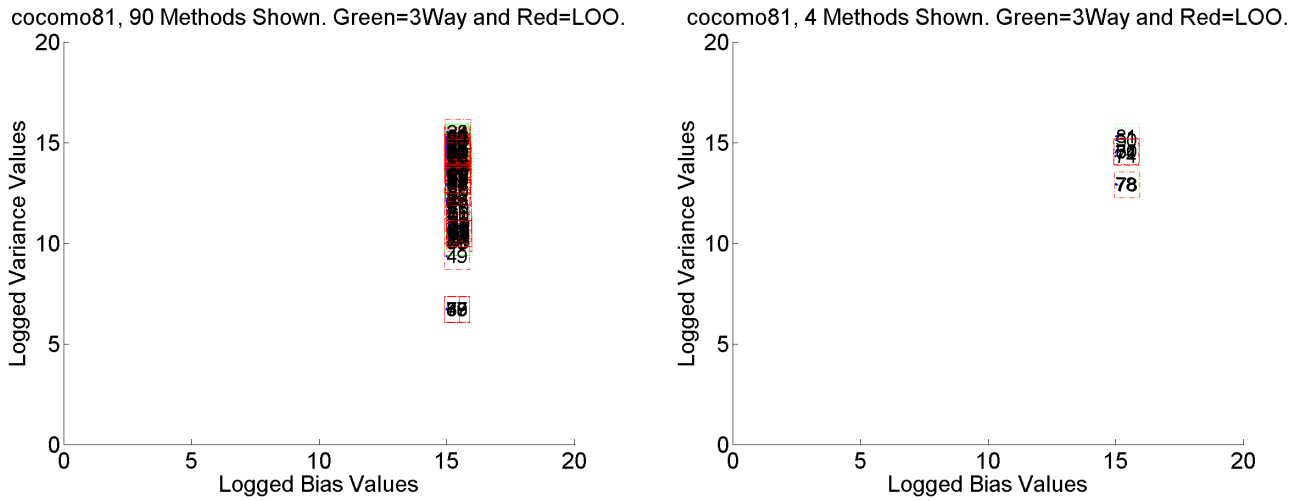
1 BIAS AND VARIANCE PROBLEM BETWEEN LOO AND 3WAY

1.1 Bias-Variance with and Without Statistical Test

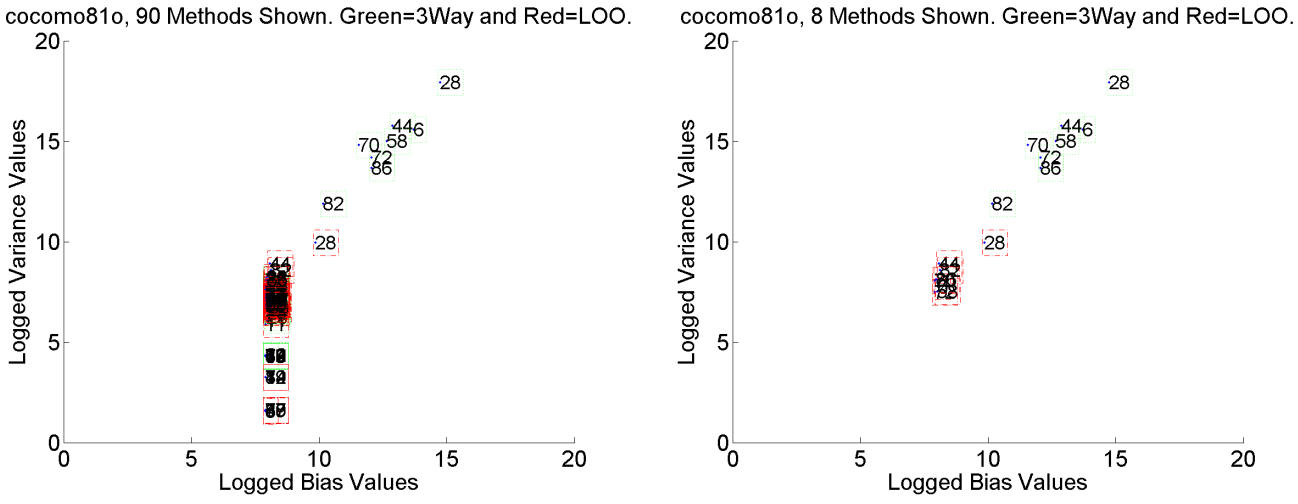
Every plot in this section for each dataset is shown with no statistical test and a statistical test applied. On the plots only those methods are shown, whose bias/variance values are statistically different to one another.

When we apply a statistical test (Wilcoxon) to compare the results of 90 methods on each dataset, we see that only a few (sometimes none) of the methods attain statistical difference. Most of the time, whether subject to LOO or 3Way, methods produce statistically same estimations.



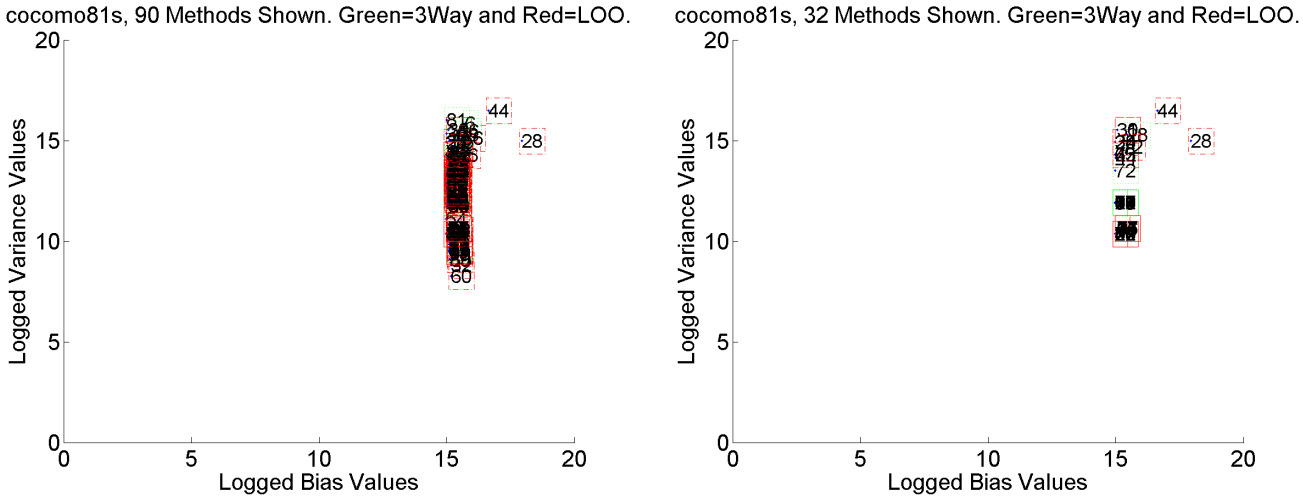
(a) Without Stat. Test (b) With Stat. Test

Fig. 1. cocomo81 Bias-Var for LOO and 3-Way



(a) Without Stat. Test (b) With Stat. Test

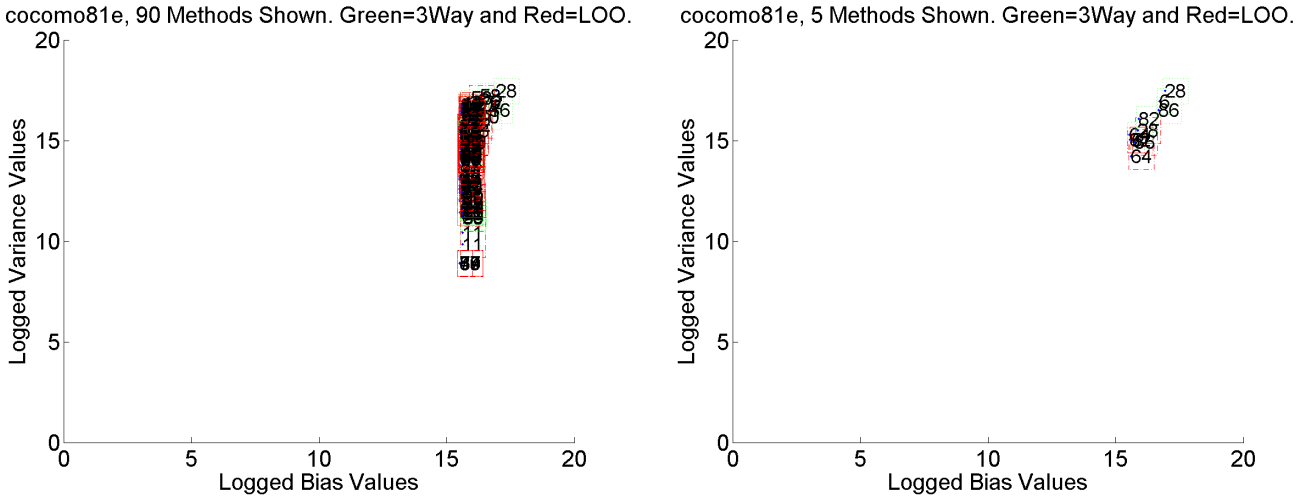
Fig. 2. cocomo81o Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

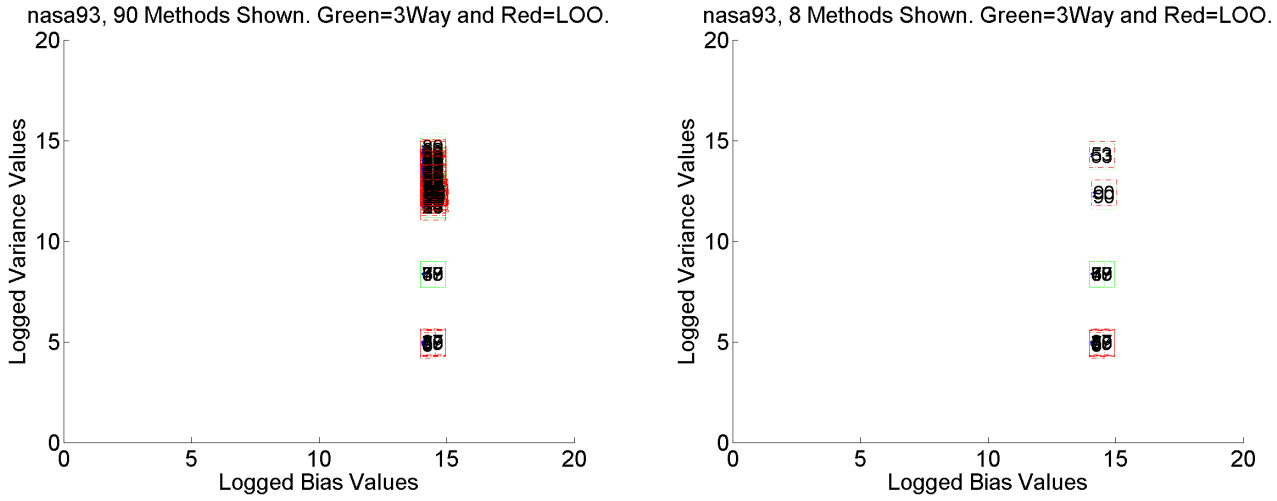
Fig. 3. cocomo81s Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

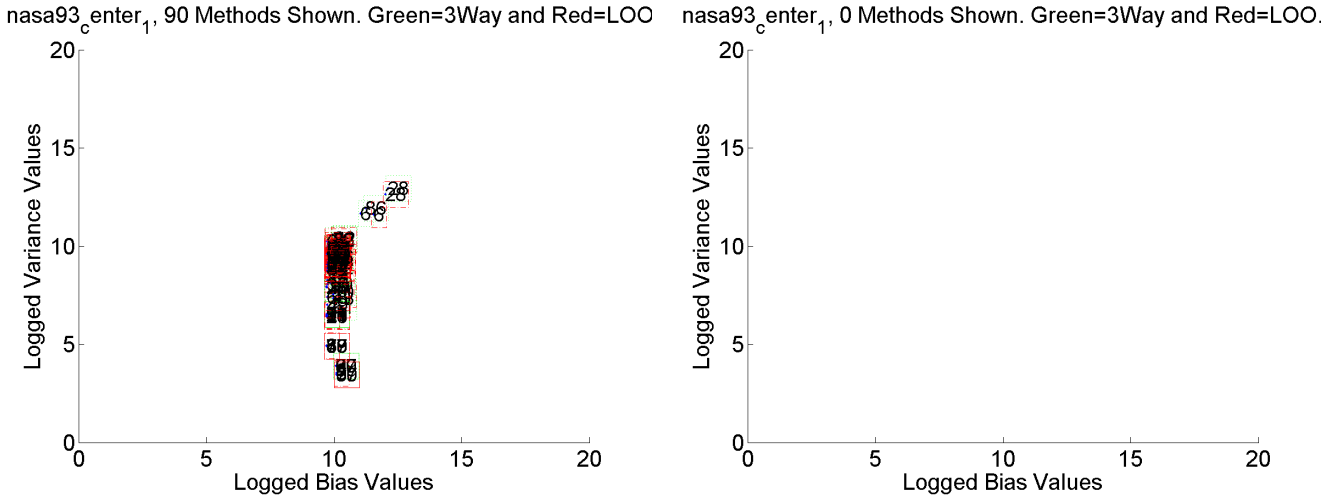
Fig. 4. cocomo81e Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

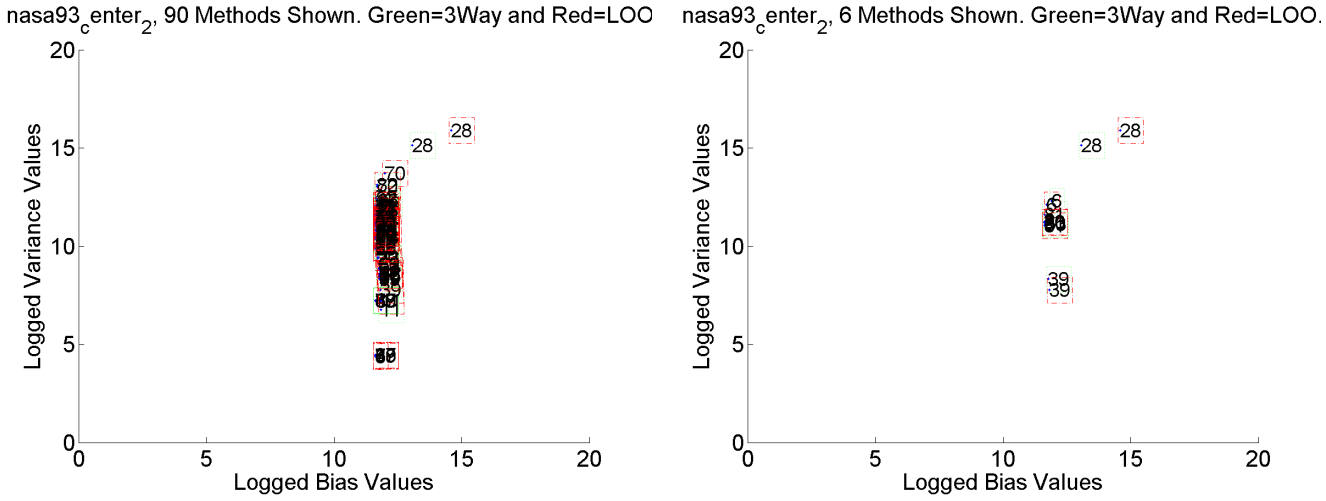
Fig. 5. nasa93 Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

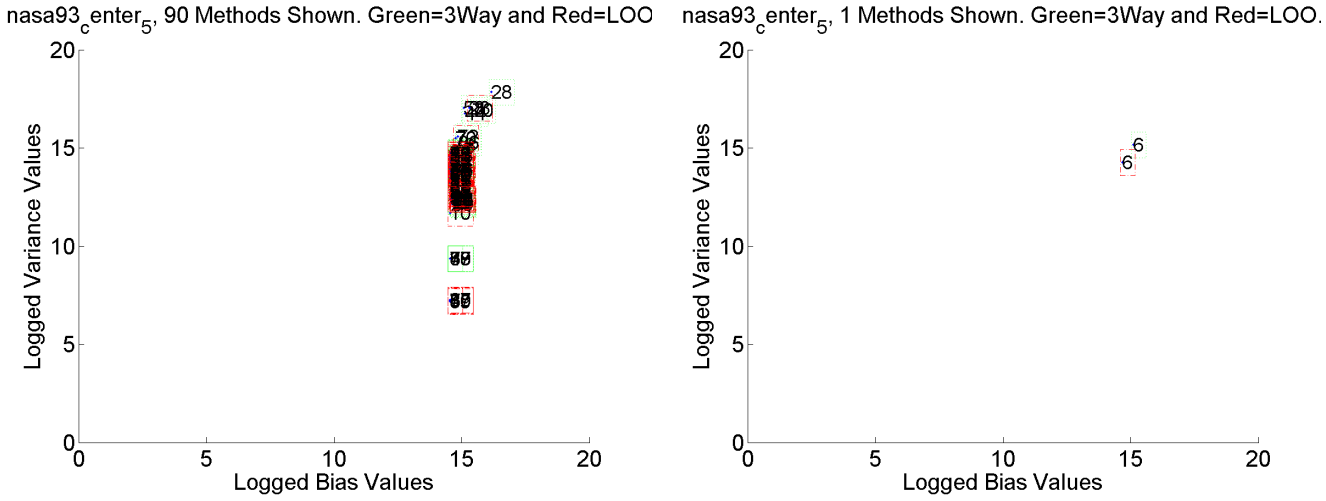
Fig. 6. nasa93_center_1 Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

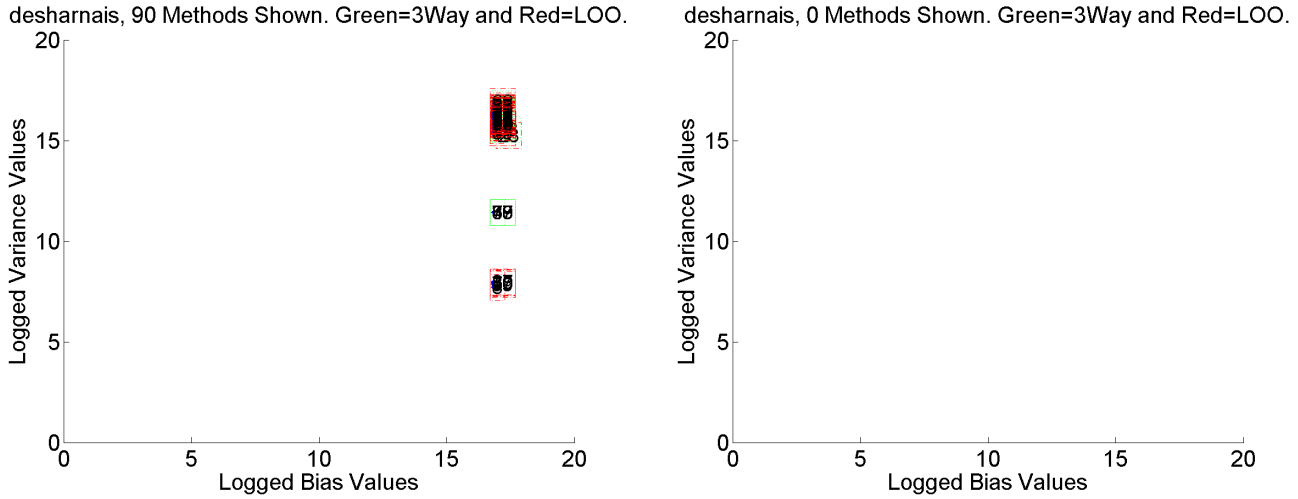
Fig. 7. nasa93_center_2 Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

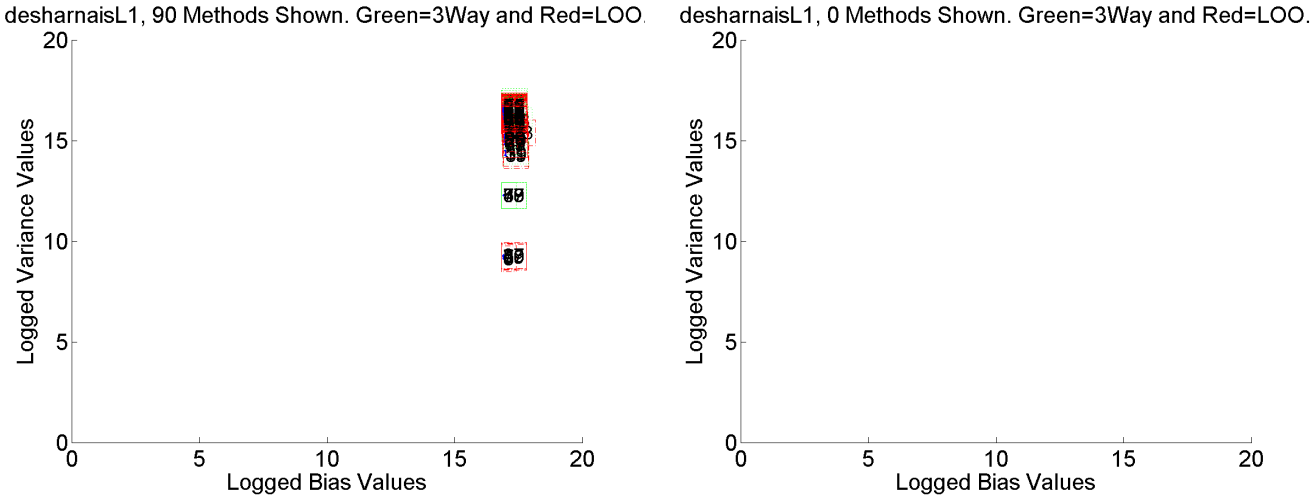
Fig. 8. nasa93_center_5 Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

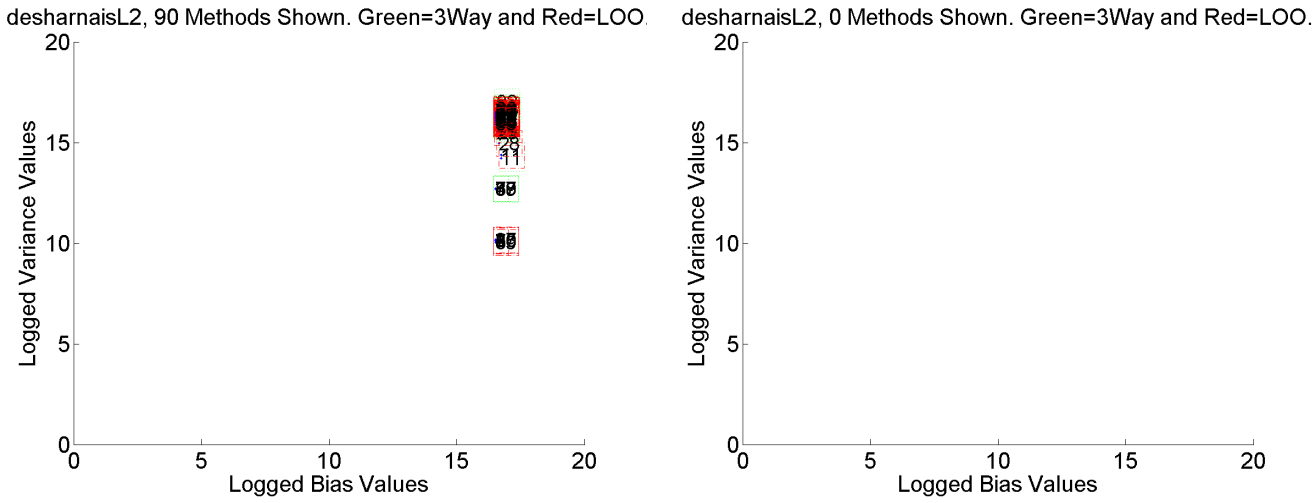
Fig. 9. desharnais Bias-Var for LOO and 3-Way



(a) Without Stat. Test

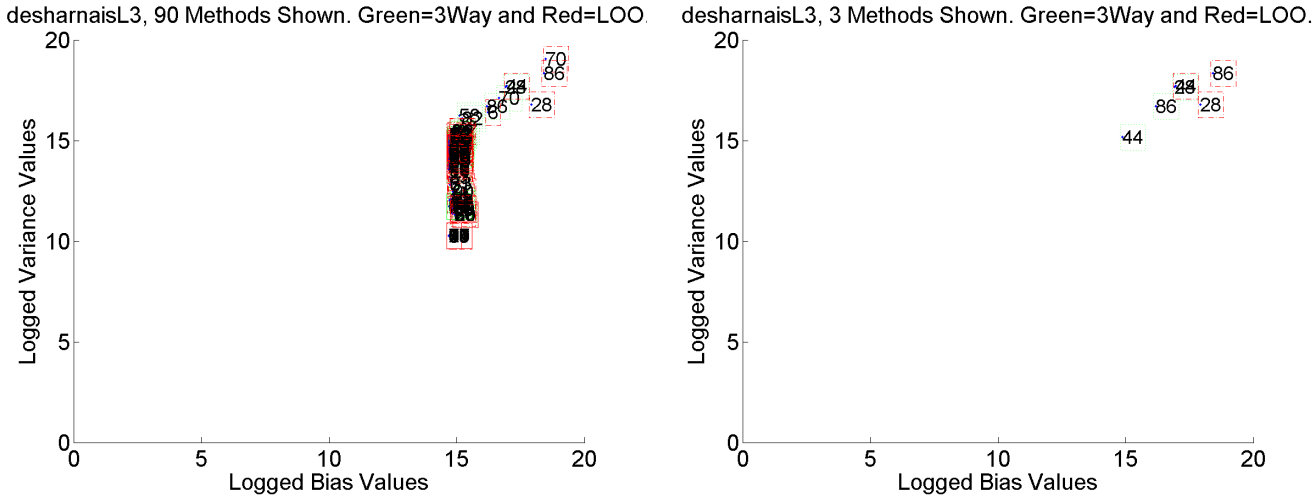
(b) With Stat. Test

Fig. 10. desharnaisL1 Bias-Var for LOO and 3-Way



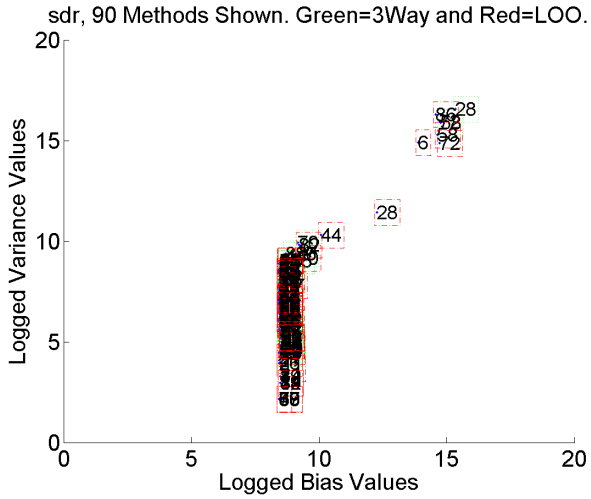
(a) Without Stat. Test (b) With Stat. Test

Fig. 11. desharnaisL2 Bias-Var for LOO and 3-Way

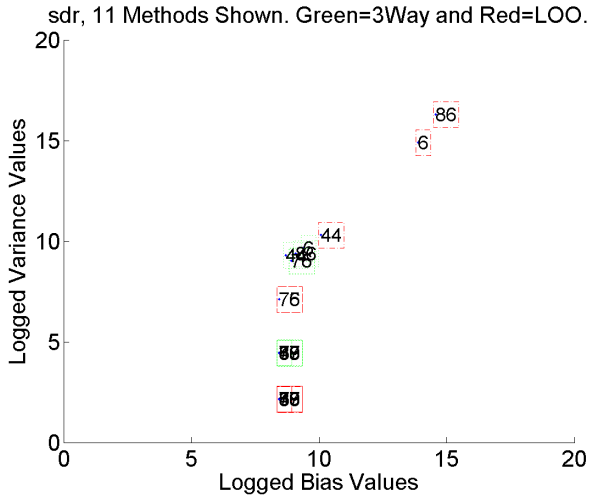


(a) Without Stat. Test (b) With Stat. Test

Fig. 12. desharnaisL3 Bias-Var for LOO and 3-Way

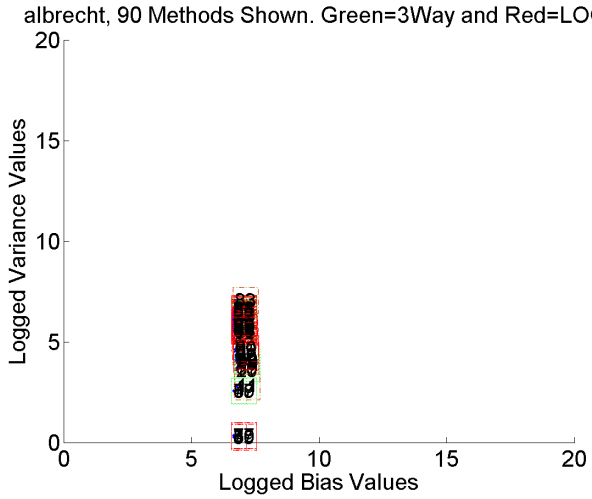


(a) Without Stat. Test

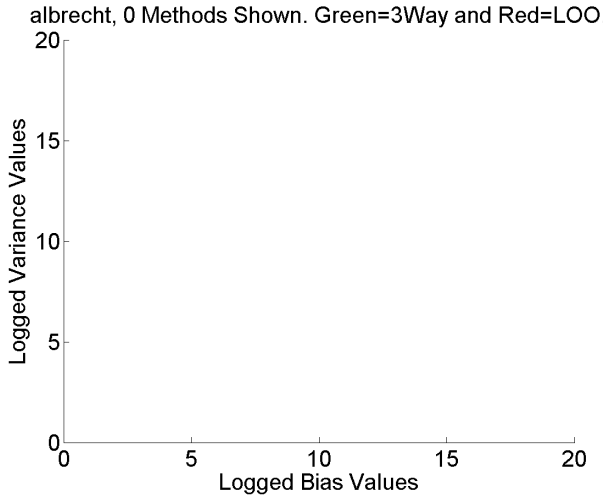


(b) With Stat. Test

Fig. 13. sdr Bias-Var for LOO and 3-Way

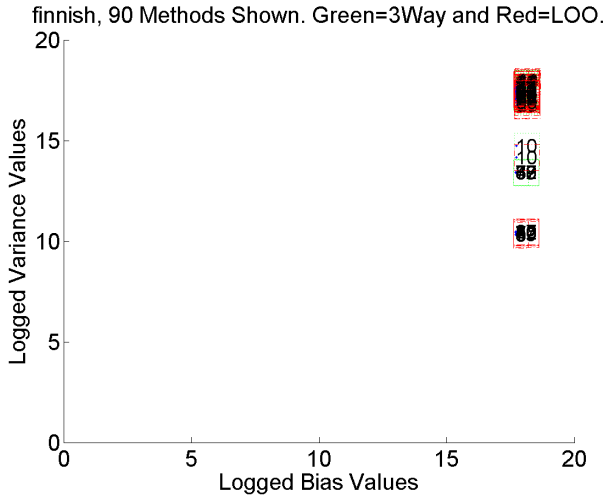


(a) Without Stat. Test

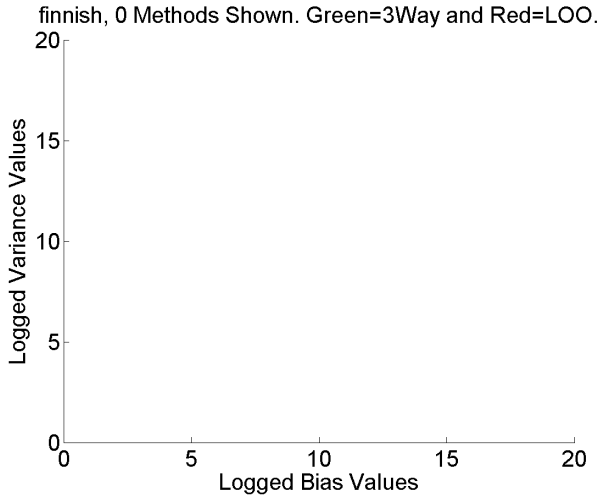


(b) With Stat. Test

Fig. 14. albrecht Bias-Var for LOO and 3-Way

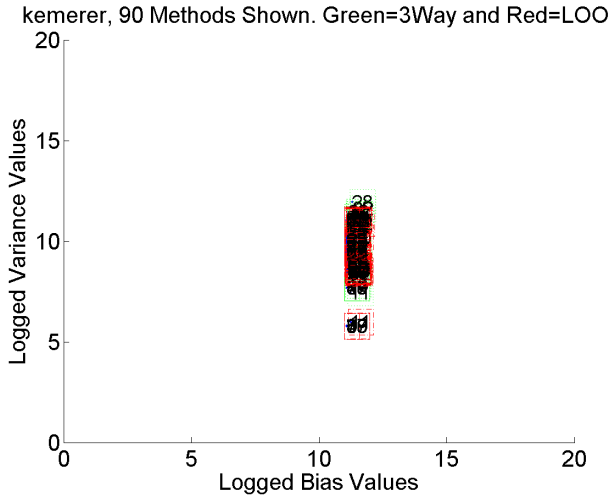


(a) Without Stat. Test

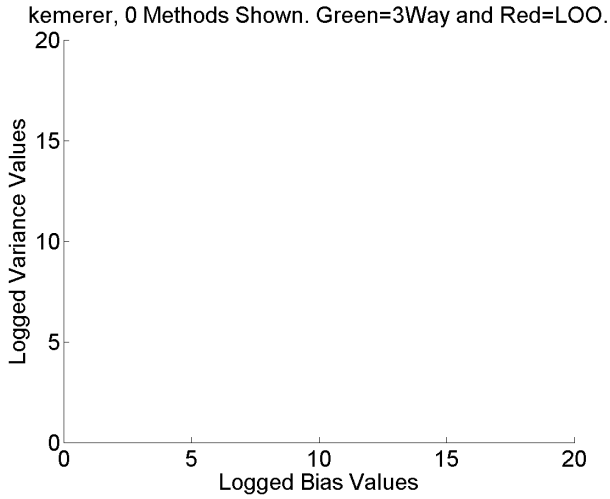


(b) With Stat. Test

Fig. 15. finnish Bias-Var for LOO and 3-Way

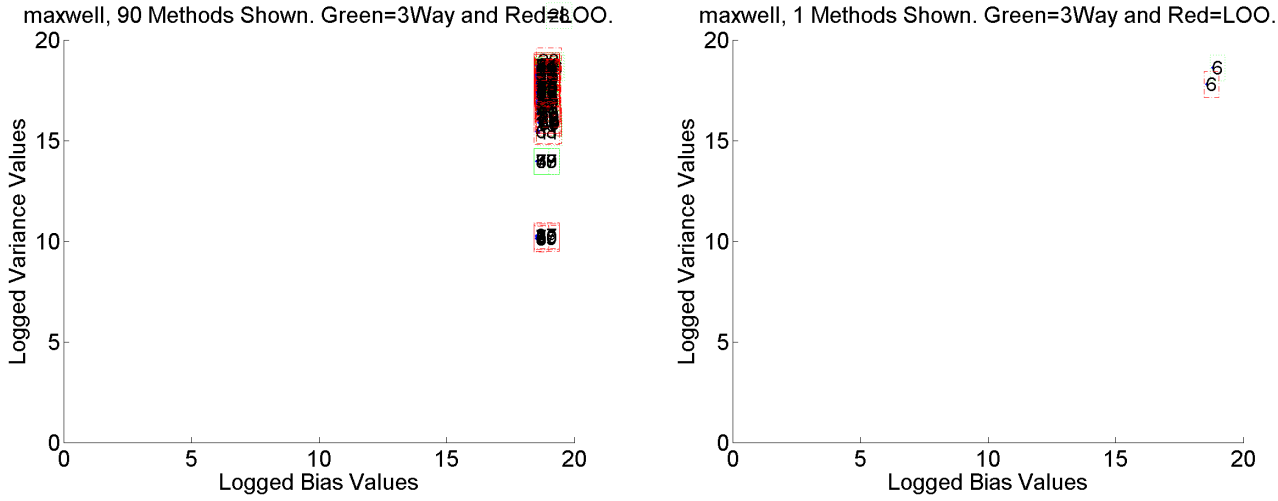


(a) Without Stat. Test



(b) With Stat. Test

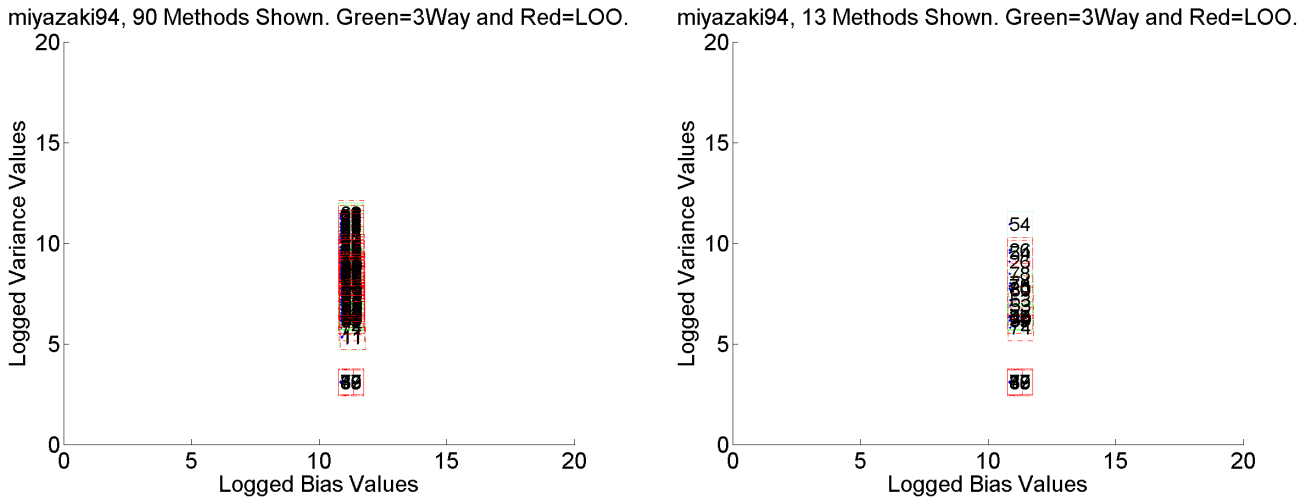
Fig. 16. kemerer Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

Fig. 17. maxwell Bias-Var for LOO and 3-Way



(a) Without Stat. Test

(b) With Stat. Test

Fig. 18. miyazaki94 Bias-Var for LOO and 3-Way

2 JIGGLE PROBLEM: HOW MUCH DOES A METHOD DEVIATE?

Our aim is to find the deviations of methods between two testing strategies (LOO and CV). When doing that our fix ordering is the ICSE submission, fig.4, where ordering was based on loss-values-ordering across all datasets and all error measures subject to LOO.

There are two dimensions to that problem:

1) How much does a method deviate across datasets for different error measures?
2) How much variation per method do we have across error measures when we fix the dataset?

For both problems, there are intermediary analysis and summary analysis:

- **Intermediary:** What is rank delta (absolute rank change) for each method?
- **Intermediary:** What is the rank-order for each method when we keep icse order fixed in one axis and plot new order in the other axis?
- **Summary:** What is the occurrence count of a method in top/bottom-x across?

2.1 Deviation Across Datasets for Each Error Measure

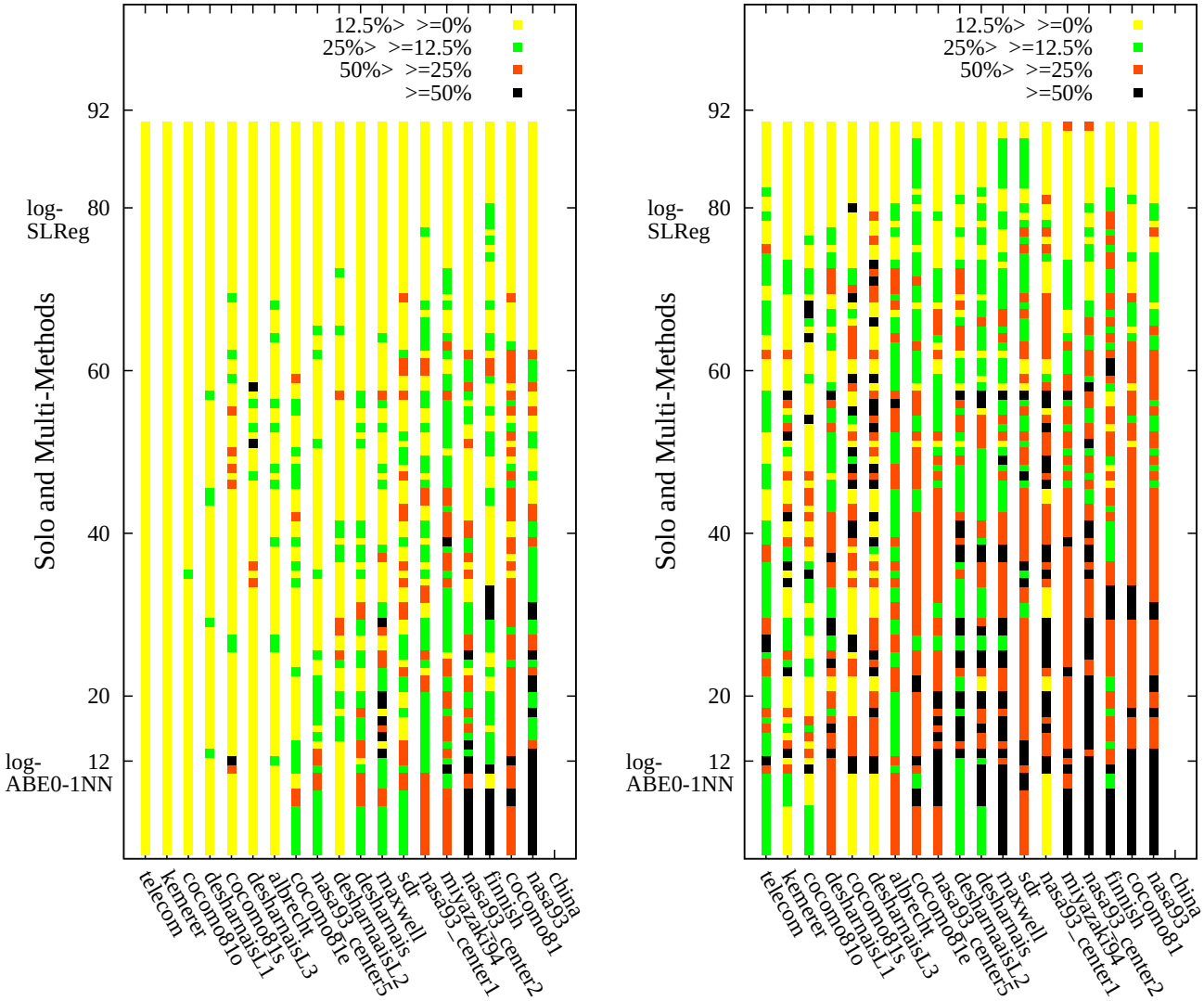
To answer the first item, I searched for top-16 and bottom-16 methods in all datasets for each error measure separately. Whenever a method appeared in top/bottom-16 according to an error measure I incremented its occurrence value by 1. Since we have 19 datasets, a method may have the highest number of 19. The complete list of occurrence numbers (**summary**) is here and the rank-delta-plots (**intermediary**) are here. The formula for rank-delta plots is: *abs(my Index acc. to win/loss/win-loss - myIcseOrder)*. Note in summary plots that 3-Way jiggles mostly behave like an envolope for LOO jiggles, i.e. 3-Way mostly has higher variation w.r.t. icse sorting.

The rank-order is another way of looking at stability of ranking. On rank-order plots, the x-axis is the methods in order of Icse, whereas y-axis is the order of corresponding methods according to win/loss or win-loss count under current error measure. Rank-orders are here.

2.2 Deviation Across Error Measures for Each Dataset

Another dimension to the jiggle problem is the deviations across error measures for every dataset. The absolute rank-order for that dimension can be found here. It is difficult to observe a general trend between LOO and 3-Way on dataset level when absolute rank deltas are concerned.

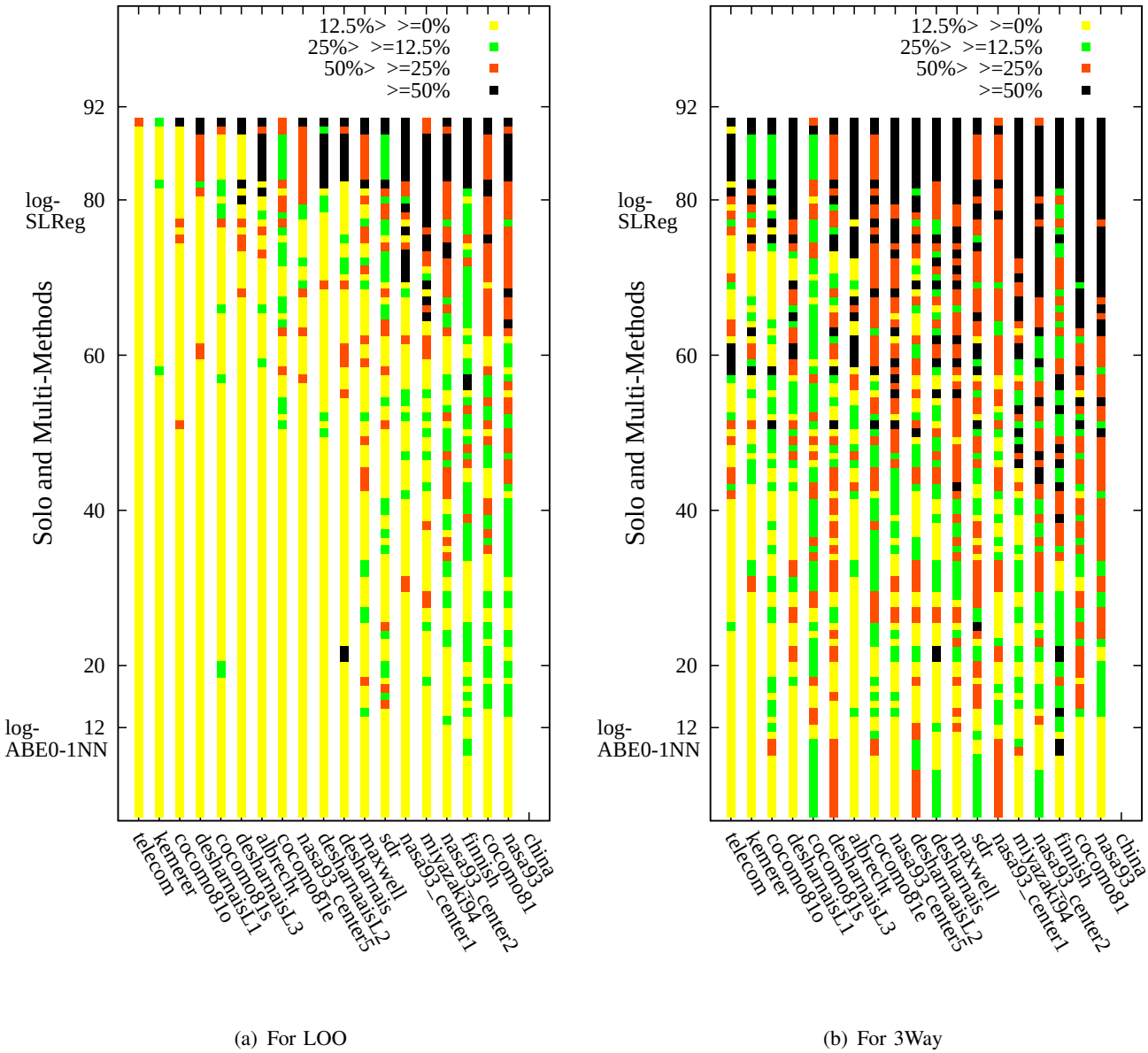
3 ICSE PLOTS



(a) For LOO

(b) For 3Way

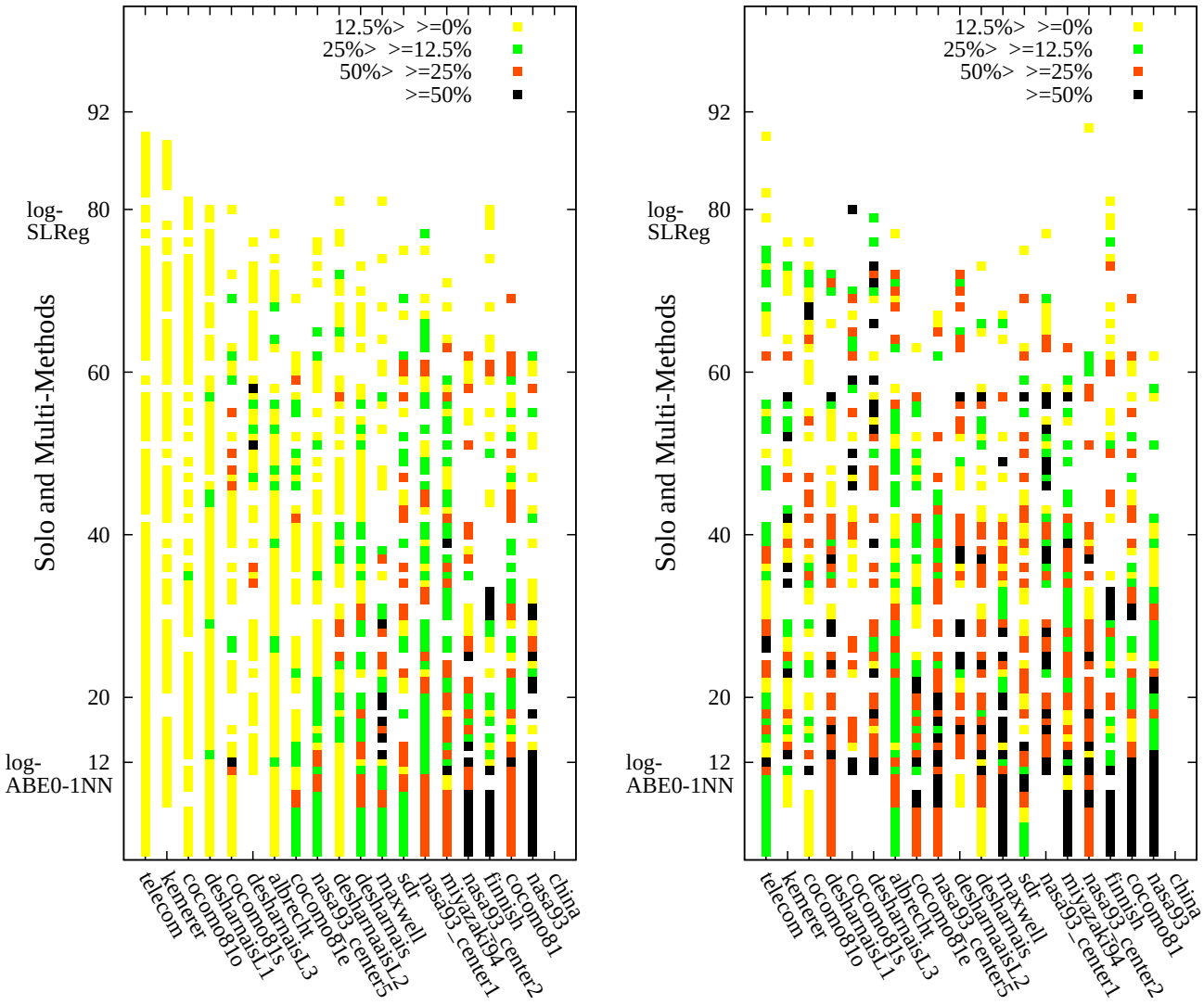
Fig. 19. Win Percentages



(a) For LOO

(b) For 3Way

Fig. 20. Loss Percentages



(a) For LOO

(b) For 3Way

Fig. 21. Win-Loss Percentages