Supplemental Material
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 the following supplemental material along with the paper "A Taxonomy of Visual Cluster Separation Factors":

- Appendix A: Mathematical details about the measures used and the extensions we made
- Appendix B: Parameterization of the dimension reduction (DR) techniques we used
- Appendix C: A list of all datasets we analyzed in the qualitative data study
- Appendix D: Condensed list of codes resulting from the open coding process
- Appendix E: Plots of further grid size analysis
- Video 1: Lookup table of all 816 scatterplot representations we inspected in our study (AVI format, tested on VLC 1.1.12, no audio)
- Video 2: The interactive 3D data viewer we used in our study (MP4 format, tested on VLC 1.1.12, no audio)

Appendix A: Mathematical Details

This section provides the mathematical definitions of the cluster separation measures we used [SNLH09], as well as the extensions we made.

A.1. Original Definitions by Sips et al.

A.1.1. Centroid Measure

In their original work Sips et al. [SNLH09] explain the *centroid measure* for 2D scatterplots, or *Distance Consistency* (DSC) as they call it, as follows:

> "Given a data space $X \subseteq R_n$ and a class structure $C(X)$ defining m classes. Let c_i be a class and $centr(c_i)$ its centroid, and let x be $x \in X$ with $clabel(x) = i$. **CD** describes the property of class members that the distance $d(x, centr(ci))$ to its class centroid should be always minimal in comparison to the distance to all other centroids, thus
>
> $$d(x, centr(c_i)) < d(x, centr(c_j)) \forall j : 1 \leq j \leq m; j \neq i \tag{1}$$
>
> and d denotes a metric defined in X. $\mathbf{CD}(x, centr(c_i)) = true$ denotes that the centroid property for x and its centroid $centr(c_i)$ is fulfilled." [SNLH09]

Based on the **CD** property for each single point Sips et al. define the *Distance Consistency* measure **DSC** (= centroid measure):

> "Let $X \subseteq R_n$ be a n-D data set with k data points. Let $C(X)$ be a class structure of X defining m classes $C(X) = \{c_1, ..., c_m\}$. Let c_i be a class and $centr(c_i)$ its centroid in $C(X)$. Let $clabel(x)$ be the class label of a point $x \in X$. Let $v(X)$ be a 2-D view of X, then distance consistency $DSC(v(C))$ is defined as the classification error
>
> $$\mathbf{DSC} = \frac{|x' \in v(X) : \mathbf{CD}(x', centr'(c_{clabel(x)})) \neq true|}{k} \tag{2}$$
>
> with x' is the 2-D projection of the data point x and $centr'(c_i)$ is the 2-D projection of the centroid of class c_i." [SNLH09]

A.1.2. Grid Measure

Sips et al. [SNLH09] describe the *grid measure* for 2D scatterplots, or *Distribution Consistency* (DC) as they call it, as follows:

> "Let $C(X) = \{c_1, ..., c_m\}$ be a class structure of a high-dimensional data space $X \subseteq R_n$ describing m classes. Calling $p_c \equiv p_c c(x, y)$ as the number of data points of class $c \in C(X)$ in the region centered at

screen location x, y, the entropy of the class data probability density within the region

$$H(x,y) = -\sum_{c \in C(X)} \frac{p_c}{\sum p_c} \log_2(\frac{p_c}{\sum p_c}) \tag{3}$$

is a measure of consistency violation, having minimum value zero if the region contains data from only one class [...], and maximum value $\log_2 m$ if all m classes are mixed equally [...]. " [SNLH09]

Based on the **H** property for a single grid cell they define the *Distribution Consistency* measure **DC** (= grid measure):

"Let $C(X) = \{c_1, ..., c_m\}$ be a class structure of a high-dimensional data space $X \subseteq R_n$ describing m clusters. Let $v(X)$ be a 2-D view of X then distribution consistency $DC(v(X))$ is a integrated and weighted measure with

$$\mathbf{DC} = 100 - \frac{1}{Z} \sum_{x,y} p(x,y) H(x,y) \tag{4}$$

The $1/Z$ is a normalizing constant chosen to improve interpretability. We choose $100/\log_2(m) \sum_{x,y} \sum pc$ to give a score between 0 and 100." [SNLH09]

A.2. Usage and Extensions

In the data study presented, we use and extend these measures to judge three different visual encoding techniques, 2D scatterplots (2D), 3D scatterplots (3D), and scatterplot matrices (SPLOMs). For all of them we compute:

1. m one-against-all *class-wise* measures, where m equals the number of classes
2. one *overall* measure

Based on the definitions from Sips et al. (see above), we use and extend the centroid and grid measure as follows:

A.2.1. Centroid Measure

2D Scatterplot: For all classes $c \in C(X)$, we compute a *class-wise* value $\mathbf{Cent_{2D_c}}$ as follows:

$$\mathbf{Cent_{2D_c}} = \frac{|x'_c \in v(X) : \mathbf{CD}(x'_c, centr'(c_{clabel(x)})) \neq true|}{k_c} \tag{5}$$

with $\{x'_c | \forall x' : clabel(x) = c\}$ and k_c as the number of the data points in class c.

For the *overall* measure we use the original measure as described in (2):

$$\mathbf{Cent_{2D}} = \mathbf{DSC} \tag{6}$$

Note that there is an alternative way to derive the *overall* value from the m *class-wise* values as follows:

$$\mathbf{Cent_{2D}} = \frac{\sum_{\mathbf{c} \in \mathbf{C(X)}} \mathbf{Cent_{2D_c} k_c}}{\mathbf{k}} = \mathbf{DSC} \tag{7}$$

3D Scatterplot: We simply extend the euclidean distance measure used in (1) from 2D to 3D and compute the *class-wise* and *overall* measures as for the 2D Scatterplots: (5) and (6).

SPLOM: Let $V = \{v_1, ..., v_n\}$ be all $n = d(d-1)/2$ 2D Scatterplot views of a d-dimensional SPLOM ($d \times d$ SPLOM) and $\mathbf{Cent_{2D_c}}(v_i)$ the *class-wise* value of class c in the 2D Scatterplot view v_i as defined in (5). For all classes $c \in C(X)$, we define the *class-wise* measure of a SPLOM as the highest score of all 2D views v_i:

$$\mathbf{Cent_{SPLOM_c}} = \max(\mathbf{Cent_{2D_c}}(v_i)) \tag{8}$$

We define the *overall* SPLOM measure as the weighted sum of all *class-wise* values:

$$\mathbf{Cent_{SPLOM}} = \frac{\sum_{c \in C(X)} \mathbf{Cent_{SPLOM_c}} k_c}{k} \tag{9}$$

A.2.2. Grid Measure

2D Scatterplot: For each class $c \in C(X)$, we compute the *class-wise* measure as follows. Let δ_c be the grid cell at position x, y with $\forall \delta_c \exists x' : clabel(x) = c$. We then define the *class-wise* 2D grid measure as:

$$\mathbf{Grid_{2D_c}} = 100 - \frac{1}{Z} \sum_{\delta_c} p(\delta_c) H(\delta_c) \tag{10}$$

using the definition of H as given in (3). The class-wise measure of a class c is therefore the entropy measure H applied to all grid cells that at least contain one point of class c.

For the *overall* measure we use the original measure as described in (4):

$$\mathbf{Grid_{2D}} = \mathbf{DC} \tag{11}$$

Note that based on this definition there is no obvious way to derive the *overall* value from the *class-wise* values, as there is for the centroid measure.

We use a dynamic grid size $g \times g$ derived from k, the number of points of the dataset:

$$g = \lfloor \sqrt[2]{k} \rfloor \tag{12}$$

3D Scatterplot: For 3D *class-wise* measures we change the definition of δ_c to be the grid cell at position x, y, z with $\forall \delta_c \exists x' : clabel(x) = c$ and use the formula given in (10).

For the *overall* measure, we extend the formulas given in (3) and (4) as follows:

$$H(x,y,z) = -\sum_{c \in C(X)} \frac{p_c}{\sum p_c} \log_2(\frac{p_c}{\sum p_c}) \tag{13}$$

$$\mathbf{Grid_{3D}} = 100 - \frac{1}{Z} \sum_{x,y,z} p(x,y,z) H(x,y,z) \tag{14}$$

choosing $100/\log_2(m) \sum_{x,y,z} \sum pc$ for $1/Z$.

For 3D Scatterplots we derive the grid size $g \times g$ as follows:

$$g = \lceil \sqrt[3]{k} \rceil \tag{15}$$

assuring that the number of grid cells is equal or slightly larger as the 2D grid size defined in (12).

SPLOM: Similarly, we define the *class-wise* value $\mathbf{Grid_{SPLOM_c}}(v_i)$ as the best *class-wise* score of all 2D projections v_i as defined in (10):

$$\mathbf{Grid_{SPLOM_c}} = \max(\mathbf{Grid_{2D_c}}(v_i)) \tag{16}$$

We define the *overall* SPLOM measure as the weighted sum of all *class-wise* values:

$$\mathbf{Grid_{SPLOM}} = \frac{\sum_{c \in C(X)} \mathbf{Grid_{SPLOM_c}} k_c}{k} \tag{17}$$

For all SPLOM computations, we use grid sizes as defined in (12).

Appendix B: DR Parameterization

In our data study, we use four different dimension reduction (DR) techniques, which we instantiated and parameterized as follows:

PCA [Jol02]: We use R's [R11] standard PCA implementation `princomp` with default parameterization.

MDS [BG05]: For performance reasons we used the Glimmer MDS implementation provided courtesy of Ingram et al. [IM09]. We used their Java CPU version, with the following parameterization:
`Near and Random Set Size = 10`
`Termination Threshold = 1e-4`

RobPCA [TF09]: We use R's robust PCA implementation `PcaCov` from the `rrcov` package with `cov.control=CovControlMest()`.

t-SNE [vdMH08]: We use R's t-SNE implementation `tsne` from the package `tsne`. We set the maximum number of iterations to perform:
`max_iter = 500`. Except for this, we used the default parameterization.

Appendix C: Full list of datasets

ID	Name	Points	Dimensions	Classes	Provenance
			real		
1	abalone	4154	7	28	uci [FA10]
2	bbdm13	200	13	5	umass [Uni11]
3	bostonHousing	155	13	3	uci [FA10]
4	breastCancer-diagnostic	569	30	2	uci [FA10]
5	breastCancer-original	454	9	2	uci [FA10]
6	cars-1	7404	22	2	colleagues [TAE*09]
7	cars-2	7404	22	53	colleagues [TAE*09]
8	cars-3	7404	22	12	colleagues [TAE*09]
9	cereal	77	12	7	xmdv [War11]
10	ecoliProteins	332	7	8	visumap [Vis11]
11	eFashion	3272	4	8	sap [SAP10]
12	fisheriesEscapementTarget	121	12	11	colleagues [HB11]
13	fisheriesHarvestRule	121	12	11	colleagues [HB11]
14	hiv	78	159	6	colleagues [SNLH09]
15	industryIndices	102	6	13	uci [FA10]
16	ionosphere	351	34	2	visumap [Vis11]
17	iris	147	4	3	uci [FA10]
18	musicNetGroups	171	9	6	visumap [Vis11]
19	olive	572	8	3	colleagues [SNLH09]
20	pageBlocks	5473	10	5	uci [FA10]
21	parkinson	195	11	2	uci [FA10]
22	shuttle-big	43500	9	7	uci [FA10]
23	shuttle-small	14500	9	7	uci [FA10]
24	spamBase	4601	57	2	uci [FA10]
25	swanson	1875	6	3	xmdv [War11]
26	tse300	244	49	8	visumap [Vis11]
27	wine	178	13	3	uci [FA10]
28	world-10d	151	10	5	visumap [Vis11]
29	world-12d	151	12	5	visumap [Vis11]
30	worldMap	192	3	13	visumap [Vis11]
31	yeast	1452	8	10	uci [FA10]

Table 1: Real datasets. In order to make all dimension reduction techniques we used in our study work, we had to preprocess some of the original data sources, e. g., deleting duplicated data points, or deleting non-numeric dimensions. The table shows the data as used in the study.

ID	Name	Points	Dimensions	Classes
	synthetic-entangled			
32	entangled1-3d-3cl-separate	600	3	3
33	entangled1-3d-4cl-separate	400	3	4
34	entangled1-3d-5cl-separate	500	3	5
35	entangled2-10d-adjacent	1490	10	10
36	entangled2-10d-overlap	1479	10	10
37	entangled2-15d-adjacent	2049	15	15
38	entangled2-15d-overlap	2318	15	15
39	entangled2-3d-adjacent	1098	3	3
40	entangled2-3d-overlap	857	3	3
41	entangled2-4d-adjacent	1254	4	4
42	entangled2-4d-overlap	538	4	4
43	entangled2-5d-adjacent	741	5	5
44	entangled2-5d-overlap	696	5	5
45	entangled2-6d-adjacent	837	6	6
46	entangled2-6d-overlap	1034	6	6
47	entangled3-l-3d-bigOverlap	571	3	3
48	entangled3-l-3d-smallOverlap	496	3	3
49	entangled3-m-3d-adjacent	309	3	3
50	entangled3-m-3d-bigOverlap	325	3	3
51	entangled3-m-3d-smallOverlap	292	3	3
52	entangled3-s-3d-adjacent	185	3	3
53	entangled3-s-3d-bigOverlap	205	3	3
54	entangled3-xl-3d-adjacent	1821	3	3
55	entangled3-xl-3d-bigOverlap	1892	3	3
	synthetic-gaussian			
56	gauss-n100-10d-3largeCl	100	10	3
57	gauss-n100-10d-3smallCl	100	10	3
58	gauss-n100-10d-5largeCl	100	10	5
59	gauss-n100-10d-5smallCl	100	10	5
60	gauss-n100-5d-3largeCl	100	5	3
61	gauss-n100-5d-3smallCl	100	5	3
62	gauss-n100-5d-5largeCl	100	5	5
63	gauss-n100-5d-5smallCl	100	5	5
64	gauss-n500-10d-3largeCl	500	10	3
65	gauss-n500-10d-3smallCl	500	10	3
66	gauss-n500-10d-5largeCl	500	10	5
67	gauss-n500-10d-5smallCl	500	10	5
68	gauss-n500-5d-3largeCl	500	5	3
69	gauss-n500-5d-3smallCl	500	5	3
70	gauss-n500-5d-5largeCl	500	5	5
71	gauss-n500-5d-5smallCl	500	5	5
	synthetic-grid			
72	grid-3d	1000	3	8
73	grid-4d	1296	4	16
74	twoSquare	968	3	4
75	unevenDensity	905	3	2

Table 2: Synthetic datasets we generated.

Appendix D: Merged Codesets

After both open coding passes, we merged the codesets from the two investigators into a single list, one for visual separation factors in datasets after coding pass 1, and one for failure causes of the measures after coding pass 2. The two merged codeset lists can be found bellow:

Code	Description
3D-move	Movement of points in 3D helps to detect a cluster (Gestalt law: Common fate)
adjacent	Adjacent classes: no physical distance between classes
bad	Clusters heavily intermixed
bg-noise	The background noise in 3D makes detectability of clusters harder
entangled	Dataset seems to have entangled structures
gaussian	Clusters look gaussian
going-high	Going to higher dimensionality in SPLOMs does not seem to add a lot more info (higher than 5x5)
good	Clusters nicely separable
equidistant-mixed	Dataset with partly or fully overlapped classes and equidistant point structure
inner-class	Inner clusters usually okay in 2D and SPLOM but not in 3D
interesting	Example with interesting data characteristics
periphery	Clusters at the periphery easier to spot
mental	Mental model helps understanding the class structure
more-views	Different views in a SPLOM help to identify different classes
outlier	Outliers exist
shape	The shape of a cluster is important for its detection
sparse	Data of a class is very sparsely distributed in view
stringy	Data and/or clusters have a stringy shape
layery	Data and/or clusters form a layer in 3D
tsne-not-good	t-SNE does not work well for this example
tsne-great	t-SNE successfully untangles some structure that was not visible by using linear techniques
unbalanced-classes	Classes differ strongly in no. of points / class
validation	Validation of class structure in other views of a SPLOM is helpful
varying-density	Clusters have different densities
z-depth	Z-depth might influence your decision

Table 3: Merged list of codes from open coding phase 1, where we coded dataset instances for general separability factors we observed.

Code	Description	Error Type	Measure
adjacent-classes	Non-round clusters can cause false negatives for both measures	FN	grid
bg-noise	Bg noise in 3D hinders that you see a certain class	TP	both
big-class	Big class is overshadowing small classes	FP	centroid
clumpy	Clumpy class leads to an awkward position of the centroid / Entropy cannot detect the connection between points within a class	FP / FN	both
equidistant-mixed	Equidistant layouts of overlapping classes are counterproductive for the grid measure because they can easily lead to false positives	FP	grid
grid-too-coarse	Measure artifact: it seems that having a finer grid will give a better result	FP	grid
grid-too-fine	Measure artifact: it seems that having a coarser grid will give a better result	FN	grid
identical-classes	Nearly Identical or completely identical classes lead to very similar centroids	FP	centroid
many-classes	If there are a lot of classes that are not perfectly separated it usually is hard to see structure; even though the measure indicates that there is structure	FP	both
mixed-classes	Both, centroid and grid measure can have severe problems with detecting strongly overlapping classes	FP	both
outliers	Outliers can influence measures in a disadvantageous way	FP / FN	both
overlaid-shapes	Visually well separable based on Gestalt perception; overlapping shapes cannot be detected by measures	FN	both
periphery	Classes with centroids at the periphery but strongly mixed with other ones can lead to FPs	FP	centroid
shape	Shapes might lead to strange centroids: this can lead to not properly detecting the shape (FN)	FN	centroid
similar-centroid	Similar centroids for some actually nicely separable classes	FN	centroid
small-class	Small classes are generally hard to spot and might be easily overshadowed by other bigger classes	FP	centroid
sparse-class	Sparse classes can lead to interfering centroids / Chance to have only one point per bin is high	FP	both
split	Classes are split by another class	FP / FN	both
splom-exacerbate	Issue of FP exacerbates with higher-dimensional SPLOMs	FP	both
splom-wrong-pick	Measure picked poor view(s) in the SPLOM	FP	both
variable-density	Classes with varying densities can influence the performance of the measure	FP	centroid
z-depth	Z-depth led to a false picture of what is really there	TP / TN	both

Table 4: Merged list of codes from open coding phase 2, where we coded failure cases for reasons why centroid and/or grid measure were not able to provide reliable results. Error types: FP = False Positive; FN = False Negative; TP = True Positive; TN = True Negative. TPs and TNs were excluded from failure cases.

Appendix E: Plots of Grid Size

To test the grid measure's robustness against varying grid sizes, we recomputed and plotted the measure values of 58 failure cases (50 false positives and 8 false negatives) with different grid size parameterizations. Here, we present the plots we used for our analysis:

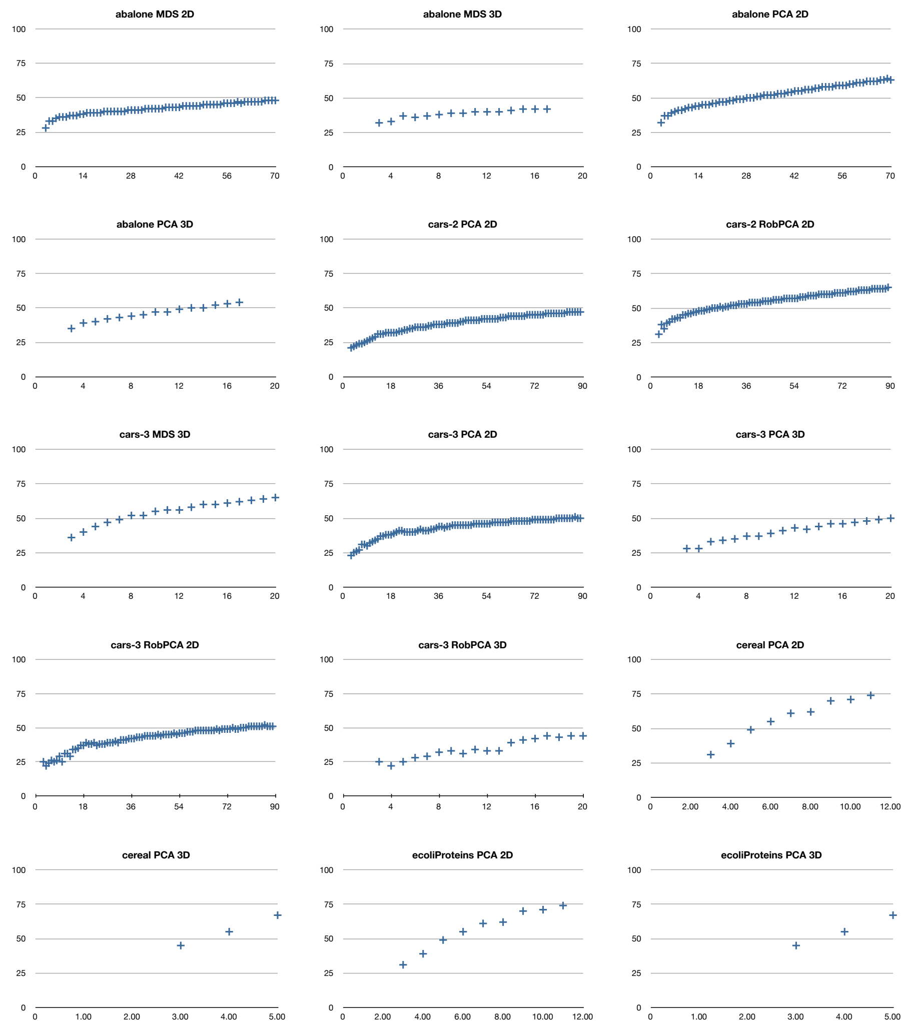

Figure 1: Plots of grid size variations for false positives (part 1): The horizontal axis shows different parameters of the grid size $g \times g$. The vertical axis shows the resulting measure values we got by computing it with these grid size parameterizations. For false positives, we expected the grid to be too fine, and therefore varied it step-wise to be more coarse. The original value we judged in the study is the right-most value in the graph.

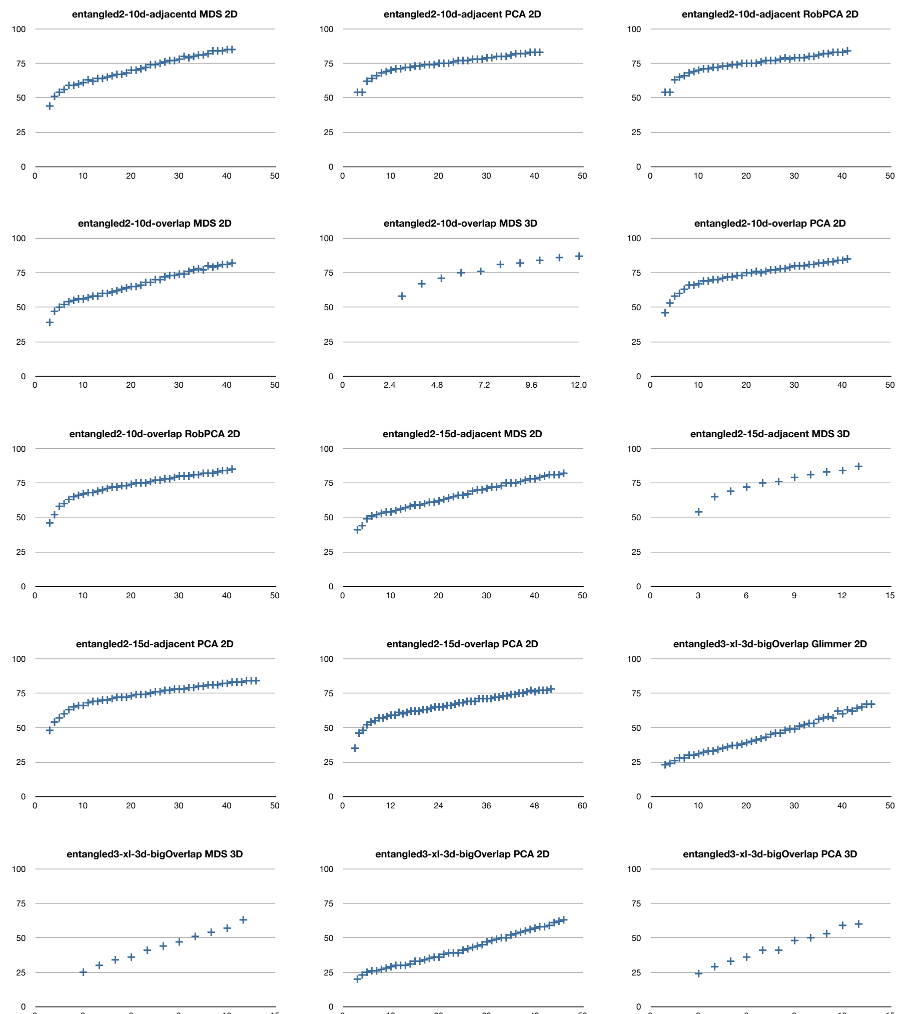

Figure 2: Plots of grid size variations for false positives (part 2)

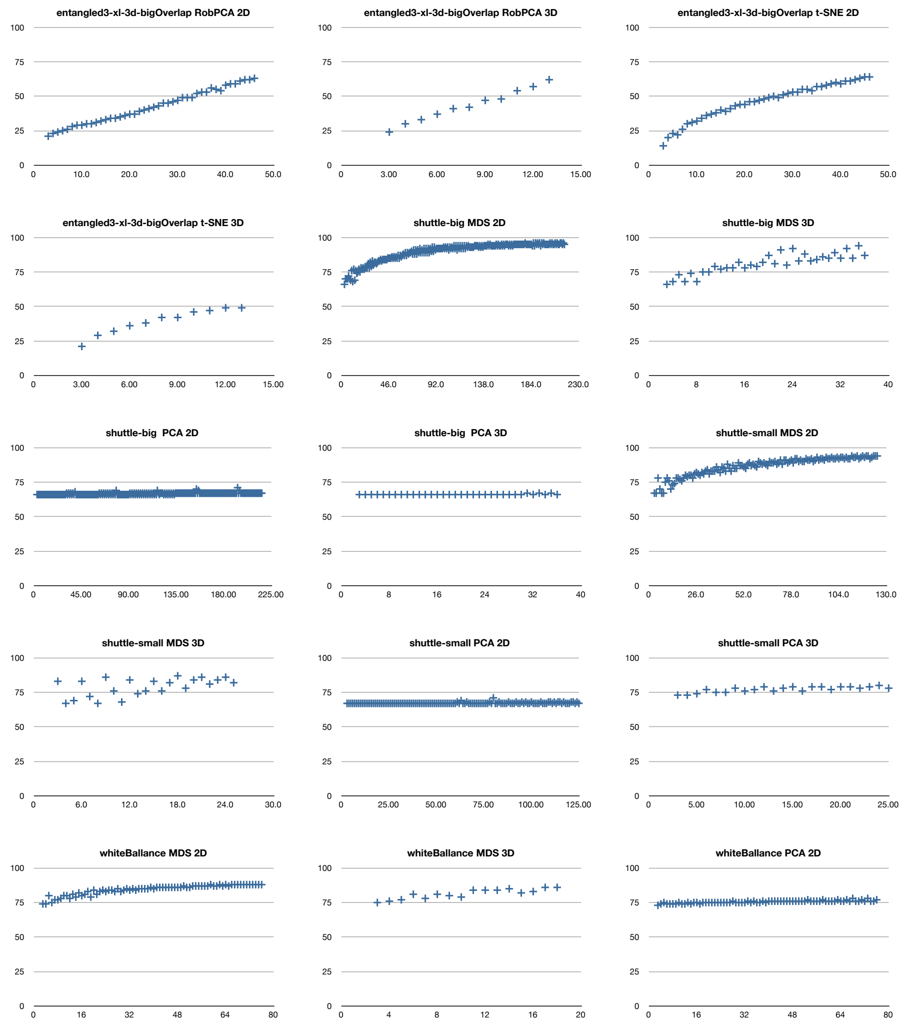

Figure 3: Plots of grid size variations for false positives (part 3)

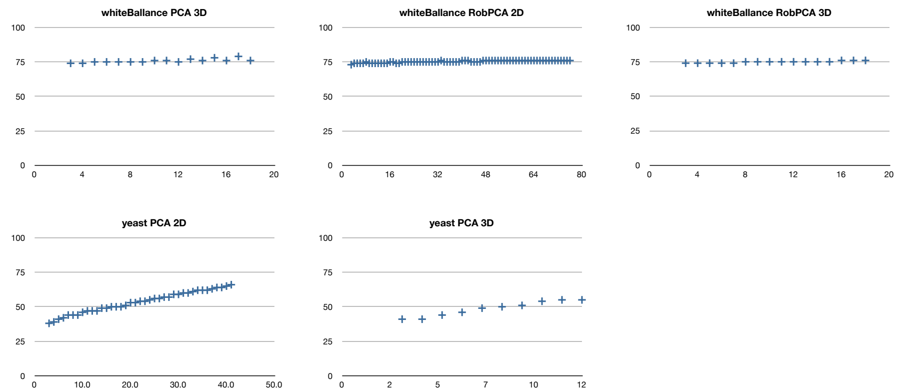

Figure 4: Plots of grid size variations for false positives (part 4)

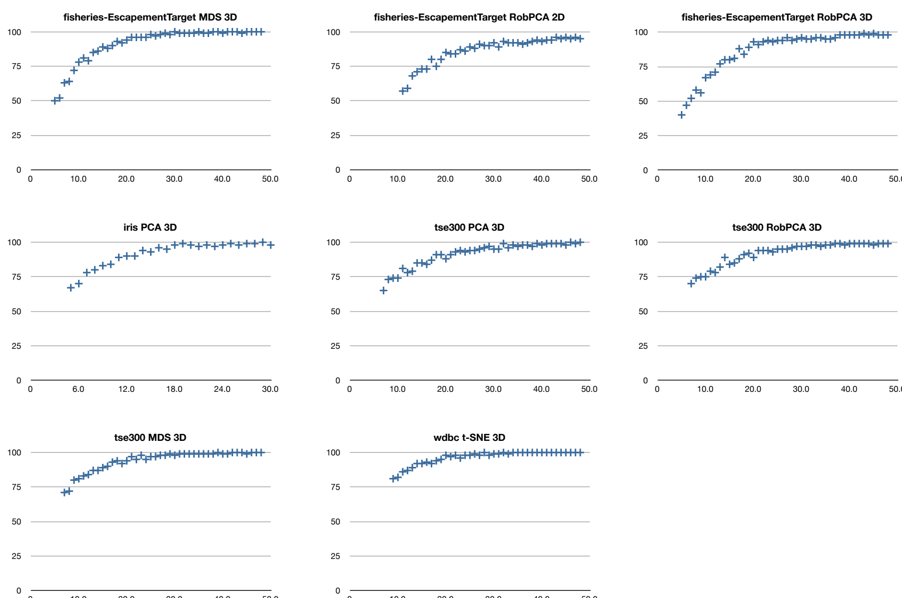

Figure 5: Plots of grid size variations for false negatives: For false negatives, we expected the grid to be too coarse, and therefore varied it step-wise to be finer. The original value we judged in the study is the left-most value in the graph.

References

[BG05] Borg I., Groenen P.: *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

[FA10] Frank A., Asuncion A.: University of California Irvine (UCI) Machine Learning Repository, 2010.

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

[IM09] Ingram S., Munzner T.: Glimmer : Multilevel MDS on the GPU. *IEEE Trans. Visualization and Computer Graphics (TVCG) 15*, 2 (2009), 249–261.

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

[R11] R: A language and environment for statistical computing, 2011. http://www.R-project.org, last accessed 12/11.

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

[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.

[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.

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

[Uni11] University of Massachusetts: Statistical Data and Software Help, 2011. http://www.umass.edu/statdata/statdata/, last accessed 11/11.

[vdMH08] van der Maaten L., Hinton G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9*, 2579-2605 (2008), 85.

[Vis11] VisuMap Technologies Inc.: VisuMap Data Repository, 2011. http://www.visumap.net/, last accessed 11/11.

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