 50% of our cases"* [STTM12]. Both argue about 50%, but with very different interpretations, positive and negative respectively.

We argue for more objectiveness in analyzing and interpreting results. One way is to be upfront about the intended use and the notion of what makes a "good" measure. In our framework, for instance, we tried to be objective with respect to different false positive (FP) / false negative (FN) settings, by integrating over all of them. Additionally, our framework can easily be tuned to evaluate other situations, for instance, when the risk associated to FP and FN is well known (restricting the AUC integration over an acceptable range).

Tie to reliable, perceptually meaningful judgments: We believe that it is valuable to tie quality measure evaluations to tasks that properly reflect the underlying perceptual phenomena. In our analysis, for instance, we sought to evaluate class separation measures based on the actual perceptual separability of single classes. In contrast, previous studies focused on integrated ranking judgments of multi-class scatterplots, disguising the actual separability of single classes [SNLH09, TBB*10].

In that respect, one current limitation of our work is that we use a binary classifier. With this approach, we can classify 'whether or not' but not 'how much' a class is separable. In future work, we thus plan to predict the actual degree of separation using ordinal regression models [O'C06]. However, such a more fine-grained approach is non-trivial and will necessitate a deeper investigation of perceptual discriminability of different degrees of class separation [LAdS12, SMT13, TBB*10]. Another interesting question is how perceptually accurate substituting multi-colored scatterplots with 1-vs-all scatterplots is. While for visual search tasks related work exists [HW12], there is, to the best of our knowledge, little known about separation tasks.

6. Conclusions

In 2010, concluding from a study on visual separation measures Tatu et al. noted that *"there is still a lot to be done until the ultimate automatic quality measure can be found"* [TBB*10]. Towards that goal, we proposed a novel evaluation framework to more efficiently and effectively evaluate such measures. The framework uses a broad set of human judgments to learn how a quality measure would predict such judgments.

We used this approach to compare 15 state-of-the-art measures for class separability in 2D scatterplots. We found the best performing measure to be *DSC* by Sips et al. [SNLH09], but, more generally, also a lot of room for future improvements. We analyzed measures for synthetic and real data separately, and found further evidence for the bias of current measures towards simple synthetic datasets. Apart from comparing different quality measures, our framework can also be used for exploring and finding proper parameter settings for parametric measures. This usage is specifically interesting, as our study has confirmed that the performance of current parametric measures is indeed strongly dependent on parameter settings [STTM12].

Our work focuses on evaluating measures based on how well they align with human judgment. Of course, there are other important factors that need to be considered when evaluating quality measures, such as, computational complexity. Also, when designing measures that seek to imitate the human perception, it is always important to bear in mind the limitations of human perception. There might be patterns in the data that a human simply cannot directly "see", or what she sees might just be an artifact of how the data has been abstracted and encoded [Aup14, KS14].

Acknowledgements

We thank Tamara Munzner and Melanie Tory for early feedback, and Stephen Ingram, Torsten Möller, Michael Oppermann, and Thomas Torsney-Weir for cross-reading and discussions. This work was partly funded by FFG project 845898 (VALID).

References

[AEM11] ALBUQUERQUE G., EISEMANN M., MAGNOR M.: Perception-based visual quality measures. In *Proc. IEEE Conf.*

Visual Analytics Science and Technology (VAST) (2011), pp. 11–18. 2

[Aup14] AUPETIT M.: Sanity check for class-coloring-based evaluation of dimension reduction techniques. In *Proc. VIS Wkshp. BEyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)* (2014), ACM, pp. 134–141. 9

[BH75] BAKER F., HUBERT L.: Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association 70*, 349 (1975), 31–38. 6

[Bis06] BISHOP C. M.: *Pattern recognition and machine learning*. Springer, 2006. 4

[BTK11] BERTINI E., TATU A., KEIM D. A.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 17*, 12 (2011), 2203–2212. 1

[Car08] CARPENDALE S.: Evaluating information visualizations. In *Information Visualization*. Springer, 2008, pp. 19–45. 2

[CH74] CALIŃSKI T., HARABASZ J.: A dendrite method for cluster analysis. *Communications in Statistics Simulation and Computation 3*, 1 (1974), 1–27. 6

[DK10] DASGUPTA A., KOSARA R.: Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 16*, 6 (2010), 1017–26. 2

[Dun74] DUNN J. C.: Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics 4*, 1 (1974), 95–104. 6

[ET93] EFRON B., TIBSHIRANI R.: *An Introduction to the Bootstrap*. Macmillan Publishers Limited, 1993. 5

[Faw06] FAWCETT T.: An introduction to ROC analysis. *Pattern Recognition Letters 27*, 8 (2006), 861–874. 4

[FT74] FRIEDMAN J. H., TUKEY J. W.: A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers 100*, 9 (1974), 881–890. 2

[Fuk90] FUKUNAGA K.: *Introduction to statistical pattern recognition*, second ed. Computer Science and Scientific Computing. Academic Press, 1990. 6

[GBNT04] GILAD-BACHRACH R., NAVOT A., TISHBY N.: Margin based feature selection – theory and algorithms. In *Proc. 21st Int. Conf. on Machine Learning (ICML)* (2004), Brodley C. E., (Ed.), ACM, pp. 43–50. 6

[HB10] HEER J., BOSTOCK M.: Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)* (2010), pp. 203–212. 9

[HW12] HAROZ S., WHITNEY D.: How capacity limits of attention influence information visualization effectiveness. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 18*, 12 (2012), 2402–2410. 9

[HYFC14] HARRISON L., YANG F., FRANCONERI S., CHANG R.: Ranking visualizations of correlation using Weber's law. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 20*, 12 (2014), 1943–1952. 8

[IIC*13] ISENBERG T., ISENBERG P., CHEN J., SEDLMAIR M., MÖLLER T.: A systematic review on the practice of evaluating visualization. *IEEE Trans. Visualization and Computer Graphics (Proc. SciVis) 19*, 12 (2013), 2818–2827. 1, 2

[JC08] JOHANSSON J., COOPER M.: A screen space quality method for data abstraction. *Computer Graphics Forum (Proc. EuroVis) 27*, 3 (2008), 1039–1046. 2

[KS14] KINDLMANN G., SCHEIDEGGER C.: An algebraic process for visualization design. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 20*, 12 (2014), 2181–2190. 9

[LAdS12] LEWIS J. M., ACKERMAN M., DE SA V.: Human cluster evaluation and formal quality measures: A comparative study. In *Proc. 34th Conf. of the Cognitive Science Society (CogSci)* (2012), pp. 1870–1875. 2, 6, 9

[LBI*12] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seven scenarios. *IEEE Trans. Visualization and Computer Graphics (TVCG) 18*, 9 (2012), 1520–1536. 2

[LvdMdS12] LEWIS J. M., VAN DER MAATEN L., DE SA V.: A behavioral investigation of dimensionality reduction. In *Proc. 34th Conf. of the Cognitive Science Society (CogSci)* (2012), pp. 671–676. 9

[MMdALO15] MOTTA R., MINGHIM R., DE ANDRADE LOPES A., OLIVEIRA M. C. F.: Graph-based measures to assist user assessment of multidimensional projections. *Neurocomputing* (2015), 583 – 598. preprint. 2, 6

[O'C06] O'CONNELL A. A.: *Logistic regression models for ordinal response variables*. Sage Publications, 2006. 9

[PVW09] PRETORIUS A. J., VAN WIJK J. J.: What does the user want to see? What do the data want to be? *Information Visualization 8*, 3 (2009), 153–166. 8

[RB10] RENSINK R., BALDRIDGE G.: The perception of correlation in scatterplots. *Computer Graphics Forum (Proc. EuroVis) 29*, 3 (2010), 1203–1210. 8

[Rou87] ROUSSEEUW P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*, 1 (1987), 53–65. 6

[SMT13] SEDLMAIR M., MUNZNER T., TORY M.: Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis) 19*, 12 (2013), 2634–2643. 4, 5, 6, 7, 8, 9

[SNLH09] SIPS M., NEUBERT B., LEWIS J. P., HANRAHAN P.: Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum (Proc. EuroVis) 28*, 3 (2009), 831–838. 1, 2, 6, 8, 9

[Str02] STREHL A.: *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, University of Texas at Austin, 2002. 6

[STTM12] SEDLMAIR M., TATU A., TORY M., MUNZNER T.: A taxonomy of visual cluster separation factors. *Computer Graphics Forum (Proc. EuroVis) 31*, 3 (2012), 1335–1344. 1, 2, 4, 7, 8, 9

[TAE*09] TATU A., ALBUQUERQUE G., EISEMANN M., SCHNEIDEWIND J., THEISEL H., MAGNOR M., KEIM D.: Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST)* (2009), pp. 59–66. 1, 2, 6, 8

[TBB*10] TATU A., BAK P., BERTINI E., KEIM D., SCHNEIDEWIND J.: Visual quality metrics and human perception: An initial study on 2D projections of large multidimensional data. In *Proc. Int. Conf. Advanced Visual Interfaces (AVI)* (2010), ACM, pp. 49–56. 1, 2, 6, 7, 8, 9

[WA05] WILKINSON L., ANAND A.: Graph-theoretic scagnostics. *Proc. IEEE Symp. Information Visualization (InfoVis)* (2005), 157–164. 1, 2

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