Eurographics Conference on Visualization (EuroVis) 2012
S. Bruckner, S. Miksch, and H. Pfister
(Guest Editors)

Volume 31 (2012), Number 3

A Taxonomy of Visual Cluster Separation Factors

M. Sedlmair[1] and A. Tatu[2] and T. Munzner[1] and M. Tory[3]

[1]University of British Columbia, Canada [2]University of Konstanz, Germany [3]University of Victoria, Canada

Abstract

We provide two contributions, a taxonomy of visual cluster separation factors in scatterplots, and an in-depth qualitative evaluation of two recently proposed and validated separation measures. We initially intended to use these measures to provide guidance for the use of dimension reduction (DR) techniques and visual encoding (VE) choices, but found that they failed to produce reliable results. To understand why, we conducted a systematic qualitative data study covering a broad collection of 75 real and synthetic high-dimensional datasets, four DR techniques, and three scatterplot-based visual encodings. Two authors visually inspected over 800 plots to determine whether or not the measures created plausible results. We found that they failed in over half the cases overall, and in over two-thirds of the cases involving real datasets. Using open and axial coding of failure reasons and separability characteristics, we generated a taxonomy of visual cluster separability factors. We iteratively refined its explanatory clarity and power by mapping the studied datasets and success and failure ranges of the measures onto the factor axes. Our taxonomy has four categories, ordered by their ability to influence successors: Scale, Point Distance, Shape, and Position. Each category is split into Within-Cluster factors such as density, curvature, isotropy, and clumpiness, and Between-Cluster factors that arise from the variance of these properties, culminating in the overarching factor of class separation. The resulting taxonomy can be used to guide the design and the evaluation of cluster separation measures.

Categories and Subject Descriptors (according to ACM CCS): H.5.0 [Information Interfaces and Presentation]: General; J.0 [Computer Applications]: General

1. Introduction

Over a century of previous work has been devoted to creating effective and efficient algorithms for dimensionality reduction (DR), where a set of points in high-dimensional space is transformed into a more compact lower-dimensional form that preserves the important aspects of its underlying structure. These techniques include the venerable principal components analysis (PCA) [Jol02], the many variants of multidimensional scaling (MDS) [IMO09], and very new approaches such as t-SNE [vdMH08] that are designed to address the failure cases of older methods. The most common visual encoding (VE) techniques for the resulting lower-dimensional data generated by these DR techniques are variants of scatterplots. Two-dimensional scatterplots are nearly ubiquitous, scatterplot matrices (SPLOMs) are also widely used, and 3D scatterplots are not uncommon.

However, little attention has been paid to the question of how to guide users through the process of choosing which DR and VE techniques to use. The DimStiller system used a workflow structure to guide users through the process of choosing DR and VE techniques [IMI*10], but it remains an open problem to develop automatic algorithms to provide such guidance. In service of this goal, we sought to use recent measures for visual cluster separation in scatterplots [SNLH09, TAE*09]. These were originally developed for selecting good views within a SPLOM, but we reasoned that they should also be applicable to providing guidance for DR and VE technique choices. A previous user study [TBB*10] had identified two particular measures as the most effective state of the art: the **centroid** measure and the **grid** measure [SNLH09, TAE*09]. (These names are our own, designed for readability. The original centroid measure name was *Distance Consistency* [SNLH09], the original grid measure names were the equivalent *Distribution Consistency* [SNLH09] and *2D-Histogram Density* [TAE*09].)

However, our initial tests showed that these measures failed to produce reliable results; that is, there was often a mismatch between the result computed by the measure and a quality judgement made by a person. These mismatches were in the form of both **false positives**, where the mea-

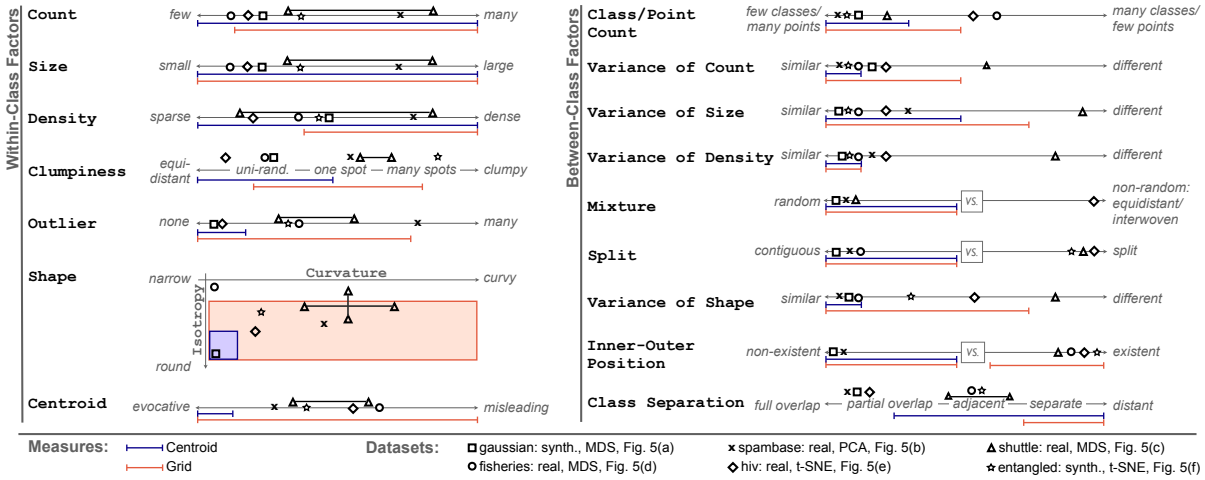


Figure 1: *Taxonomy of factors in visual cluster separation, where factor axes are marked to show the ranges where existing measures are successful; gaps represent failure cases. The centroid measure is marked in blue and the grid is marked in red. All positions are approximate estimates. The six datasets shown in Figure 5 are also marked along the factor axes. Five of them have little inter-class variance and are marked with a single representative point on the Within-Class axes, but the high-variance shuttle dataset has ranges marked with lines and boxes.*

sure value was high but the human judged the visual separation between clusters as poor, and **false negatives**, where the measure value was low but humans were indeed able to distinguish clusters in the scatterplots. We conjectured that these measures encapsulated implicit assumptions about dataset characteristics, which were not previously uncovered because they were only tested with relatively simple datasets. Therefore, we decided to systematically study the differences between computed measures and human judgement across a broad set of 75 real and synthetic datasets, using a range of four DR and three VE techniques, to untangle these assumptions and create a taxonomy of factors underlying visual cluster separation in scatterplots.

In our **qualitative data study**, two human investigators (the first two authors of this paper) visually inspected over 800 plots to determine whether or not the measures created plausible results. We found that the measures failed in over half the cases, and over two-thirds of the time for the real datasets. In addition, the investigators generated a detailed set of characteristics that influenced cluster separability in general, and specific reasons why the measures failed in the cases where they found a mismatch. Based on separability characteristics and failure reasons, we generate a higher-level taxonomy of factors, which we iteratively refined in multiple passes, not only by considering its explanatory clarity and power, but also by mapping the ranges where each measure was successful along the factor axes, and by placing some of the studied datasets along them. Figure 1 shows the measure success ranges on a simplified version of the taxonomy. The extent of the gaps which indicate measure failure ranges is readily apparent. This figure is discussed in more detail in Section 6 after the factors are presented in Section 5; the meaning of each is further explained in Figure 3.

The primary contributions of this paper are (1) a taxonomy for dataset factors that influence visual cluster separation, and (2) the systematic evaluation of two visual cluster separation measures with respect to human judgement. The taxonomy is intended to provide operational guidance in terms of how to develop and evaluate better cluster separation measures. Two smaller contributions are the extension of these two measures from 2D scatterplots to SPLOMs and 3D scatterplots, and the qualitative data study method itself.

2. Cluster Separation Measures

We chose to focus on cluster finding and verification as the main task supported by scatterplot usage when visually encoding DR data. Although finding correlation is an even more common task for general scatterplot usage [RB10, LMvW08], that task is not relevant in our case because DR techniques are designed to generate a set of dimensions that are as independent as possible; that is, not correlated. We thus chose to test measures for visual cluster separation. For the purpose of our study, we selected two measures, centroid [SNLH09] and grid [SNLH09, TAE*09] measures, which were found to outperform other visual cluster separation measures in a previous user study [TBB*10].

Our chosen measures were designed for the task of verifying clusters in classified data. Our study design required that we use classified data as if it were ground truth, to check if the class structure matched up with the visible cluster structure. However, we know that often class structure is a conjec-

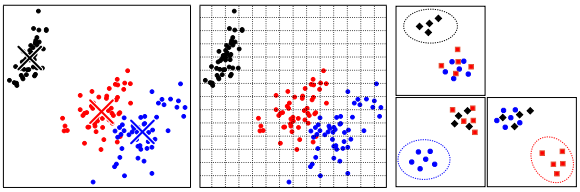


Figure 2: *(a) Computed centroids are marked with an X. (b) Grid cells used to compute class intermixtures are shown with dotted black lines. (c) Synthetic example of a SPLOM that should score well because each individual class is separated within at least one view, even though no single view shows all classes simultaneously separated.*

ture or proposal rather than a final answer, and that finding clusters in non-classified data is an important task. We thus kept this latter task in mind when designing our taxonomy.

2.1. Original Measures

The original centroid and grid measures operate on a single 2D scatterplot, and rate it on a scale of 0-100 where low values are poor and high values are good.

The centroid measure iterates over all points and determines the closest class centroid, defined as the arithmetic mean of all points of a particular class, using euclidean distance. For each class, the ratio between points closest to the centroid of their own class and those closer to the centroid of another is computed. The overall measure is computed by the weighted mean of all class-wise judgements using the number of points per class as the weight. Figure 2(a) shows an example dataset annotated by centroids drawn with an X.

The grid measure superimposes a virtual grid over a scatterplot and measures the mixture of points in each grid cell: a cell containing points of only one class is good, and one with intermixing of all classes is poor. A weighted sum of all grid cells based on the number of points in each grid cell is used for the overall measure value. Figure 2(b) shows an example dataset annotated with dotted grid lines.

2.2. Extensions

In order to include multiple scatterplot-based VE techniques in our study, we extended these measures to operate on 3D scatterplots and 2D SPLOMs. The extension to 3D was straightforward for the centroid measure, by simply substituting 3D for 2D euclidean distance. For the grid measure we used a 3D grid of cubes instead of a 2D grid with boxes, and took care to keep the number of cells roughly constant across these cases to ensure comparability between the 2D and 3D measures.

We extend the cluster separation measures from the single scatterplot case to an entire SPLOM by picking the best view independently for each class and computing a weighted mean across all classes, using number of points per class as the weight. For our purposes, a SPLOM where each class is clearly separated in at least one of its constituent views should score well, even if no single view shows all classes separated simultaneously. The original measure judged all classes simultaneously for each view, so the example in Figure 2(c) would score poorly. Instead, we compute the best view for each class using a **class-wise** measure that provides a judgment between 0-100 for each class in each view. Using a one-vs-all comparison, this extension was straightforward for the centroid measure, while we needed to make some minor adaptions for the grid measure. The mathematical details of our extensions can be found in the supplemental material.

Sips et al. claim that the grid measure is relatively insensitive to grid size [SNLH09] in terms of the number of grid cells, based on tests with grids of 20x20, 25x25, and 33x33. However, our tests showed that using a static grid size did not give reliable values across datasets with different numbers of points. We thus extended the measure to incorporate a simple dynamic grid size rule dependent on the number of points, using the square root of the point count for the 2D case and the cube root for the 3D case.

3. Related Work

The most relevant previous work is the scagnostics (scatterplot diagnostics) framework for finding interesting views within SPLOMs [WA05, WW08]. Wilkinson et al. proposed nine axes for characterizing the shapes formed by points within scatterplots: outlying, skewed, clumpy, convex, skinny, striated, stringy, straight, and monotonic. A few of these axes correspond directly to our own factors, and our names reflect this similarity (`Outlier`, `Clumpiness`). Wilkinson et al.'s goals are more broad than ours since they include structures relevant for many scatterplot usage tasks, while we focus on cluster separation and specifically exclude correlation analysis; our taxonomy is thus more tightly focused. Another difference is while their framework is clearly informed by informal exploration of many datasets, our taxonomy is rigorously grounded in systematic qualitative data gathering and analysis. A final critical difference is the angle of attack: their top-down approach has a core contribution of providing precise mathematical descriptions and efficient algorithms for computation of their nine measures; in contrast, our bottom-up approach builds up categories derived from open coding of phenomena that humans noticed, and we deliberately refrain from such top-down descriptions to avoid a bias towards what is easy to compute algorithmically or crisply define mathematically. We see our work as complementary to theirs.

We distinguish our contribution from previous work on visual cluster perception, such as the fundamental work on the Gestalt principles that describe the phenomena underlying human perception of groupings [Wer23]. Rather than

explain the mechanisms behind human perception of clusters, our goal is to provide operational guidance for computational designers who seek to build and evaluate better class separability measures.

We furthermore distinguish our work from the abundant body of research on cluster analysis in data mining and machine learning for unsupervised classification, which focuses on how to computationally extract groups of similar objects. Based on the deficiency of popular clustering techniques such as k-means [Mac67] to detect non-spherical clusters, various researchers sought more robust ways to identify arbitrarily shaped clusters [AT89, EKSX96, ZFLW02]. With this line of work, we share the critique of centroid-based approaches and the acknowledgment of `Shape` as an important factor of class separation. However, while cluster analysis research focuses on automatically finding class structure in the data, we study the human judgment of clusters in scatterplots, based on a broad set of data.

4. Method

We call our methodological approach a **qualitative data study**. Echoing a previous call for the balance of focus on both user and data [PVW09], we consider a qualitative data study approach to be the dual of a user study: rather than a few datasets observed by many people, there are many datasets observed by a few people. Data studies themselves are not new in visualization and have, for instance, been used to evaluate the scagonstics measures [WW08], or to verify novel DR techniques [IMO09]. While these studies have been usually conducted to evaluate or compare techniques based on quantitative measures or informal discussions of images, we use a data study to derive and systematize knowledge about datasets by applying qualitative coding techniques as used in the social sciences [Cha06]. Embraced in the HCI community [FBC11], in visualization coding has been successfully employed for analyzing video and audio footage of user interviews and behavior [IFM*10], and for systematic literature review [BTK11, LM10, LBI*11]. Here, we propose using coding for analyzing judgements about visually encoded data. We conjecture that our method of creating taxonomies through qualitative data studies may be applicable in a wide variety of visualization contexts.

Our study is qualitative in that a fundamental operation is human judgement, and our subsequent analysis is based on the methods of open and axial coding [Cha06]. These judgements were made by two investigators — the first and second authors of this paper. In particular, our study consisted of four stages:

1. Choosing variables for study
   a. 75 datasets: 31 real, 44 synthetic
   b. 4 DR techniques: PCA, robust PCA, MDS, t-SNE
   c. 3 scatterplot VE techniques: 2D, 3D, SPLOM
   d. 2 visual cluster separation measures: centroid, grid
2. Generating dataset instances and computing measures
3. Open coding and measure evaluation
   a. Pass 1 (all dataset instances):
      i. Open code factors affecting separability
      ii. Judge if measures worked
   b. Determine failure cases of measures
   c. Pass 2 (all failure cases):
      i. Open code reasons for failure
4. Axial coding and taxonomy building
   a. Merge together the codesets between the investigators
   b. Axial coding to create initial taxonomy from combination of separability and failure codesets
   c. Refine taxonomy through using it

We now discuss the stages in more detail. Further details (full dataset list, DR parameterization, coding results) can be found in the supplemental materials.

1 — Choosing Variables for Study: Our study encompasses a range of datasets, DR techniques, and VE techniques, which we call **variables**.

We wanted to use datasets with a broad range of characteristics, yet of course at the start of the study we did not know the factors that we would eventually construct in our final taxonomy. We experimented with several data generators to construct synthetic datasets that matched our initial intuitions about interesting cases based on our previous experience with DR and VE techniques. In particular, we generated data with random *gaussian* clusters, datasets that have *entangled* or interwoven class structure in higher dimensions that cannot easily be untangled by a linear 2D projection, and some synthetic *grids*. We also gathered complex *real* datasets both from direct contact with colleagues [HB11, SNLH09, TAE*09] and from online repositories [SAP10, FA10, Uni11, Vis11, War11], also including simpler benchmarks used for evaluation in previous work. Class structure was either given as ground truth or was provided by our colleagues using clustering algorithms. We categorize our final collection of 75 datasets into four groups:

name	*#sets*	*#points*	*#dim.*	*#classes*
real	31	77-43500	3-159	2-53
synthetic-gaussian	16	100-500	5-10	3-5
synthetic-entangled	24	185-2318	3-15	3-15
synthetic-grid	4	905-1296	3-4	2-16

We chose a mix of popular and recent DR techniques: the linear PCA technique is the first choice of most high-dimensional data analysts [Jol02]; the Robust PCA algorithm was designed to be tolerant to outliers [TF09]; nonlinear MDS [BG05], with the Glimmer algorithm designed to work well with both dense and sparse datasets [IMO09]; and t-Distributed Stochastic Neighbor Embedding (t-SNE), a recently proposed nonlinear DR method designed to separate clusters well and show multi-scale structure [vdMH08].

We chose three VE techniques, all based on scatterplots. The most commonly used technique is the 2D scatterplot. Some DR systems reduce to three dimensions and

use 3D scatterplots [KSC*10], under the logic that more information can be shown with the addition of an extra dimension (despite evidence of the perceptual difficulties of understanding point clouds in 3D space [Mun09]). Other DR systems use 2D SPLOMs to visually encode DR data when the low-dimensional space has three or more dimensions [IMI*10]. Our rationale for choosing the measures is covered in Section 2.

2 — Generate Dataset Instances and Compute Measures: We define a **dataset instance** as a particular combination of $data \times DR \times VE$. After pre-processing the data by deleting duplicate points and non-numeric dimensions, we computed all DR data using R for PCA, RobPCA and t-SNE, and Java for Glimmer. We had to exclude 28 out of 300 data-DR combinations due to computational problems ranging from scalability to singularity issues. Using the four DR techniques, we reduced each of the 75 high-dimensional datasets to 2D to create a single 2D scatterplot, to 3D to create a single 3D one, and created a sequence of $d \times d$ SPLOMs by reducing to d from 3 up to a maximum of d_{max}. We chose d_{max} based on the original dimensionality of the dataset, on whether the separability measures continued increasing with higher dimensions, and on an absolute maximum of $d = 15$. All scatterplot VEs were color-coded based on the given class structure. The centroid and grid measures were computed using code provided courtesy of Sips [SNLH09], on which we built the extensions described in Section 2. We computed both overall and class-wise values for 2D scatterplots, 3D scatterplots, and the set of SPLOMs. For each data-DR combination, each investigator individually picked one SPLOM from the set of SPLOMs, based on their judgement of when increasing the size of the SPLOM was no longer helpful for understanding cluster structure. The maximum SPLOM size selected was a 7x7 matrix. The final set inspected by each investigator was **816 dataset instances**. An accompanying video shows a fast forward of all 816 instances, where each image is annotated with both measure values.

3 — Open Coding and Failure Cases: The two investigators worked separately and coded the dataset instances in two passes. In the first pass, they looked at all 816 instances using static images for the 2D and SPLOM, and for the 3D case a custom interactive 3D viewer designed to present a similar visual appearance to the static scatterplots generated by R. For each dataset instance they judged the measure's overall performance on a three point scale of *ok*, *dubious*, or *poor*, and additionally noted *class-wise poor* cases where the overall measure was ok, yet one class-wise value was extraordinarily off (excluding special cases such as classes with only a single point). While making these judgements, the investigators also open coded factors that influenced cluster separability; that is, they gradually built up a set of codes and noted when a particular code applied to an instance. Because open coding is an iterative process, some instances needed to be reanalyzed before the final set of codes was settled upon. Multiple codes could be associated with each instance.

We combined the judgements from the two investigators and partitioned all instance-measure pairs into one of two categories: success and **failure cases**. The criterion for failure was that at least one investigator judged the measure's overall performance as *dubious* or *poor*, or at least one class as *class-wise poor*. In the second coding pass, both investigators coded their failure cases for potential reasons underlying the measure failure.

4 — Axial Coding for Taxonomy Building: The investigators merged the independent sets of codes that they had built up in their first coding pass. The combined separability codeset had 25 codes; the combined failure codeset had 22 codes. We began the taxonomy creation process with an axial coding pass to merge and categorize the combined sets of codes for both separability and failures. We often checked back to inspect the instances when necessary [Cha06]. Axial coding is an iterative process; we refined and improved the taxonomy through a sequence of over a dozen passes. We further refined the taxonomy by testing its explanatory power and clarity. We mapped the ranges of the failure cases for each measure on the axes of our proposed factors, and also some of the dataset instances. Figure 1 shows these ranges on the final taxonomy.

5. Taxonomy

Our taxonomy of factors that affect visual cluster separation is shown graphically in Figure 3. At the top level, we differentiate between Within-Class factors that are solely based on the structure or appearance of a single class, and Between-Class factors depicting interactions between two or more classes. Many Between-Class factors arise from the variance of the base Within-Class ones; all of these dependencies are encoded in the diagram by the horizontal arrow. For both the Within and Between sets of factors we group them into the high-level categories of Scale, Point Distance, Shape, and Position. These categories are ordered, as shown in the diagram by the vertical arrow that indicates that higher levels can influence those beneath them. The overarching factor of class separation in the bottom right corner is thus influenced by all other contributing factors.

5.1. Within-Class Factors

In the Scale category, we describe the `Count` factor as the number of points in a class ranging from *few* to *many*, the `Size` factor ranging from *small* to *large* as a cluster's spread in terms of 2D area or 3D volume. These `Scale` factors are less interesting on their own, but more so when their variance is considered with the corresponding Between-Class factors.

The Point Distance category contains three factors. The `Density` factor is the ratio between the Scale factors

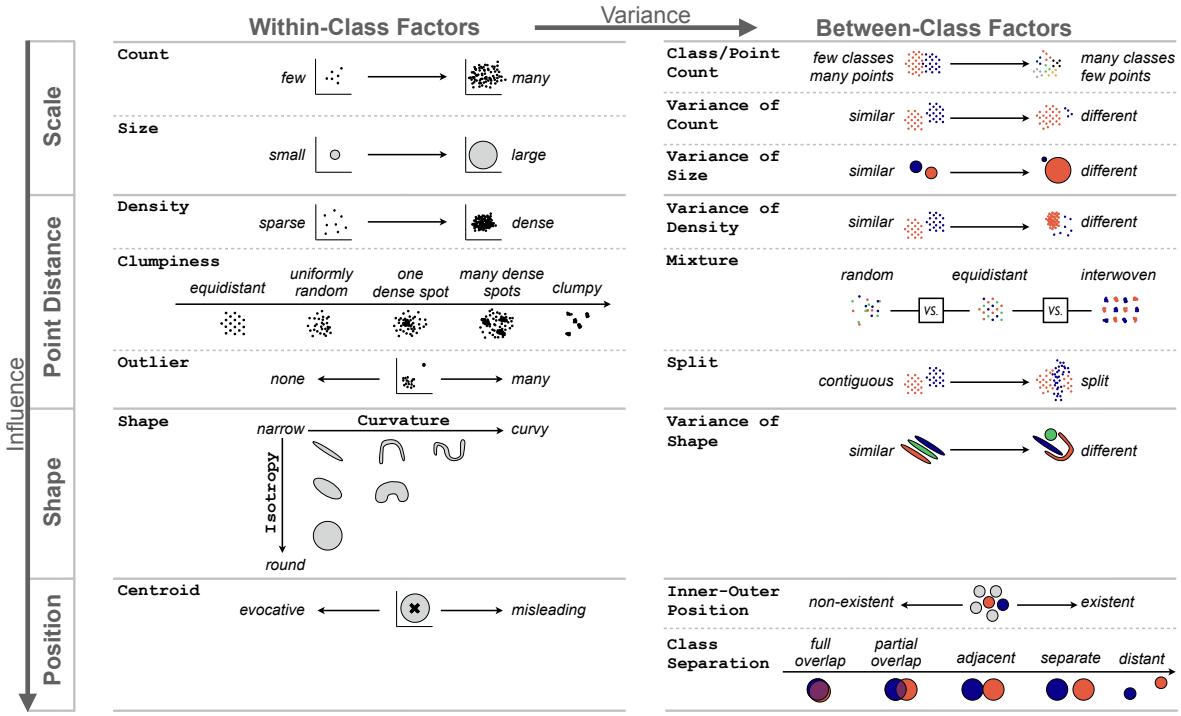


Figure 3: *A taxonomy of data characteristics with respect to class separation in scatterplots. Some factors are organized as axes (arrows) while others are binned. Between-Class factors often result from the variance of Within-Class factors (horizontal dependencies), and factors at the top can strongly influence factors below them (vertical dependencies).* `Class Separation` *is therefore dependent on all other factors.*

`Count` and `Size`, ranging from *sparse* with few points and large spread to *dense* with many points and a small spread. Following the naming conventions of scagnostics [WA05], we use `Clumpiness` to depict different patterns of inter-point distribution. We identified several landmarks on this axis: pairwise *equidistant* points (a perfect grid is the extreme case); a *uniformly random* distribution; *one dense spot* (as in a gaussian distribution); *many dense spots*; *clumpy*. The `Outlier` factor pertains to points that are distant to the collection of the rest, ranging from *none* to *many*.

The Shape category pertains to the perceived Gestalt of a point cloud [Wer23]. We describe it with two factors that are orthogonal axes. `Isotropy` describes how directional the shape of a class is, ranging from *round* to *narrow*; in 2D, narrow converges to a line, and in 3D to a planar layer or a line. `Curvature` describes nonlinearity, ranging from linear through convex to *curvy*.

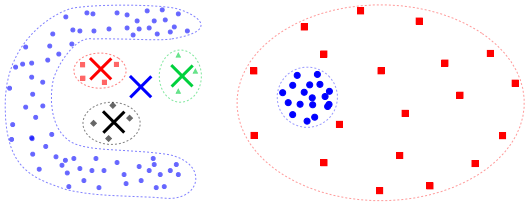
The sole factor in the Position category is the `Centroid`, an axis with *evocative* on one side and *misleading* on the other. Using the centroid to robustly indicate the position of a cluster essentially assumes that it is more or less a gaussian distribution of points; if not, the centroid can be misleading. For example, the centroid can fall completely outside the point cloud when the factors in the `Shape` category have particular values, namely a *narrow* `Isotropy` factor and a *curvy* `Curvature` factor, as shown in Figure 4(a).

5.2. Between-Class Factors

Between-Class factors encapsulate the variance and combination of the Within-Class factors across multiple classes. Our taxonomy contains the ones that we deem to be most important based on our data study; we do not list all possible combinations.

In the Scale category, the `Class-Point Count` factor is the ratio between the number of classes and the number of points of the dataset. We found that the case of *few classes, many points* was easier to perceive than the cases of *many classes, few points*. We also characterize the factors `Variance of Size`, and `Variance of Count`, which can influence the ability to perceive clusters.

The first factor in the category of Point Distance is `Variance of Densities`, the mutual product of `Variance of Size` and `Variance of Count`. Consider the example in Figure 4(b) where a big sparse class overshadows a small dense class. An overly simplistic separability measure would rate the small class as poor, but it could be that the large class is sufficiently sparse that the small one can be easily identified. The `Mixture` factor describes



(a) (b)

Figure 4: *(a) The combination of the narrow* `Isotropy` *factor and the curvy* `Curvature` *factor in the blue class lead to the* `Centroid` *factor being misleading because it falls outside the point cloud. The X's show the location of the centroids. (b) The red large sparse class overshadows the blue small dense class which remains distinguishable.*

Point Distance characteristics of *fully* or *partly overlapping* classes (see `Class Separation` axis below). We identify some specific bins as landmarks: *random* is the common case where there is no apparent structure in the overlapping area; *equidistant* describes overlapping areas where points of different classes have pairwise similar or equal distances, as we frequently found with the t-SNE DR technique in our data study; *interwoven* represents a case described as problematic in previous work [SNLH09]. The `Split` factor distinguishes between clusters that are *contiguous* in the scatterplot, and those that are *split* apart into separate regions of a plot, with another class appearing between their pieces.

The `Variance of Shape` factor ranges from *similar* shapes for all clusters to very *different* shapes across them.

In the Position category, the `Inner-Outer Position` factor describes a positional relationship between classes where inner classes near the center can be surrounded by outer classes on the periphery. We distinguish whether such a relation is *existent* or *non-existent*. Several of our study instances showed that inner classes were more difficult to identify, especially those in the *synthetic-gaussian* data family. This perceptual problem is particularly pronounced for 3D scatterplots where inner classes can be occluded by outer ones. Also, using a `Centroid` approach to determine `Inner-Outer Position` might be misleading, as shown in Figure 4(a).

Finally, the overarching Between-Class factor is `Class Separation`. While straightforward when considering two equally scaled, round, and contiguous clusters, separation can be strongly influenced by nearly all other factors that we discussed above. We illustrate the axis with this simple case: *full overlap*; *partial overlap*; *adjacent*; *separate*; *distant*. For the classified data that we studied, we found that the range of small or no overlap between clusters — *adjacent*, *separate*, or *distant* — was good enough for cluster identification. For non-classified data that cannot be color-coded, the *adjacent* case would not suffice.

Measure	Data	Total	Failures	Overall Poor	Overall Dubious	Class-wise Poor
centroid	all	816	400	22%	49%	29%
grid	all	816	416	39%	54%	7%
centroid	real	296	201	30%	33%	37%
grid	real	296	193	43%	45%	12%

Table 1: *Number of failure cases for each measure, and breakdown showing percent of failures in each category.*

6. Evaluation of Measures

We found a surprisingly high number of failure cases: 49% of the overall instances for the centroid measure and 51% for the grid measure. The performance is even worse when we restrict the analysis to the real datasets: the centroid measure failed in 68% of the cases and the grid measure in 65%. Table 1 shows the breakdown of these groups into the three kinds of failures: overall poor, overall dubious, and class-wise poor. The raw data, including all 816 dataset instances along with the measure values and the coders' judgments, can be found in the supplemental material.

Based on these findings, we cannot recommend either measure for a rigorous, reliable and robust cluster separation judgement. Overall, we found that they worked well only for very clearly separated classes, which is often not the case with real-world datasets. For more realistic data characteristics, certain assumptions in the measure design make them prone to incorrect judgements.

We further analyzed the failure cases in terms of false negatives, where the measure does not detect structure that is actually apparent in the visualization, versus false positives, where the measure yields a high value when clusters are in fact not visually separated. We found a major disparity in their distribution: 68% of failures for the centroid measure and 85% of failures for the grid measure were false positives. We conclude that the measure designers may have successfully reduced false negatives but neglected to carefully consider the problem of false positives.

We now provide details about the reasons for measure failures, explain them by using our taxonomy, and illustrate them with examples of both real and synthetic datasets from our study. We show the 2D scatterplot cases for clarity, but these failures also appeared in 3D and SPLOM cases.

6.1. Centroid Measure Results

We found many examples where the position of one or more centroids did not accurately reflect the position of a class because of other factors, resulting in false positives or false negatives for both class-wise and overall measure values. Figure 5(a) shows an example of a false positive: the red class is fully overlapped with the blue and the black but has a high value of 77 — where 100 is best and 0 worst, and our judgment of high/low is adjusted to the original work on these measures [SNLH09, TAE*09]. Considering how the centroid measure splits the image into Voronoi cells, it becomes clear how it fails to account for `Variation of`

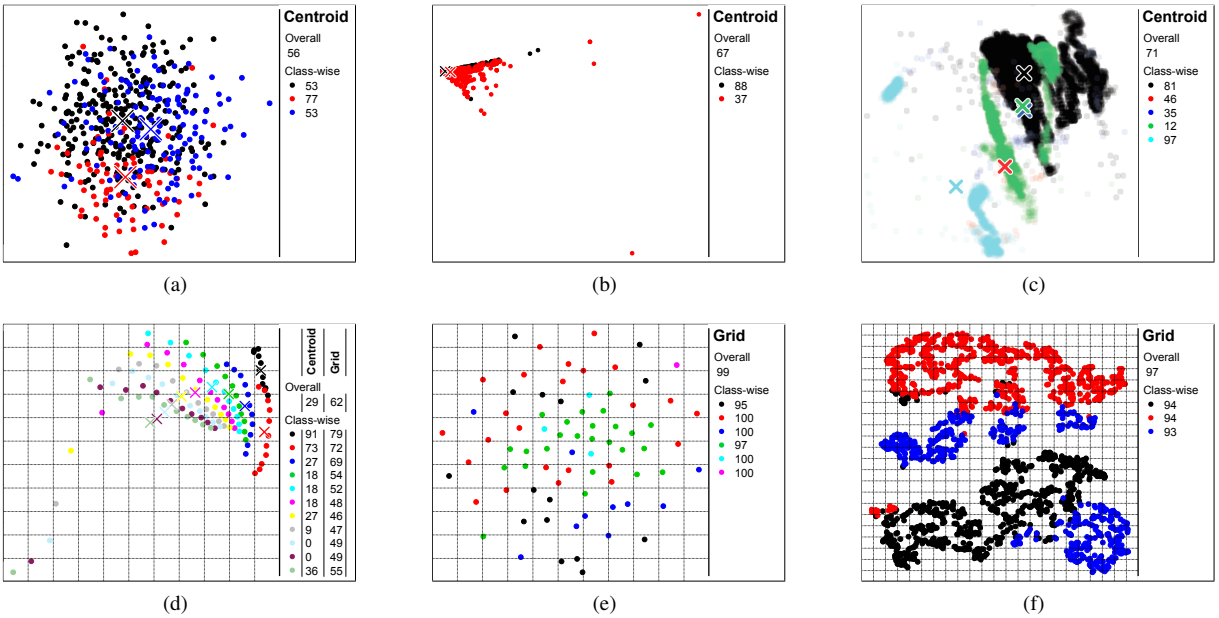


Figure 5: *2D scatterplot examples for six example datasets, enriched with centroids and/or virtual grid used for the measures computation. Attached to each plot are the relevant overall and class-wise measure values. (a) A synthetic gaussian with MDS. False positive for centroid. (b) The real* `spamBase` *dataset with PCA. False positive for centroid. (c) The real* `shuttle-big` *dataset with MDS (detail view). False negative for centroid. (d) The real* `fisheriesHarvestRule` *dataset with MDS. False negative for centroid and grid. (e) The real* `hiv` *dataset with t-SNE. False positive for grid. (f) A synthetic entangled dataset with t-SNE. False positive for grid.*

`Size` in this example. Figure 5(b) shows how the `Outliers` factor can influence the centroid of class that is nearly fully overlapped, resulting in an unreasonably high class-wise value of 88 for the black class. Figure 5(c) shows a zoomed-in detail view of the shuttle dataset, where the `Clumpiness` and `Variance of Shapes` factors can misleadingly influence the centroid: the green class is given an extremely low score of 12, which we consider a false negative since useful structure is indeed visible. This example also shows the tendency of classes with *inner* centroids towards false negatives, a phenomenon that we observed in many dataset instances. The centroid measure is also not robust against non-convex classes, as was stated in previous work [SNLH09]; Figure 5(d) shows an example of a real dataset where classes with a *narrow* shape have visible structure, but the overall measure yields a false negative with 29. This value is also influenced by `Outliers`.

Our taxonomy implies that the `Centroid` factor can be influenced by many other factors; these examples illustrate that the centroid measure alone is definitely not a reliable indicator of class separation. The centroid measure is vulnerable with respect to `Shape`, `Clumpiness`, `Outliers`, `Variance of Count`, `of Size`, or `of Density`, and `Inner-Outer Position`. It is reliable only under the assumptions of *round*-ish clusters with no more than *one dense spot*, *no outliers*, and *similar sizes*, as shown in Figure 1. This combination of assumptions is rarely fulfilled for real datasets.

6.2. Grid Measure Results

We also found many examples of problems with the grid measure. False negatives arise from *narrow*, *adjacent* classes, as shown in Figure 5(d). The black and red classes have suitably good class-wise values, but most classes had overly low values in the range of 46 to 55. Points of different narrow, adjacent classes fall into one grid cell and are judged as poor, while the overall contiguity of the strings is not detected.

Figures 5(e) and 5(f) show two false positive examples where the grid measure yields 99 and 97 respectively, which would not be judged as high by a human. The underlying issue with Figure 5(e) is that the overlapping *equidistant* `Mixture` layout produced by t-SNE led to many points falling into their own grid cell. They are therefore judged as good without considering their surroundings more globally. Figure 5(f) shows how the grid approach fails to detect *split* classes.

We were particularly surprised that the grid measure performed even more poorly than the centroid measure, given the claims in previous work that it was the more powerful and robust of the two [SNLH09]. We knew from our

data study that a poor choice of grid size was part of the problem, because the investigators had noted many cases of overly small and overly large grid sizes in their coding. We wondered if the measure could be improved with a straightforward automatic way to compute an appropriate grid size. We investigated further by recomputing the measure values of the 58 affected failure cases with a variety of different grid size parameterizations. Despite the argument in previous work that the measure is "relatively insensitive to the choice [of the grid size]" [SNLH09], we found that it was relatively constant in only 14% of the cases and was sensitive to grid size in 86% of the cases. We plotted all of these examples to check if distinct features such as a knee, peak or plateau appeared in these curves to hint at a viable algorithmic approach, but we did not see any such features (see supplemental material). Thus, there is no obvious solution to the problem by simply adapting the grid size.

While relatively robust against false negatives, our data study revealed severe problems with the grid measure in terms of false positives. As shown in Figure 1, the grid measure, in particular, is vulnerable to the combination of spatial overlap on the `Class Separation` axis and certain inter-point characteristics stemming from `Variance of Density` and `Mixture`, most critically *equidistant* structures. Even when classes overlap completely, the measure might yield very good values depending on how the points fall into grid cells. It also fails in detecting *split* classes and might not be able to correctly detect *narrow* ones. For these reasons, the grid measure is not robust for overlapping classes and can only give reliable results for nicely *separated* or *distant*, *contiguous* classes.

7. Discussion and Limitations

The goal of our taxonomy is operational guidance for measure developers and evaluators. As an example, we used the taxonomy for explaining reasons for failures of two state-of-the-art separability measures. Inherent with the inductive qualitative research approach we chose, our measure evaluation proceeded in parallel with the taxonomy development. Conducting a qualitative data study is very time-consuming – in our case data gathering and analysis took over six months – so our goal was to abstract the findings into a more general taxonomy, which then can be applied broadly by others without having the overhead of a full-blown qualitative data study.

In particular, we suggest that the taxonomy can be directly used as a descriptive checklist device to inform the design of more reliable separability measures. It can also be used in the evaluation of measures. First, it can guide the choice of datasets to study, both in selecting real datasets and in generating synthetic ones. Whether the goal is to carry out a qualitative study as done here or a quantitative setup as done by Tatu et al. [TBB*10], mapping the chosen datasets onto the taxonomy axes as we did in Figure 1 will provide useful information about the coverage of relevant factors. Once the datasets are mapped to the axes, analyzing which datasets worked or failed for a particular measure allows the measure's effective range to be mapped onto the axes as well, helping untangle the underlying assumptions. We hope that such taxonomy use will accelerate the development of more reliable measures, ultimately leading to algorithms that provide sophisticated user guidance by automatically selecting appropriate DR/VE combinations.

In this study we tested the centroid and grid measures based on the findings of previous work [TBB*10]. Many other measures exist, including an older centroid-based approach by Dhillon et al. [DMS98] and a recent proposal from Albuquerque et al. [AEM11] created by machine learning on the results of perceptual studies. We conjecture that these measures may show a similar pattern of results, since a close reading of these papers shows that they were designed and validated with similarly simple datasets. In other words, we suspect that mapping these datasets onto our taxonomy's axes would result in a similarly restricted coverage of factors. After further investigation, we did find a promising measure that was designed and tested on real data [LZVB06]; it would be interesting to evaluate this k-nearest neighbor approach as future work.

Our taxonomy does have limitations. It was constructed and validated only with DR data, so while we also believe that our taxonomy might be applicable for non-DR projects as well, this conjecture has not been verified. We have focused on the visible appearance of DR data, which might show artifacts of the visual encoding or the projection rather than what is true or interesting about dataset structure. While we strove for good coverage in our selection of datasets, we are well aware that many possible dataset characteristics were not covered by our study, and hope that others will build on our findings and extend the taxonomy.

In service of replicability, extensive supplemental materials are available at `http://www.cs.ubc.ca/labs/imager/tr/2012/VisClusterSep`.

8. Conclusion

We have presented a taxonomy that characterizes cluster separability factors of DR data in scatterplots. It is based on a qualitative study of 75 datasets, four DR techniques, three scatterplot-based VE techniques, and two cluster separation measures. The two studied measures failed to provide a robust and reliable judgement in nearly 50% of our cases; we use the taxonomy to explain the reasons for this outcome. We offer the taxonomy in hopes that it will guide others in designing, using, and evaluating cluster separability measures.

9. Acknowledgements

This project was funded by NSERC STPGP 350540-07 with additional support from DFG-664/11. We thank Matt Brehmer, Jessica Dawson, Stephen Ingram, and Torsten Möller for their helpful comments on this project and paper.

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. 9

[AT89] AHUJA N., TUCERYAN M.: Extraction of early perceptual structure in dot patterns: Integrating region, boundary, and component gestalt. *Computer Vision, Graphics, and Image Processing 48*, 3 (1989), 304–356. 4

[BG05] BORG I., GROENEN P.: *Modern multidimensional scaling: Theory and applications*. Springer, 2005. 4

[BTK11] BERTINI E., TATU A., KEIM D. A.: Quality metrics in high-dimensional data visualization: an overview and systematization. *IEEE Trans. on Visualization and Computer Graphics 17*, 12 (2011), 2203–2212. 4

[Cha06] CHARMAZ K.: *Constructing grounded theory. A practical guide through qualitative analysis*. Sage Publications, Inc, 2006. 4, 5

[DMS98] DHILLON I., MODHA D., SPANGLER W.: Visualizing Class Structure of Multidimensional Data. *Proc. Symp. Interface: Computing Science and Statistics* (1998), 488–493. 9

[EKSX96] ESTER M., KRIEGEL H., SANDER J., XU X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. AAAI Conf. Knowledge Discovery and Data mining* (1996), vol. 1996, pp. 226–231. 4

[FA10] FRANK A., ASUNCION A.: University of California Irvine (UCI) Machine Learning Repository, 2010. 4

[FBC11] FURNISS D., BLANDFORD A., CURZON P.: Confessions from a grounded theory PhD: experiences and lessons learnt. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)* (2011), pp. 113–122. 4

[HB11] HOLT C., BRADFORD M.: Evaluating benchmarks of population status for Pacific salmon. *North American Journal of Fisheries Management 31*, 2 (2011), 363–378. 4

[IFM*10] ISENBERG P., FISHER D., MORRIS M., INKPEN K., CZERWINSKI M.: An Exploratory Study of Co-located Collaborative Visual Analytics around a Tabletop Display. In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST)* (2010), pp. 179–186. 4

[IMI*10] INGRAM S., MUNZNER T., IRVINE V., TORY M., BERGNER S., MÖLLER T.: DimStiller: Workflows for dimensional analysis and reduction. In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST)* (2010). 1, 5

[IMO09] INGRAM S., MUNZNER T., OLANO M.: Glimmer: Multilevel MDS on the GPU. *IEEE Trans. Visualization and Computer Graphics (TVCG) 15*, 2 (2009), 249–261. 1, 4

[Jol02] JOLLIFFE I. T.: *Principal Component Analysis, 2nd ed.* Springer, 2002. 1, 4

[KSC*10] KIM H., SCHULZE J., CONE A., SOSINSKY G., MARTONE M.: Dimensionality reduction on multi-dimensional transfer functions for multi-channel volume data sets. *Information Visualization 9*, 3 (2010), 167. 5

[LBI*11] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: *Seven guiding scenarios for information visualization evaluation*. Tech. rep., 2011-992-04, University of Calgary, 2011. 4

[LM10] LAM H., MUNZNER T.: *A Guide to Visual Multi-Level Interface Design From Synthesis of Empirical Study Evidence*. Synthesis Lectures on Visualization Series. Morgan Claypool, 2010. 4

[LMvW08] LI J., MARTENS J.-B., VAN WIJK J. J.: Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization 9*, 1 (2008), 13–30. 2

[LZVB06] LEBAN G., ZUPAN B., VIDMAR G., BRATKO I.: VizRank: Data Visualization Guided by Machine Learning. *Data Mining and Knowledge Discovery 13*, 2 (2006), 119–136. 9

[Mac67] MACQUEEN J. B.: Some methods for classification and analysis of multivariate observations. In *Proc. Berkeley Symposium on Mathematical Statistics and Probability* (1967), vol. 1, University of California Press, pp. 281–297. 4

[Mun09] MUNZNER T.: Visualization (Chapter 27). In *Fundamentals of Graphics*, 3rd ed. AK Peters, 2009, pp. 675–707. 5

[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. 4

[RB10] RENSINK R., BALDRIDGE G.: The perception of correlation in scatterplots. *Computer Graphics Forum (Proc. EuroVis 2010) 29*, 3 (2010), 1203–1210. 2

[SAP10] SAP: HANA, 2010. http://www.sap.com/hana/, last accessed 01/10. 4

[SNLH09] SIPS M., NEUBERT B., LEWIS J. P., HANRAHAN P.: Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum 28*, 3 (2009), 831–838. 1, 2, 3, 4, 5, 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, 4, 7

[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. ACM Advanced Visual Interfaces (AVI)* (2010), pp. 49–56. 1, 2, 9

[TF09] TODOROV V., FILZMOSER P.: An object oriented framework for robust multivariate analysis. *Journal of Statistical Software 32*, 3 (2009), 1–47. 4

[Uni11] UNIVERSITY OF MASSACHUSETTS: Statistical Data and Software Help, 2011. http://www.umass.edu/statdata/statdata/, last accessed 11/11. 4

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9*, 2579-2605 (2008), 85. 1, 4

[Vis11] VISUMAP TECHNOLOGIES INC.: VisuMap Data Repository, 2011. http://www.visumap.net/, last accessed 11/11. 4

[WA05] WILKINSON L., ANAND A.: Graph-theoretic scagnostics. *Proc. IEEE Symp. Information Visualization (InfoVis)* (2005), 157–164. 3, 6

[War11] WARD M. O.: Xmdv data repository, 2011. http://davis.wpi.edu/xmdv/datasets.html, last accessed 11/11. 4

[Wer23] WERTHEIMER M.: Untersuchungen zur lehre von der gestalt. ii. *Psychological Research 4*, 1 (1923), 301–350. 3, 6

[WW08] WILKINSON L., WILLS G.: Scagnostics Distribution. *Journal of Computational and Graphical Statistics 17*, 2 (2008), 473–491. 3, 4

[ZFLW02] ZAÏANE O., FOSS A., LEE C., WANG W.: On data clustering analysis: Scalability, constraints, and validation. *Advances in Knowledge Discovery and Data Mining* (2002), 28–39. 4