

Boosting Diversity in Regression Ensembles

M. Bourel J. Cugliari Y. Goude Jean-Michel Poggi



Montevideo 2025

This research benefited from the support of the FMJH "Program Gaspard Monge for optimization and operations research and their interactions with data science", and from the support from EDF and Thales

Analyze ensemble aggregation through diversity-based MSE decomposition

(Krogh, Vedelsby, 1995), (Brown *et al.*, 2005)

- ▶ Regression $y_t = f(x_t) + \epsilon_t$, consider f_1, \dots, f_M different individual predictors
- ▶ Let $\hat{y}_t = \bar{f}(t) = \sum_{m=1}^M c_m f_m(t)$ be the aggregated predictor (weighted mean)

Analyze ensemble aggregation through diversity-based MSE decomposition

(Krogh, Vedelsby, 1995), (Brown *et al.*, 2005)

- ▶ Regression $y_t = f(x_t) + \epsilon_t$, consider f_1, \dots, f_M different individual predictors
- ▶ Let $\hat{y}_t = \bar{f}(t) = \sum_{m=1}^M c_m f_m(t)$ be the aggregated predictor (weighted mean)
- ▶ Then: squared error = weighted average error of the predictors - diversity term

$$(\bar{f}(t) - y_t)^2 = \sum_{m=1}^M c_m (f_m(t) - y_t)^2 - \sum_{m=1}^M c_m (f_m(t) - \bar{f}(t))^2$$

Analyze ensemble aggregation through diversity-based MSE decomposition

(Krogh, Vedelsby, 1995), (Brown *et al.*, 2005)

- ▶ Regression $y_t = f(x_t) + \epsilon_t$, consider f_1, \dots, f_M different individual predictors
- ▶ Let $\hat{y}_t = \bar{f}(t) = \sum_{m=1}^M c_m f_m(t)$ be the aggregated predictor (weighted mean)
- ▶ Then: squared error = **weighted average error of the predictors** - **diversity term**

$$(\bar{f}(t) - y_t)^2 = \sum_{m=1}^M c_m (f_m(t) - y_t)^2 - \sum_{m=1}^M c_m (f_m(t) - \bar{f}(t))^2$$

- ▶ Decomposes an instantaneous error (no expectation taken)
- ▶ **Adding relatively accurate diverse predictors reduces the error**

A modified cost function to incorporate diversity in the boosting sequence

- ▶ Starting from the *L_2 -gradient boosting* for regression problems $y = f(x) + \epsilon$
Buhlmann and Yu, 2003, Friedman et al., 2000, Mason et al., 2000
- ▶ we propose to modify the cost function able to *enhance diversity* during the learning iterations generating the individual experts

A modified cost function to incorporate diversity in the boosting sequence

- ▶ Starting from the *L_2 -gradient boosting* for regression problems $y = f(x) + \epsilon$
Buhlmann and Yu, 2003, Friedman et al., 2000, Mason et al., 2000
- ▶ we propose to modify the cost function able to *enhance diversity* during the learning iterations generating the individual experts
- ▶ Take uniform weights $c_m = 1/M$ for \bar{f}

$$C(y, f) = \frac{1}{2}(y - f)^2 - \frac{\kappa}{2}(f - \bar{f})^2$$

where:

- ▶ κ modulates the importance given to the diversity of the predictor to the mean of the previous.
- ▶ \bar{f} is seen as a constant (in practice it is the mean of past individual predictors)

BOosting Diversity (BoDi algorithm)

$\mathcal{L} = \{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$ a sample, \mathcal{F} a family of functions, $\kappa > 0$ and $\delta > 0$.

Randomly split the data in two disjoint parts $I = I_1 \cup I_2$ (not mandatory)

1. Fit an initial learner $\hat{f}_0 = \hat{F}_0 \in \mathcal{F}$ over I_1 such that $\hat{F}_0 = \underset{f \in \mathcal{F}}{\text{Argmin}} \sum_{i \in I_1} (y_i - f(\mathbf{x}_i))^2$.

Set $\hat{F}_0^*(\mathbf{x}) = \hat{F}_0(\mathbf{x})$.

BOosting Diversity (BoDi algorithm)

$\mathcal{L} = \{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$ a sample, \mathcal{F} a family of functions, $\kappa > 0$ and $\delta > 0$.

Randomly split the data in two disjoint parts $I = I_1 \cup I_2$ (not mandatory)

1. Fit an initial learner $\hat{f}_0 = \hat{F}_0 \in \mathcal{F}$ over I_1 such that $\hat{F}_0 = \underset{f \in \mathcal{F}}{\text{Argmin}} \sum_{i \in I_1} (y_i - f(\mathbf{x}_i))^2$.

Set $\hat{F}_0^*(\mathbf{x}) = \hat{F}_0(\mathbf{x})$.

2. For $m \in \{1, \dots, M\}$:

2.1 $\forall i \in I_2$, evaluate the negative diversity gradient of the cost function at $\hat{F}_{m-1}(\mathbf{x}_i)$:

$$u_i = (y_i - \hat{F}_{m-1}(\mathbf{x}_i)) + \kappa_m \left(\hat{F}_{m-1}(\mathbf{x}_i) - \hat{F}_{m-1}^*(\mathbf{x}_i) \right)$$

with $\kappa_m = \kappa \left(1 - \frac{1}{m}\right)$ if $m > 1$, $\kappa_1 = \kappa$ and $\hat{f}_m = \underset{f \in \mathcal{F}}{\text{Argmin}} \sum_{i \in I_2} (u_i - f(\mathbf{x}_i))^2$

BOosting Diversity (BoDi algorithm)

$\mathcal{L} = \{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$ a sample, \mathcal{F} a family of functions, $\kappa > 0$ and $\delta > 0$.

Randomly split the data in two disjoint parts $I = I_1 \cup I_2$ (not mandatory)

1. Fit an initial learner $\hat{f}_0 = \hat{F}_0 \in \mathcal{F}$ over I_1 such that $\hat{F}_0 = \underset{f \in \mathcal{F}}{\text{Argmin}} \sum_{i \in I_1} (y_i - f(\mathbf{x}_i))^2$.

Set $\hat{F}_0^*(\mathbf{x}) = \hat{F}_0(\mathbf{x})$.

2. For $m \in \{1, \dots, M\}$:

2.1 $\forall i \in I_2$, evaluate the negative diversity gradient of the cost function at $\hat{F}_{m-1}(\mathbf{x}_i)$:

$$u_i = (y_i - \hat{F}_{m-1}(\mathbf{x}_i)) + \kappa_m \left(\hat{F}_{m-1}(\mathbf{x}_i) - \hat{F}_{m-1}^*(\mathbf{x}_i) \right)$$

with $\kappa_m = \kappa \left(1 - \frac{1}{m}\right)$ if $m > 1$, $\kappa_1 = \kappa$ and $\hat{f}_m = \underset{f \in \mathcal{F}}{\text{Argmin}} \sum_{i \in I_2} (u_i - f(\mathbf{x}_i))^2$

2.2 Update boosting predictor as $\hat{F}_m(\mathbf{x}) = \hat{F}_{m-1}(\mathbf{x}) + \delta \hat{f}_m(\mathbf{x})$.

Compute $\hat{F}_m^*(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^m \hat{F}_i(\mathbf{x})$ and update $I_2 = I \setminus I_1$ with a new subsample I_1 of I

Outputs: family of experts $\hat{f}_0, \hat{f}_1, \hat{f}_2, \dots, \hat{f}_M$ and aggregated predictors \hat{F}_m and \hat{F}_m^*

Comments and a theoretical result

- ▶ With $\kappa = 0$ we get L_2 -gradient boosting

Comments and a theoretical result

- ▶ With $\kappa = 0$ we get L_2 -gradient boosting
- ▶ If κ is non negative and $\kappa \leq 1$ the new cost function is strongly convex and Lipschitz (with associated regularity L)

Comments and a theoretical result

- ▶ With $\kappa = 0$ we get L_2 -gradient boosting
- ▶ If κ is non negative and $\kappa \leq 1$ the new cost function is strongly convex and Lipschitz (with associated regularity L)
- ▶ Under these conditions, a result from Biau and Cadre (2019) warranties that:

Theorem

*If δ the step of BoDi is such that $0 < \delta < 1/(2L)$, then **the optimisation strategy converges to a global optimum***

$$\lim_{t \rightarrow \infty} \mathbb{E}(C(y, F_t)) = \inf_{F \in \text{lin}(\mathcal{F})} \mathbb{E}(C(y, F))$$

where $\text{lin}(\mathcal{F})$ is the linear span of the family of functions \mathcal{F} we reach (typically, the collection of all CART trees with k terminal nodes)

Comments and a theoretical result

- ▶ With $\kappa = 0$ we get L_2 -gradient boosting
- ▶ If κ is non negative and $\kappa \leq 1$ the new cost function is strongly convex and Lipschitz (with associated regularity L)
- ▶ Under these conditions, a result from Biau and Cadre (2019) warranties that:

Theorem

*If δ the step of BoDi is such that $0 < \delta < 1/(2L)$, then **the optimisation strategy converges to a global optimum***

$$\lim_{t \rightarrow \infty} \mathbb{E}(C(y, F_t)) = \inf_{F \in \text{lin}(\mathcal{F})} \mathbb{E}(C(y, F))$$

where $\text{lin}(\mathcal{F})$ is the linear span of the family of functions \mathcal{F} we reach (typically, the collection of all CART trees with k terminal nodes)

- ▶ This is an **optimization warranty** but **not** a **statistical** one

Numerical experiments on simulated data

- ▶ Well-known [simulated data set](#) Friedman 1991 used in Breiman 1996 for bagging
- ▶ 10 independent variables uniformly distributed on $[0,1]$, only 5 are active:

$$y_i = 10 \sin(\pi x_{1,i} x_{2,i}) + 20(x_{3,i} - 0.5)^2 + 10x_{4,i} + 5x_{5,i} + \varepsilon_i$$

where ε_i is $N(0, \sigma^2)$

- ▶ As in Breiman 1996 we simulated a learning set of size $n_0 = 200$ and a test set of size $n_1 = 1000$ observations with $\sigma = 1$

Numerical experiments on simulated data

- ▶ Well-known **simulated data set** Friedman 1991 used in Breiman 1996 for bagging
- ▶ 10 independent variables uniformly distributed on $[0,1]$, only 5 are active:

$$y_i = 10 \sin(\pi x_{1,i} x_{2,i}) + 20(x_{3,i} - 0.5)^2 + 10x_{4,i} + 5x_{5,i} + \varepsilon_i$$

where ε_i is $N(0, \sigma^2)$

- ▶ As in Breiman 1996 we simulated a learning set of size $n_0 = 200$ and a test set of size $n_1 = 1000$ observations with $\sigma = 1$
- ▶ Three **base-learners** (see Hastie, Tibshirani, Friedman 2009) offering increasing "diversifiability"):
 - ▶ **stumps** which are very simple CART trees
 - ▶ two types of Random Forests **RF** the Breiman's original ones and **PRF** a purely random variant

About CART trees (figure from The Elements of Statistical Learning, HTF)

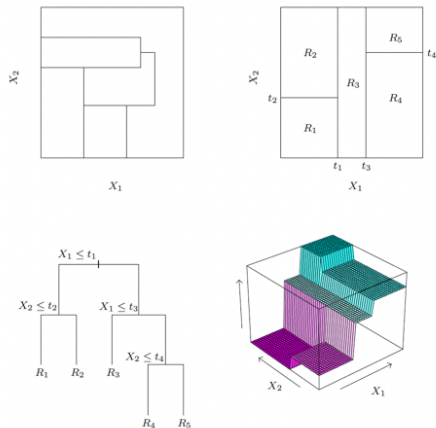
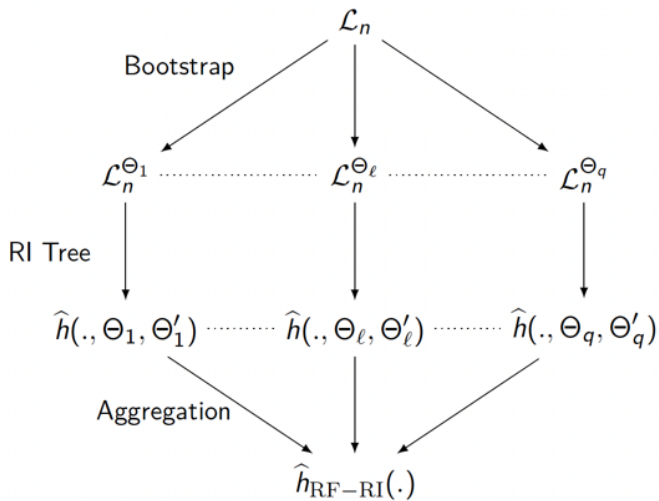


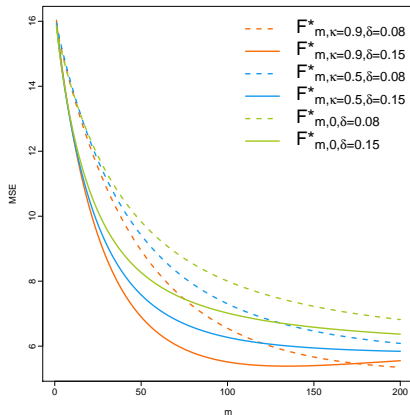
FIGURE 9.2. Partitions and CART. Top right panel shows a partition of a two-dimensional feature space by recursive binary splitting, as used in CART, applied to some fake data. Top left panel shows a general partition that cannot be obtained from recursive binary splitting. Bottom left panel shows the tree corresponding to the partition in the top right panel, and a perspective plot of the prediction surface appears in the bottom right panel.

About RF



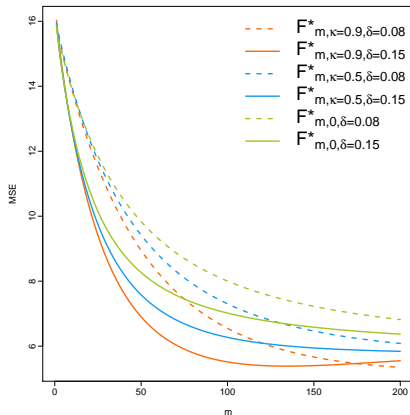
Influence of the diversity weight and the gradient step (for PRF)

- MSE as a function of m for diversity weights ($\kappa = 0$ (green); 0.5 (blue); 0.9 (red)) and gradient step $\delta = 0.08$ (dotted); 0.15 (solid) with **PRF** as base learner



Influence of the diversity weight and the gradient step (for PRF)

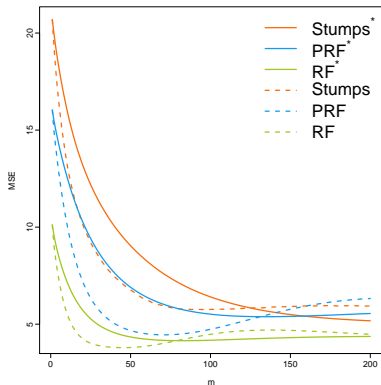
- MSE as a function of m for diversity weights ($\kappa = 0$ (green); 0.5 (blue); 0.9 (red)) and gradient step $\delta = 0.08$ (dotted); 0.15 (solid) with **PRF** as base learner



- Best results for $\kappa = 0.9$ showing the interest of encouraging diversity
- Influence of δ on the rate of convergence as in classical boosting

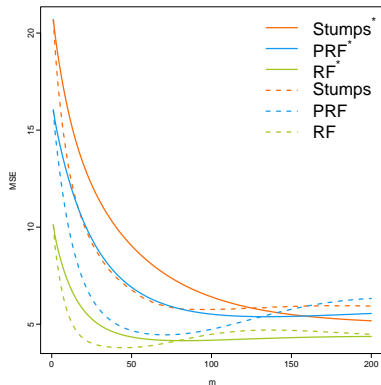
Influence of the base learner

- MSE of aggregated predictors F (dotted) and F^* (solid) as a function of boosting steps for 3 base learners: **Stumps** (red), **PRF** (blue) and **RF** (green)



Influence of the base learner

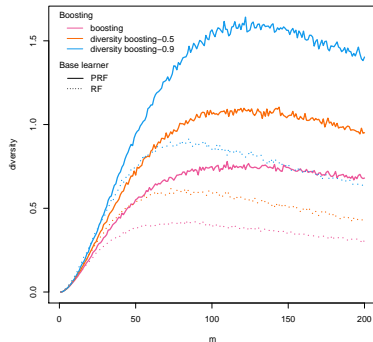
- MSE of aggregated predictors F (dotted) and F^* (solid) as a function of boosting steps for 3 base learners: **Stumps** (red), **PRF** (blue) and **RF** (green)



- Best results for RF. But the relative improvement is far more important for PRF which can generate more diversity than RF
- Good convergence of the algorithm and its robustness regarding the choice of m

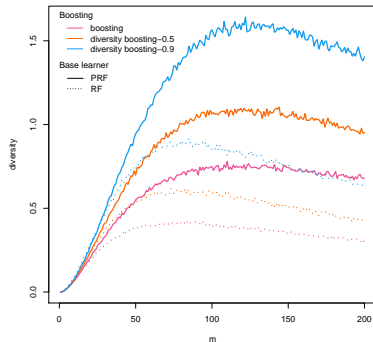
More about the diversity term

- Diversity term $\frac{1}{m} \sum_{k=1}^m (F_k(x_i) - F_k^*(x_i))^2$ averaged for $i \in I_1$, as a function of m for **PRF** (solid) and **RF** (dotted) for classical boosting ($\kappa = 0$) and diversity boosting ($\kappa = 0.5; 0.9$)



More about the diversity term

- Diversity term $\frac{1}{m} \sum_{k=1}^m (F_k(x_i) - F_k^*(x_i))^2$ averaged for $i \in I_1$, as a function of m for **PRF** (solid) and **RF** (dotted) for classical boosting ($\kappa = 0$) and diversity boosting ($\kappa = 0.5; 0.9$)



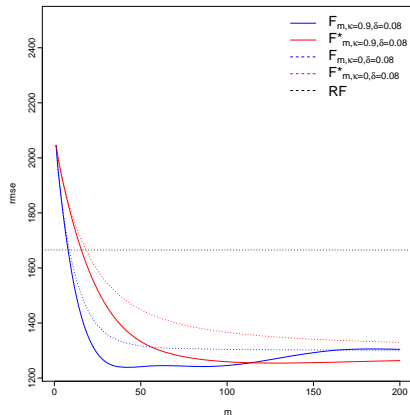
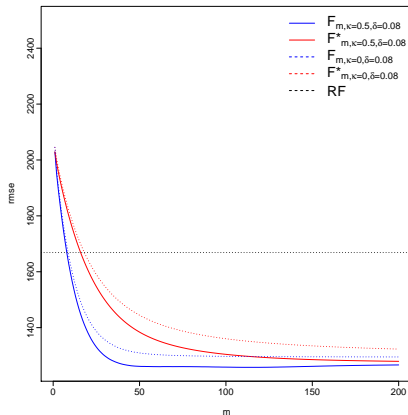
- Diversity increases quickly with m before decreasing slowly, increases with κ
- PRF allows to generate **more diversity** than RF

Numerical experiments on electricity consumption data

- ▶ **French electricity consumption** provided by the system operator RTE (Réseau de Transport d'Electricité)
- ▶ from the 1st of January 2012 to the 15th of March 2020 with a 30 minutes sampling period
- ▶ we add a covariate: the **national averaged temperature** from the French weather forecaster Meteo-France
- ▶ We **train** the models on historical data from January 2012 to the end of August 2019 and **test** on the last year.
- ▶ To avoid outliers we drop the holidays periods and bank holidays

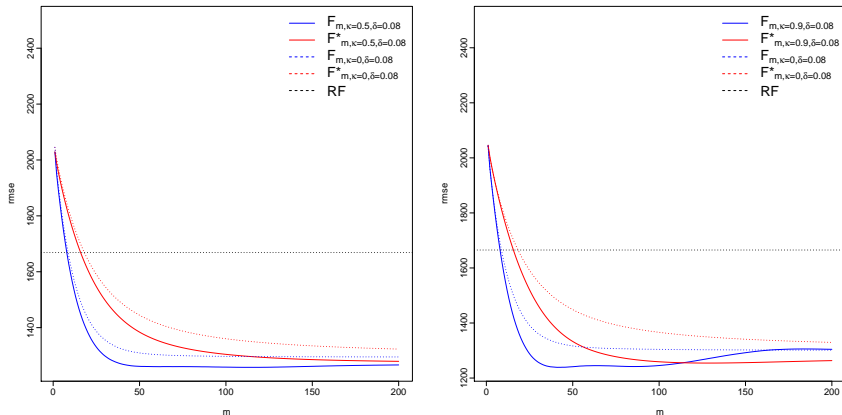
RF as base learner

- RMSE of F and F^* as a function of m for $\kappa = 0.5$ (left) $\kappa = 0.9$ (right) and for a gradient step ($\delta = 0.08$) with **RF** (ntree=100 and mtry=3) as base learner



RF as base learner

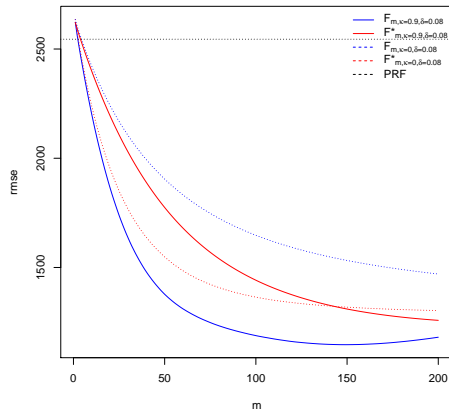
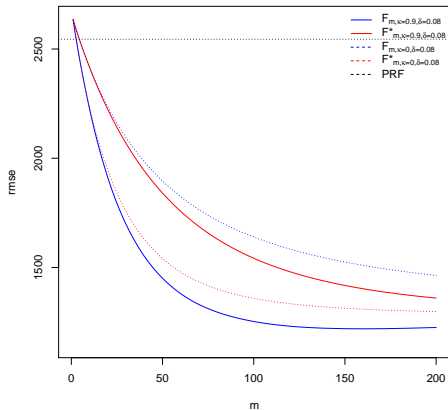
- ▶ RMSE of F and F^* as a function of m for $\kappa = 0.5$ (left) $\kappa = 0.9$ (right) and for a gradient step ($\delta = 0.08$) with **RF** (ntree=100 and mtry=3) as base learner



- ▶ The best RMSE are obtained for F_κ , followed by F^*_κ
- ▶ κ close to 1 improves forecasting performance as the learner can generate diversity

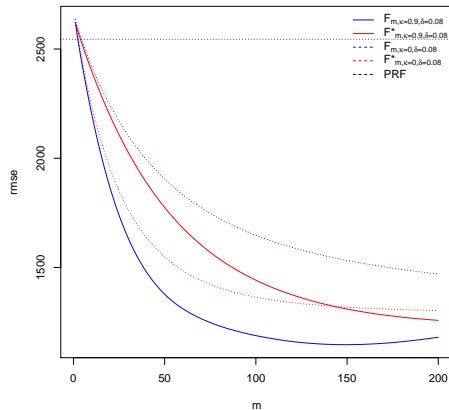
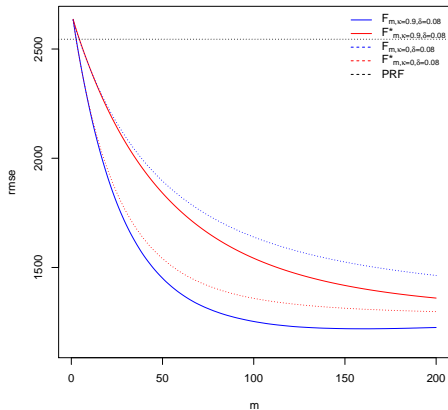
PRF as base learner: more diversity

- RMSE of F and F^* as a function of m for $\kappa = 0.5$ (left) $\kappa = 0.9$ (right) and for a gradient step ($\delta = 0.08$) with **PRF** (ntree=100) as base learner



PRF as base learner: more diversity

- RMSE of F and F^* as a function of m for $\kappa = 0.5$ (left) $\kappa = 0.9$ (right) and for a gradient step ($\delta = 0.08$) with **PRF** (ntree=100) as base learner



- Best RMSE for $F_{\kappa=0.9}$ with PRF a good base learner for diversity boosting

More comparisons on benchmark datasets

Data set	No. of observations	No. of expl. variables	Source and reference
Carseats	400	10	ISLR2
Ozone	366	12	mlbench
College	777	17	ISLR2
Airquality	153	5	ISLR2
Boston Housing	506	13	mlbench
Bikeshare	8645	12	ISLR2
Hitters	322	19	ISLR2

Median RMSE of the different learners +boosted +BoDi

	Ozone	Bikeshare	Hitters	Airquality	BostonHousing	Carseats	College
Stump	5.6	107.47	353.35	26.85	7.12	2.43	2330.58
+boost	5.03	90.75	327.86	19.13	5.53	2	1906.95
+BoDi	5.08	87.91	336	19.36	5.14	1.94	1919.95
CART	4.91	67.11	337.35	21.06	4.68	2.12	1415.45
+boost	4.74	54.06	307.02	17.36	3.78	1.72	1273.76
+BoDi	4.92	53.76	311.68	17.39	3.98	1.74	1247.74
PRF	4.28	70.89	301.82	17.41	4.95	2.27	1460.54
+boost	3.96	60.13	272.66	15.35	3.46	1.72	1130.4
+BoDi	3.91	52.26	268.21	15.65	3.2	1.5	1058.55
RF	3.99	47.73	267.89	15.02	3.1	1.71	994.73
+boost	3.97	42.66	267.46	14.8	2.92	1.43	1027.18
+BoDi	4.04	37.98	273.86	15.22	2.9	1.36	1010.88

Median RMSE of the different learners +boosted +BoDi

	Ozone	Bikeshare	Hitters	Airquality	BostonHousing	Carseats	College
Stump	5.6	107.47	353.35	26.85	7.12	2.43	2330.58
+boost	5.03	90.75	327.86	19.13	5.53	2	1906.95
+BoDi	5.08	87.91	336	19.36	5.14	1.94	1919.95
CART	4.91	67.11	337.35	21.06	4.68	2.12	1415.45
+boost	4.74	54.06	307.02	17.36	3.78	1.72	1273.76
+BoDi	4.92	53.76	311.68	17.39	3.98	1.74	1247.74
PRF	4.28	70.89	301.82	17.41	4.95	2.27	1460.54
+boost	3.96	60.13	272.66	15.35	3.46	1.72	1130.4
+BoDi	3.91	52.26	268.21	15.65	3.2	1.5	1058.55
RF	3.99	47.73	267.89	15.02	3.1	1.71	994.73
+boost	3.97	42.66	267.46	14.8	2.92	1.43	1027.18
+BoDi	4.04	37.98	273.86	15.22	2.9	1.36	1010.88

Table: Winners: +boosted or +BoDi?

Median RMSE of the different learners +boosted +BoDi

	Ozone	Bikeshare	Hitters	Airquality	BostonHousing	Carseats	College
Stump	5.6	107.47	353.35	26.85	7.12	2.43	2330.58
+boost	5.03	90.75	327.86	19.13	5.53	2	1906.95
+BoDi	5.08	87.91	336	19.36	5.14	1.94	1919.95
CART	4.91	67.11	337.35	21.06	4.68	2.12	1415.45
+boost	4.74	54.06	307.02	17.36	3.78	1.72	1273.76
+BoDi	4.92	53.76	311.68	17.39	3.98	1.74	1247.74
PRF	4.28	70.89	301.82	17.41	4.95	2.27	1460.54
+boost	3.96	60.13	272.66	15.35	3.46	1.72	1130.4
+BoDi	3.91	52.26	268.21	15.65	3.2	1.5	1058.55
RF	3.99	47.73	267.89	15.02	3.1	1.71	994.73
+boost	3.97	42.66	267.46	14.8	2.92	1.43	1027.18
+BoDi	4.04	37.98	273.86	15.22	2.9	1.36	1010.88





These experiments exhibit two situations:

1. if the dataset and the learner enable the generation of diversity, diversity boosting performs better than boosting and better than the base learners
2. if the dataset or the learner allow only limited diversity generation, the diversity boosting is very close to classical boosting

CRAN package Bodi

Bodi: Boosting Diversity in Regression Ensembles

A gradient boosting-based algorithm by incorporating a diversity term to guide the gradient boosting iterations, see Bourel, Cugliari, Goude, Poggi (2021) <

Version: 0.1.0
Imports: [mgcv](#), [ranger](#), [rpart](#), [gbm](#), [opera](#)
Published: 2022-03-23
Author: Yannig Goude  [aut, cre], Mathias Bourel  [aut], Jairo Cugliari  [aut], Jean-Michel Poggi  [aut]
Maintainer: Yannig Goude <yannig.goude at edf.fr>
License: [MIT](#) + file [LICENSE](#)
NeedsCompilation: no
Materials: [README](#)
CRAN checks: [Bodi results](#)

Documentation:

Reference manual: [Bodi.pdf](#)

Some perspectives

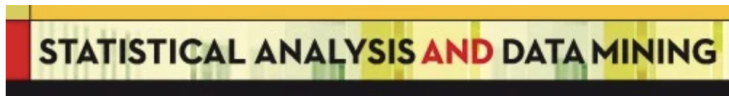
1. Which **base learners**?

- ▶ Enlarge the family: RF, stumps, GAM, ...
- ▶ Consider biased experts as base learners? (e.g. qGAM)

Some perspectives

1. Which **base learners**?
 - ▶ Enlarge the family: RF, stumps, GAM, ...
 - ▶ Consider biased experts as base learners? (e.g. qGAM)
2. In the **forecasting** context, explore **sequential aggregation strategies** to combine the sequence of boosting predictors
3. Explore **ensemble strategies** used
 - ▶ in **meteo** (different scenarios)
 - ▶ in **optimization** (different tuning parameters or initialization) to generate diversity

For more details, see the paper



RESEARCH ARTICLE

Boosting diversity in regression ensembles

Mathias Bourel, Jairo Cugliari , Yannig Goude, Jean-Michel Poggi

First published: 30 December 2023 | <https://doi.org/10.1002/sam.11654>

Thank you for your attention!

Questions?

Contact:

jean-michel.poggi@universite-paris-saclay.fr