Towards Perceptual Optimization of the Visual Design of Scatterplots

Luana Micallef, Gregorio Palmas, Antti Oulasvirta, Tino Weinkauf

1 Calibration Study: Sub-Studies

In this supplement, we describe each calibration sub-study separately. For further details, refer to the paper. All data will be made available on our project homepage.

1.1 General

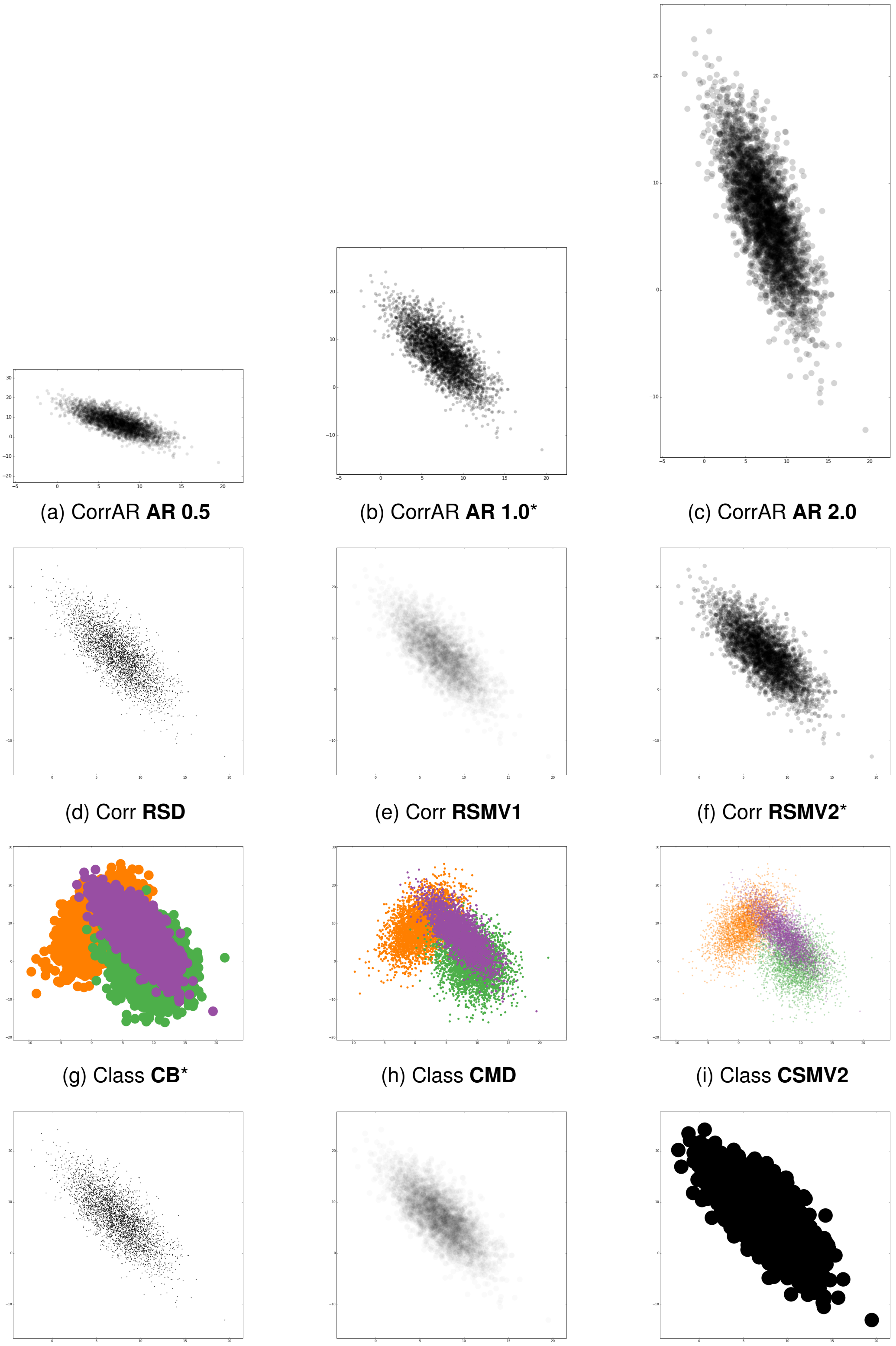
Table 1 reports the candidate weight sets in each sub-study. The candidates were chosen by three co-authors through an extensive visualization of the weight space. We generated candidate weight sets that produce designs with different characteristics. Different weight sets produce different designs for the same data. Hence, certain weight sets support a particular data analysis task better than others (Figure 1). We tested this empirically by conducting four sub-studies each handling a specific task:

- **CorrAR**: Correlation estimation (with aspect ratio only)
- **Corr**: Correlation estimation (without aspect ratio)
- **Class**: Class Separation
- **Outl**: Outlier detection

For statistical testing, we preferred non-parametric tests (Friedman rank sum) due to non-normality of our dependent variables. Because we were interested in the *generally* best weight settings, our statistical testing focused on weight set as the factor and collapsed data over all other conditions. Non-parametric statistics is a widely accepted scientific practice [3]. However, recent work [1] raises issues against practices in inferential statistics, so we also report box plots.

Task /	**Weight set candidate**	**Weight variables**										**Empirical results**			
Sub-study	(with its design characteristics)	w_α	w_r	w_μ	w_σ	$w_{\bar{\mu}}$	$w_{\bar{\sigma}}$	w_ℓ	w_p	w_c	w_o	**N. Obs.**	**Success**	**Error**	**Time (s)**
CorrAR	Aspect ratio **0.5**	n/a	n/a	-½	0	½	½	-½	½	0	0	1206	0.47	0.16	4.05
	Aspect ratio **1.0***	n/a	n/a	-½	0	½	½	-½	½	0	0	1229	0.58	0.12	3.62
	Aspect ratio **2.0**	n/a	n/a	-½	0	½	½	-½	½	0	0	1194	0.45	0.17	4.12
Corr	**RSD**: classic design with small, dark dots	½	1	-1	0	½	½	½	0	0	0	1505	0.60	0.12	3.97
	RSMV1: smooth design with merging light dots	½	1	½	0	½	½	-½	½	0	0	1523	0.60	0.11	3.92
	RSMV2*: smooth design with merging dark dots	½	1	-½	0	½	½	-½	½	0	0	1531	0.65	0.10	3.98
Class	**CB***: blob design with large, dark dots	½	1	-½	0	0	-½	-½	0	½	0	930	0.87	-	7.02
	CMD: blob design with medium, dark dots	½	1	0	0	0	-½	1	-½	½	0	932	0.83	-	7.56
	CSMV2: smooth design with merging dark dots	½	1	-½	0	½	½	-½	½	½	0	948	0.83	-	7.57
Outl	**OSD**: classic design with small, dark dots	0	0	-1	0	½	½	½	0	0	0	1171	0.90	0.12	5.18
	OSMV1: smooth design with merging light dots	0	0	0	0	½	½	-½	½	0	0	1142	0.76	0.49	5.81
	OSMV1_OP*: smooth design with merging light dots & outlier perceivability	0	0	0	0	½	½	-½	½	0	1	1187	0.93	0.09	3.31

Table 1: Weight factors were calibrated empirically over four crowdsourced experiments. This table reports the weight set candidates and results. * - winner (selected for final optimization)



(a) CorrAR **AR 0.5**

(b) CorrAR **AR 1.0***

(c) CorrAR **AR 2.0**

(d) Corr **RSD**

(e) Corr **RSMV1**

(f) Corr **RSMV2***

(g) Class **CB***

(h) Class **CMD**

(i) Class **CSMV2**

(j) Outl **OSD**

(k) Outl **OSMV1**

(l) Outl **OSMV1_OP***

Figure 1: Plots for the same data (with the exception that plots a-f and j-l show only one of the three classes in the data) optimized for different tasks using the candidate weight sets in Table 1. * - winner (selected for final optimization)

1.2 *CorrAR* Sub-study: Correlation Estimation–Aspect Ratio

1.2.1 Method

Participants: 69 workers were recruited from 31 countries (45 male; mean age 34).

Experimental Design: The study followed a 3 x 3 x 6 within-subjects design with 3 dataset sizes (100, 1,000, and 10,000), 3 aspect ratios (0.5, 1, 2), and 6 correlation coefficients (*r*s = {-0.75, -0.5, -0.25, 0.25, 0.5, 0.75}) as factors.

Procedure, Task and Materials: In this sub-study, users were shown a plot generated from a single distribution on left, and they had to mark the correlation using buttons: *“What linear correlation is shown by this scatterplot?* Choices: -1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1, No clue”.

1.2.2 Results

One of the aspect ratios (1.0) was superior to others in all dependent variables. There was a statistically significant effect of aspect ratio on success rate, $\chi^2(2) = 15.12, p < 0.01$, error, $\chi^2(2) = 16.88, p < 0.01$, and on time ($\chi^2(2) = 1.19.83, p < 0.01$. Post hoc tests using Wilcoxon Signed-rank test (Holm’s adjustment) showed significant differences to the two aspect ratios in all three dependent variables. This shows that an aspect ratio of 1 is beneficial for estimating correlation and justifies the definition of our measures E_α, E_r. Box plots of the measured dependent variables for each candidate weight set are shown in Figure 2.

1.3 *Corr* Sub-study: Correlation Estimation–no Aspect Ratio

1.3.1 Method

Participants: 86 workers were recruited from 33 countries (72 male; mean age 33).

Experimental Design: The study followed a 3 x 3 x 6 within-subjects design with 3 weight set candidates, 3 dataset sizes (100, 1,000, and 10,000), and 6 correlation coefficients (*r*s = {-0.75, -0.5, -0.25, 0.25, 0.5, 0.75}) as factors.
Other aspects were the same as in CorrAR sub-study.

1.3.2 Results

One of the weight sets (RSMV2) was superior to others in both success rate and error but not in time. There was a statistically significant effect of weight set on success rate, $\chi^2(2) = 11.65, p < 0.01$, as well as on error, $\chi^2(2) = 11.21, p = 0.02$. Wilcoxon Signed-rank test (Holm’s) showed that RSMV2 had significantly lower error and better success rate than the two other sets. There was no significant effect on time ($\chi^2(2) = 1.32, p = 0.52$. Box plots of the measured dependent variables for each candidate weight set are shown in Figure 3.

1.4 *Class* Sub-study: Class Separation

1.4.1 Method

Participants: 100 workers were recruited from 30 countries (65 male; mean age 33.5).

Experimental Design: The study followed a 3 x 3 x 6 within-subjects design with 3 weight set candidates, 3 dataset sizes (100, 1,000, and 10,000), and 4 class counts (2, 3, 4, 5) as factors.

Procedure, Task, and Materials: On each page, a multi-class scatterplot was shown. The task was to choose a representation on the right that best matched the shown scatterplot: *“Which of these images matches the scatterplot on the left?”* The choices were: 1) image with covariance ellipses corresponding to the scatterplot, 2) image with covariance ellipses corresponding to the scatterplot but with 1 class removed,

3) image with covariance ellipses corresponding to the scatterplot but with 1 class replaced, 4) "No clue" button. In other respects, the method was the same as in CorrAR sub-study.

1.4.2 Results

One of the weight sets (CB) was superior to others in both success rate and time. There was a statistically significant effect of weight set on success rate, $\chi^2(2) = 21.29, p < 0.01$, as well as on time: $\chi^2(2) = 35.63, p < 0.01$. Post hoc tests using Wilcoxon Signed-rank test (Holm's) showed significant differences to the two other conditions in both dependent variables. Box plots of the measured dependent variables for each candidate weight set are shown in Figure 4.

1.5 *Outl* Sub-study: Outlier Detection

1.5.1 Method

Participants: 82 workers were recruited from 34 countries (72 male; mean age 35.7).

Experimental Design: The study followed a 3 x 3 x 6 within-subjects design with 3 weight set candidates, 3 dataset sizes (100, 1,000, and 10,000), and 4 outlier counts (2, 3, 4, 5) as factors.

Procedure, Task, and Materials: Outliers were defined using modified Z-score > 2.5 [2] as threshold. On each page, a multi-class scatterplot was shown. User was asked to mark how many outliers it showed: *"How many outliers are shown in the scatterplot?* Choices: 0, 1, 2, 3, 4, 5, No clue".

1.5.2 Results

One of the weight sets (OSMV1_OP) was superior to others in all three dependent variables. There was a statistically significant effect of weight set on success rate, $\chi^2(2) = 55.50, p < 0.01$, on error, $\chi^2(2) = 56.93, p < 0.01$, as well as on time: $\chi^2(2) = 145.47, p < 0.01$. Post hoc tests using Wilcoxon Signed-rank test (Holm's) showed significant differences to the two other sets across the three dependent variables. Box plots of the measured dependent variables for each candidate weight set are shown in Figure 5.

1.6 Summary

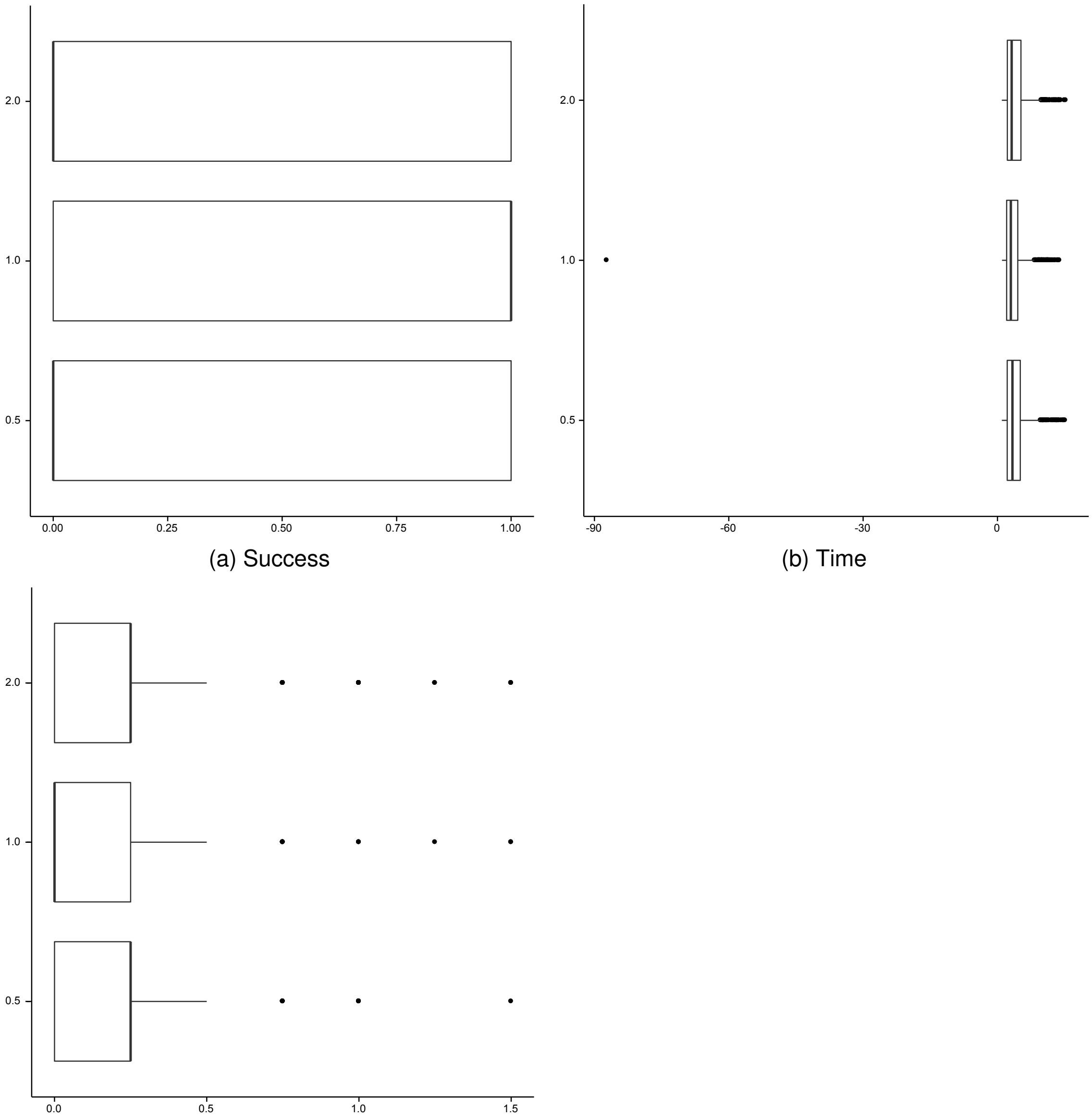
Superior weight sets were found for each task. The weight sets we select for the optimizer are superior for at least two of the dependent variables. The selected sets are marked with asterisks in Table 1.

References

[1] P. Dragicevic. Fair statistical communication in hci. In *Modern Statistical Methods for HCI*, pages 291–330. Springer, 2016.

[2] B. Iglewicz and D. C. Hoaglin. *How to detect and handle outliers*, volume 16. Asq Press, 1993.

[3] C. Jacob. The earth is round (p<. 05). *American Psychologist*, pages 997–1003, 1994.

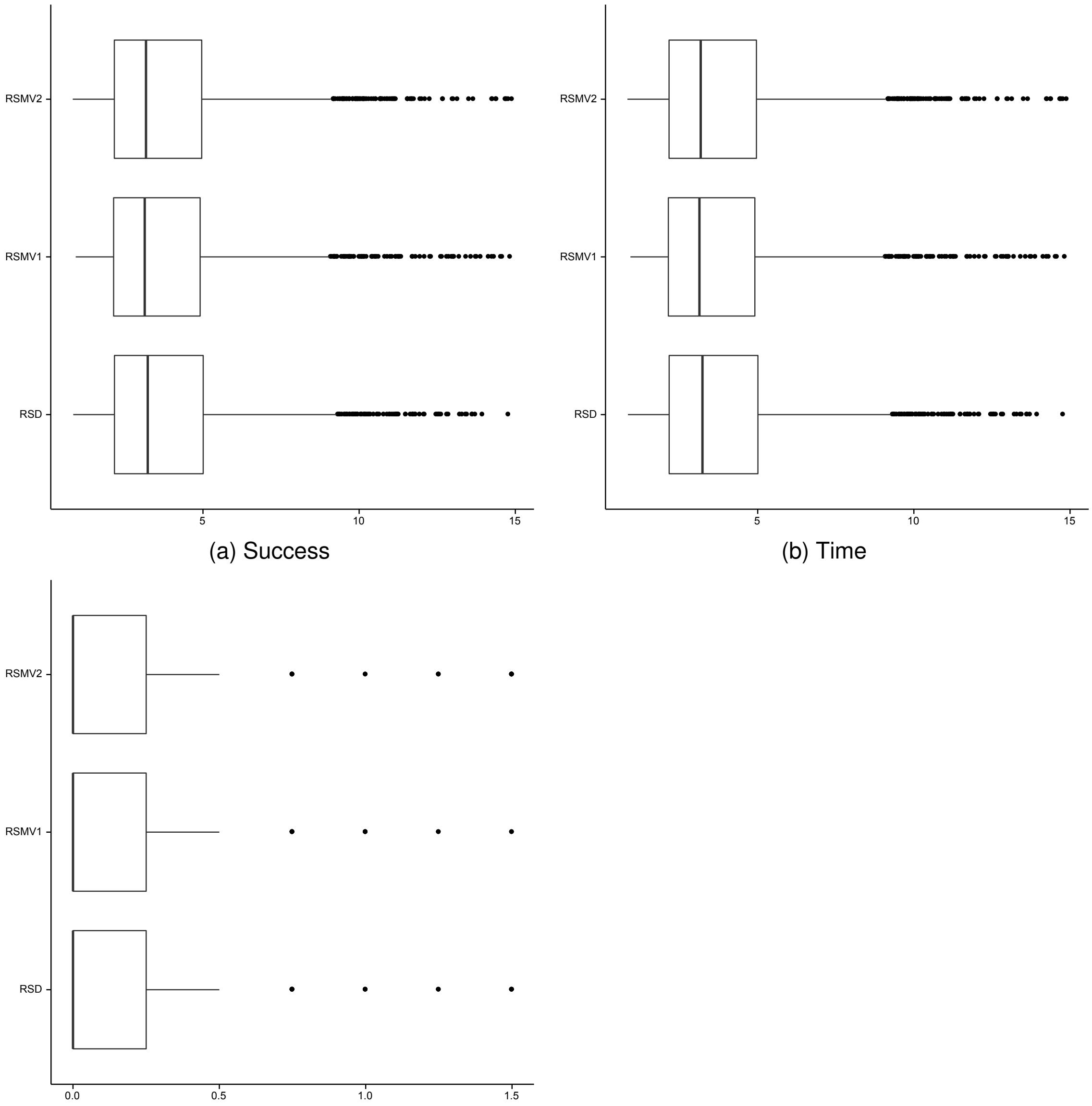


(a) Success

(b) Time

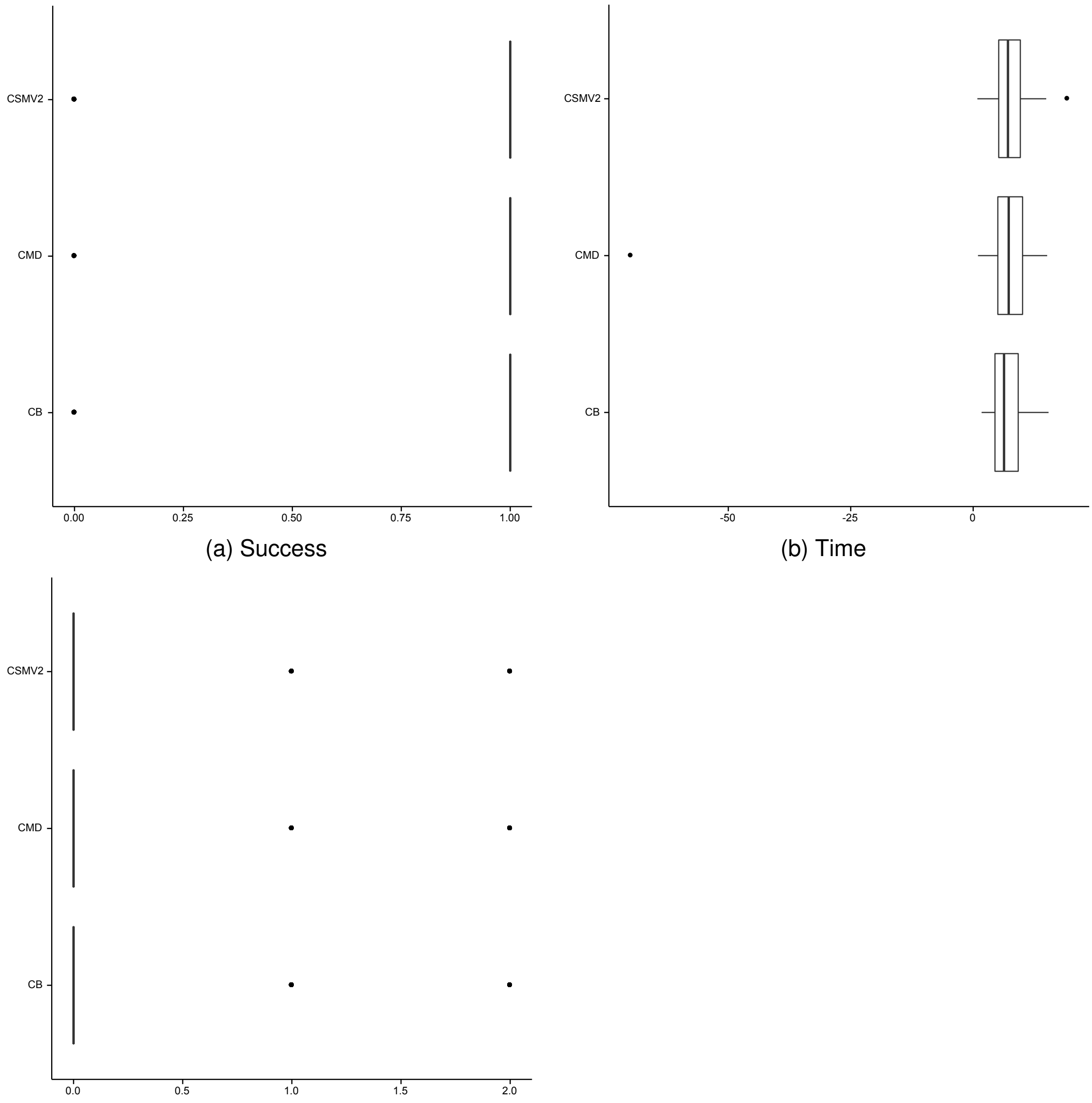
(c) Error

Figure 2: Box plot of the measured dependent variables (a) Success, (b) Time and (c) Error of each candidate weight set in the CorrAR sub-study.



(c) Error

Figure 3: Box plot of the measured dependent variables (a) Success, (b) Time and (c) Error of each candidate weight set in the Corr sub-study.

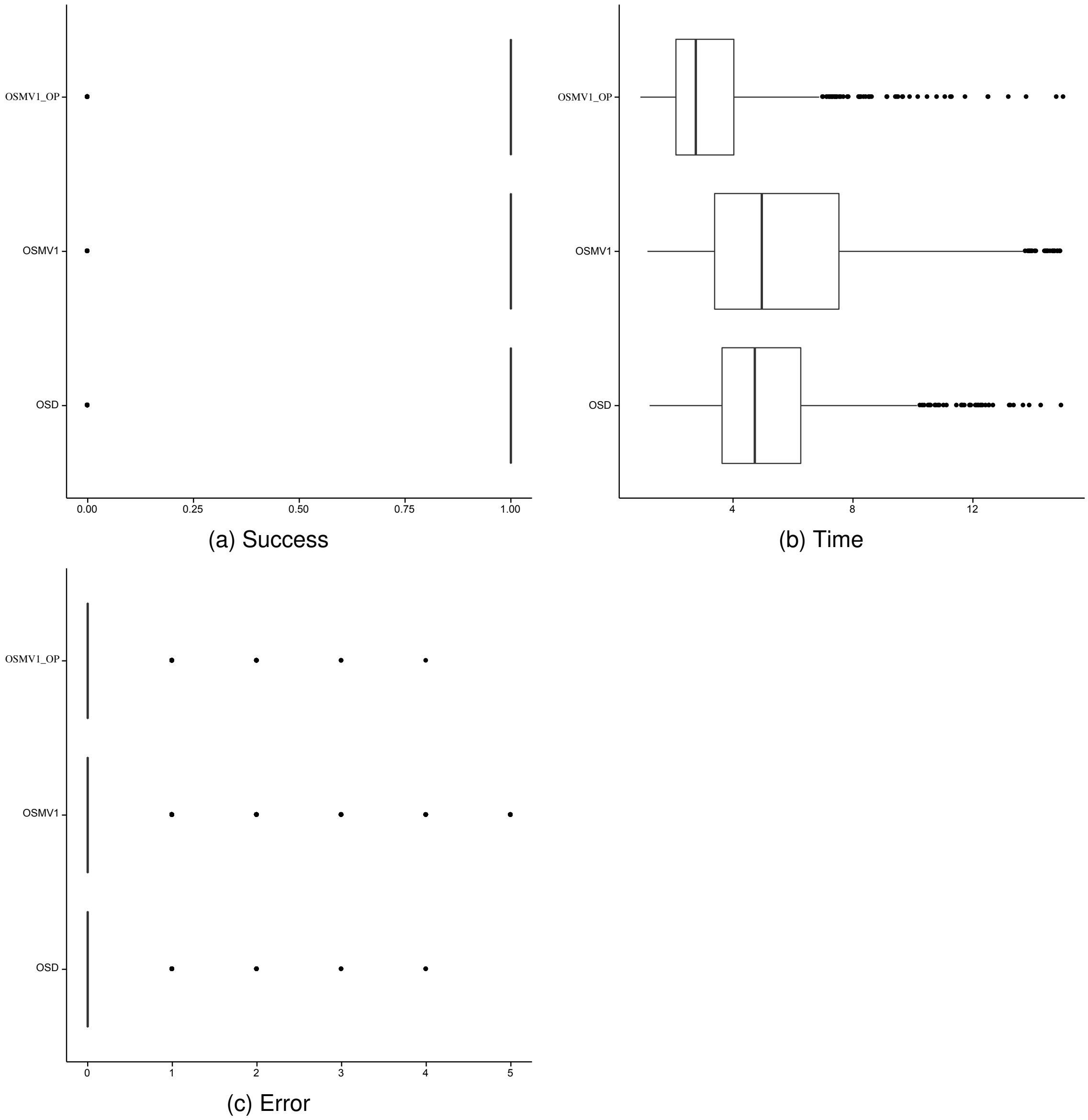


(a) Success

(b) Time

(c) Error

Figure 4: Box plot of the measured dependent variables (a) Success, (b) Time and (c) Error of each candidate weight set in the Class sub-study.



(a) Success

(b) Time

(c) Error

Figure 5: Box plot of the measured dependent variables (a) Success, (b) Time and (c) Error of each candidate weight set in the Outl sub-study.