Linear combiner training on random subsets of different sizes

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(scales)
```

```
## Warning: package 'scales' was built under R version 4.0.3
```

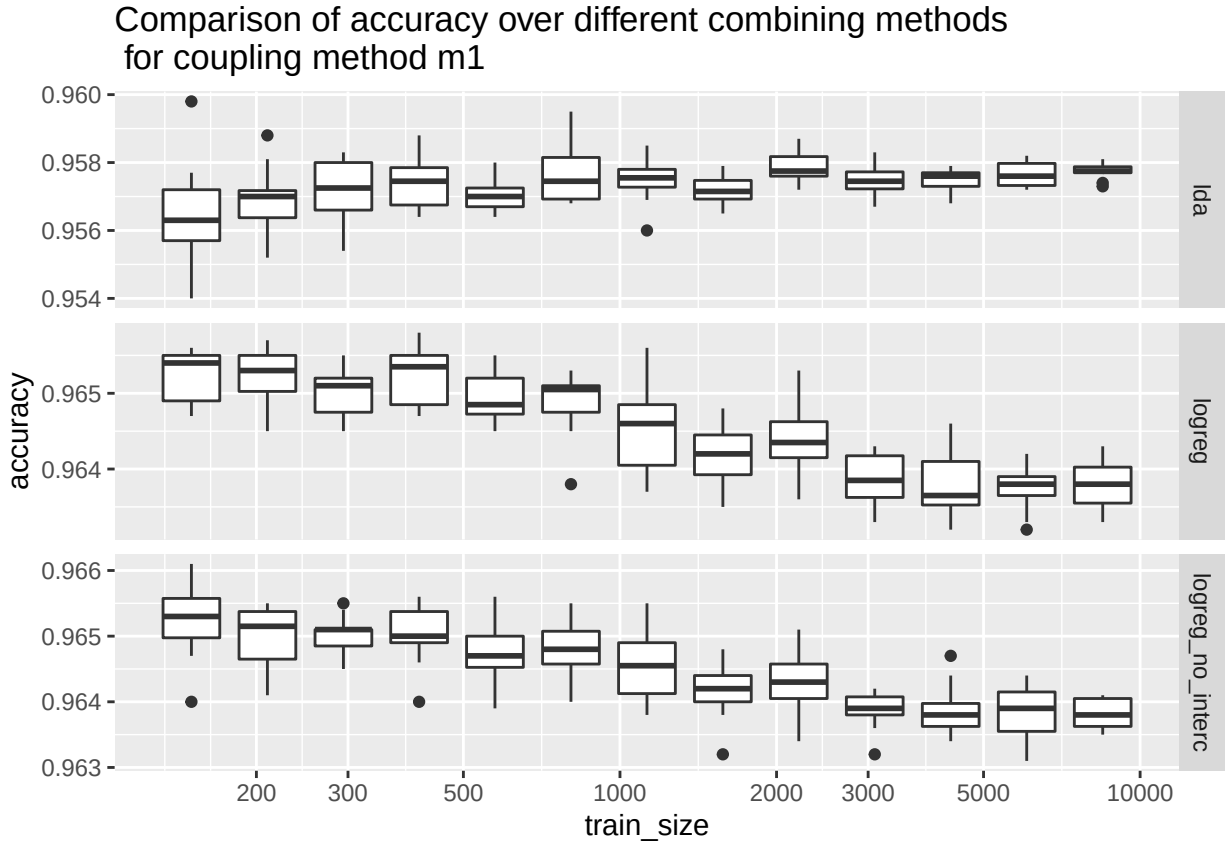
Experiment code is in the file training_subsets_sizes_experiment.py. Experiment on both CIFAR10 and CIFAR100 datasets. This experiment trains WeightedLinearEnsemble on various subsets of different sizes of data on which neural networks were trained. Four different coupling methods are used: method one and two (m1 and m2) from (Wu, Lin, and Weng 2004), Bayes covariant method (bc) from (Šuch and Barreda 2016) and a coupling method by Such, Benus and Tinajova (sbt) from (Šuch, Benuš, and Tinajová 2015). Three combining methods are used: lda, logreg and logreg_no_interc which employ linear discriminant analysis or logistic regression individually on each pair of classes. Goal of this experiment is to determine, for which size of the combiner training set, the ensemble achieves the best performance. The requirement of lda, that the predictors are normally distributed is not met by more than half of the class pairs. This may be reflected in the worsened prediction capacity of ensembles using lda combining method and we will consider these ensembles as less influential on the decision of the training set size.

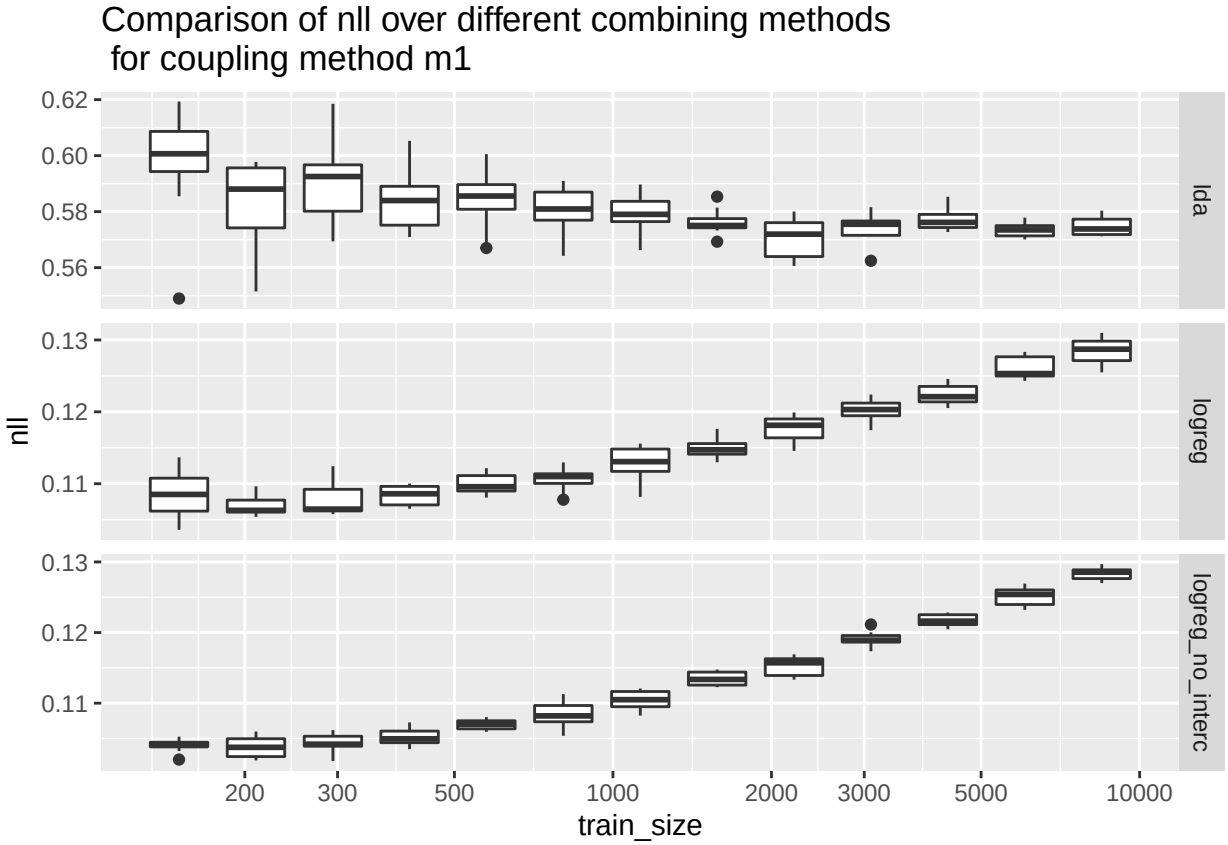
```
metrics <- c("accuracy", "nll", "ece")
```

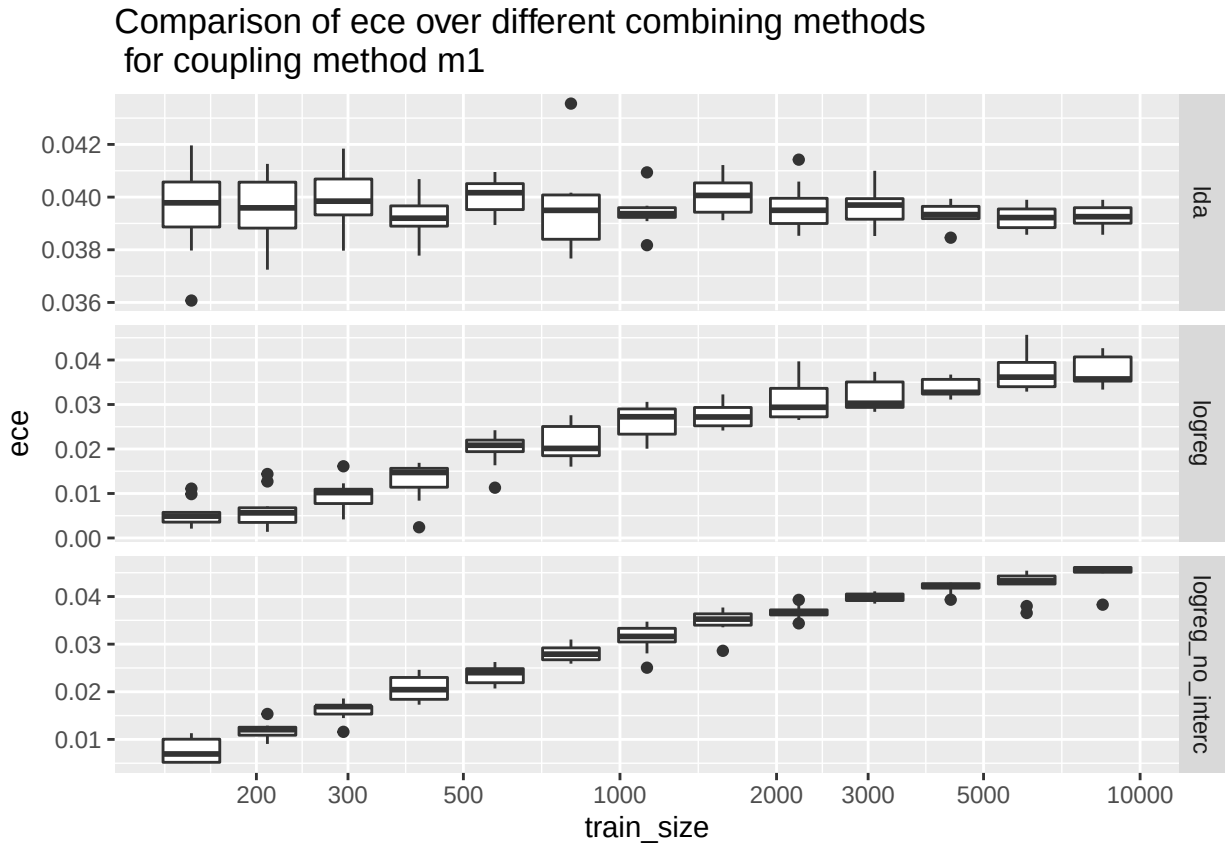
CIFAR-10

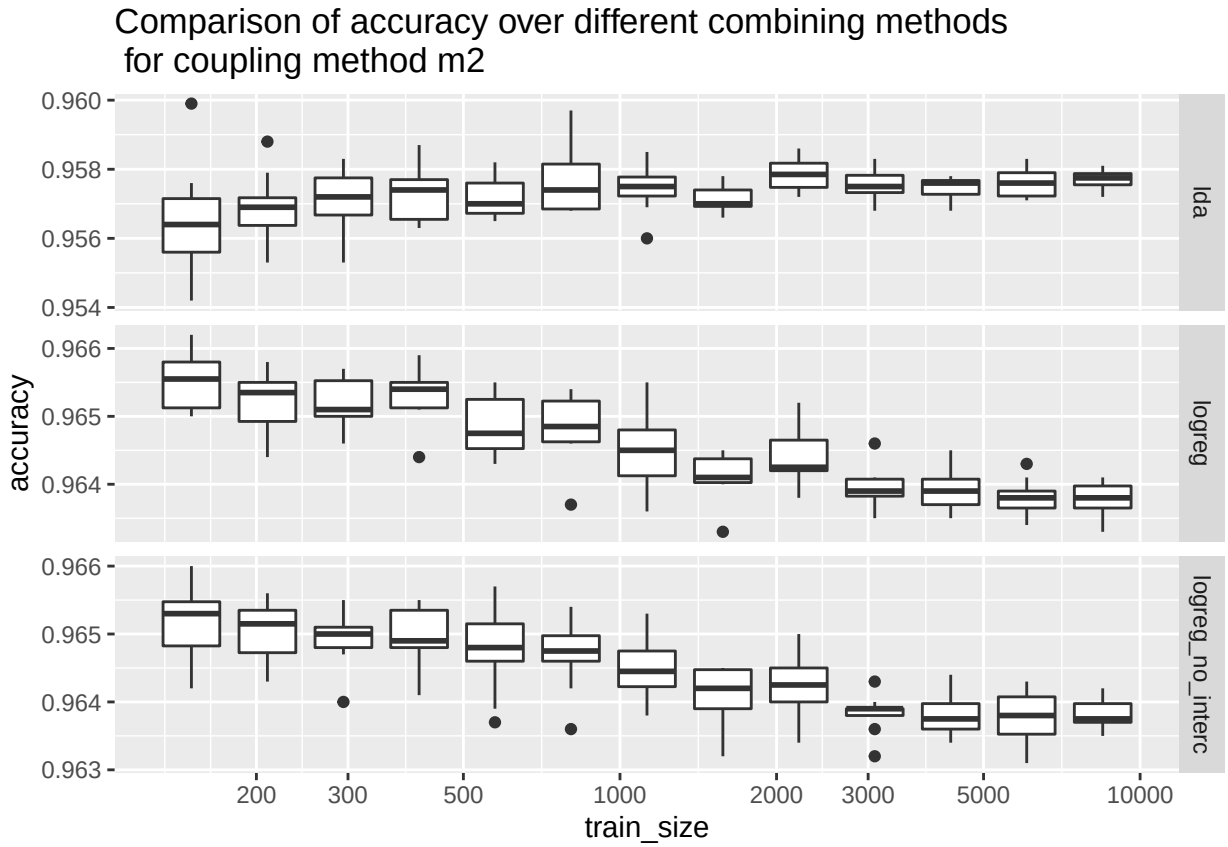
```
ens_metrics_c10 <- read.csv("../data/data_tv_5000_c10/0/exp_subsets_sizes_train_outputs/ens_metrics.csv
nets_metrics_c10 <- read.csv("../data/data_tv_5000_c10/0/exp_subsets_sizes_train_outputs/net_metrics.cs
```

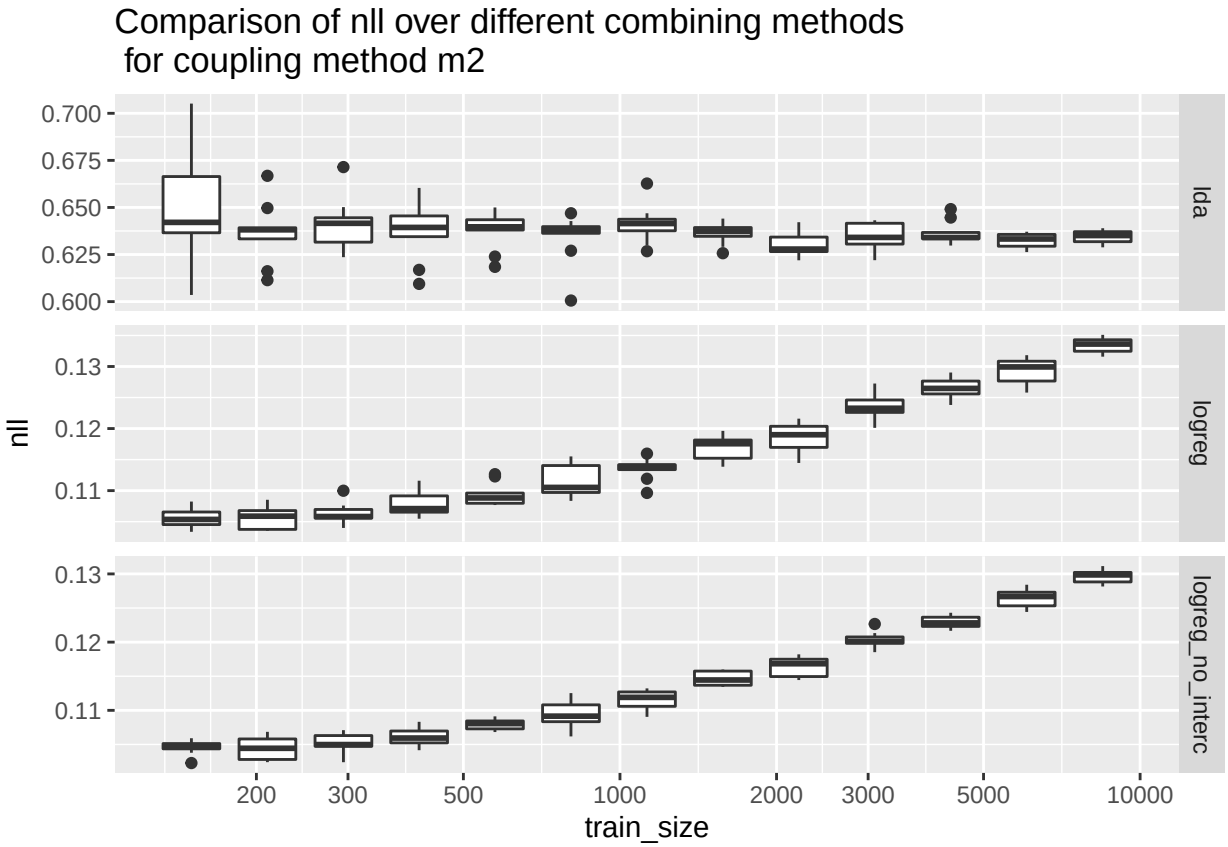
```
for (cp_m in unique(ens_metrics_c10$coupling_method))
{
  for (metric in metrics)
  {
    box_plt <- ens_metrics_c10 %>% filter(coupling_method==cp_m) %>%
      ggplot() +
      geom_boxplot(mapping=aes_string(x="train_size", y=metric, group="train_size")) +
      facet_grid(rows=vars(combining_method), scales="free") +
      scale_x_log10(breaks=log_breaks(n=10)) +
      ggtitle(paste0("Comparison of ", metric, " over different combining methods\n for coupling method

    print(box_plt)
  }
}
```

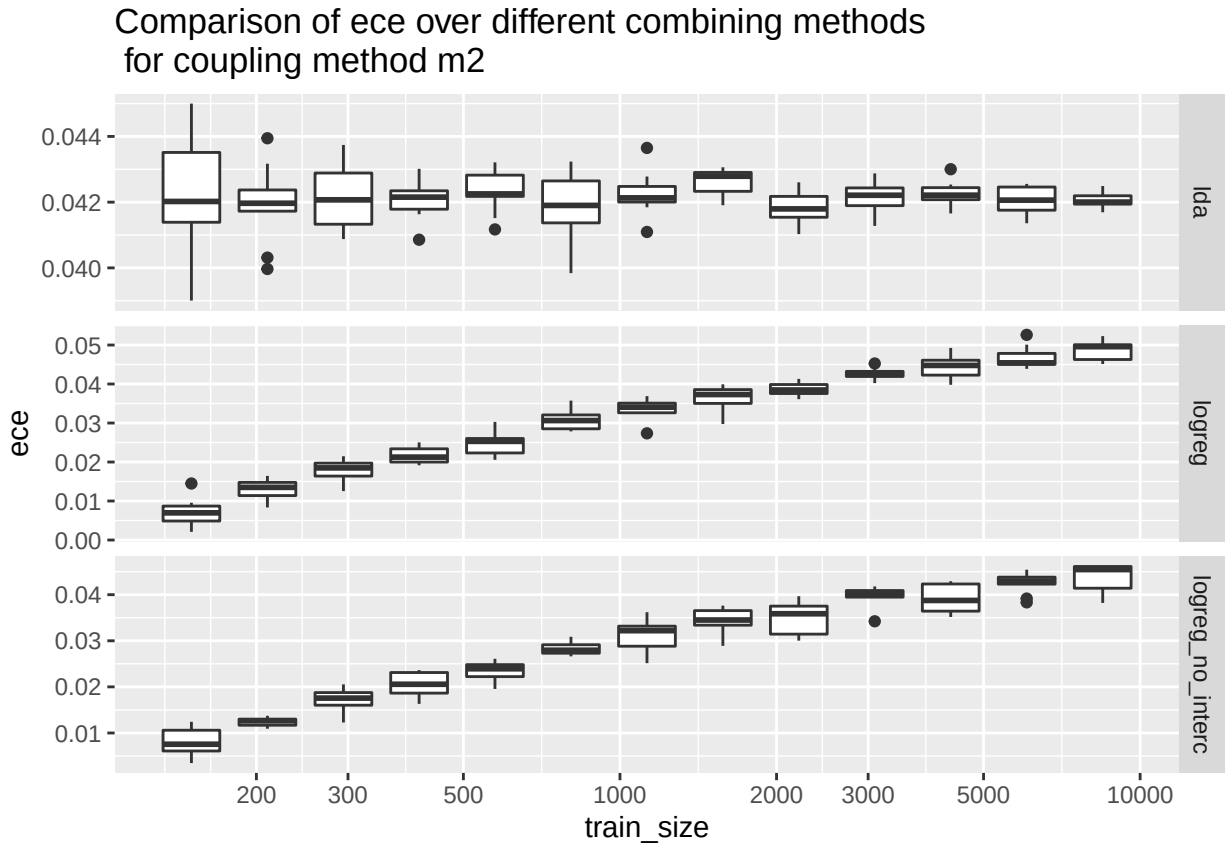


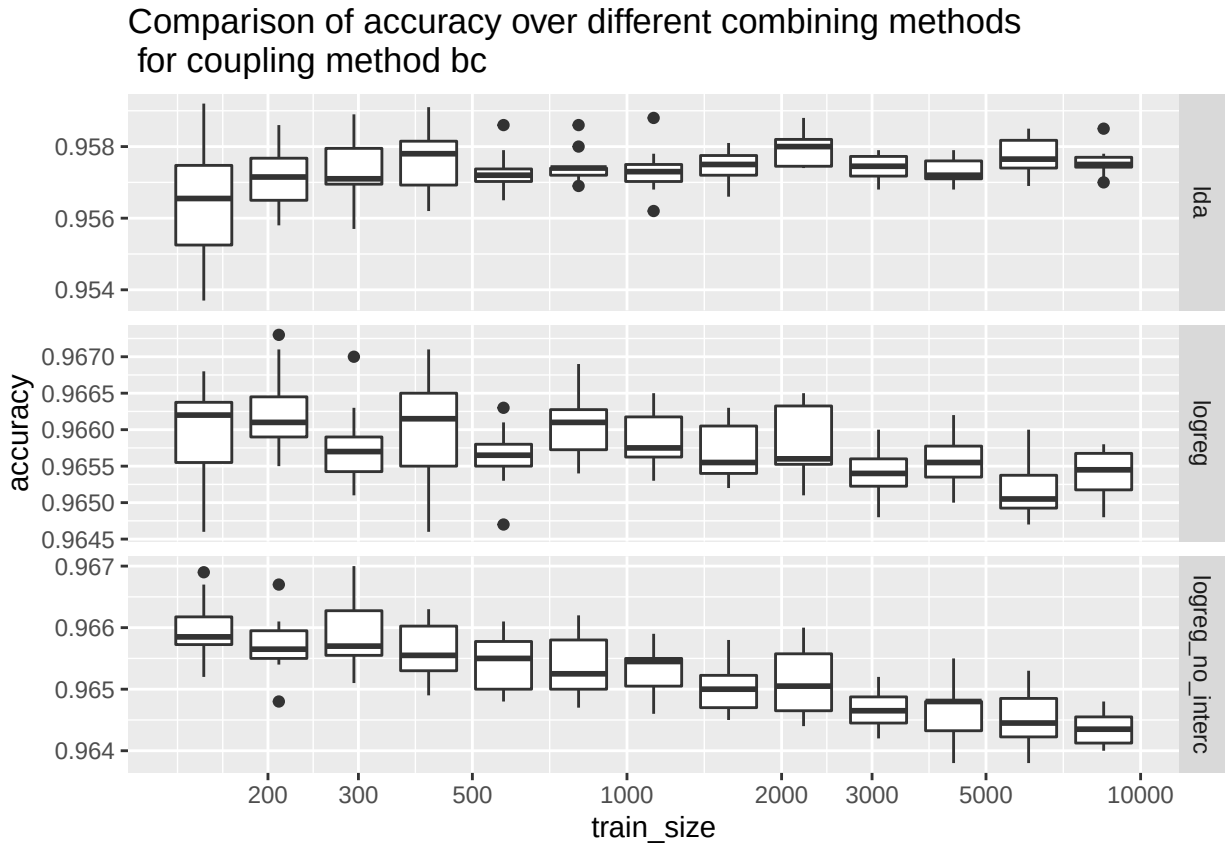

Comparison of nll over different combining methods
for coupling method m1
nll
0.62
0.60
0.58
0.56
0.13
0.12
0.11
0.13
0.12
0.11
lda
logreg
logreg_no_interc
200
300
500
1000
2000
3000
5000
10000
train_size

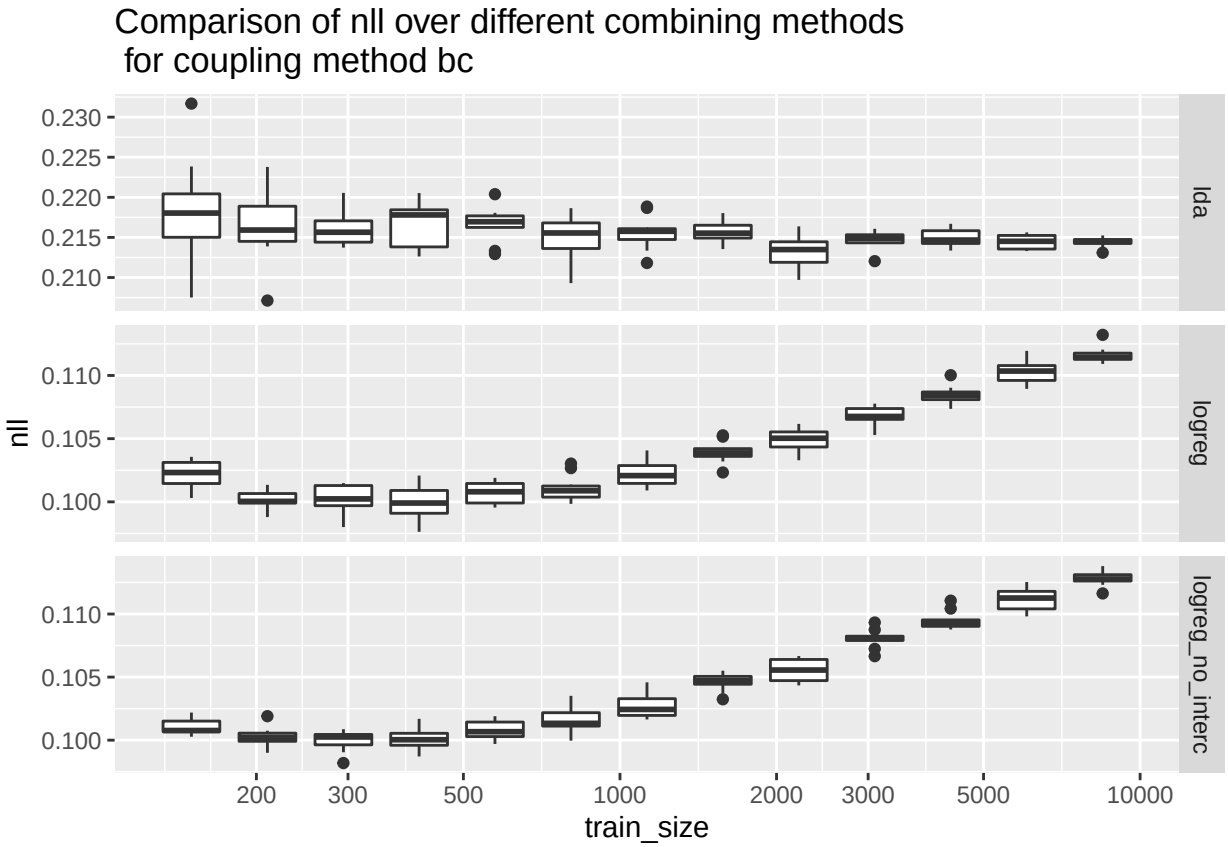

Comparison of ece over different combining methods
for coupling method m1
ece
train_size
lda
logreg
logreg_no_interc
0.042
0.040
0.038
0.036
0.04
0.03
0.02
0.01
0.00
200
300
500
1000
2000
3000
5000
10000

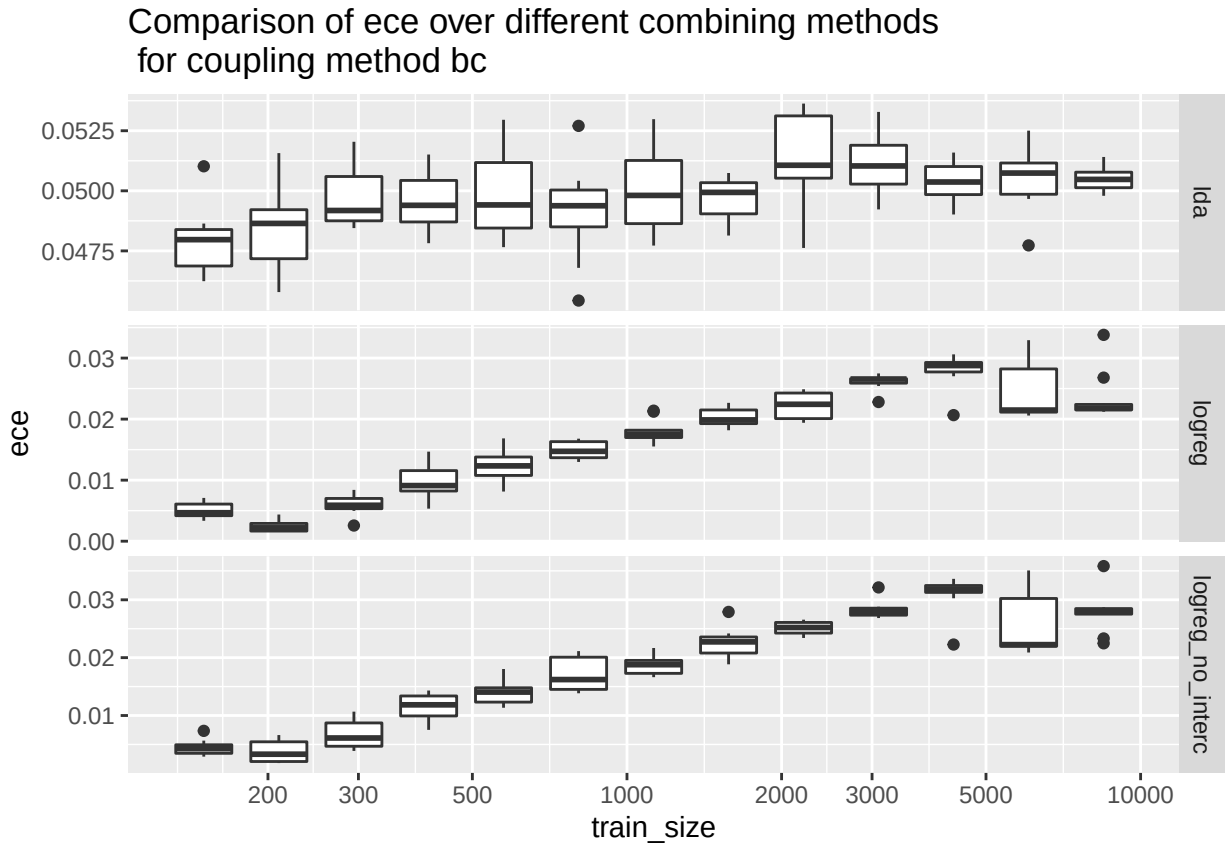

Comparison of accuracy over different combining methods
for coupling method m2
accuracy
0.960
0.958
0.956
0.954
0.966
0.965
0.964
0.966
0.965
0.964
0.963
lda
logreg
logreg_no_interc
200
300
500
1000
2000
3000
5000
10000
train_size

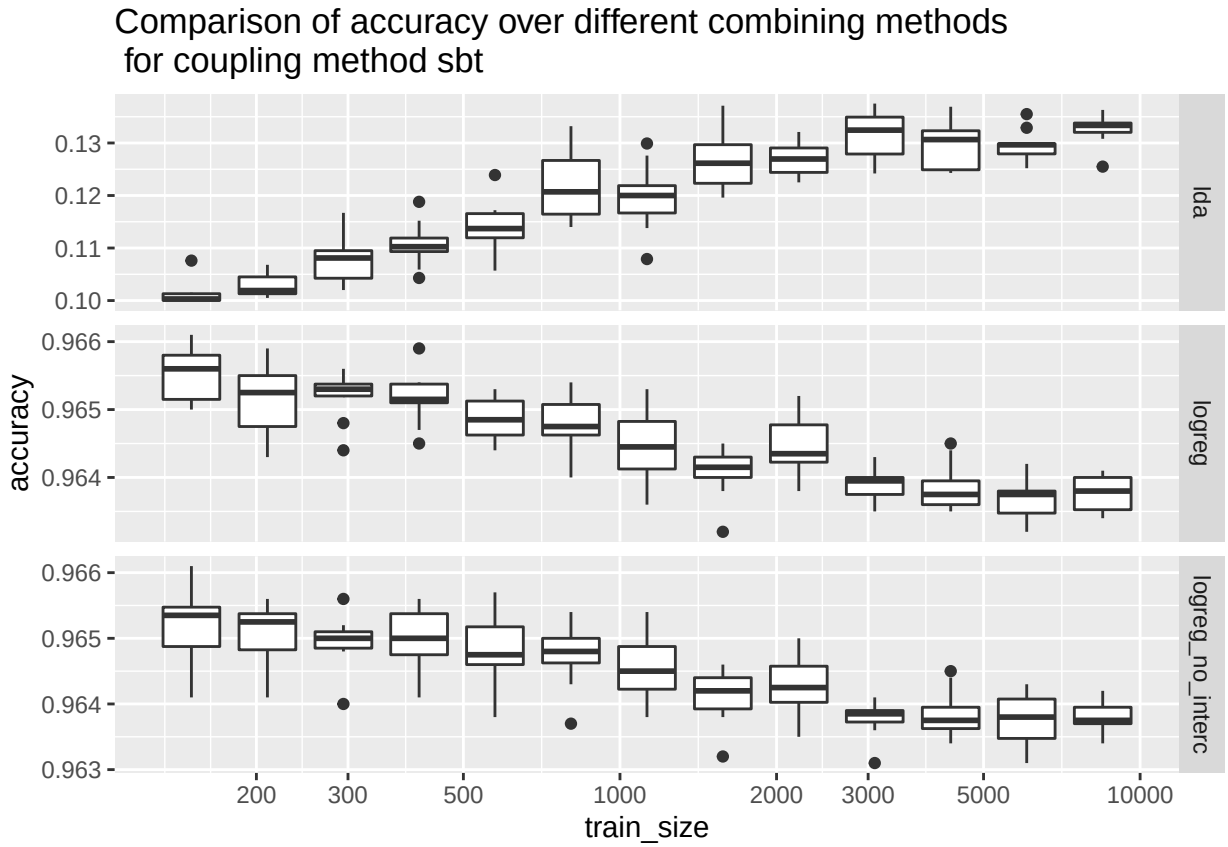

Comparison of nll over different combining methods
for coupling method m2
nll
0.700
0.675
0.650
0.625
0.600
0.13
0.12
0.11
0.13
0.12
0.11
lda
logreg
logreg_no_interc
200
300
500
1000
2000
3000
5000
10000
train_size


Comparison of ece over different combining methods
for coupling method m2
ece
train_size
lda
logreg
logreg_no_interc
0.044
0.042
0.040
0.05
0.04
0.03
0.02
0.01
0.00
200
300
500
1000
2000
3000
5000
10000

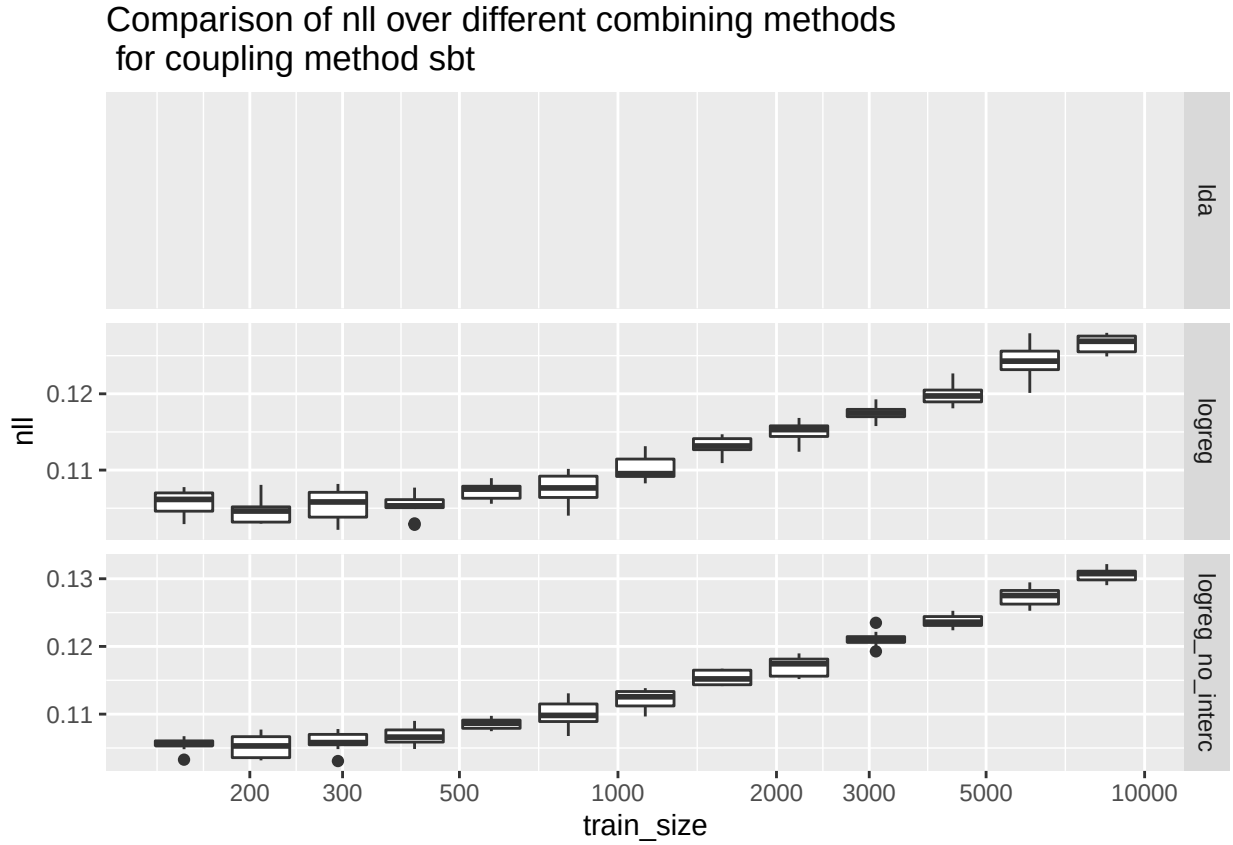

Comparison of accuracy over different combining methods
for coupling method bc
accuracy
0.958
0.956
0.954
0.9670
0.9665
0.9660
0.9655
0.9650
0.9645
0.967
0.966
0.965
0.964
lda
logreg
logreg_no_interc
200
300
500
1000
2000
3000
5000
10000
train_size


Comparison of nll over different combining methods
for coupling method bc
nll
0.230
0.225
0.220
0.215
0.210
0.110
0.105
0.100
lda
logreg
logreg_no_interc
200
300
500
1000
2000
3000
5000
10000
train_size

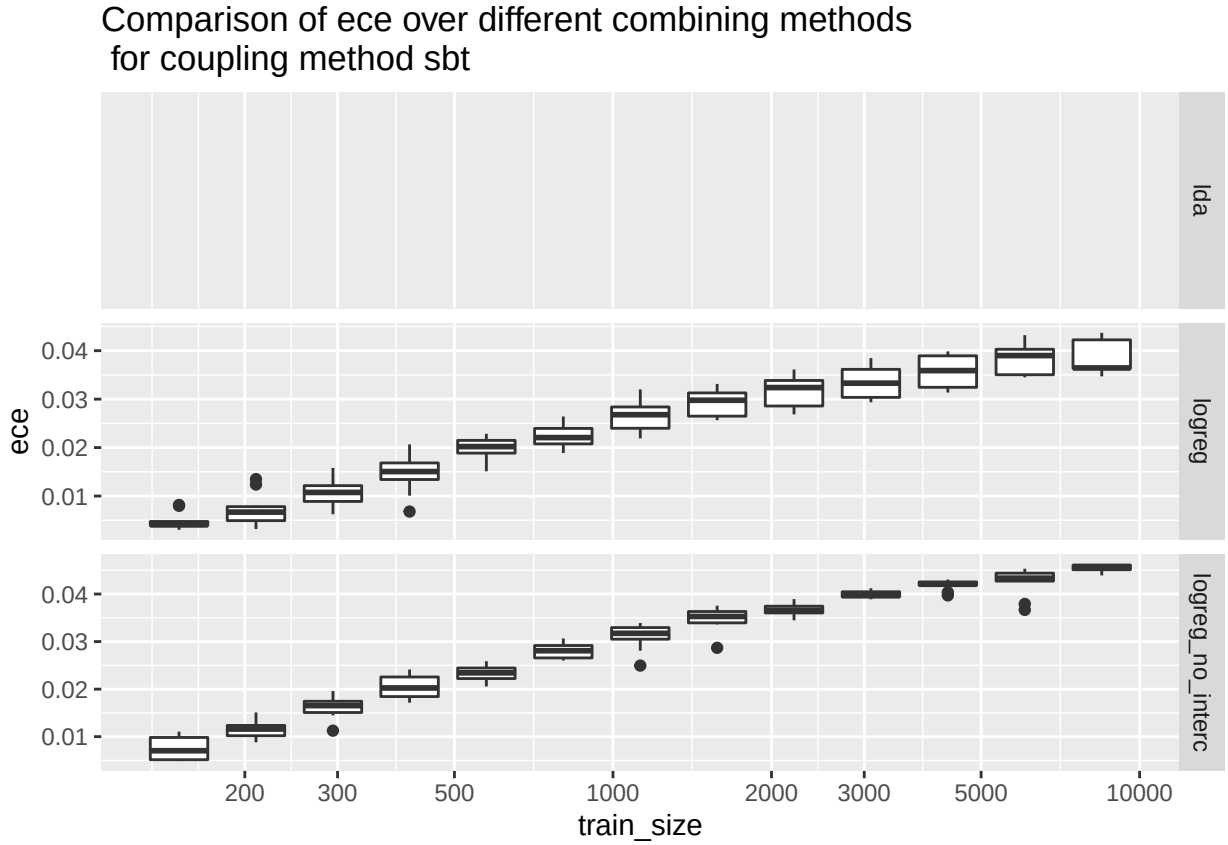

Comparison of ece over different combining methods
for coupling method bc
ece
0.0525
0.0500
0.0475
0.03
0.02
0.01
0.00
lda
logreg
logreg_no_interc
200
300
500
1000
2000
3000
5000
10000
train_size




```
## Warning: Removed 130 rows containing non-finite values (stat_boxplot).
```




```
## Warning: Removed 130 rows containing non-finite values (stat_boxplot).
```



As we can see on these plots, accuracy varies only slightly over different training set sizes. The differences among median accuracies are only about 0.05% for logistic regression combining methods and a little more for lda. The difference in accuracy of 0.05% represents only 5 samples of the test set of size 10000, therefore we don't consider these differences to provide conclusive information for the choice of training set size.

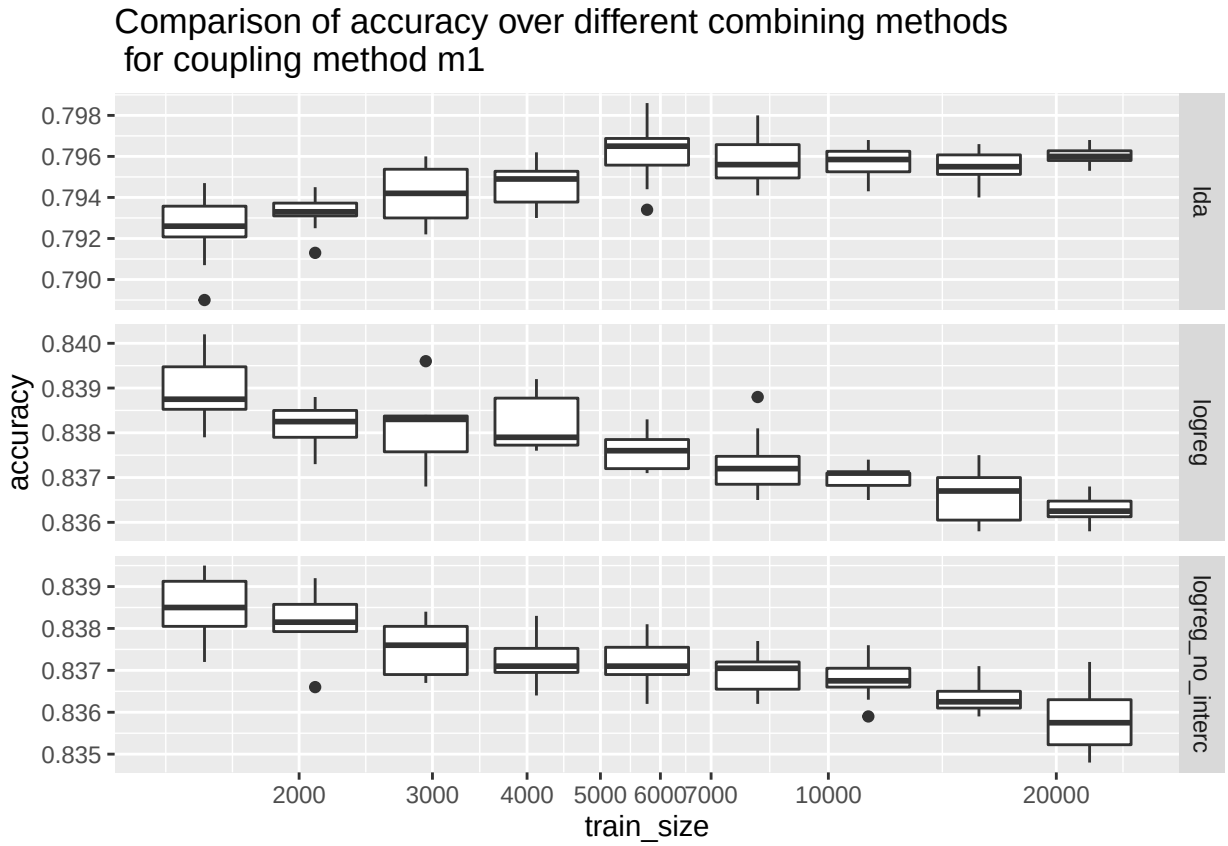
However, if we look at the metrics negative log likelihood (nll) and estimated calibration error (ece) we can see a clear increasing trend for logreg combining methods and mostly stagnating trend for combining method lda. For logreg combining methods, we can observe a stagnation or even a slight decrease in nll for several smaller training set sizes. This decrease, or stagnation changes into increase at training set size of about 500.

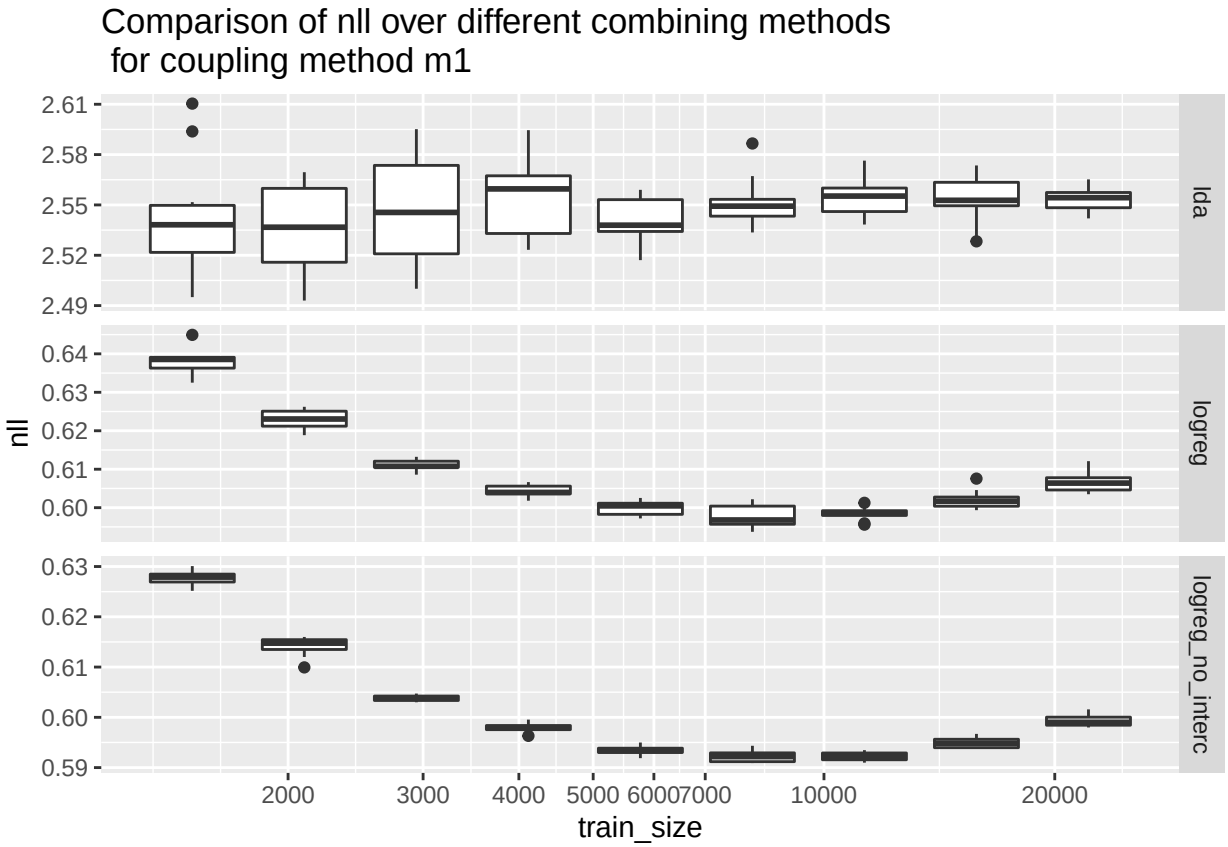
CIFAR-100

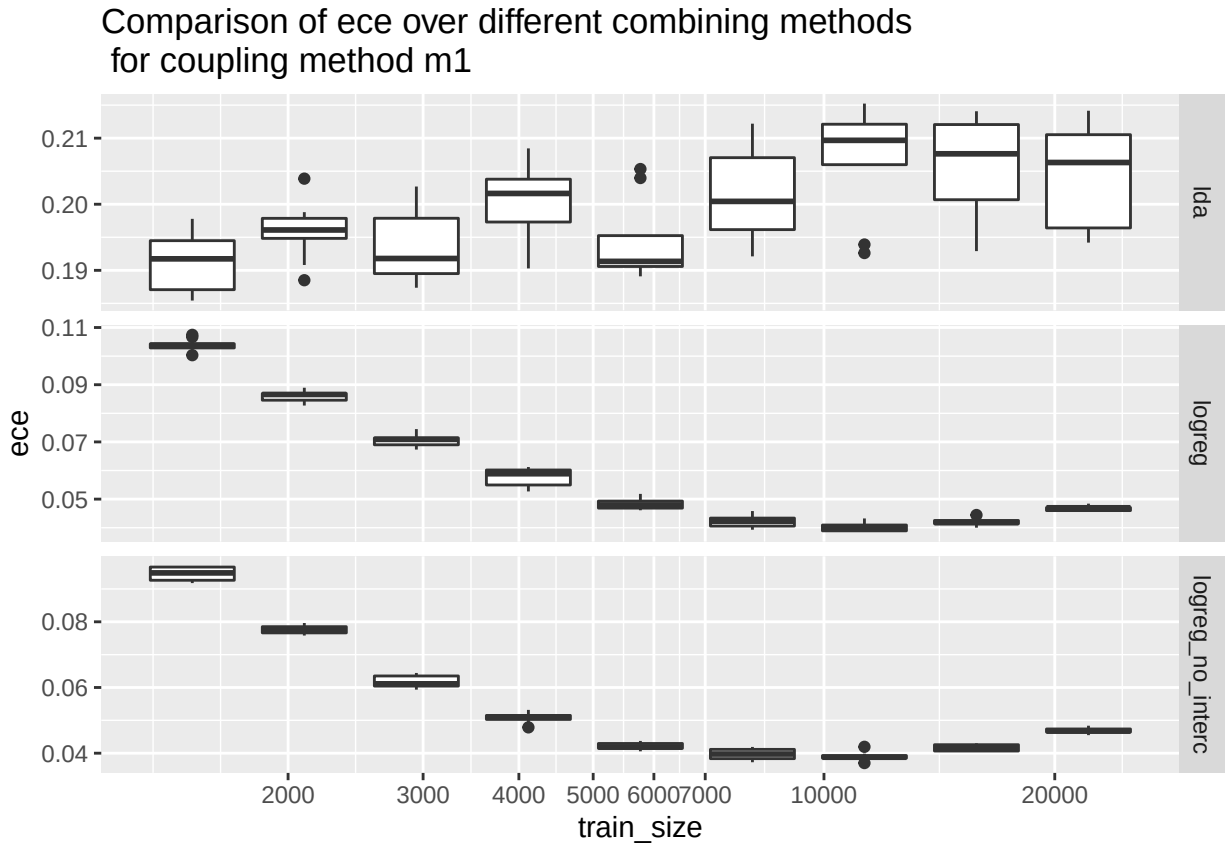
```
ens_metrics_c100 <- read.csv("../data/data_tv_5000_c100/0/exp_subsets_sizes_train_outputs/ens_metrics.cs
nets_metrics_c100 <- read.csv("../data/data_tv_5000_c100/0/exp_subsets_sizes_train_outputs/net_metrics.c
```

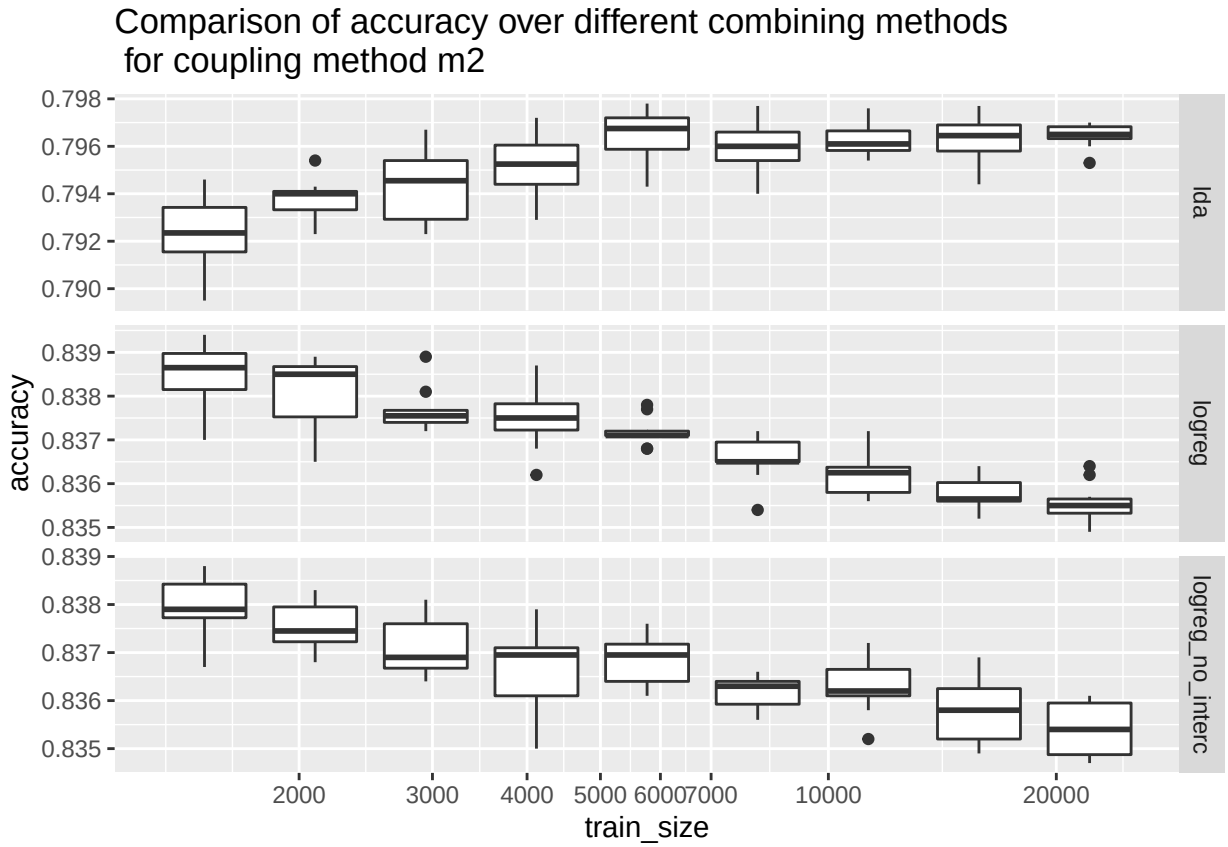
```
for (cp_m in unique(ens_metrics_c10$coupling_method))
{
  for (metric in metrics)
  {
    box_plt <- ens_metrics_c100 %>% filter(coupling_method==cp_m) %>%
      ggplot() +
      geom_boxplot(mapping=aes_string(x="train_size", y=metric, group="train_size")) +
      facet_grid(rows=vars(combining_method), scales="free") +
```

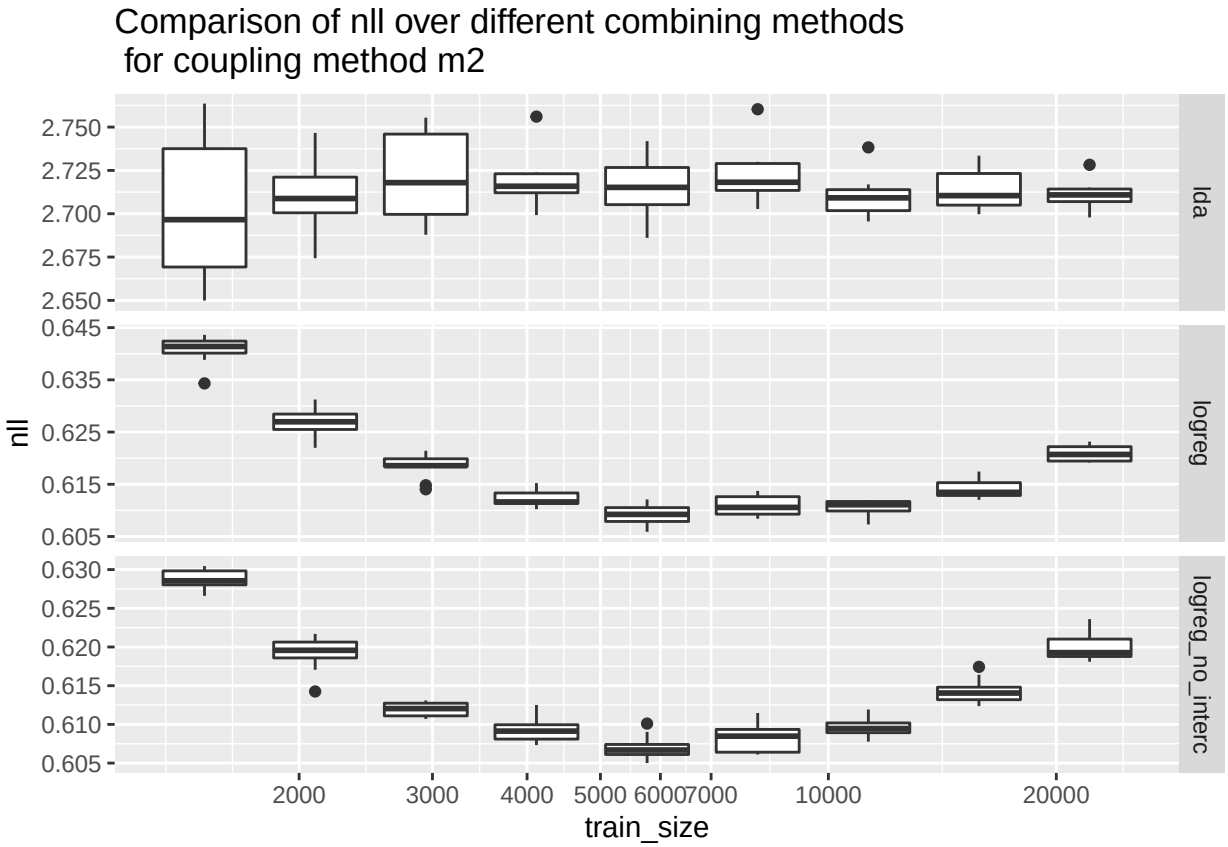
```
      scale_x_log10(breaks=log_breaks(n=10)) +
      ggtitle(paste0("Comparison of ", metric, " over different combining methods\n for coupling method
    print(box_plt)
  }
}
```

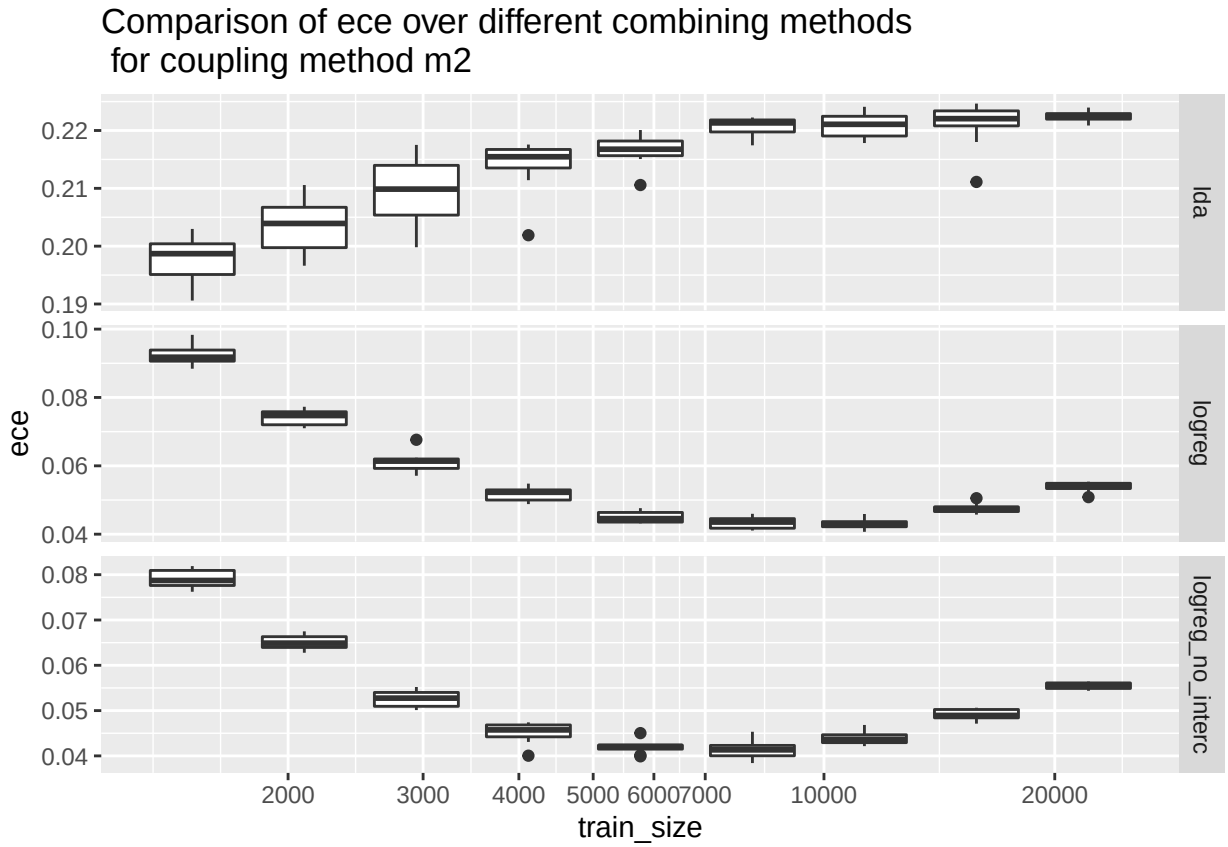


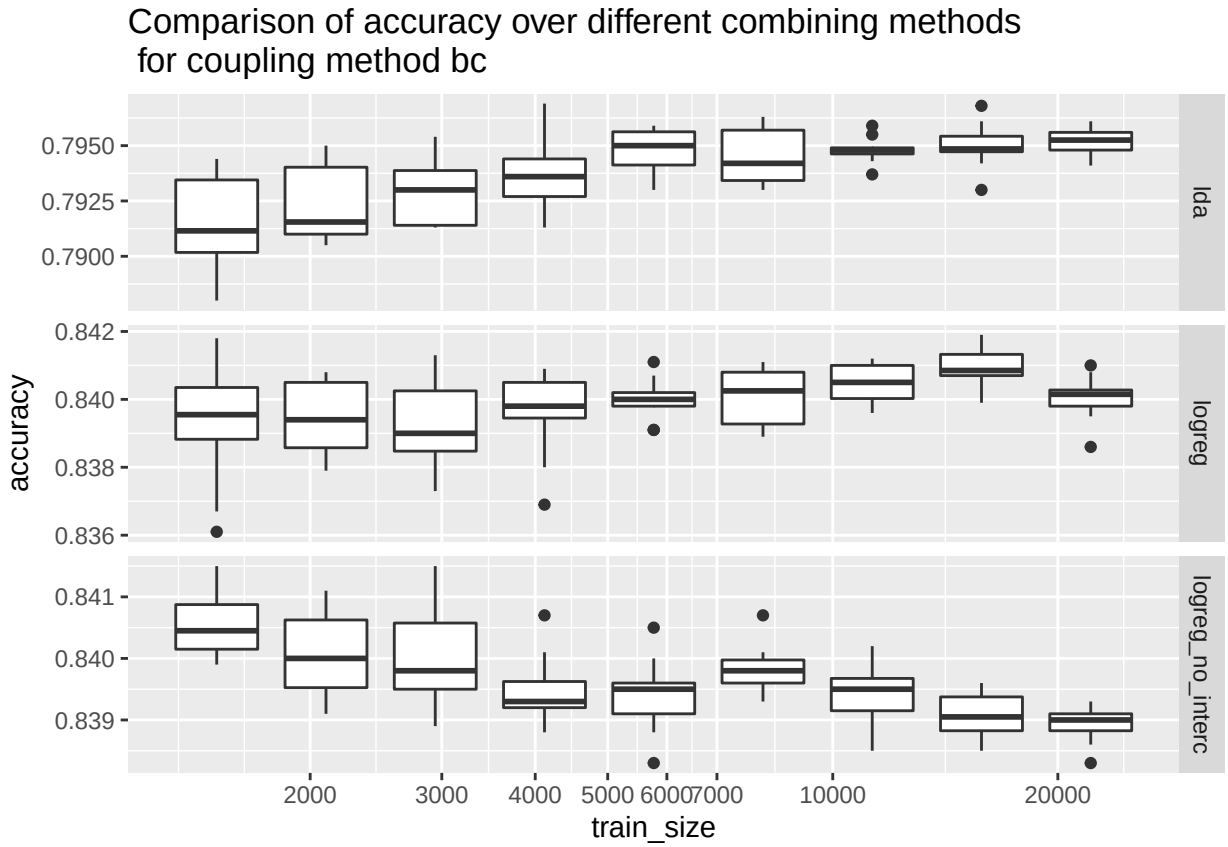

Comparison of nll over different combining methods
for coupling method m1
nll
train_size
lda
logreg
logreg_no_interc
2.61
2.58
2.55
2.52
2.49
0.64
0.63
0.62
0.61
0.60
0.59
2000
3000
4000
5000
6000
7000
10000
20000

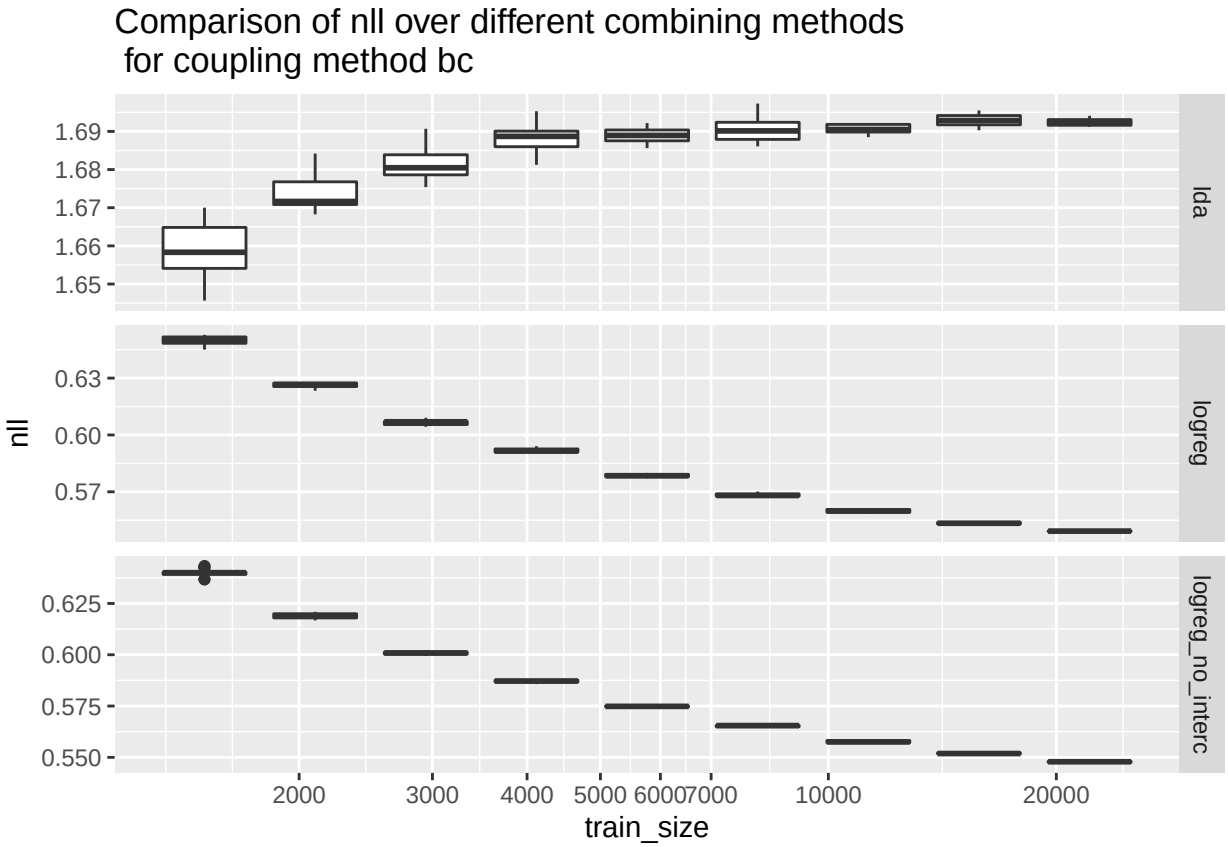

Comparison of ece over different combining methods
for coupling method m1
ece
0.21
0.20
0.19
0.11
0.09
0.07
0.05
0.08
0.06
0.04
lda
logreg
logreg_no_interc
2000
3000
4000
5000
6000
7000
10000
20000
train_size

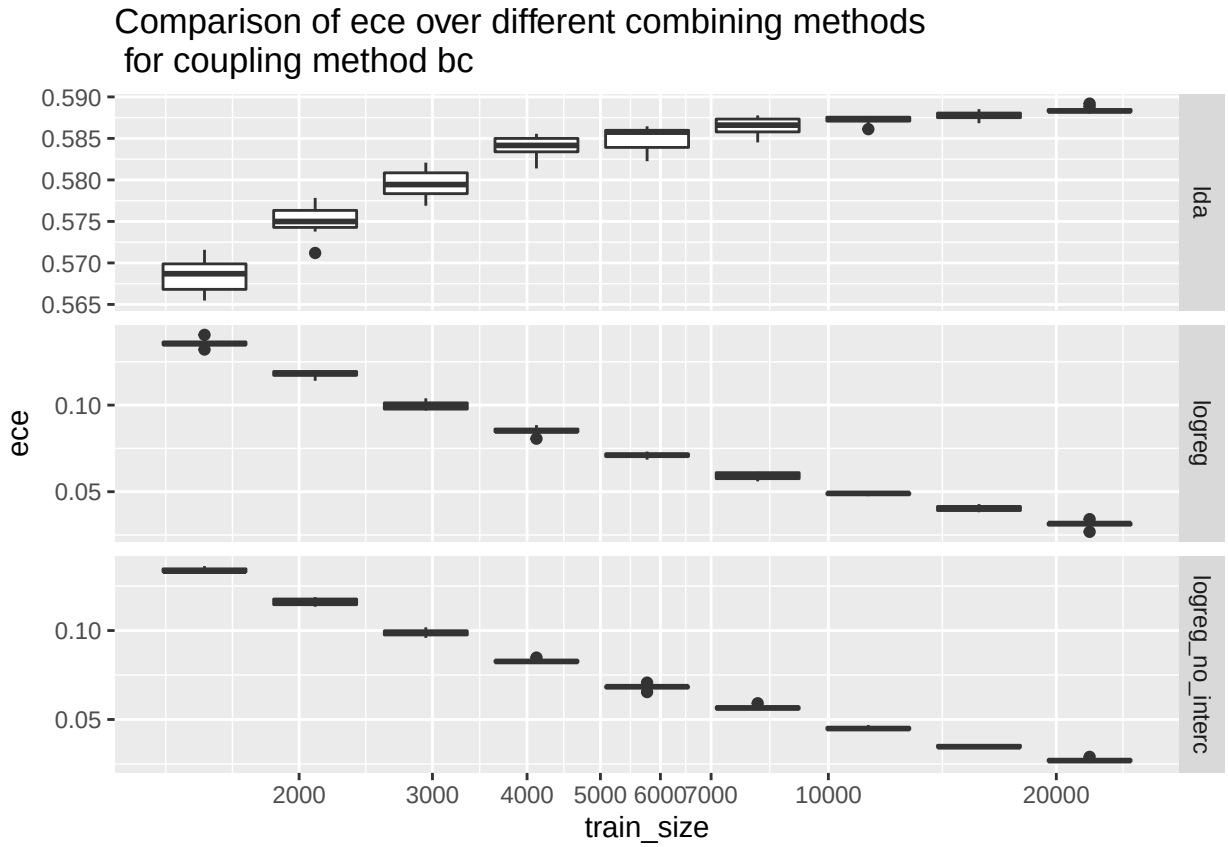

Comparison of accuracy over different combining methods
for coupling method m2
accuracy
0.798
0.796
0.794
0.792
0.790
0.839
0.838
0.837
0.836
0.835
lda
logreg
logreg_no_interc
2000
3000
4000
5000
6000
7000
10000
20000
train_size

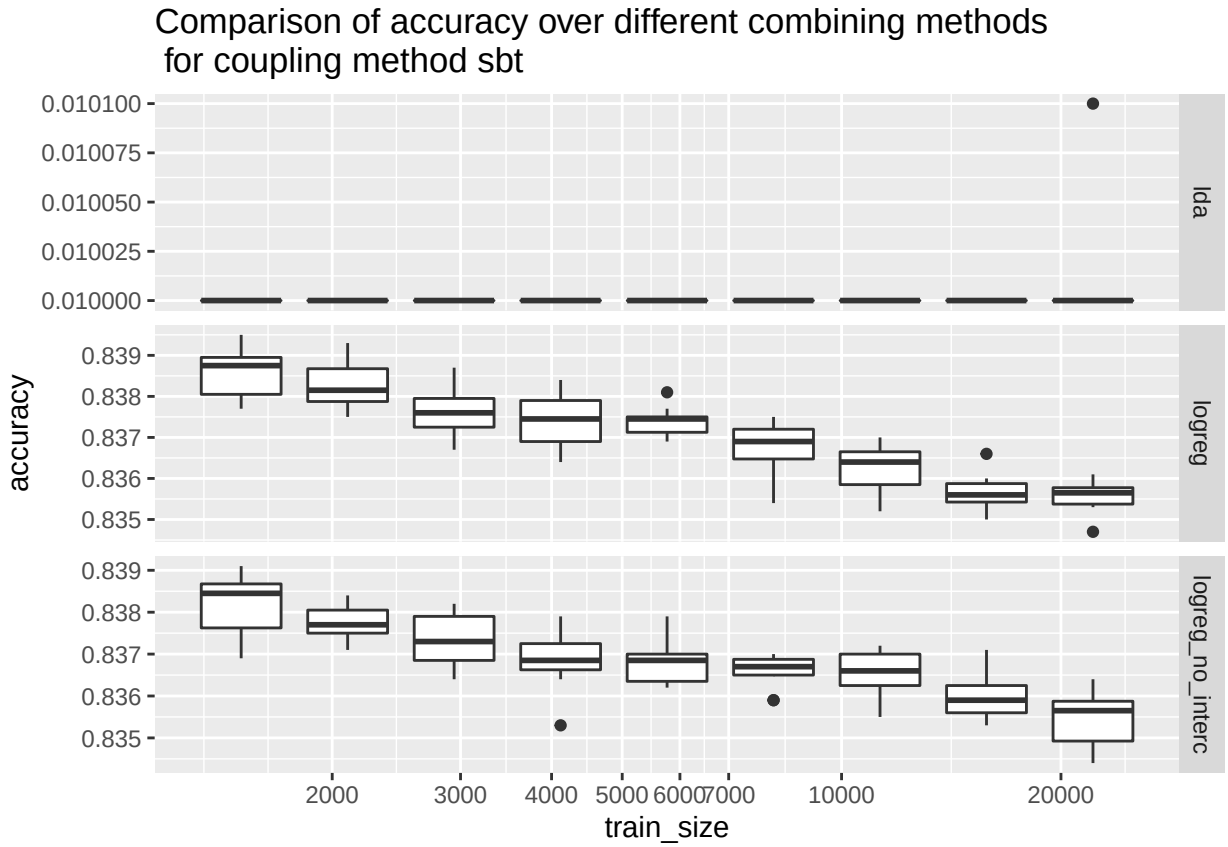

Comparison of nll over different combining methods
for coupling method m2
nll
train_size
lda
logreg
logreg_no_interc
2.750
2.725
2.700
2.675
2.650
0.645
0.635
0.625
0.615
0.605
0.630
0.625
0.620
0.615
0.610
0.605
2000
3000
4000
5000
6000
7000
10000
20000


Comparison of ece over different combining methods
for coupling method m2
ece
train_size
lda
logreg
logreg_no_interc
0.22
0.21
0.20
0.19
0.10
0.08
0.06
0.04
0.08
0.07
0.06
0.05
0.04
2000
3000
4000
5000
6000
7000
10000
20000

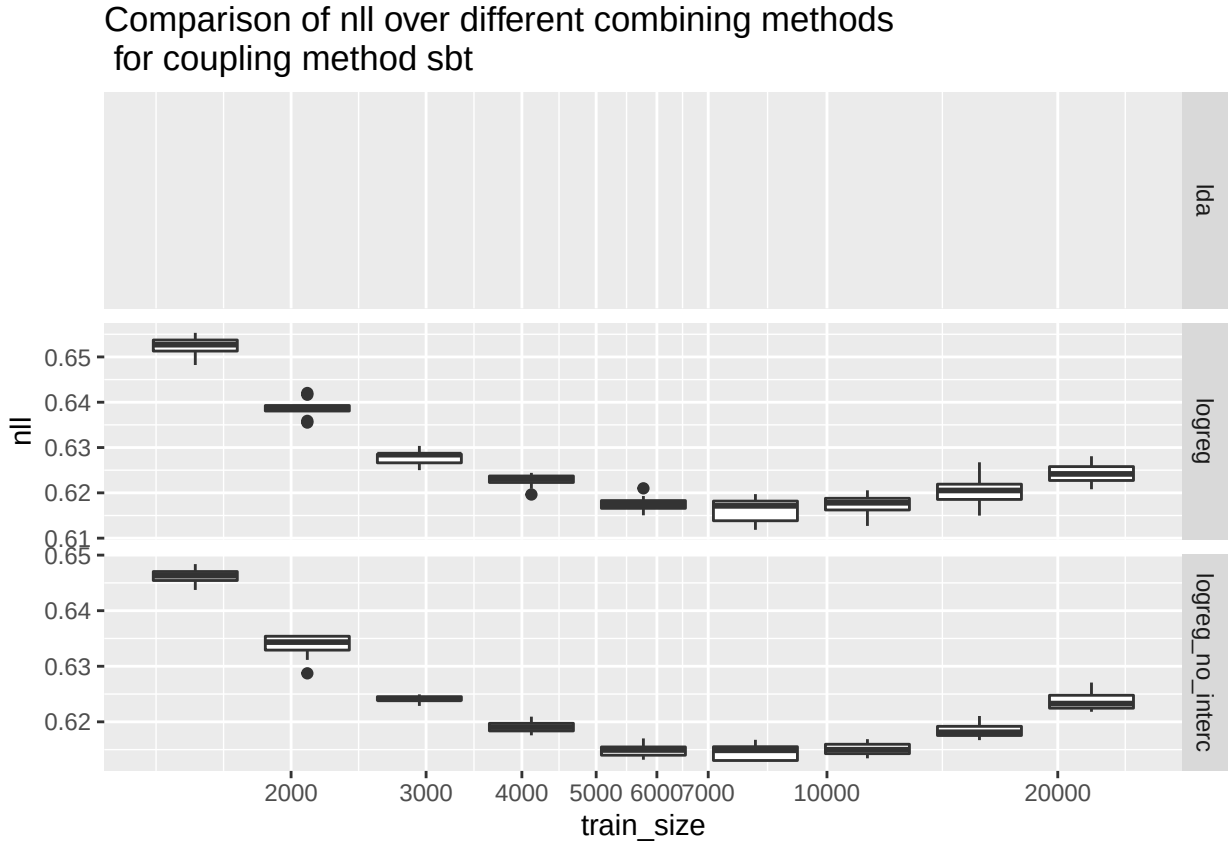

Comparison of accuracy over different combining methods
for coupling method bc
accuracy
0.7950
0.7925
0.7900
0.842
0.840
0.838
0.836
0.841
0.840
0.839
lda
logreg
logreg_no_interc
2000
3000
4000
5000
6000
7000
10000
20000
train_size


Comparison of nll over different combining methods
for coupling method bc
nll
1.69
1.68
1.67
1.66
1.65
0.63
0.60
0.57
0.625
0.600
0.575
0.550
lda
logreg
logreg_no_interc
2000
3000
4000
5000
6000
7000
10000
20000
train_size

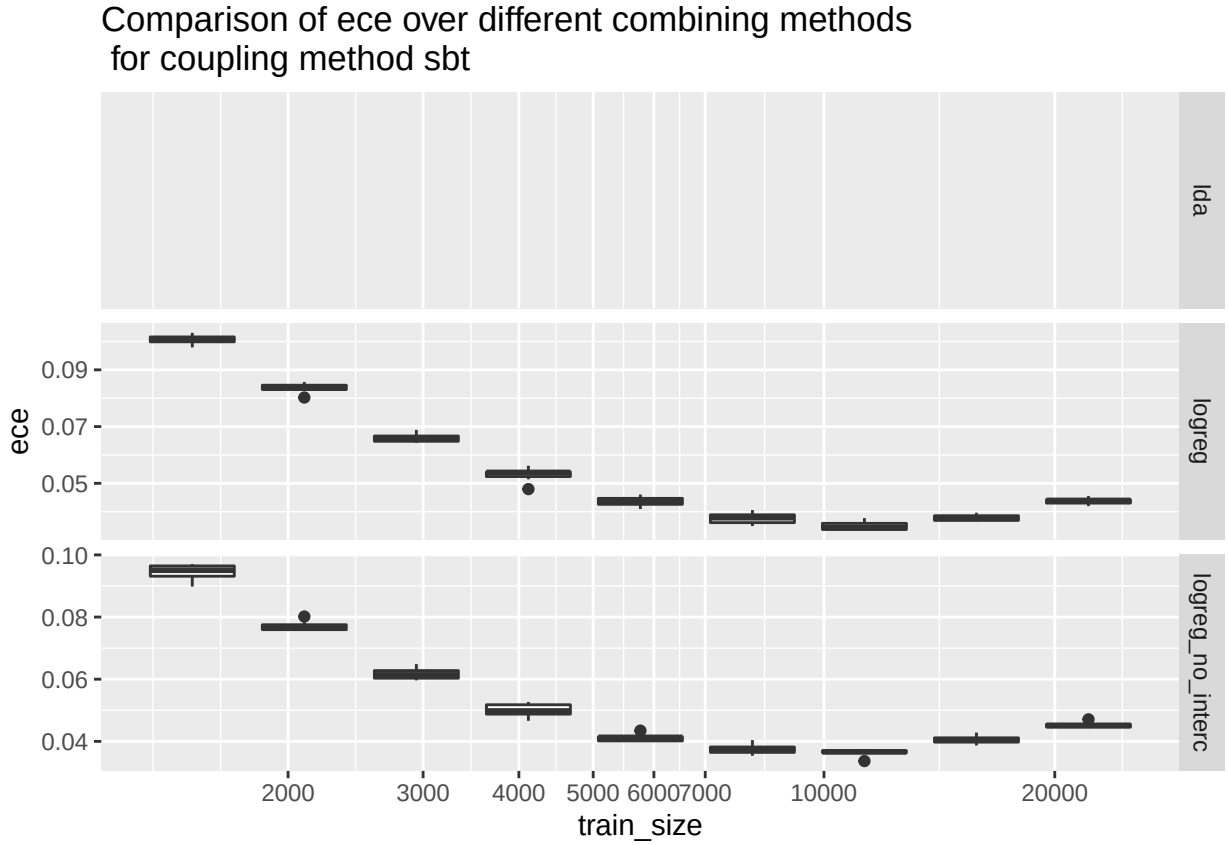

Comparison of ece over different combining methods
for coupling method bc
ece
0.590
0.585
0.580
0.575
0.570
0.565
0.10
0.05
0.10
0.05
lda
logreg
logreg_no_interc
2000
3000
4000
5000
6000
7000
10000
20000
train_size




```
## Warning: Removed 90 rows containing non-finite values (stat_boxplot).
```




```
## Warning: Removed 90 rows containing non-finite values (stat_boxplot).
```



Šuch, Ondrej, and Santiago Barreda. 2016. "Bayes Covariant Multi-Class Classification." *Pattern Recognition Letters* 84: 99–106.

Šuch, Ondrej, Štefan Benuš, and Andrea Tinajová. 2015. "A New Method to Combine Probability Estimates from Pairwise Binary Classifiers." *Rmj* 1: 12.

Wu, Ting-Fan, Chih-Jen Lin, and Ruby C Weng. 2004. "Probability Estimates for Multi-Class Classification by Pairwise Coupling." *Journal of Machine Learning Research* 5 (Aug): 975–1005.