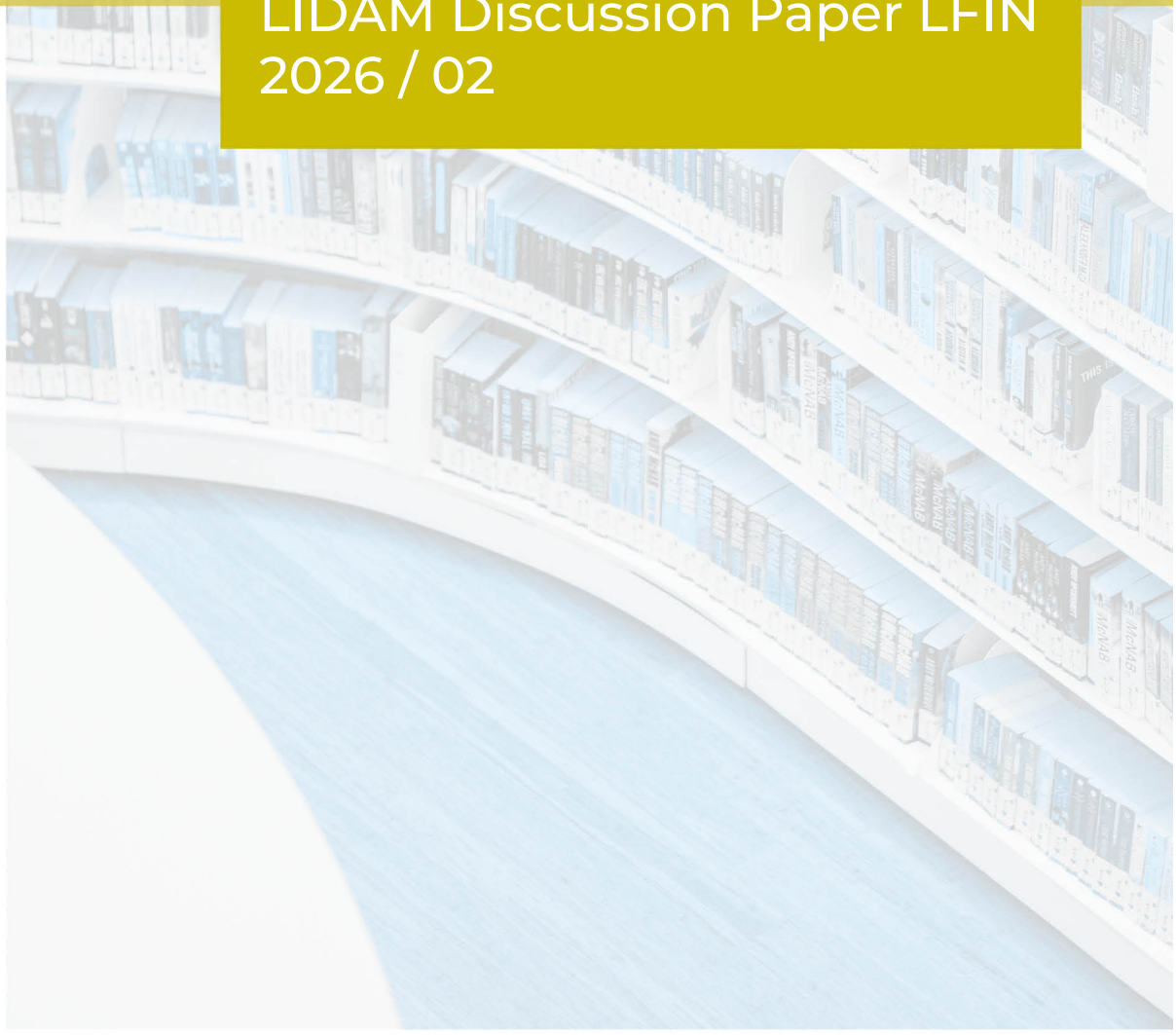
CLAGGING: AN EFFICIENT ALTERNATIVE TO BAGGING

Arnaud Germain, Frédéric Vrins

LIDAM Discussion Paper LFIN
2026 / 02



UCLouvain

LFIN | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

LFIN

Voie du Roman Pays 34, L1.03.01 B-1348 Louvain-la-Neuve

Tel (32 10) 47 43 04

Email: lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/lfin/publications

Clagging: An efficient alternative to bagging

Arnaud Germain, Frédéric Vrins

UCLouvain, Louvain Institute of Data Analysis in economics and statistics, LIDAM, LFIN

Voie du Roman Pays 34, 1348 Louvain-la-Neuve, Belgium

Abstract

We introduce a new forecast combination strategy called clagging (for *cluster aggregating*), which consists in combining models fitted on different clusters. First, we perform K clustering tasks of the same training set, increasing the number of clusters from 1 to K. Next, we fit a model on each of those $1 + 2 + \ldots + K$ clusters. Finally, the aggregate forecast for a new observation is obtained by combining the forecasts of the corresponding models using the distance of the new observation to the clusters' centroids. We perform an extensive horse race study where we benchmark clagging on 20 datasets using 7 prediction models, considering both regression and classification tasks. Our results suggest that clagging outperforms bagging, where a bootstrapped sample is traditionally created by drawing observations with replacement until the size of the bootstrapped sample coincides with the size of the original training set. Clagging also improve the performance compared to a standard fit on the whole training set.

1 Introduction

Forecast combination consists in relying on multiple models $\varphi_1, \ldots, \varphi_J$ – called *individual models* hereafter – and aggregating the vector of forecasts $\hat{\boldsymbol{y}}(X) := (\varphi_1(X), \ldots, \varphi_J(X))$ associated with a new observation X to form an aggregate forecast $\hat{y}(X) = f(\hat{\boldsymbol{y}}(X))$. Its main purpose is to reduce the variance of the forecast error $y(X) - \hat{y}(X)$ compared to that of $\hat{y}(X) = \varphi(X)$ of a single individual model φ fitted on the full training set $\mathcal{D}$. This approach has been shown to outperform individual models on a wide range of tasks, especially when the set of individual models φ_js exhibits some diversity (Kuncheva and Whitaker, 2003; Brown et al., 2005). Throughout, we use forecast in the broad out-of-sample sense, including continuous outcomes (regression) and binary events (classification).

Several routes can be considered to foster diversity among individual models. A first class of approaches is to rely on different predictive models (e.g., linear, non-linear, rule-based, tree-based, ...) fitted on the same training set $\mathcal{D}$ (Atiya, 2020; Roccazzella et al., 2022). Another class of approaches is to choose a model φ from a given family and fit the latter on various subsets $\mathcal{D}_1, \ldots, \mathcal{D}_J$ of $\mathcal{D}$. By doing so, all the φ_js belong to the same family of models (e.g., are decision trees), but have distinct configurations or parameters. The most popular example in this class is perhaps *bagging*, where the $\mathcal{D}_j$s are built using a bootstrapping procedure, that is, by drawing samples from $\mathcal{D}$, usually with replacement. It has been shown to work both theoretically and empirically (Breiman, 1996; Bauer and Kohavi, 1999; Bühlmann and Yu, 2002). In standard bagging, the size of each subset $\mathcal{D}_j$ coincides with that of the original training set: $\sharp\mathcal{D}_j = \sharp\mathcal{D}$. However, this choice is quite arbitrary and might be suboptimal out-of-sample. For instance, it is shown empirically that the optimal choice for the bootstrap size in terms of out-of-bag estimate of the generalization error is generally lower than the size of the original training set, i.e., $\sharp\mathcal{D}_j < \sharp\mathcal{D}$ (Martínez-Muñoz and Suárez, 2010). This is because smaller samples tend to increase the diversity of individual models. Enhancing diversity can also be achieved by forcing the training sets $\mathcal{D}_j$s to be disjoint. Empirical evidence suggests that a committee of $J > 1$ individual models fitted on random partitions of $\mathcal{D}$ can outperform a committee created using the same number and size of bootstrap aggregates (Chawla et al., 2001).

Interestingly, non-random partitioning is expected to foster diversity even further: Forming the disjoint sets using clustering, for instance, would generate $\mathcal{D}_j$s – hence φ_js – tending to be intrinsically different from each other. This is the approach followed in Ramchandran et al. (2021), and the aggregate forecast $\hat{y}(X)$ is a weighted average of the $\varphi_j(X)$ where the weights w_js are found by linear regression on a disjoint validation set. In the same vein, one could run several clustering tasks, thereby leading to a set of clusters $\mathcal{D}_j$s being not necessarily disjoint, and form the aggregate forecast $\hat{y}(X)$ by averaging only the forecasts associated with the models φ_js that are trained on clusters $\mathcal{D}_j$s including X. This is the approach adopted in Trivedi et al. (2015). These two approaches can be seen as "special combination" schemes. The first one features weights that do not depend on the new observation X, the second features binary weights, depending on whether X belongs to the cluster on which the corresponding individual model was trained. Clearly, these schemes are rather extreme, and none of them is expected to be optimal.

Our contribution is fourfold. First, we introduce a generalized version of Trivedi et al. (2015) called *clagging*, which consists of three steps: (i) Form a list of J – possibly overlapping– clusters by

running multiple clustering tasks, (ii) fit a forecasting model on each cluster, and (iii) aggregate the corresponding forecasts using some combination scheme. Second, we consider a new combination scheme particularly relevant for cluster-based partitioning, relying on distance-to-centroid. This can be viewed as a soft version of the hard assignment scheme used in Trivedi et al. (2015). Third, we extend the aforementioned cluster-based ensemble methods in order for clagging to accommodate categorical data. Fourth, we analyze the performance of our approach via an extensive horse race study on 20 datasets with 7 prediction models, considering both regression and classification tasks (that is, on 140 configurations). This study is thus substantially larger than what has been considered in previous papers. We use as benchmarks not only a standard model fitted on the whole training set, but also a bagging model that contains a similar number of individual models. We also apply clagging using various combination schemes, such as equally-weighted, inverse-variance, minimum-variance and distance-based. Our evidence suggests that relying on clustering to build non-random partitions of the training set boosts the performance when compared to the standard bagging procedure with a similar number of individual models. Moreover, we find that clagging can improve the performance of any prediction model compared to a standard fit on the whole training set. This is especially true when relying on the distance-based combination scheme that we introduce. In this setup, clagging displays better performance in 76% of configurations across all datasets. This is a striking result given the remarkable performance of bagging, which remains one of the most widely used combination models.

The paper is organized as follows. In Section 2, we review the existing approaches that exploit clustering for prediction tasks and then specifically in the context of forecast combination, namely Ramchandran et al. (2021) and Trivedi et al. (2015), and compare how they differ from ours. Next, in Section 3, we introduce clagging and detail how the clusters are formed, how the individual models are fitted and how the forecasts are aggregated. Section 4 presents the empirical setup and gives an overview of the different datasets, prediction models and performance metrics considered. Finally, Section 5 presents and discusses the empirical results.

2 Literature review

In this section, we first review how clustering has been used for prediction tasks and then specifically to form several subsets $\mathcal{D}_j$ in the context of forecast combination.

Relying on clustering for prediction tasks has been investigated when the data can be naturally partitioned. Bouwmeester et al. (2013) propose prediction models for patients treated by different

anesthesiologists. They compare random effect and standard logistic regression models. Adding the cluster effect in the prediction model increases the amount of predictive information, resulting in improved overall discriminative ability and calibration-in-the-large within clusters. Deodhar and Ghosh (2010) develop a regime in which clustering and model fitting are performed simultaneously to iteratively improve both cluster assignment and fit of the models when the data can be naturally partitioned into two groups. They focus on generalized linear models as individual models. In this work, we remove these requirements by allowing the clustering algorithm to find any substructure in the data. Van den Berg et al. (2008) form optimal country clusters instead of using a traditional panel dataset in the context of financial crisis prediction. By doing so, they avoid the implicit assumption that crises are homogeneously caused by identical factors by fitting a different model per cluster. They consider a data-driven segmentation using Hausman's test, geographical clusters and a global model and find that data-driven segmentation has the best performance considering in-sample goodness-of-fit indicators.

The idea of using clustering in the specific context of forecast combination has been previously investigated by Ramchandran et al. (2021). The authors build the $\mathcal{D}_j$s by clustering $\mathcal{D}$ in the specific context of Random Forests in regression tasks. They fix a number J of (disjoint) clusters $\mathcal{D}_j$, fit a forest on each of them, evaluate the J corresponding forecasts at a new observation X to get $\hat{y}_j(X) := \varphi_j(X)$, and aggregate the forecasts using a linear combination scheme:

$$\hat{y}(X) = \sum_{j=1}^{J} w_j \hat{y}_j(X) = \boldsymbol{w}^T \hat{\boldsymbol{y}}(X) \quad \text{where} \quad \hat{\boldsymbol{y}}(X) := (\hat{y}_1(X), \dots, \hat{y}_J(X)) \ , \quad \boldsymbol{w} \in \mathbb{R}^J \ .$$

The vector of combination weights $\boldsymbol{w}$ is found using stacked regression on a validation set that is disjoint from the training and the test sets. As a consequence, $\boldsymbol{w}$ is the same vector for any inputs X in the test set. A different approach is considered in Trivedi et al. (2015), where the authors compute $\hat{y}(X)$ by considering multiple clustering tasks. First, they set a number K of layers. For each layer $k = 1, \dots, K$, the training set $\mathcal{D}$ is partitioned into k disjoint clusters, $\mathcal{D}_1^k, \dots, \mathcal{D}_k^k$, and a model φ is fitted on each of them, leading to k individual models $\varphi_1^k, \dots, \varphi_k^k$. Next, for each layer k, they identify the model $\varphi_{i(X)}^k$ associated with the cluster $\mathcal{D}_{i(X)}^k$ which the new observation X belongs to. Finally, the aggregate forecast is given by the simple average of the local forecasts $\hat{y}_j^k(X) := \varphi_j^k(X)$ across the layers:

$$\hat{y}(X) = \sum_{k=1}^{K} \sum_{i=1}^{k} \frac{\mathbb{1}_{\{X \in \mathcal{D}_i^k\}}}{K} \hat{y}_i^k(X) = \frac{1}{K} \sum_{k=1}^{K} \hat{y}_{i(X)}^k(X) \ .$$

Note that the first layer ($k = 1$) coincides with the standard fit approach where a single model φ is trained on the whole training set $\mathcal{D}$. In contrast with Chawla et al. (2001), observe that the clusters $\mathcal{D}^1_{i(X)}, \ldots, \mathcal{D}^K_{i(X)}$ involved in the combination are related to different clustering tasks (layers) but to the same test point X, hence, do overlap. An important difference with Ramchandran et al. (2021) is that the combination weights used to form the aggregate forecast $\hat{y}(X)$ depend on X. The authors provide theoretical insights about why clustering could be a useful step prior to prediction. The reason for averaging over K layers is to avoid relying on an arbitrary number k of clusters, thereby reducing the sensitivity of the results with respect to K.

3 Clagging

Clagging, the method we introduce, is a compromise between the last two clustering-based approaches detailed in Section 2. It adopts the multi-layer approach proposed in Trivedi et al. (2015) leading to $J = 1 + 2 + \ldots + K = K(K + 1)/2$ clusters but, as in Ramchandran et al. (2021), all the J models intervene when forming the aggregate forecast to any new observation X. It proceeds in 3 steps: forming the clusters $\mathcal{D}_j$, training a model on each of them, and finally aggregating the forecasts.

3.1 Forming the J clusters

The first step of clagging is similar to Trivedi et al. (2015): It consists in setting the number K of layers and, for each layer $k = 1, \ldots, K$, in applying a clustering algorithm splitting the training set $\mathcal{D}$ into k clusters. By doing so, one obtains K partitions of $\mathcal{D}$ made of an increasing number of sets $\{\mathcal{D}^k_1, \ldots, \mathcal{D}^k_k\}$, from $k = 1$ to $k = K$. We denote by $\mathcal{S}$ the resulting list of the $J := K(K+1)/2$ subsets of $\mathcal{D}$. In this paper, we focus on k-means, which has become one of the most popular clustering algorithms due to its computational speed, convergence, ease of interpretation and scaling ability. Moreover, it can create clusters of different shapes and sizes. K-means is designed for numerical data. There are extensions of k-means to deal with other types of data. For instance, k-modes (Huang and Ng, 1999) is designed for categorical data. Similarly, k-prototypes can handle datasets with both numerical and categorical data. Ramchandran et al. (2021) find that the choice of the clustering algorithm does not seem to have a significant impact on the results.

3.2 Training the J models

The second step consists in fitting a forecasting model φ_i^k on each subset $\mathcal{D}_i^k \in \mathcal{S}$. By doing so, we have for each new observation X a set of J forecasts $\hat{y}_1^1(X), \hat{y}_1^2(X), \hat{y}_2^2(X), \hat{y}_1^3(X), ..., \hat{y}_K^K(X)$ that we need to combine. For clarity, we re-order these J models according to the index mapping $(i,k) \to j(i,k) \coloneqq k(k-1)/2 + i$ for $k = 1, \ldots, K$ and $i = 1, \ldots, k$, such that $\mathcal{D}_i^k = \mathcal{D}_{j(i,k)}$ and $\varphi_{j(i,k)}(\cdot) = \varphi_i^k(\cdot) = \hat{y}_i^k(\cdot) = \hat{y}_{j(i,k)}(\cdot)$.

3.3 Aggregating the J forecasts

The third step consists in aggregating the forecasts $\hat{\boldsymbol{y}}(X)$ computed above into a single response, $\hat{y}(X)$. Although non-linear schemes could be considered, too, we restrict ourselves to discussing linear combinations:

$$\hat{y}(X) = \sum_{k=1}^{K} \sum_{i=1}^{k} w_i^k(X) \hat{y}_i^k(X) \quad \text{or, using our new coordinates,} \quad \hat{y}(X) = \boldsymbol{w}^T(X) \hat{\boldsymbol{y}}(X) \ , \quad \boldsymbol{w}(X) \in \mathbb{R}^J.$$

The two cluster-based methods introduced earlier are special cases of this setup in the sense that Ramchandran et al. (2021) considers a fixed $\boldsymbol{w}$ that does not depend on the test point X, while Trivedi et al. (2015) considers as vector of weights $\boldsymbol{w}^{TPH}(X)$ defined as $w_j^{TPH}(X) \coloneqq \mathbb{1}_{\{X \in \mathcal{D}_j\}}/K$.

Obviously, several natural alternatives can be considered, such as the equally-weighted strategy, $\boldsymbol{w}^{EW}$ where $w_j^{EW} \coloneqq 1/J$. Another possibility is to weight the local outputs according to the inverse of the forecast error's standard deviation as suggested in Bates and Granger (1969) or in Samuels and Sekkel (2017), i.e., to choose $\boldsymbol{w}^{\sigma}$ defined as $w_j^{\sigma} \propto 1/\sigma_j$ where σ_j^2 is the variance of the forecast error $v_j \coloneqq y - \hat{y}_j$ computed within each of the J clusters. Another possibility is to adopt the minimum-variance strategy. For instance, Granger and Ramanathan (1984) showed that minimizing $\mathbb{E}\|y - \boldsymbol{w}^T \hat{\boldsymbol{y}}\|_2^2$ under the constraint $\boldsymbol{w}^T \mathbf{1} = 1$ is equivalent to setting:

$$\boldsymbol{w}^{MV} \coloneqq \underset{\boldsymbol{w}}{\operatorname{argmin}} \ \boldsymbol{w}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{w}, \tag{1}$$

where $\hat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix of models' forecast errors v_j. Clearly, $\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}$ is the variance of $v \coloneqq y - \hat{y}$ when the $\hat{y}_j$s are unbiased estimators of y.[1] It has been shown in Roccazzella et al. (2022) that choosing a robust estimator of the covariance matrix can outperform the robust

[1] This is naturally defined for the regression context where $y, \hat{y} \in \mathbb{R}$. In the binary classification context, we have $y \in \{0, 1\}$ and $\hat{y} \in [0, 1]$. To compute some performance metrics, one will need to assume a threshold to transform $\hat{y}$ into binary forecasts (see Section 4.2). We define $v_j \coloneqq y - \hat{y}_j$ in the binary classification context without proving that the proposition of Granger and Ramanathan (1984) still holds. Yet, the empirical performance of this combination scheme in the binary classification context is similar to the regression context as displayed in Section 5.

equally-weighted scheme. This yields a convex combination of the individual forecasts and ensures that the weight vector is defined on the unit simplex, which is a desirable property from a bias-variance perspective (Jaeger and Shawe-Taylor, 2005; Gönen and Alpaydın, 2011).

Intuitively, a relevant choice for clagging would be to set the weights according to the "distance" $d(X, \mathcal{D}_j)$ between X and each of the J clusters, that is, to choose $\boldsymbol{w}^d(X)$ where $w_j^d(X) \propto 1/d(X, \mathcal{D}_j)$. More specifically, denoting by p the number of explanatory variables in the dataset, the aggregate forecast associated with a new observation X is given by the combination scheme defined as:

$$w_j^d(X) := \frac{\frac{1}{d(X,\mathcal{D}_j)}^{p-1}}{\sum_{j=1}^{J} \frac{1}{d(X,\mathcal{D}_j)}^{p-1}}. \tag{2}$$

Observe that Trivedi et al. (2015) can be intuitively regarded as a special case associated with a "membership distance".

In what follows, we refer to these schemes as EW (equally-weighted), INV (inverse-variance), MV (minimum-variance with the sample covariance estimator), MVLW (minimum-variance with the Ledoit–Wolf shrinkage estimator), TPH (the hard-assignment scheme of Trivedi et al., 2015) and DIST (our soft assignment scheme using distance-based weights).

4 Empirical Setup

4.1 Datasets and algorithms

We consider 20 different datasets, 10 for binary classification tasks and 10 for regression tasks that come from the UCI repository (Asuncion and Newman, 2007), displayed in Table 1. We rely on 7 different prediction models for each task, displayed in Table 2. This leads to a total of 140 configurations: 70 for regression and 70 for classification. To avoid any cherry-picking, we fix the hyperparameters of each model for all the datasets *a priori* using default values from the `R` package `Caret` and report those in Appendix A for reproducibility purposes. For the same reasons, we do a very light data cleaning and preprocessing for each dataset, which is detailed in Appendix B. For each of the 20 datasets and using the same random seed, we allocate 75% of all observations to a training set to train individual models. If the minimum-variance combination scheme is considered, we split the 25% into two parts: 10% is used to estimate the sample covariance matrix, and the remaining 15% is used as testing set. In this case, we consider two estimators for the covariance matrix: the sample estimator and the shrinkage estimator introduced in Ledoit and Wolf (2004),

both with a non-negativity constraint. In all the other cases, the remaining 25% is used as testing set. For bagging and clagging, we consider all the schemes discussed in Section 3.3, namely: equally-weighted, inverse-variance and the minimum-variance schemes. For clagging, we also consider two cluster-based schemes: the hard assignment scheme of Trivedi et al. (2015) and our soft assignment with a distance-based approach.

ID	Dataset	# variables	# observations	Type of variables	Target
1	Spambase	58	4,601	Numerical	Binary
2	Musk	167	6,598	Numerical	Binary
3	Ringnorm	21	7,376	Numerical	Binary
4	Twonorm	21	7,400	Numerical	Binary
5	HTRU2	9	17,898	Numerical	Binary
6	MAGIC	11	19,020	Numerical	Binary
7	Online Shoppers	18	12,330	Mixed	Binary
8	Credit card	24	30,000	Mixed	Binary
9	Adult	15	30,162	Mixed	Binary
10	Bank Marketing	16	45,211	Mixed	Binary
1	Communities&Crime	100	1,994	Numerical	Continuous
2	Parkinsons	17	5,875	Numerical	Continuous
3	Extreme Weather	22	7,752	Numerical	Continuous
4	Electrical Grid	13	10,000	Numerical	Continuous
5	Appliances Energy	28	19,735	Numerical	Continuous
6	Superconductivity	82	21,263	Numerical	Continuous
7	Protein	10	45,730	Numerical	Continuous
8	Abalone	9	4,177	Mixed	Continuous
9	Seoul	12	8,465	Mixed	Continuous
10	Bike Sharing	13	17,379	Mixed	Continuous

Table 1: Description of the 20 UCI datasets used. *Mixed* type of variables refers to a dataset containing both numerical and categorical features. *Binary* target indicates classification task and *continuous* target indicates regression task.

ID	Model	`Caret` name	Task
a	Decision Tree (CART)	`rpart`	Classification and Regression
b	Partial Least Squares	`pls`	Classification and Regression
c	Boosted Generalized Linear Model	`glmboost`	Classification and Regression
d	Extreme Gradient Boosting	`xgbDART`	Classification and Regression
e	Multivariate Adaptive Regression Spline	`earth`	Classification and Regression
f	Penalized Logistic Regression	`multinom`	Classification
g	Boosted Logistic Regression	`LogitBoost`	Classification
h	Linear Regression	`lm`	Regression
i	Least Angle Regression	`lars`	Regression

Table 2: List of the prediction models used depending on the task. Models are fitted using `R` package `Caret`.

We aim for J to be on the order of 100. This choice follows common practice in forecast combination (Breiman, 1996; Hastie et al., 2009; Pedregosa et al., 2011). Therefore, we set $K = 14$ for clagging and consider 100 bags for bagging. For bagging, we consider the most widely used

setup in the literature: We create bootstrapped samples by selecting observations from the original training data set with replacement so that the size of the bootstrapped samples coincide with the size of the original training set.

4.2 Performance metrics

The performance metrics considered are obviously different whether the target is binary or continuous. When the target is continuous, we consider the Mean Squared Error (MSE), the Mean Absolute Error (MAE) and the out-of-sample (OOS) R^2 (see Hawinkel et al. (2024)) as performance metrics. To assess the significance of the results, we consider a paired t-test for MSE as well as for MAE. When the target is binary, we consider the Accuracy, the AUC and the H-measure. The H-measure improves on the AUC by replacing its arbitrary, problem-dependent weighting of misclassification severities with a coherent, explicitly defined distribution of misclassification costs, making performance comparisons more consistent and interpretable (Hand, 2009). To assess the significance of the results, we consider a McNemar test for the sensitivity and a DeLong test (DeLong et al., 1988) as recommended in Rainio et al. (2024). For each performance metric, we also define the Median Gain (MG) as the median change in percentage of the performance metric from a competitor to a reference model across all configurations. Similarly, we define respectively the Median Conditional Gain (MCG) and the Median Conditional Loss (MCL) as the median change percentage of the performance metric from a competitor to a reference model across all configurations where the change is positive and all configurations where the change is negative.

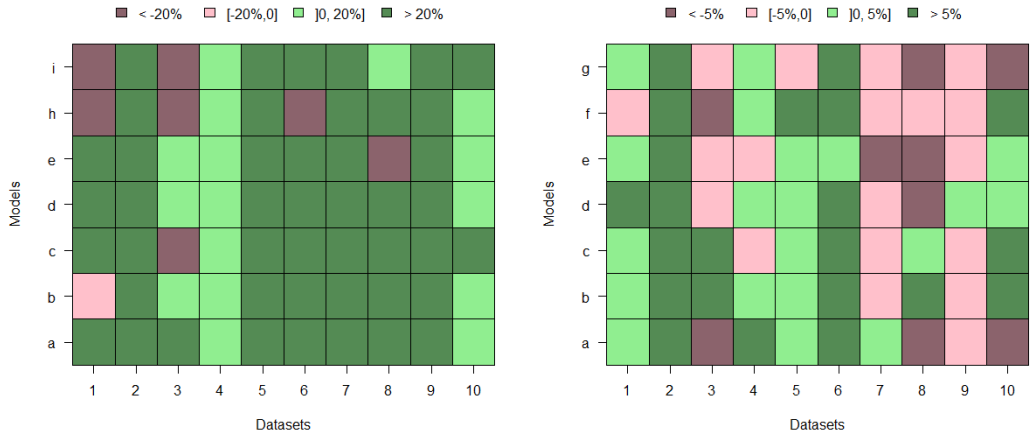
5 Results and Discussion

In this section, we assess the empirical performance of clagging. We first compare our proposed distance-based aggregation to the hard-assignment strategy of Trivedi et al. (2015) as well as to other natural alternatives. Next, we benchmark clagging against a standard fit on the full training set and against bagging featuring ensembles of similar sizes. Finally, we contrast clagging and bagging under equally-weighted averaging to disentangle the effect of the combination scheme from that of the fitting of the individual models. Results are reported across 20 datasets and multiple prediction models using the performance metrics and tests described in Section 4.2.

5.1 Impact of the combination scheme in clagging

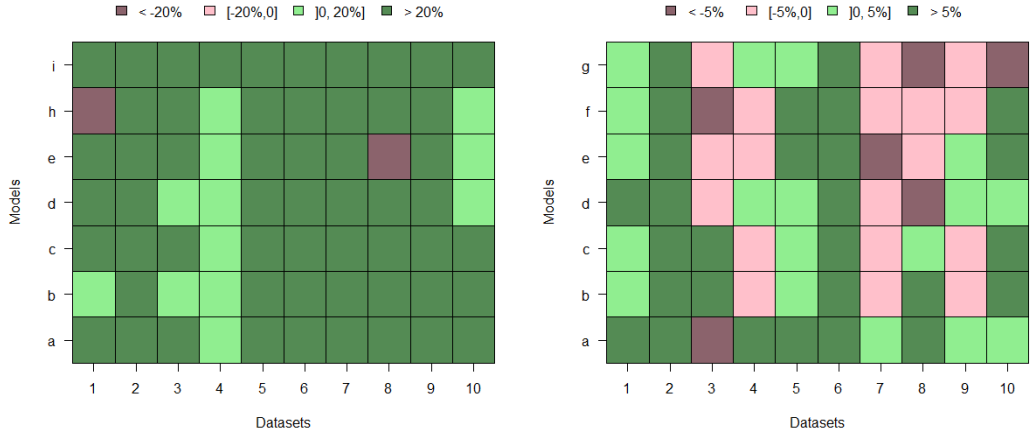
The first question we address is whether the information that a new observation might belong to a cluster should be regarded when cluster aggregating is considered. Trivedi et al. (2015) proposed a hard assignment combination scheme to take into account this information. We propose instead in Equation 2 a soft assignment using a distance-based combination scheme. Nevertheless, one could also disregard this information and use combination schemes such as the equally-weighted, the inverse-variance and the minimum-variance ones. In Figures 1 and 2, we compare clagging with various combination schemes (EW, INV, TPH, MV and MVLW) to clagging with the distance-based combination scheme both in regression tasks (left panels, OOS R^2 metric) and classification (right panels, H-measure metric) tasks. For each of the 20 datasets and for each of the 7 prediction models (which may differ, depending on the task), the panels report the relative change in the metric when switching from the chosen combination scheme to the distance-based combination scheme.

One can see from the left panels of Figure 1 that, in regression tasks, there is an increase in OOS R^2 in more than 88%, 97% and 68% of configurations by choosing our distance-based combination scheme for clagging instead of the equally-weighted scheme, the inverse-variance scheme and the scheme proposed by Trivedi et al. (2015). The right panels show that the H-measure is increased in 60%, 65% and 68% of configurations in classification tasks. Similarly in Figure 2, the left panels display that, in regression tasks, there is an increase in OOS R^2 in more than 97% and 98% of configurations by relying on our distance-based combination scheme for clagging instead of the minimum-variance scheme with sample estimator and with Ledoit and Wolf (2004) estimator, respectively. The right panels display that there is an increase in H-measure in more than 62% and 64% of configurations in classification tasks. For the sake of conciseness, we report in Appendix C the results for other performance metrics (MAE, MSE, AUC and accuracy). The conclusions remain similar and all point to the superiority of our distance-based combination scheme.



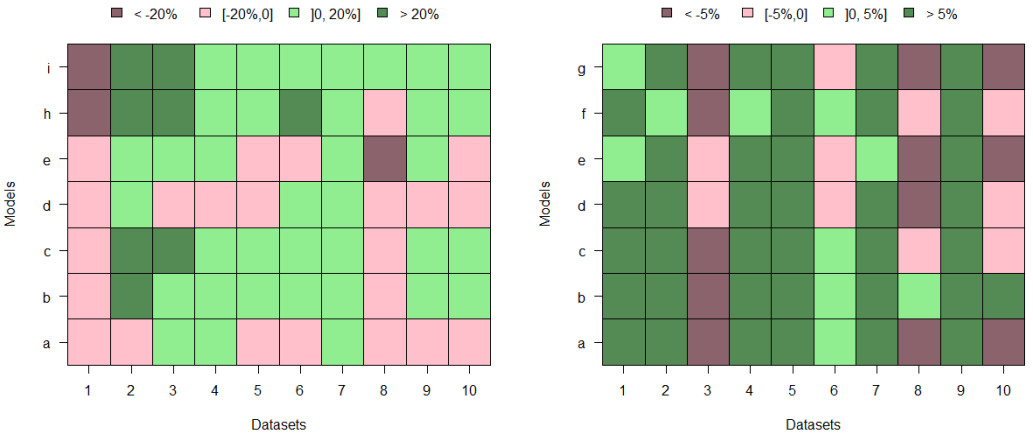
(a) OOS R^2 - EW

(b) H-measure - EW



(c) OOS R^2 - INV

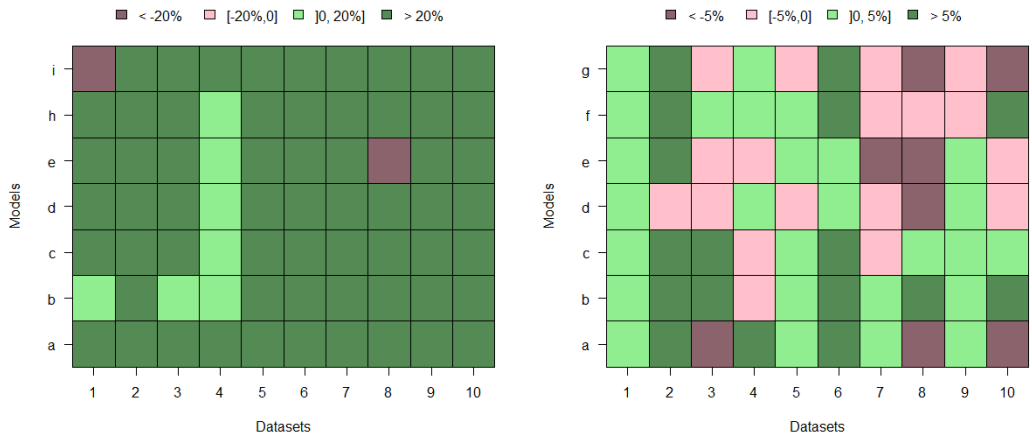
(d) H-measure - INV



(e) OOS R^2 - TPH

(f) H-measure - TPH

Figure 1: Change in percentage of performance metric using clagging when switching from a benchmark combination scheme (indicated in the captions) to our distance-based combination scheme. The panels report the OOS R^2 in regression tasks (left) and the H-measure in classification tasks (right). The considered benchmark schemes are the equally-weighted (top), the inverse-variance (middle) and the hard assignment scheme of Trivedi et al. (2015) (bottom).



(a) OOS R^2 - MV

(b) H-measure - MV

(c) OOS R^2 - MVLW

(d) H-measure - MVLW

Figure 2: Change in percentage of performance metric using clagging when switching from a benchmark combination scheme (indicated in the captions) to our distance-based combination scheme. The panels report the OOS R^2 in regression tasks (left) and the H-measure in classification tasks (right). The considered benchmark schemes are the minimum-variance scheme with sample estimator (top), and the minimum-variance scheme with Ledoit and Wolf (2004) estimator (bottom).

To assess the significance of the performance gains, we rely on the statistical tests described in Section 4.2. Table 3 suggests that the distance-based combination scheme with clagging outperforms all other benchmarks both in regression and in classification. For instance, one can see from the last row of Table 3 that, compared to the hard assignment of Trivedi et al. (2015), clagging with our distance-based scheme is significantly better in 40 out of the 70 configurations, and is significantly worse only in 16 configurations. Both schemes lead to comparable results, statistically speaking, in 14 configurations. Looking at every row, we observe that we often win and rarely lose by choosing distance-based scheme instead of any other scheme.

On the one hand, in classification, the highest Median Gain from the distance-based combination

scheme is with the hard assignment of Trivedi et al. (2015): 1.29%. On the other hand, it is also the highest Median Conditional Loss compared to any other scheme. This might indicate a high volatility in the results by using Trivedi et al. (2015) scheme and therefore a potential lack of robustness of this scheme. In regression, we observe a Median Gain from the distance-based combination scheme compared with the minimum-variance scheme with sample estimator of 89.9%, meaning that we are close to halving the MSE.

For the sake of conciseness, we report in Appendix C the results for other performance metrics and tests (MAE using t-test and sensitivity using McNemar test). The conclusions remain similar and all point to the superiority of our distance-based combination scheme.

	Reference Model: Clagging (DIST)					
$\mathcal{M}$	RM $>$ $\mathcal{M}$	RM $\approx$ $\mathcal{M}$	RM $<$ $\mathcal{M}$	MG	MCG	MCL
Clagging (EW)	65	4	1	58.57%	60.03%	-30.87%
Clagging (MV)	63	6	1	89.9%	109.79%	-98.29%
Clagging (MVLW)	63	6	1	90.13%	104.46%	-97.36%
Clagging (INV)	60	10	0	78.47%	85.28%	-99.6%
Clagging (TPH)	27	37	6	1.39%	9.01%	-4.52%
Clagging (EW)	33	24	13	0.45%	1.73%	-0.99%
Clagging (MV)	27	32	11	0.26%	1.01%	-0.38%
Clagging (MVLW)	26	33	11	0.22%	0.97%	-0.38%
Clagging (INV)	35	25	10	0.79%	2.12%	-0.61%
Clagging (TPH)	40	14	16	1.31%	3.13%	-1.45%

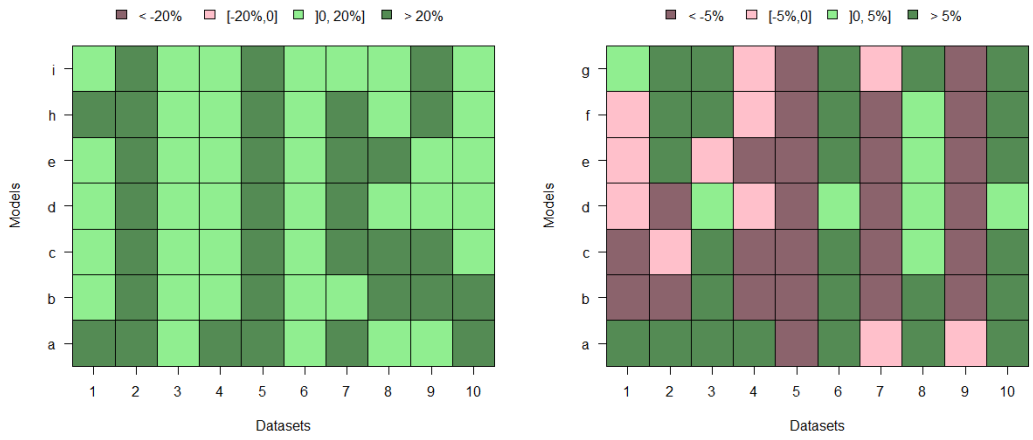
Table 3: Comparison between clagging with distance-based combination scheme and each forecast combination strategy. Columns 1 to 3 respectively report the number of configurations where the reference model is significantly better than, has no significant difference with, and is significantly worse than a baseline model $\mathcal{M}$. Columns 4 to 6 respectively report the Median Gain, the Median Conditional Gain and the Median Conditional Loss when switching from the models mentioned in the left column to our distance-based clagging. Top half is related to regression and considers MSE and t-test whereas bottom half is related to classification and considers AUC and DeLong test, both at 95% confidence level.

We conclude from this section that the "cluster information" conveyed by a new observation should not be disregarded when cluster aggregating is considered. Moreover, our soft assignment scheme provides better results than the hard assignment one of Trivedi et al. (2015), suggesting that the distance conveys more information than the membership variable.

5.2 Comparison with standard fit

Having concluded from the previous section that the best combination to consider with cluster aggregating is the distance-based scheme, we take the latter as reference to assess the performance of alternative methods. The first natural competitor is obviously a standard fit on the whole dataset. Note that clagging always considers a standard fit on the whole dataset as one of its

individual models; a standard fit could be seen in clagging as a case where $K = 1$. In other words, the question we address in this section is the following: Is there an interest in combining $K > 1$ models in clagging? Trivedi et al. (2015) provides empirical evidence that clagging with their hard assignment performs better than a standard fit. However, they only analyzed regression tasks and only for 33 configurations (3 models and 11 datasets). In this section, we extend the analysis to more configurations (140) and classification tasks. For each of the 20 datasets and for each of the 7 prediction models (appropriate for each task), Figure 3 displays the change in percentage of a performance metric from a standard fit to clagging with the hard assignment scheme of Trivedi et al. (2015). Panel a suggests an increase in OOS R^2 in every configuration by relying on the hard assignment combination scheme for clagging instead of a standard fit on the whole training set considering regression tasks. This extends the results observed by Trivedi et al. (2015). However, panel b displays that there is an increase in H-measure only in 49% of configurations by relying on the hard assignment scheme for clagging instead of a standard fit considering classification tasks.

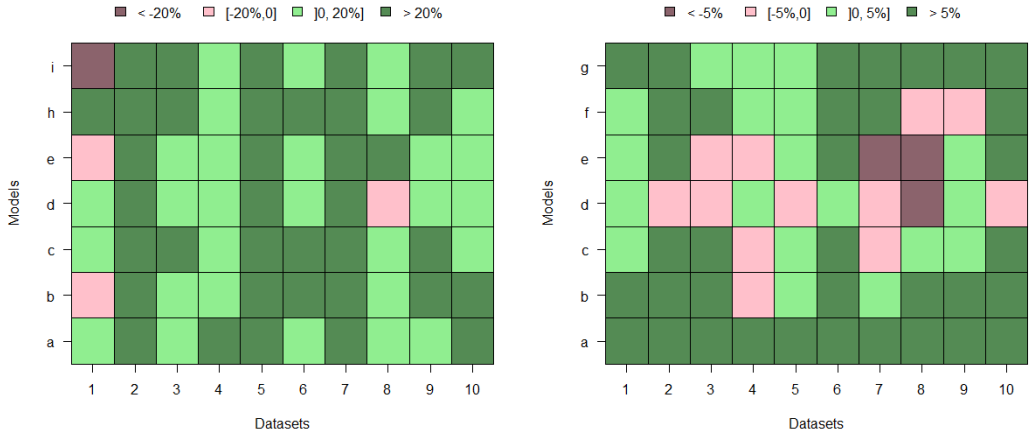


(a) OOS R^2 (b) H-measure

Figure 3: Change in percentage of a performance metric from a standard fit to clagging with Trivedi et al. (2015) combination scheme. The panels report the OOS R^2 in regression tasks (left) and the H-measure in classification tasks (right).

Let us now present a similar analysis for our distance-based scheme. For each of the 20 datasets and for each of the 7 prediction models (depending on the task), Figure 4 displays the change in percentage of a performance metric from a standard fit to clagging with our distance-based scheme. Panel a suggests an increase in OOS R^2 in more than 94% of configurations by relying on our distance-based combination scheme for clagging instead of a standard fit on the whole training set considering a regression task. Similarly, panel b displays that there is an increase in H-measure in more than 78% of configurations by relying on our distance-based combination scheme for clagging

instead of a standard fit considering a classification task. Our combination scheme provides better results both for classification and regression whereas the hard assignment scheme only provides better results in regression tasks. For the sake of conciseness, we report in Appendix D the results for other performance metrics (MAE, MSE, AUC and accuracy) and all conclusions remain similar.



(a) OOS R^2 (b) H-measure

Figure 4: Change in percentage of a performance metric from a standard fit to clagging with our distance-based combination scheme. The panels report the OOS R^2 in regression tasks (left) and the H-measure in classification tasks (right).

One can assess the significance of those results using statistical tests described in Section 4.2. Table 4 confirms that the distance-based combination scheme with clagging outperforms a standard fit on the whole training set both in regression and in classification. For instance, one can see from the last row of Table 4 that, compared to clagging with our distance-based combination scheme, a standard fit is significantly better in only 5 out of the 70 configurations, and is significantly worse in 40 configurations, considering classification tasks. Results are broadly similar for regression tasks. By looking at the second row of the table, one can observe a Median Gain of 21.11% in MSE by using clagging with distance-based combination scheme instead of a standard fit. Table 4 also confirms that the hard assignment scheme performs better than a standard fit in regression but not in classification. One can see from the third row of Table 4 that, compared to clagging with a hard assignment scheme, a standard fit is significantly better in only 31 out of the 70 configurations, and is significantly worse only in 30 configurations. This leads to a Median Gain very close to 0.

For the sake of conciseness, we report in Appendix D the results for other performance metrics and tests (MAE using t-test and sensitivity using McNemar test) and all conclusions remain similar: Clagging can outperform a standard fit on the whole training set if forecasts are aggregated using a soft assignment scheme. A hard assignment scheme will only provide better results in regression

but not in classification.

	Reference Model: standard fit					
$\mathcal{M}$	RM $>\mathcal{M}$	RM $\approx\mathcal{M}$	RM $<\mathcal{M}$	MG	MCG	MCL
Clagging (TPH)	1	10	59	-20.35%	2638.9%	-20.74%
Clagging (DIST)	2	14	54	-21.11%	1168.78%	-22.72%
Clagging (TPH)	31	9	30	-0.02%	4.3%	-3.46%
Clagging (DIST)	5	25	40	-1.19%	0.31%	-2.3%

Table 4: Comparison between a standard fit on the whole training set and clagging with both hard and soft assignment. Columns 1 to 3 respectively report the number of configurations where the reference model is significantly better than, has no significant difference with, and is significantly worse than a baseline model $\mathcal{M}$. Columns 4 to 6 respectively report the Median Gain, the Median Conditional Gain and the Median Conditional Loss. Top half is related to regression and considers MSE and t-test whereas bottom half is related to classification and considers AUC and DeLong test, both at 95% confidence level.

The above results confirm that clagging outperforms standard fit, and that opting for the soft assignment scheme delivers better results than the hard assignment one. In regression, both TPH and DIST significantly outperform standard fit in most of the cases, in 59 and 54 configurations out of 70, respectively. In classification, however, the comparison is more favorable to DIST. Indeed, TPH performs significantly better than standard fit in 30 configurations, but performs worse than standard fit in 31 configurations, suggesting mixed results. DIST behaves much better in this regard, since standard fit outperforms significantly DIST in only 5 configurations.

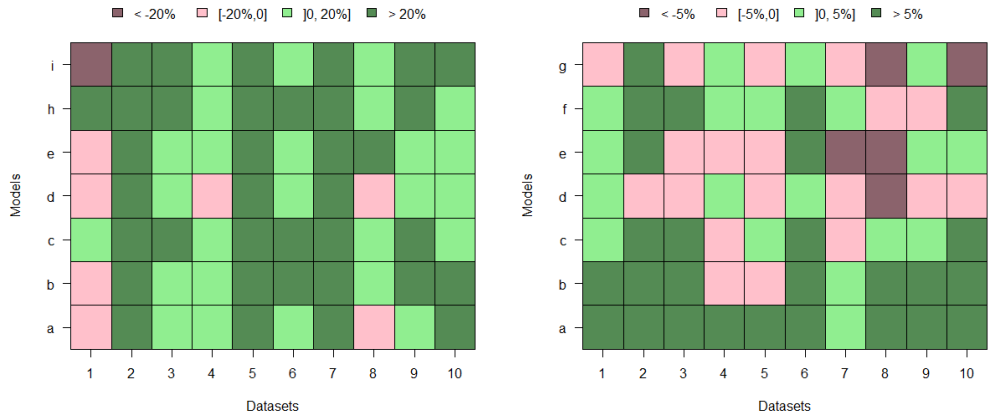
5.3 Comparison with bagging

The above results highlight that clagging can outperform a standard fit on the whole training set. However, one might argue that, in order for clagging to be considered as a competitor to state-of-the-art prediction models, it should be compared to other forecast combination techniques, and not merely to a global fit on the whole training set, as in Trivedi et al. (2015). The most natural competitor is perhaps bagging, due to the similarities between the two techniques. We therefore compare in this section the performance of clagging and bagging. We consider three combination schemes for bagging: equally-weighted strategy, inverse-variance strategy and minimum-variance strategy. For each of the 20 datasets and for each of the 7 prediction models (depending on the task), the panels report the relative change in the metric when switching from bagging with the chosen combination scheme to clagging with the distance-based combination scheme.

One can see from the left panels of Figure 5 that, in regression tasks, there is an increase in OOS R^2 in more than 88% and 90% of configurations by relying on our distance-based combination scheme for clagging instead of bagging with equally-weighted scheme and minimum-variance scheme

with sample estimator, considering a regression task. The right panels show that the H-measure is increased in more than 64% and 65% of configurations in classification tasks. Similarly in Figure 6, the left panels display that, in regression tasks, there is an increase in OOS R^2 in more than 88% and 90% of configurations by relying on our distance-based combination scheme for clagging instead of bagging with inverse-variance scheme and minimum-variance scheme with Ledoit and Wolf (2004) estimator, respectively. The right panels display that there is an increase in H-measure in more than 65% and 66% of configurations in classification tasks.

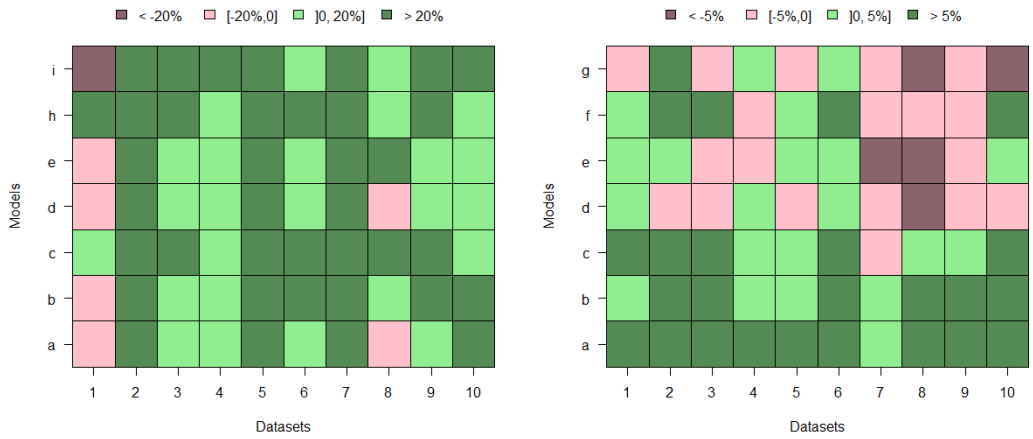
This evidence suggests that clagging has a superior performance compared to bagging when combined with our distance-based combination scheme. For the sake of conciseness, we report in Appendix E the results for the other performance metrics (MAE, MSE, AUC and accuracy). They remain similar and all point to the superiority of clagging over bagging.



(a) OOS R^2 - EW (b) H-measure - EW

(c) OOS R^2 - INV (d) H-measure - INV

Figure 5: Change in percentage of performance metric when switching from bagging with a benchmark combination scheme to clagging with our distance-based combination scheme. The panels report the OOS R^2 in regression tasks (left) and the H-measure in classification tasks (right). The considered benchmark scheme is the equally-weighted scheme (top), and the inverse-variance scheme (bottom).



(a) OOS R^2 - MV

(b) H-measure - MV

(c) OOS R^2 - MVLW

(d) H-measure - MVLW

Figure 6: Change in percentage of performance metric when switching from bagging with a benchmark combination scheme to clagging with our distance-based combination scheme. The panels report the OOS R^2 in regression tasks (left) and the H-measure in classification tasks (right). The considered benchmark schemes are the minimum-variance scheme with sample estimator (top), and the minimum-variance scheme with Ledoit and Wolf (2004) estimator (bottom).

As previously explained, we assess the significance of those results using statistical tests described in Section 4.2. Table 5 confirms that the distance-based combination scheme with clagging outperforms bagging whatever the combination scheme considered both in regression and in classification. For instance, one can see from the first row of Table 5 that, compared to bagging with equally-weighted scheme, clagging with our distance-based scheme is significantly better in 51 out of the 70 configurations, and is significantly worse only in 4 configurations. Both schemes lead to comparable results, statistically speaking, in 15 configurations. We observe a Median Gain in MSE around 25% for every scheme. Looking at top 4 rows, we observe that, in regression, we often win and very rarely lose by using clagging with distance-based scheme instead of bagging, whatever

combination scheme is considered.

Regarding classification, one can see from the last row of Table 5 that, compared to bagging with equally-weighted scheme, clagging with our distance-based scheme is significantly better in 31 out of the 70 configurations, and is significantly worse only in 10 configurations, with a Median Gain in AUC of 0.41%. Results are comparable for the other schemes. We continue to observe that, in classification, we often win and rarely lose by using clagging with distance-based scheme instead of bagging, whatever the combination scheme considered.

For the sake of conciseness, we report in Appendix E the results for other performance metrics and tests (MAE using t-test and sensitivity using McNemar test). The conclusions remain similar and all point to the superiority of clagging with our distance-based combination scheme compared to bagging.

	Reference Model: Clagging (DIST)					
$\mathcal{M}$	RM $>$ $\mathcal{M}$	RM $\approx$ $\mathcal{M}$	RM $<$ $\mathcal{M}$	MG	MCG	MCL
Bagging (EW)	51	15	4	24.98%	29.92%	-32.41%
Bagging (MV)	53	15	2	26.07%	29.38%	-28.66%
Bagging (MVLW)	53	15	2	25.76%	28.98%	-30.72%
Bagging (INV)	51	17	2	27.39%	32.34%	-34.94%
Bagging (EW)	33	24	13	0.35%	1.56%	-0.44%
Bagging (MV)	30	26	14	0.33%	1.86%	-0.5%
Bagging (MVLW)	30	26	14	0.34%	1.86%	-0.49%
Bagging (INV)	31	29	10	0.41%	1.49%	-0.37%

Table 5: Comparison between bagging with each forecast combination strategy and clagging with distance-based scheme. Columns 1 to 3 respectively report the number of configurations where the reference model is significantly better than, has no significant difference with, and is significantly worse than a baseline model $\mathcal{M}$. Columns 4 to 6 respectively report the Median Gain, the Median Conditional Gain and the Median Conditional Loss. Top half is related to regression and considers MSE and t-test whereas bottom half is related to classification and considers AUC and DeLong test, both at 95% confidence level.

5.4 Where does the gain come from?

One could wonder where the gain of clagging compared to bagging comes from: Is it because of the way we form the subsets on which the models are trained (bootstrap vs. clusters) or because of the way we combine the forecasts (simple average vs. distance-based weights)? To address this question, we compare clagging to bagging both with the equally-weighted scheme. Figure 7 displays a superiority of bagging in this case. Indeed, when switching from clagging (EW) to bagging, we observe a decrease in performance in 60% of configurations in regression and in more than 45% of configurations in classification. However, recall from Section 5.3 and top panels of Figure 5 that clagging (DIST) outperformed bagging EW in 88% and 64% of the configurations in regression and

classification, respectively. This suggests that the increase in performance of clagging over bagging comes essentially from the combination scheme rather than the way individual models are fitted. Observe, however, that our distance-based combination scheme is of course tailored to the subsets considered in clagging. In particular, it would be pointless to consider a distance-based combination scheme in bagging.

(a) OOS R^2 - EW (b) H-measure - EW

Figure 7: Change in percentage of a performance metric from bagging to clagging both with equally-weighted scheme. The panels report the OOS R^2 in regression tasks (left) and the H-measure in classification tasks (right).

6 Conclusion

Forecast combination has been shown to outperform individual models, especially when there is some diversity in the set of models. One of the most widely used forecast combination model is *bagging*. Usually, a bootstrapped sample is created by drawing samples from the original training dataset with replacement so that the size of the bootstrapped sample coincides with the size of the original training set. In this work, we propose to aggregate clusters instead of bootstraps.

The general idea of cluster aggregating (clagging) is to rely on clustering to divide the training set into clusters on which forecasting models are fitted. Forecasts of the different models are subsequently combined and the information that an observation might belong to a cluster should not be disregarded. A combination scheme based on the distance of a new observation to each cluster is likely to outperform standard benchmarks such as equally-weighted and minimum-variance strategies. Clagging can be used both for regression and classification tasks and can deal with categorical features. Our evidence obtained through a horse race study with 20 datasets and 7 prediction

models suggests that clagging outperforms the standard random sampling with replacement widely used for bagging. Moreover, clagging outperforms a standard fit on the whole training set.

Acknowledgment

The work of Arnaud Germain is funded by the ING chair. Frédéric Vrins benefits from the financial support of the Fonds de la Recherche Scientifique F.S.R.-FNRS (grant J.0225.24) as well as of the Belgian Federal Science Policy Office (grant ARC 18-23/089).

References

Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, 36(1):197–200.

Bates, J. M. and Granger, C. W. (1969). The combination of forecasts. *Journal of the operational research society*, 20(4):451–468.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, 36:105–139.

Bouwmeester, W., Twisk, J. W., Kappen, T. H., van Klei, W. A., Moons, K. G., and Vergouwe, Y. (2013). Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC medical research methodology*, 13:1–10.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Brown, G., Wyatt, J., Harris, R., and Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information fusion*, 6(1):5–20.

Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *The annals of Statistics*, 30(4):927–961.

Chawla, N., Moore, T. E., Bowyer, K. W., Hall, L. O., Springer, C., and Kegelmeyer, P. (2001). Bagging is a small-data-set phenomenon. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Deodhar, M. and Ghosh, J. (2010). Scoal: A framework for simultaneous co-clustering and learning from complex data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):1–31.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268.

Granger, C. W. and Ramanathan, R. (1984). Improved methods of combining forecasts. *Journal of Forecasting*, 3(2):197–204.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, 77(1):103–123.

Hastie, T., Tibshirani, R., Friedman, J., et al. (2009). The elements of statistical learning.

Hawinkel, S., Waegeman, W., and Maere, S. (2024). Out-of-sample r 2: estimation and inference. *The American Statistician*, 78(1):15–25.

Huang, Z. and Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE transactions on Fuzzy Systems*, 7(4):446–452.

Jaeger, S. A. and Shawe-Taylor, J. (2005). Generalization bounds and complexities based on sparsity and clustering for convex combinations of functions from random classes. *Journal of Machine Learning Research*, 6(3).

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411.

Martínez-Muñoz, G. and Suárez, A. (2010). Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition*, 43(1):143–152.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Rainio, O., Teuho, J., and Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14(1):6086.

Ramchandran, M., Mukherjee, R., and Parmigiani, G. (2021). Cross-cluster weighted forests. *arXiv preprint arXiv:2105.07610*.

Roccazzella, F., Gambetti, P., and Vrins, F. (2022). Optimal and robust combination of forecasts via constrained optimization and shrinkage. *International Journal of Forecasting*, 38(1):97–116.

Samuels, J. D. and Sekkel, R. M. (2017). Model confidence sets and forecast combination. *International Journal of Forecasting*, 33(1):48–60.

Trivedi, S., Pardos, Z. A., and Heffernan, N. T. (2015). The utility of clustering in prediction tasks. *arXiv preprint arXiv:1509.06163*.

Van den Berg, J., Candelon, B., and Urbain, J.-P. (2008). A cautious note on the use of panel models to predict financial crises. *Economics Letters*, 101(1):80–83.

A Set of hyperparameters

Decision Tree (CART):

- $cp = 0.05$

Partial Least Squares:

- $ncomp = 3$

Boosted Generalized Linear Model:

- $mstop = 250$
- $prune = \text{F}$

Extreme Gradient Boosting:

- $nrounds = 50$
- $max\_depth = 3$
- $eta = 0.3$
- $gamma = 0$
- $colsample\_bytree = 0.8$
- $min\_child\_weight = 1$
- $subsample = 1$
- $skip\_drop = 0.05$
- $rate\_drop = 0.01$

Multivariate Adaptive Regression Spline:

- $nprune = 50$
- $degree = 2$

Penalized Logistic Regression:

- $decay = 0.0001$

Boosted Logistic Regression:

- $nIter = 50$

Least Angle Regression:

- $fraction = 0.5$

Linear Regression: no hyperparameters

B Data cleaning and preprocessing

We start with a very light data cleaning for some datasets that is detailed hereafter. In the rare event that observations with missing values remain, they are deleted from the dataset. Note that the remaining number of observations is at least 90% of the original number of observations for every dataset and no observations are deleted for the vast majority of datasets. All numeric variables are independently normalized using the `scale` function in `R`.

Rice: no preprocessing

Spambase: no preprocessing

Musk: removing *molecule_name*, *conformation_name* and *ID* features

Ringnorm: no preprocessing

Twonorm: no preprocessing

HTRU2: no preprocessing

MAGIC: no preprocessing

Online Shoppers: no preprocessing

Credit card: no preprocessing

Adult: no preprocessing

Bank Marketing: removing *duration* feature

Communities&Crime: removing the features containing missing values

Parkinsons: removing the first five features

Extreme Weather: removing *station*, *Date* and *Next_Tmin* features

Electrical Grid: removing *stabf* feature

Appliances Energy: removing *Date* feature

Superconductivity: no preprocessing

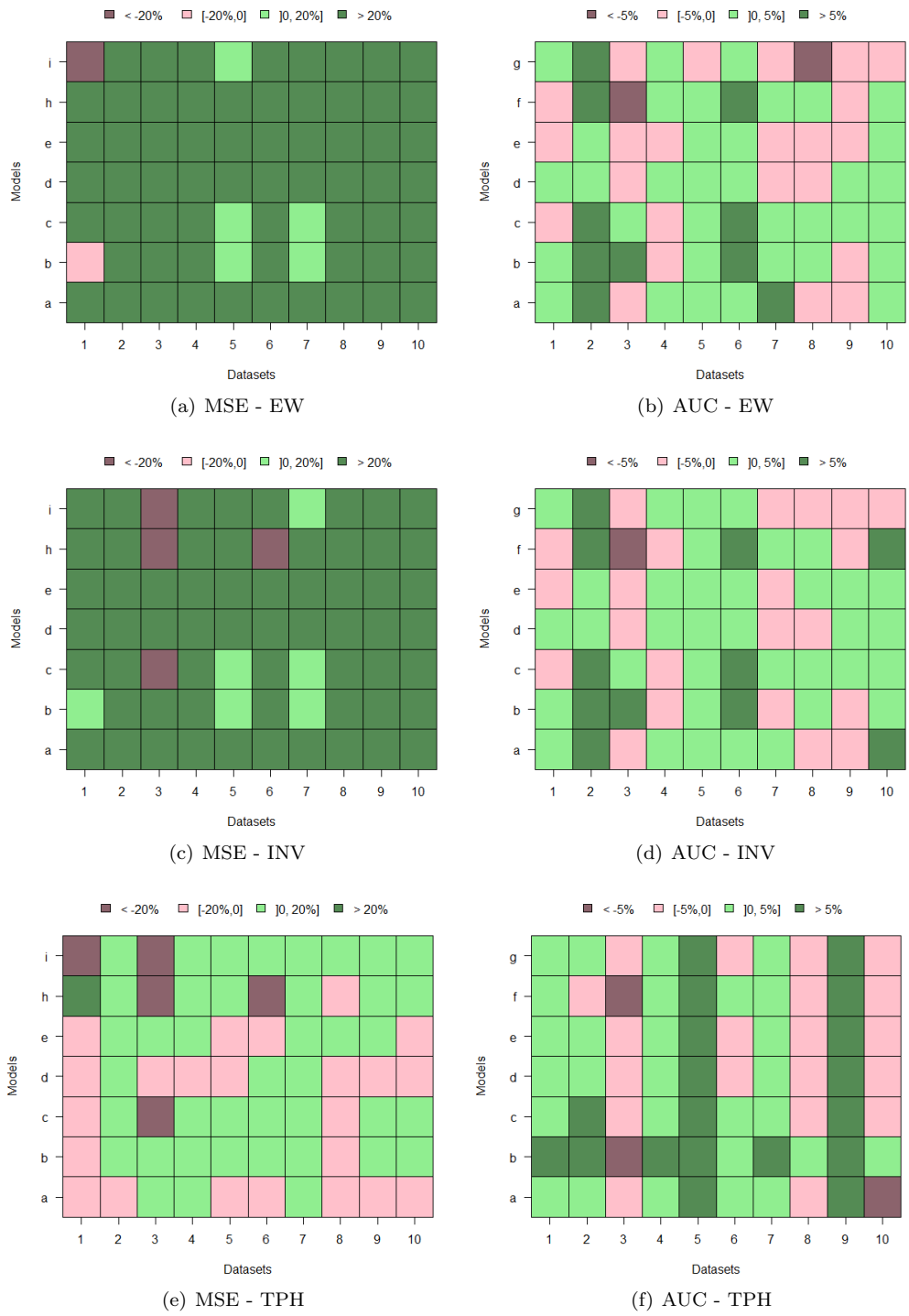
Protein: no preprocessing

Abalone: no preprocessing

Seoul: removing all the observations for which the feature *Functioning Day* is "No" and removing *Date* and *Functioning Day* features

Bike Sharing: removing *instant*, *dteday*, *casual* and *registered* features

C Combination scheme



(a) MSE - EW

(b) AUC - EW

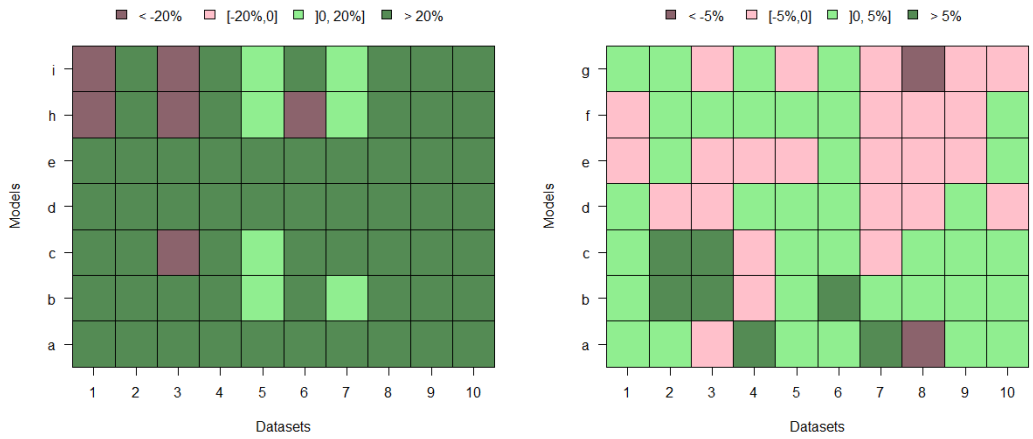
(c) MSE - INV

(d) AUC - INV

(e) MSE - TPH

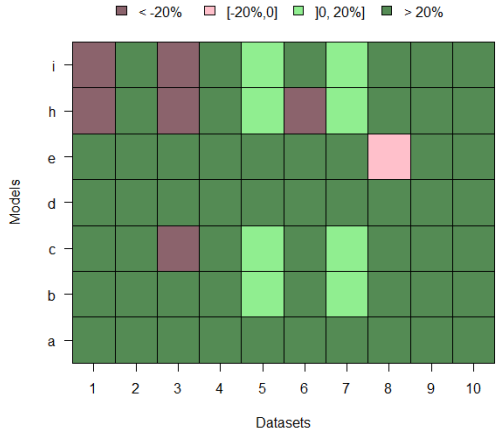
(f) AUC - TPH

Figure 8: Change in percentage of performance metric using clagging when switching from a benchmark combination scheme (indicated in the captions) to our distance-based combination scheme. The panels report the MSE in regression tasks (left) and the AUC in classification tasks (right). The considered benchmark schemes are the equally-weighted (top), the inverse-variance (middle) and the hard assignment scheme of Trivedi et al. (2015) (bottom).

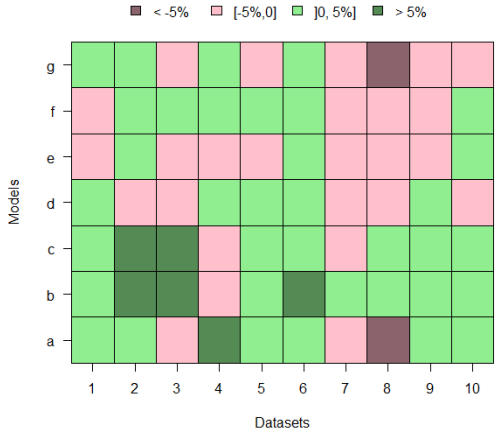


(a) MSE - MV

(b) AUC - MV

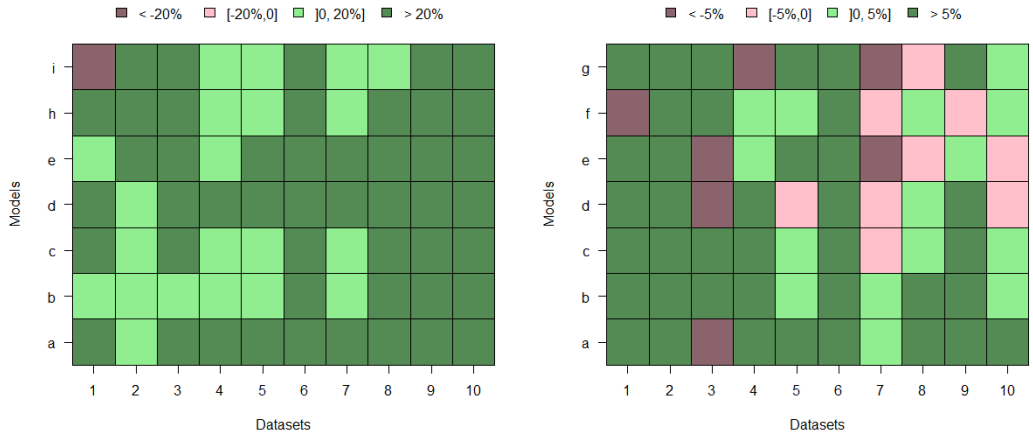


(c) MSE - MVLW



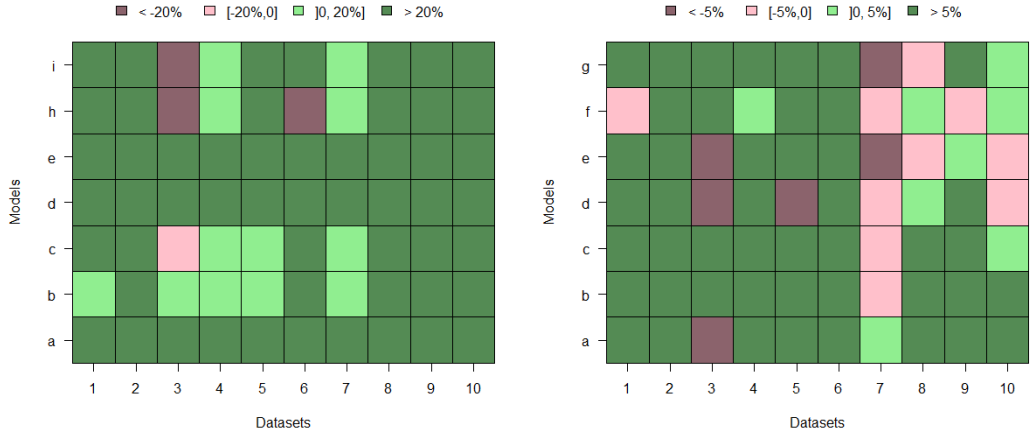
(d) AUC - MVLW

Figure 9: Change in percentage of performance metric using clagging when switching from a benchmark combination scheme (indicated in the captions) to our distance-based combination scheme. The panels report the MSE in regression tasks (left) and the AUC in classification tasks (right). The considered benchmark schemes are the minimum-variance scheme with sample estimator (top), and the minimum-variance scheme with Ledoit and Wolf (2004) estimator (bottom).



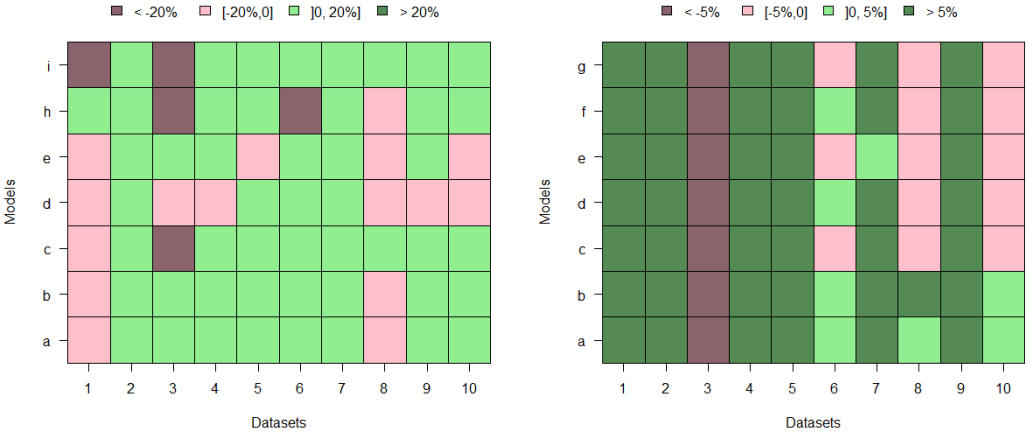
(a) MAE - EW

(b) Accuracy - EW



(c) MAE - INV

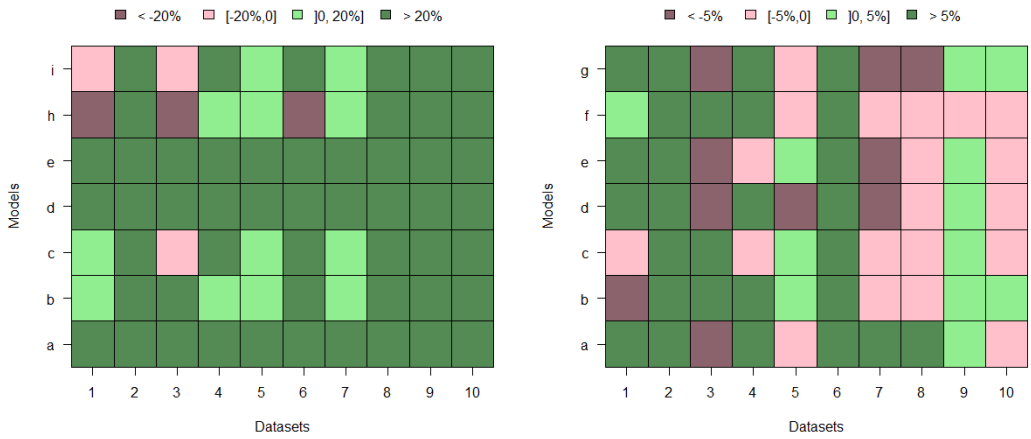
(d) Accuracy - INV



(e) MAE - TPH

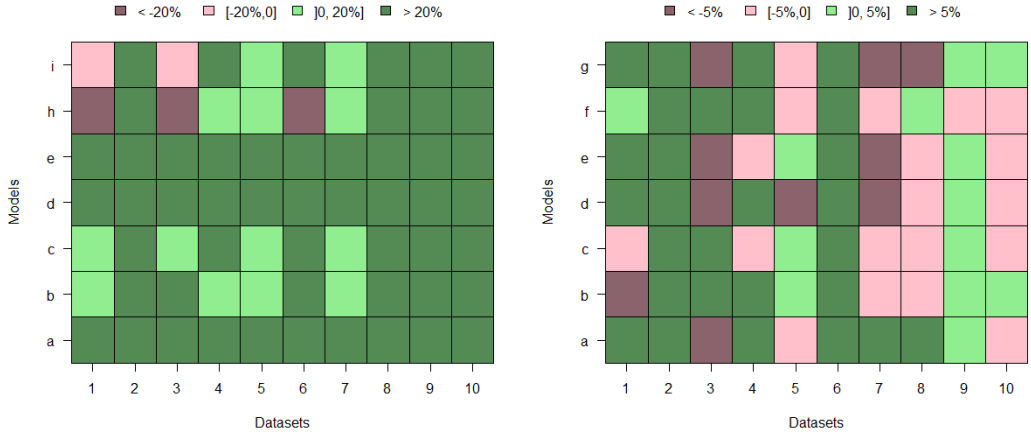
(f) Accuracy - TPH

Figure 10: Change in percentage of performance metric using clagging when switching from a benchmark combination scheme (indicated in the captions) to our distance-based combination scheme. The panels report the MAE in regression tasks (left) and the Accuracy in classification tasks (right). The considered benchmark schemes are the equally-weighted (top), the inverse-variance (middle) and the hard assignment scheme of Trivedi et al. (2015) (bottom).



(a) MAE - MV

(b) Accuracy - MV



(c) MAE - MVLW

(d) Accuracy - MVLW

Figure 11: Change in percentage of performance metric using clagging when switching from a benchmark combination scheme (indicated in the captions) to our distance-based combination scheme. The panels report the MAE in regression tasks (left) and the Accuracy in classification tasks (right). The considered benchmark schemes are the minimum-variance scheme with sample estimator (top), and the minimum-variance scheme with Ledoit and Wolf (2004) estimator (bottom).

	Reference Model: Clagging (DIST)					
$\mathcal{M}$	RM $>$ $\mathcal{M}$	RM $\approx$ $\mathcal{M}$	RM $<$ $\mathcal{M}$	MG	MCG	MCL
Clagging (EW)	68	1	1	42.63%	42.64%	-23.55%
Clagging (MV)	63	6	1	58.92%	62.49%	-27.46%
Clagging (MVLW)	62	6	2	58.03%	63.38%	-37.04%
Clagging (INV)	66	4	0	43.16%	47.07%	-34.27%
Clagging (TPH)	35	27	8	1.79%	4.31%	-2.37%
Clagging (EW)	33	32	5	7%	24.45%	-4.19%
Clagging (MV)	22	43	5	1.73%	24.59%	-3.67%
Clagging (MVLW)	20	45	5	1.68%	19.05%	-3.7%
Clagging (INV)	35	31	4	10.07%	25.47%	-2.87%
Clagging (TPH)	42	20	8	55.61%	117.86%	-3.1%

Table 6: Comparison between clagging with distance-based combination scheme and each forecast combination strategy. Columns 1 to 3 respectively report the number of configurations where the reference model is significantly better than, has no significant difference with, and is significantly worse than a baseline model $\mathcal{M}$. Columns 4 to 6 respectively report the Median Gain, the Median Conditional Gain and the Median Conditional Loss. Top half is related to regression and considers MAE and t-test whereas bottom half is related to classification and considers Sensitivity and McNemar test, both at 95% confidence level.

D Standard fit

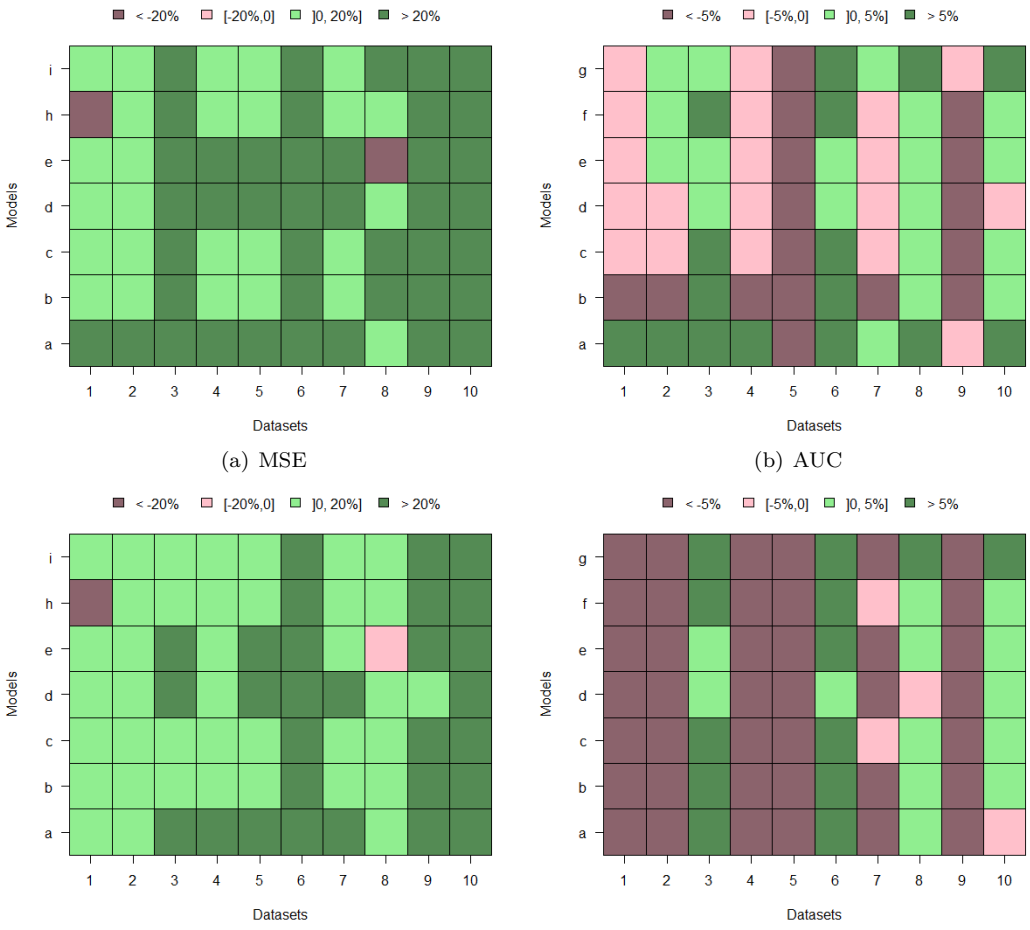


Figure 12: Change in percentage of performance metric from a standard fit to clagging with hard assignment scheme of Trivedi et al. (2015). The panels report the MSE and MAE in regression tasks (left) and the AUC and Accuracy in classification tasks (right).

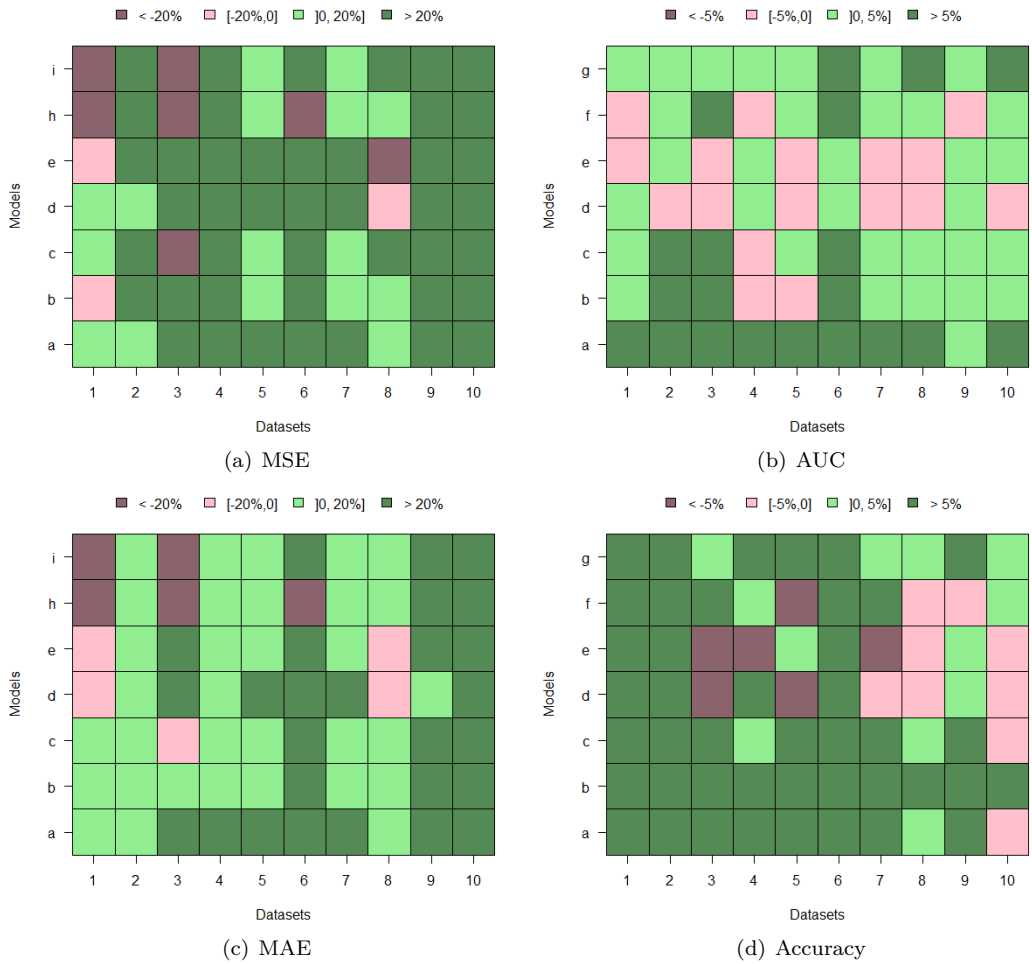
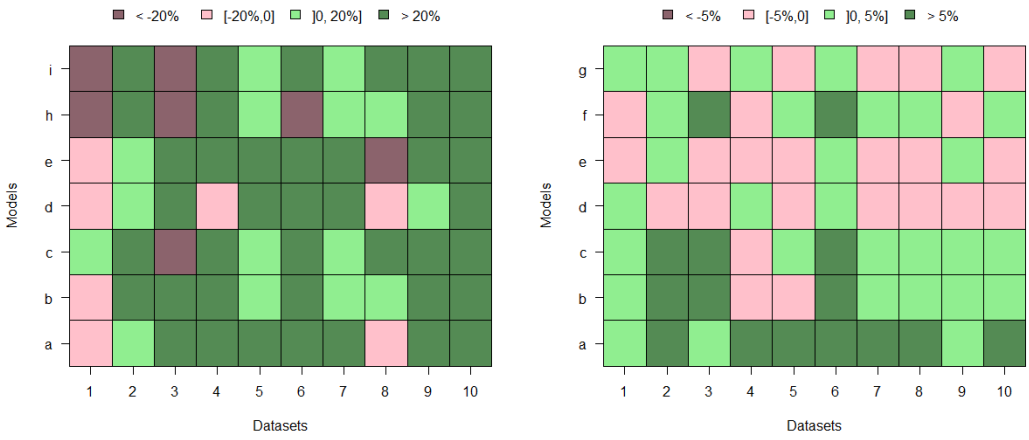


Figure 13: Change in percentage of performance metric from a standard fit to clagging with our distance-based combination scheme. The panels report the MSE and MAE in regression tasks (left) and the AUC and Accuracy in classification tasks (right).

	Reference Model: standard fit					
$\mathcal{M}$	RM $>$ $\mathcal{M}$	RM $\approx$ $\mathcal{M}$	RM $<$ $\mathcal{M}$	MG	MCG	MCL
Clagging (TPH)	1	6	63	-11.35%	124.32%	-11.88%
Clagging (DIST)	3	11	56	-12.76%	21.38%	-15.21%
Clagging (TPH)	37	19	14	25.36%	79.09%	-5.8%
Clagging (DIST)	4	31	35	-7.49%	3.7%	-12.91%

Table 7: Comparison between a standard fit on the whole training set and clagging with both hard and soft assignment. Columns 1 to 3 respectively report the number of configurations where the reference model is significantly better than, has no significant difference with, and is significantly worse than a baseline model $\mathcal{M}$. Columns 4 to 6 respectively report the Median Gain, the Median Conditional Gain and the Median Conditional Loss. Top half is related to regression and considers MAE and t-test whereas bottom half is related to classification and considers Sensitivity and McNemar test, both at 95% confidence level.

E Bagging



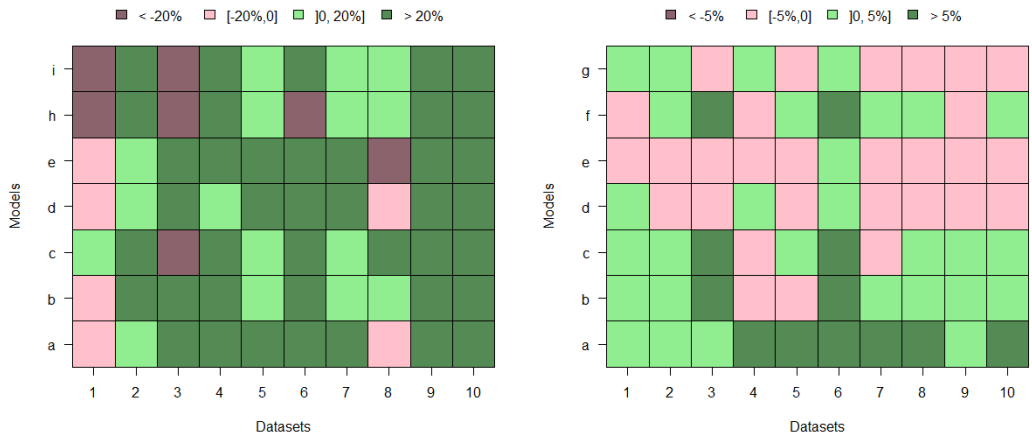
(a) MSE - EW

(b) AUC - EW

(c) MSE - INV

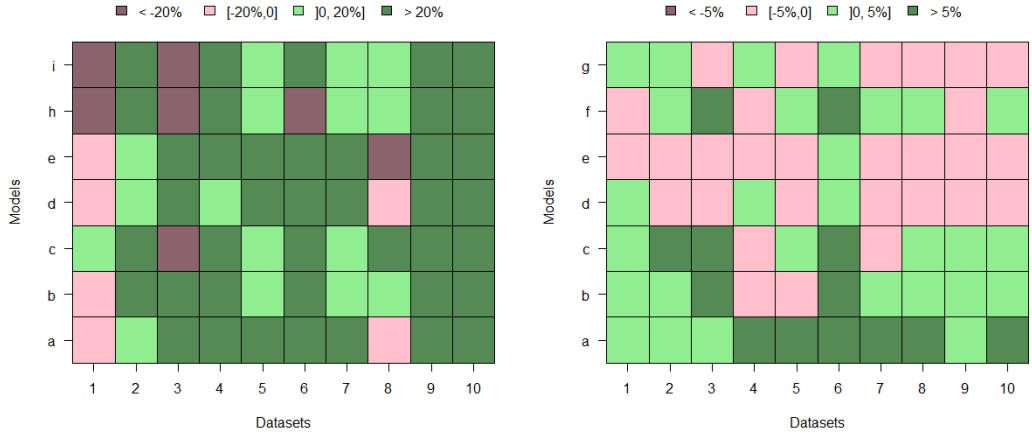
(d) AUC - INV

Figure 14: Change in percentage of performance metric when switching from bagging with a benchmark combination scheme to clagging with our distance-based combination scheme. The panels report the MSE in regression tasks (left) and the AUC in classification tasks (right). The considered benchmark scheme is the equally-weighted scheme (top), and the inverse-variance scheme (bottom).



(a) MSE - MV

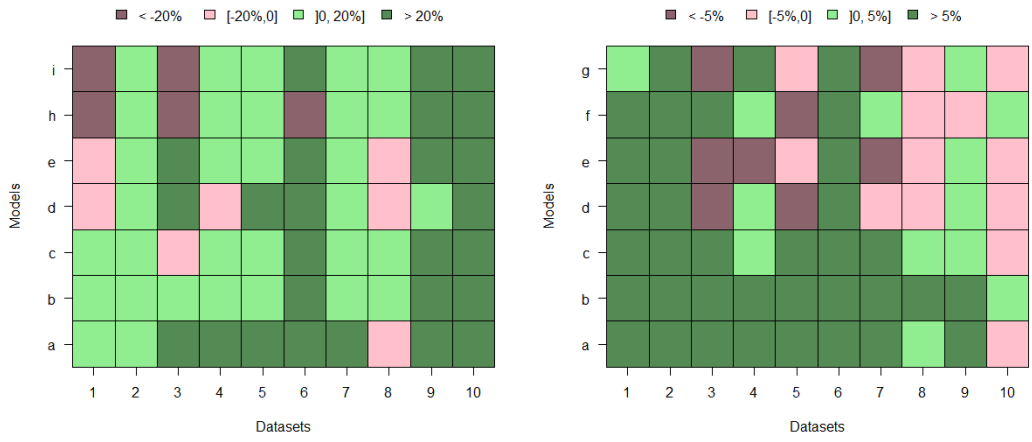
(b) AUC - MV



(c) MSE - MVLW

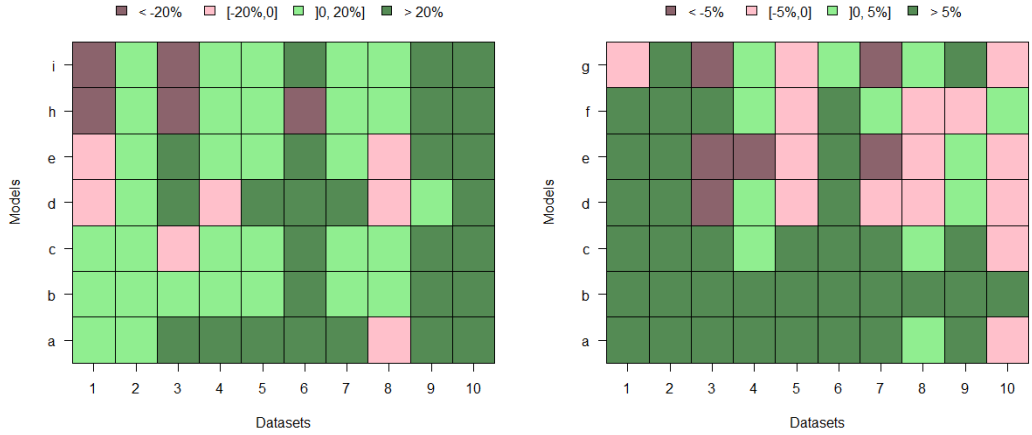
(d) AUC - MVLW

Figure 15: Change in percentage of performance metric when switching from bagging with a benchmark combination scheme to clagging with our distance-based combination scheme. The panels report the MSE in regression tasks (left) and the AUC in classification tasks (right). The considered benchmark schemes are the minimum-variance scheme with sample estimator (top), and the minimum-variance scheme with Ledoit and Wolf (2004) estimator (bottom).



(a) MAE - EW

(b) Accuracy - EW



(c) MAE - INV

(d) Accuracy - INV

Figure 16: Change in percentage of performance metric when switching from bagging with a benchmark combination scheme to clagging with our distance-based combination scheme. The panels report the MAE in regression tasks (left) and the Accuracy in classification tasks (right). The considered benchmark scheme is the equally-weighted scheme (top), and the inverse-variance scheme (bottom).

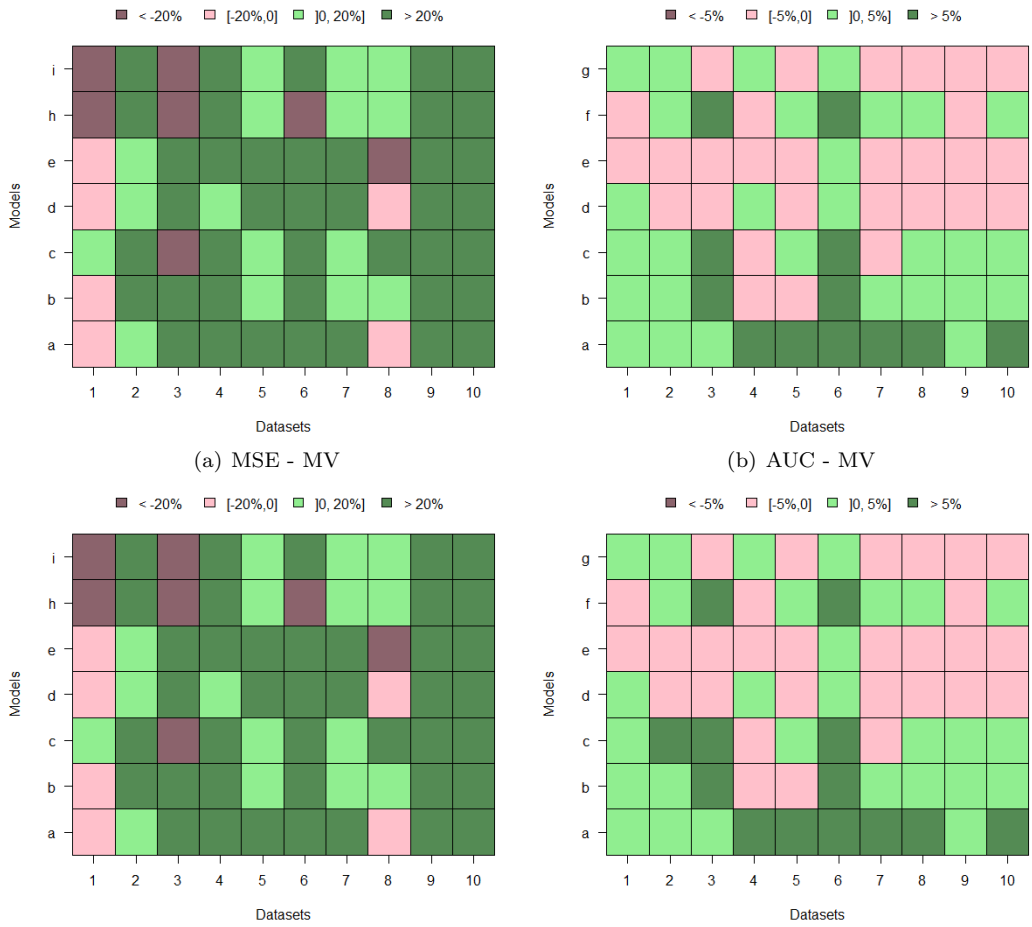


Figure 17: Change in percentage of performance metric when switching from bagging with a benchmark combination scheme to clagging with our distance-based combination scheme. The panels report the MAE in regression tasks (left) and the Accuracy in classification tasks (right). The considered benchmark schemes are the minimum-variance scheme with sample estimator (top), and the minimum-variance scheme with Ledoit and Wolf (2004) estimator (bottom).

	Reference Model: Clagging (DIST)					
$\mathcal{M}$	RM $>$ $\mathcal{M}$	RM $\approx$ $\mathcal{M}$	RM $<$ $\mathcal{M}$	MG	MCG	MCL
Bagging (EW)	54	12	4	14.08%	17.81%	-11.95%
Bagging (MV)	55	13	2	14.61%	17.61%	-11.76%
Bagging (MVLW)	55	12	3	14.54%	17.61%	-11.76%
Bagging (INV)	55	11	4	14.58%	17.98%	-11.09%
Bagging (EW)	27	36	7	5.59%	10.61%	-3.22%
Bagging (MV)	26	38	6	5.01%	19.09%	-3.33%
Bagging (MVLW)	26	38	6	4.96%	15.09%	-3.21%
Bagging (INV)	28	36	6	5.62%	10%	-3.33%

Table 8: Comparison between bagging with each forecast combination strategy and clagging with distance-based scheme. Columns 1 to 3 respectively report the number of configurations where the reference model is significantly better than, has no significant difference with, and is significantly worse than a baseline model $\mathcal{M}$. Columns 4 to 6 respectively report the Median Gain, the Median Conditional Gain and the Median Conditional Loss. Top half is related to regression and considers MAE and t-test whereas bottom half is related to classification and considers Sensitivity and McNemar test, both at 95% confidence level.