Supplementary Material for RFpredInterval: An R Package for Prediction Intervals with Random Forests and Boosted Forests

by Cansu Alakuş, Denis Larocque and Aurélie Labbe

Data preprocessing

In this section, we present the steps to prepare Ames Housing data set for the analysis. We use the processed version of the data set from the **AmesHousing** package.

```
library("AmesHousing")
AmesHousing <- make_ordinal_ames()

## Data preprocessing
# remove observations with missing values
AmesHousing <- AmesHousing[complete.cases(AmesHousing), ]

# convert the response variable in thousands
AmesHousing$Sale_Price <- AmesHousing$Sale_Price/1000

# convert the ordered factors to numeric to preserve the ordering of the factors
ord_vars <- vapply(AmesHousing, is.ordered, logical(1))
nam_ord <- names(ord_vars)[ord_vars]
AmesHousing[, nam_ord] <- data.frame(lapply(AmesHousing[, nam_ord], as.numeric))

# group together levels with less than 30 observations,
# we use the combineLevels() function from "rockchalk" package for this step
library("rockchalk")
fac_vars <- vapply(AmesHousing, is.factor, logical(1))
AmesHousing[, fac_vars] <- data.frame(
  lapply(AmesHousing[, fac_vars],
       function(x, nmin) combineLevels(x, levs = names(table(x))[table(x)<nmin],
                                       newLabel=c("combinedLevels")),
         nmin=30)
)
```

Mean PI length results for the simulated and real data sets

Figures S1 to S7 present the mean PI length of each method for each simulated data set. Similarly, figures S8 to S10 show the mean PI length of each method for each real data set. The list of methods compared in the simulation study and real data analyses are as follows.

PIBF	Prediction intervals with boosted forests (the proposed method)
$\widehat{PI}_\alpha$	Conditional α-level prediction interval
OOB	Out-of-Bag approach
LM	Classical method with LS splitting rule
Quant	Quantiles with LS splitting rule
SPI	Shortest prediction interval with LS splitting rule
HDR	Highest density region with LS splitting rule
CHDR	Contiguous highest density region with LS splitting rule
SC	Split conformal
QRF	Quantile regression forest
GRF	Generalized random forest

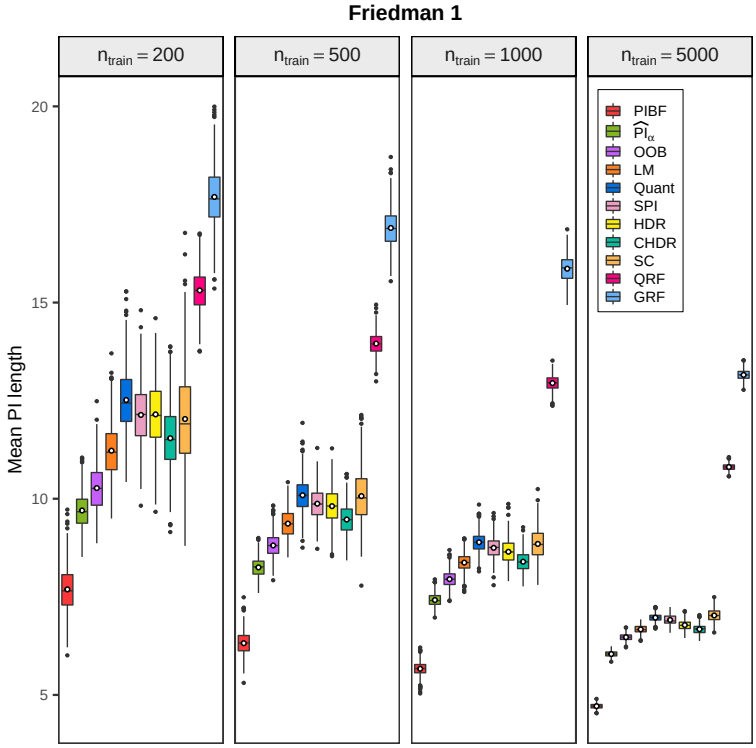

Figure S1: Distributions of the mean PI length over the test set across 500 replications for Friedman Problem 1.

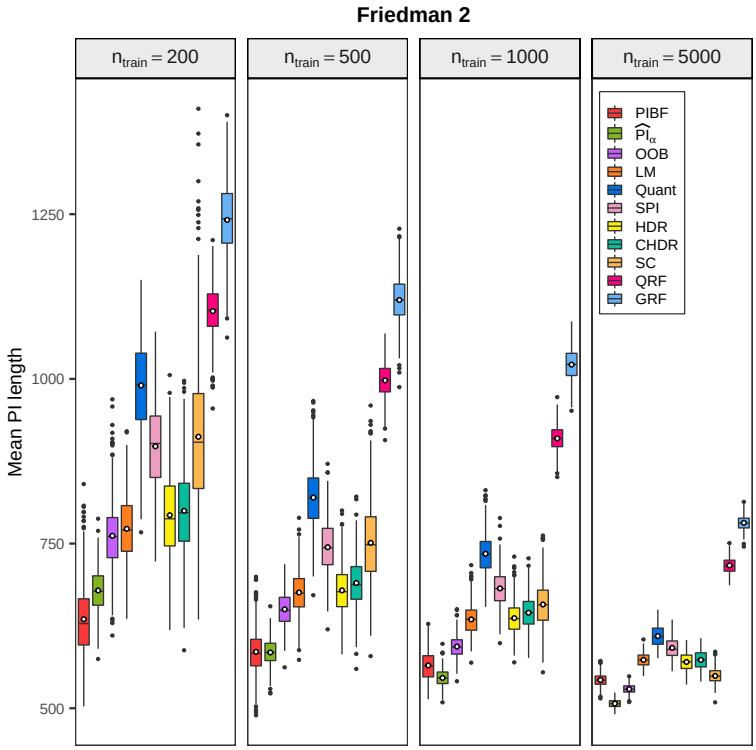

Figure S2: Distributions of the mean PI length over the test set across 500 replications for Friedman Problem 2.

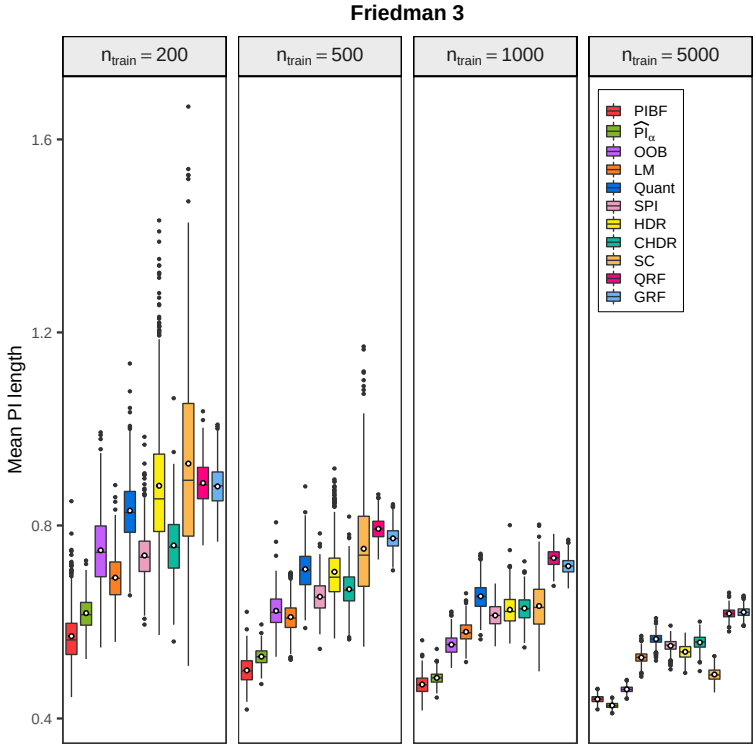

Figure S3: Distributions of the mean PI length over the test set across 500 replications for Friedman Problem 3.

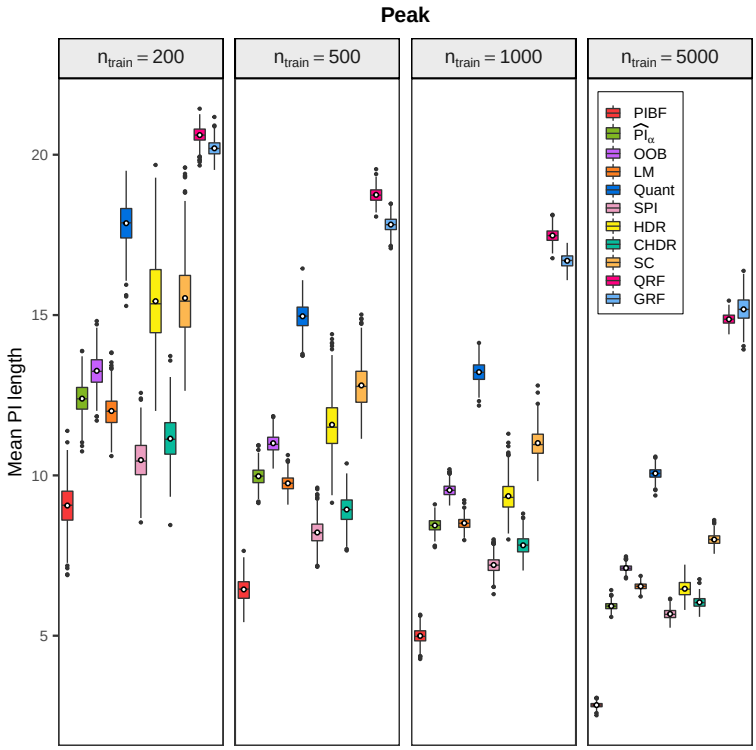

Figure S4: Distributions of the mean PI length over the test set across 500 replications for Peak Benchmark Problem.

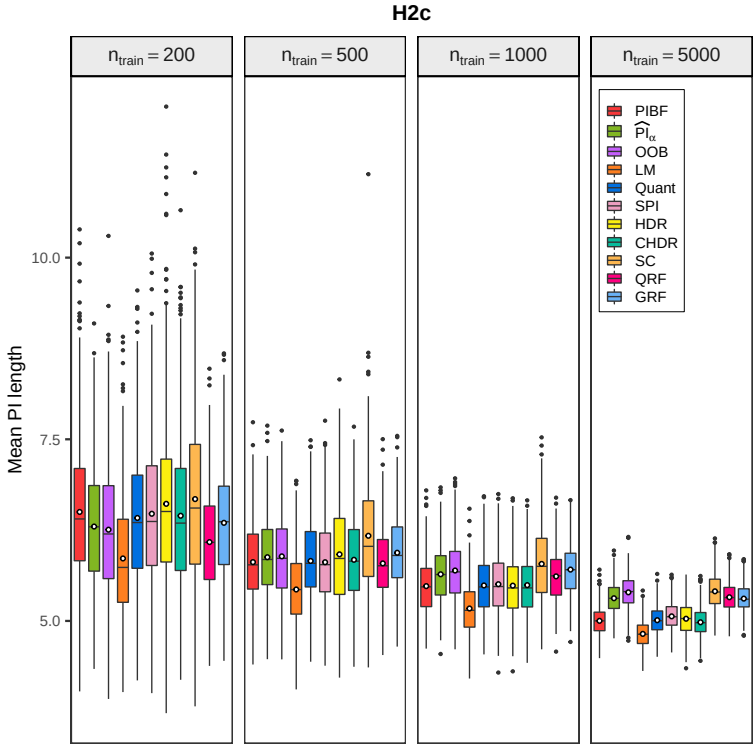

Figure S5: Distributions of the mean PI length over the test set across 500 replications for the H2c setup.

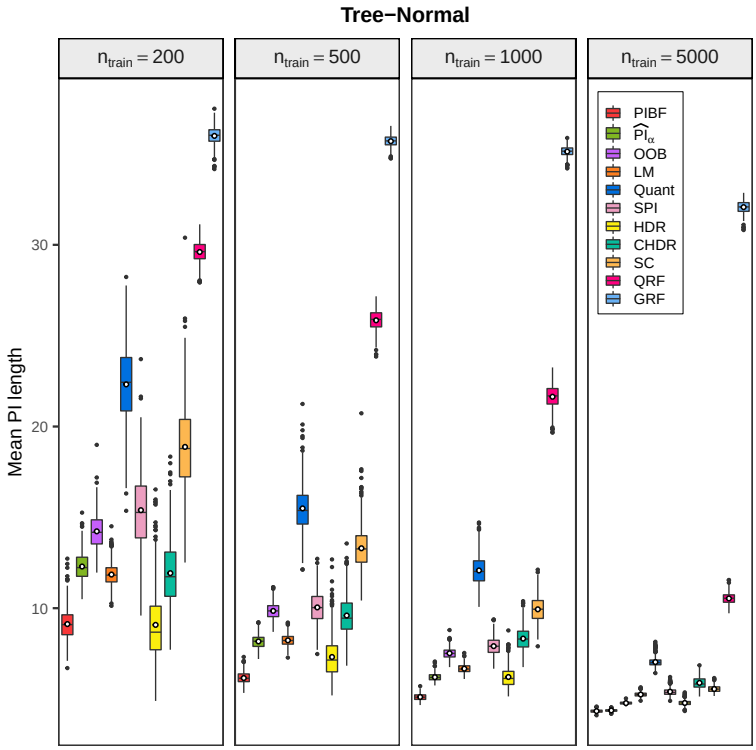

Figure S6: Distributions of the mean PI length over the test set across 500 replications for the tree-based problem with normally distributed error.

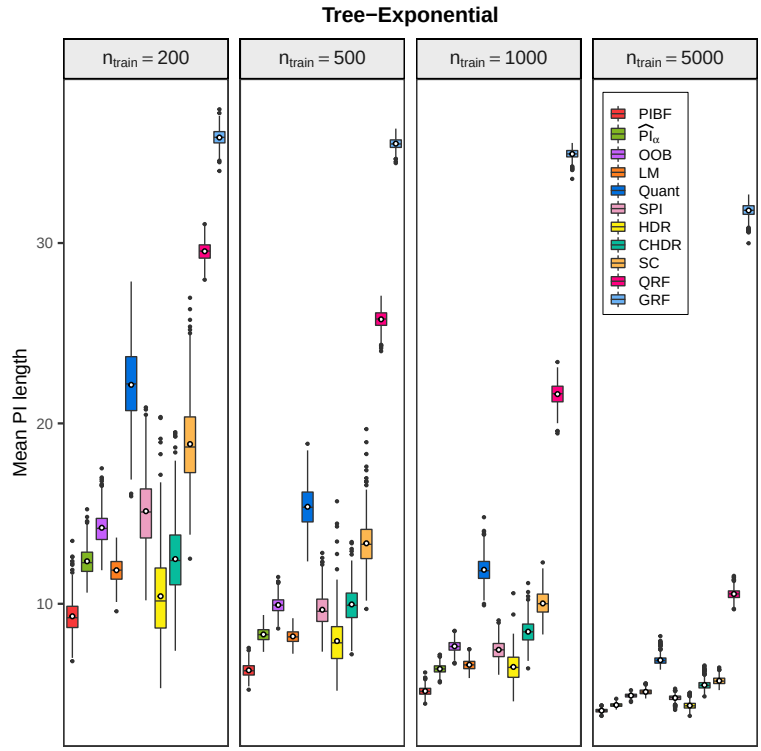

Figure S7: Distributions of the mean PI length over the test set across 500 replications for the tree-based problem with exponentially distributed error.

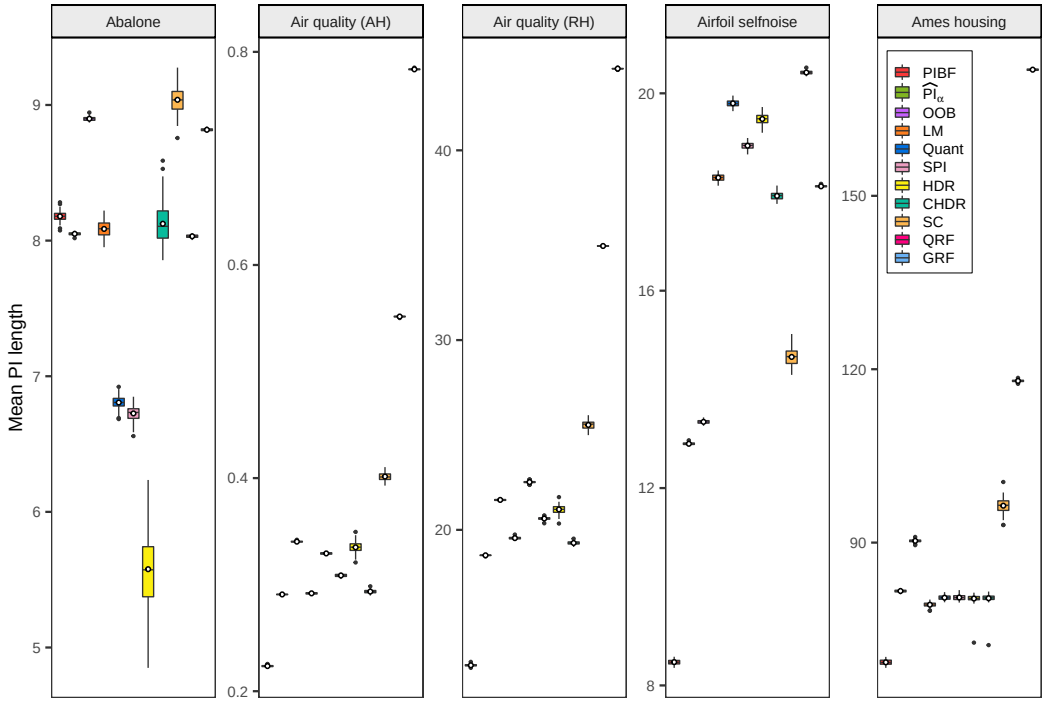

Figure S8: Distributions of the mean PI length across 100 repetitions for Abalone, Air quality with absolute and relative humidity, Airfoil selfnoise, and Ames housing data sets.

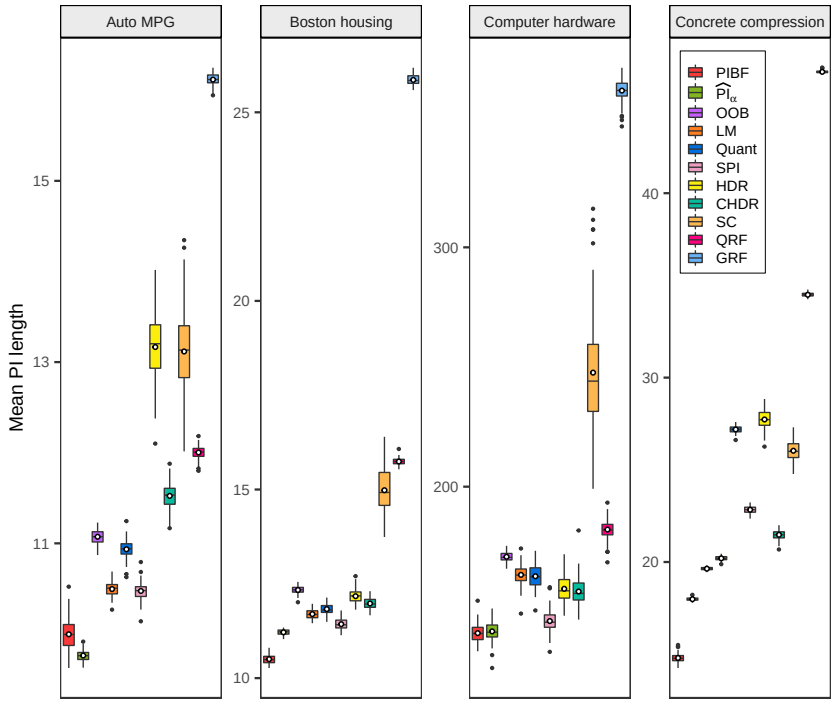

Figure S9: Distributions of the mean PI length across 100 repetitions for Auto MPG, Boston housing, Computer hardware, and Concrete compression data sets.

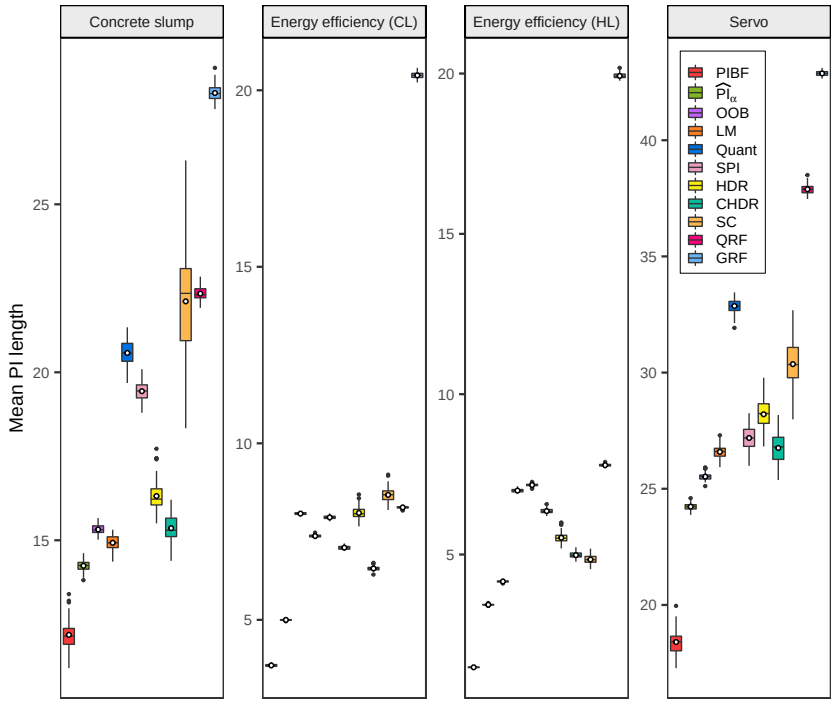

Figure S10: Distributions of the mean PI length across 100 repetitions for Concrete slump, Energy efficiency with cooling and heating load, and Servo data sets.