Dynamic and Heterogeneous Ensembles for Time Series Forecasting

Vitor Cerqueira
LIAAD-INESCTEC &
University of Porto
Porto, Portugal
Email: vmac@inesctec.pt

Luis Torgo
LIAAD-INESCTEC &
University of Porto
Porto, Portugal
Email: ltorgo@inesctec.pt

Mariana Oliveira
LIAAD-INESCTEC &
University of Porto
Porto, Portugal
Email: mariana.r.oliveira@inesctec.pt

Bernhard Pfahringer
University of Auckland
New Zealand
Email: b.pfahringer@auckland.ac.nz

***Abstract*—This paper addresses the issue of learning time series forecasting models in changing environments by leveraging the predictive power of ensemble methods. Concept drift adaptation is performed in an active manner, by dynamically combining base learners according to their recent performance using a non-linear function. Diversity in the ensembles is encouraged with several strategies that include heterogeneity among learners, sampling techniques and computation of summary statistics as extra predictors. Heterogeneity is used with the goal of better coping with different dynamic regimes of the time series. The driving hypotheses of this work are that (i) heterogeneous ensembles should better fit different dynamic regimes and (ii) dynamic aggregation should allow for fast detection and adaptation to regime changes. We extend some strategies typically used in classification tasks to time series forecasting. The proposed methods are validated using Monte Carlo simulations on 16 real-world univariate time series with numerical outcome as well as an artificial series with clear regime shifts. The results provide strong empirical evidence for our hypotheses. To encourage reproducibility the proposed method is publicly available as a software package.**

***Keywords*-dynamic ensembles; time series forecasting**

I. Introduction

This paper addresses the task of time series forecasting, which is an important topic with vast applicability across several domains. The focus is on univariate series where the observable is numerical. The problem is approached by ensemble learning methods, which have been proved to surpass single model learning on a variety of tasks. As was explained by Brown [1], the superior predictive performance of ensembles is in great part due to the diversity among the individual learners comprising them.

Dynamic ensembles for classification tasks is a well studied topic, for example [2], [3]. However, while there is a vast research in ensemble methods for regression tasks (e.g. [4]), the literature regarding the application of these methods to changing environments, such as univariate time series, is limited. In this context, this paper presents a new dynamic ensemble for time series series forecasting tasks.

The ensemble method we present settles on individually pre-trained models which are dynamically combined at run-time to make a prediction. The combination rule is reactive to changes in the environment, rendering an online combined model.

This paper explores new techniques for encouraging ensemble diversity in time series forecasting tasks. Moreover, we use model weighting schemes to adapt the learning device to the presence of concept drift [5]. Essentially, concept drift occurs when the underlying distribution of the data changes over time, disturbing the learning process. The main properties of our proposal are:

- **heterogeneity**: Heterogeneous ensembles are those comprised of different types of base learners. By employing models that follow different learning strategies, use different features and/or data observations we expect that individual learners will disagree with each other, introducing a natural diversity into the ensemble that helps in handling different dynamic regimes in a time series forecasting setting;
- **responsiveness**: We promote greater responsiveness of heterogeneous ensembles in time series tasks by making the aggregation of their members' predictions time-dependent. By tracking the loss of each learner over time, we weigh the predictions of individual learners according to their recent performance using a non-linear function. This strategy may be advantageous for better detecting regime changes and also to quickly adapt the ensemble to new regimes.

Our main contribution is combining these two properties to tackle numerical time series prediction tasks. We expect that, due to the heterogeneity of the ensemble, some individual models will perform better than others in particular data-spaces. We hypothesise that the combination of such an ensemble with a time-dependent aggregation function that rewards the best recent models will improve time series forecasting results. The prediction made at each new observation is produced by a committee, which is a subset of the best recent performing models.

The methods proposed in this paper were evaluated on 16 real world univariate time series with numerical outcome. Numerical experiments reveal that our approach is competitive with different forms of ensemble learning methods as well as other state-of-art methods for the adaptive combinination of forecasting models. To further improve our argument we also evaluate the proposed method in an artificial environment with marked regimes. In this setting the adaptability of our model becomes clear. In order to encourage reproducibility our methods are publicly available as an *R* package.

We start by addressing the related work in Section II; the proposed methodology for dynamic and heterogeneous ensembles is presented in Section III, along with a formal explanation of our contributions; the experiments and respective results are presented in Section IV, followed by a discussion in Section V; the final remarks are drawn in Section VI.

II. Related Work

The related work to this paper originates from two research branches: (i) dynamic ensembles (or Dynamic Combiners) and (ii) combination of forecasters. In this section we briefly revise typical dynamic combination methods used in classification and regression tasks. Then, we examine the state-of-art approaches for combination of forecasting models, pointing out their drawbacks as well as the main contributions of our work.

A. Dynamic Combiners

In this paper we focus our study on dynamic ensembles for numerical time series forecasting tasks. Building adaptable models is important in dynamic real-world environments in which data is constantly changing over time due to several factors, for example seasonality.

Heterogeneity among base learners of an ensemble has been reported to increase the predictive ability of such models in many settings (e.g., stacking [6]). Moreover, the stream mining community has put an effort towards creating models that are able to cope with changing environments. One common approach to this problem is the use of what Kuncheva [2] denominates *Dynamic Combiners* – this strategy involves training the ensemble's base models in advance and then somehow dynamically combining them to make a prediction. Several examples of Dynamic Combiners geared toward classification tasks can be found in the literature (e.g. [7], [3], [8], [9]). In [7], an ensemble of classifiers is trained and dynamically weighted to adapt to concept drift in a data stream. In [3], an homogeneous ensemble of incremental learners is trained and weighted, with experts added or removed as needed. In [8], an heterogeneous ensemble of incremental learners (each using its own feature subset) are trained and combined, adding new experts when necessary. In [9], the authors use classifier chains to dynamically select a subset of models. One of the most popular strategies for weighting expert advice used in online learning is regret minimization [10, Chapter 2]. Regret is the average loss incurred w.r.t. the best prediction we could have obtained.

We differentiate our approach in two ways:

1. We focus on numerical time series forecasting tasks, while the related work is mostly built towards classification or regression tasks. We aim at adapting work developed in these scenarios to improve the adaptive combination of forecasting models;
2. We introduce a novel combination formula for aggregating base learners. This formula computes the weight of a learner applying the complementary Gaussian error function to its recent loss which is quantified by the moving average of its squared error. In this strategy, the weight of a given model decays exponentially as its error increases.

On top of these two differentiating factors, we represent the dynamics of a series using only the sequence of measurements of the same collected variable. The typical approach involves using different predictors of the target variable.

B. Combining Individual Forecasters

Combining forecasters has been proved successful before. In [11] the authors use an ensemble of bagged trees, specially designed for time series forecasting tasks. Moreover, they encourage diversity across trees by exploring different representations of the recent dynamics of the time series.

Timmermann has proved in his seminal work [12] that combining forecasters using the simple average is a robust method. In [13], the authors use a similar approach, but trim the 20% worst performing models in all past data. Another method is introduced in [14], where the linear weights of forecasters are determined using recent performance. A variance-based combination scheme was proposed in [15]. The authors cluster forecasters by past performance and the predictions of the best performing group are averaged for prediction. In [16], the authors use a combination strategy that uses the number of times a method performed best in the past.

AEC is a method for adaptively combining forecasters presented in [17]. It uses an exponential re-weighting strategy to combine forecasters according to their past performance. It includes a forgetting factor to give more importance to recent values. In [18] it is argued that for the prediction of stock returns models have only short-lived periods of predictability. An adaptive combination is proposed based on the recent R^2 (coefficient of determination) of forecasters. If all models have poor explained variance (low R^2) in the recent observations then the forecast is set to the mean value of those observations. Otherwise, the base-learners are combined by averaging their predictions with the arithmetic mean.

The originality of this work with respect to these approaches is two-fold:

1. These approaches are based on typical time series analysis models such as ARIMA [19]. On the other hand we focus on heterogeneous machine learning models. Our hypothesis is that these should better fit the different dynamic regimes of the time series given their distinct inductive biases;
2. We use a non-linear function based on the complementary Gaussian error function to combine subsets of best recent performing models, yielding a more reactive combined model.

III. Dynamic Ensemble for Time Series Forecasting

A univariate time series is a time-ordered sequence of values $y_1, y_2, \ldots, y_n$ from an observable Y measured at regular time intervals, where y_i is the value of Y at time i. At this stage it is important to remark that we use the term *time series* throughout the paper assuming Y is a numeric variable (i.e. $y_i \in \mathbb{R}, \forall y_i \in Y$).

To tackle the problem of time series forecasting the proposed methodology follows the ideas from [20] regarding time-delay embedding. In this context, a time series is reconstructed into a higher dimensional space with embedding dimension K. Effectively, we generate the following matrix:

$$Y_{[N,K]} = \begin{bmatrix} y_1 & y_2 & \cdots & y_{K-1} & y_K \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{i-K+1} & y_{i-K+2} & \cdots & y_{i-1} & y_i \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ y_{n-K+1} & y_{n-K+2} & \cdots & y_{n-1} & y_n \end{bmatrix} \tag{1}$$

Each row denotes an embedding vector $v_r, \forall\, r \in \{1, \ldots, t-K+1\}$. The assumption is that there are no long term time dependencies in the series and thus the embedding vectors are deemed as essentially uncorrelated.

This representation of the time series allows the use of any regression technique available in the literature. Therefore, in the proposed methodology a set of heterogeneous regression models is individually pre-trained in the available data. At run-time we combine the models according to their recent performance to predict unseen observations.

A. Dynamic Heterogeneous Ensemble

In [11], diversity in an ensemble of bagged trees is encouraged by exploring different representations of the recent observations of a time series. This is achieved by the use of different embedding dimensions along with data summary statistics. However, learning models of the same type are prone to behave similarly across similar data-spaces. For example, tree-based models are bound to learn the data space in the form of hyper-rectangles. Combining learning algorithms with different inductive bias encourages a natural diversity in the ensemble, given their different assumptions on the unknown regression function. In this context, heterogeneity among base learners leads to an improvement of the overall predictive ability of the ensemble (e.g. [6], [1], [21]).

We propose a dynamic ensemble geared towards time series forecasting where base learners are weighted according to their recent performance. As opposed to related work in combination of forecasters that weight models according to their past performance (e.g. [14]) we evaluate the performance of each model in recent observations instead of the whole past data. This achieves a reactive combined model self-adaptable to concept drift.

We weight each base learner in the proposed model by tracking its residuals. Specifically, we introduce a metric called EMASE (for Erfc Moving Average Squared Error) to quantify the recent performance of a model. This heuristic is formalized in Equation 2:

$$EMASE_s = \frac{\operatorname{erfc}(MASE_s)}{\sum_{s \in S} \operatorname{erfc}(MASE_s)}, \forall s \in S \tag{2}$$

where $MASE$ is the moving average squared error (normalized to a 0–1 scale with a *max-min* normalization) computed over a window of P periods. S is the set of heterogeneous base learners and erfc is the Gaussian complementary error function formalized as follows:

$$Erfc(x) = \frac{4}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt \tag{3}$$

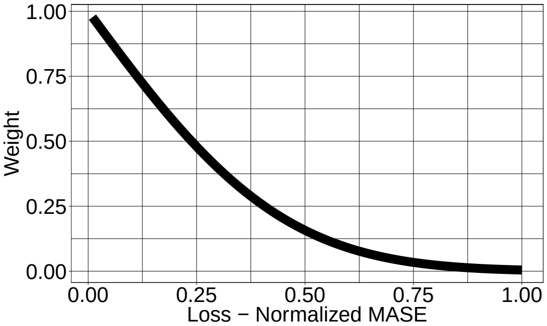

Fig. 1: Dynamics of the function Erfc. As the loss of a given model increases its weight decreases exponentially.

Figure 1 provides an intuition for the Erfc function. The weight of a given model decays exponentially as its loss increases. The original formulation of the Erfc function uses a 2 scalar instead of a 4 in the fraction of Equation 3. This change is motivated by the interest of smoothing the exponential decay of the weights with respect to the loss.

The dynamics of the moving average yield a flexibility to the combined model, in the sense that it is self-adaptable when concept drift occurs. Moreover, the number of periods P to average over, when calculating the EMASE, controls the reactiveness of the learning system to such events. A smaller value of P leads to greater reactiveness, but also makes the ensemble susceptible to be deceived by outliers. Conversely, higher values of P lead to greater stability, while losing some responsiveness. This trade-off is known as the stability-plasticity dilemma [22].

In this setting, models are pre-trained on the available data and their predictions aggregated according to the models' EMASE. This strategy culminates into a committee of models, dynamically extracted from the pool of learners. The committee is a subset of the base learners formed by the top λth percentile of models w.r.t. EMASE. In other words, at prediction time and after weighting forecasters with respect to EMASE we trim the worst models and combine the top λth percentile to make a prediction.

For a new observation d in the series the predictions of the models in the committee are combined by an *Aggregate* function:

$$\hat{y_d} = \sum_{s' \in S'} \hat{y_{s'}} \times \omega_{s'} \tag{4}$$

where $\hat{y_d}$ is the ensembles' prediction for d; $\hat{y_{s'}}$ is the prediction of learner s' in the committee S'; S' is the subset of the learners whose EMASE score is on the top λth percentile of the scores of all models in S; and $\omega_{s'}$ is the weight for model s' which is determined by EMASE and is formalized in the following equation:

$$\omega_{s'} = \frac{EMASE_{s'}}{\sum_{s' \in S'} EMASE_{s'}} \tag{5}$$

Essentially, after selecting the best recent performing models (according to EMASE and the cut-point λ) the *Aggregate* function combines these models according to their EMASE score.

Although the individual learners are trained in a batch (and hence static) fashion, their aggregation changes over time, making the ensemble dynamic and online as a whole. Learners with poor predictive performance in recent observations have their importance in the aggregation decreased or are even temporarily discarded from the committee S'. On the other hand, well-fitting models in recent observations are given more relevance and may become part of the committee S'.

We hypothesize that this strategy renders an effective mechanism against concept drift which is a common issue in dynamic environments.

B. Further Encouragements for Diversity

Besides base learner heterogeneity and dynamic aggregation of models we also encourage diversity in the ensemble by exploring different representations of the recent dynamics of a time series. We use different parameter settings of the base learners, embedding dimensions and training windows for each base learner.

Following the ideas presented by [11], groups of learners are trained using different predictors according to the embedding dimension and also using different training window sizes.

- **embedding size diversity**: Each base learner is trained using embedding dimensions K, $K/2$ and $K/4$, where K is the maximum embed size of the original data;
- **training window diversity**: We split the training set window in a similar fashion. Suppose that the time series has N observations. We train each learner in the last N, $N/2$ and $N/4$ observations.

Overall, this leads to 9 different data combinations available for training. On top of these diversity measures each learner may comprise different parameter settings. Moreover, each learner includes two extra predictors Δ, mean and standard deviation of the embedded values, in order to augment the information about the recent dynamics of the series.

In summary, we build an ensemble composed by S different models, varying both in parameter settings and in the data used for learning. Algorithm 1 summarises the proposed Dynamic Heterogeneous Ensemble.

IV. Empirical Experiments

This section describes the experiments carried out to validate the proposed methodology of dynamic and heterogeneous ensembles for time series forecasting tasks. These were especially designed to answer the following three research questions:

Q1: Is it beneficial to use heterogeneous base learners? That is, we want to see if using models with different assumptions about the underlying process causing the series is helpful to make better predictions in the future.

Q2: Is it beneficial to dynamically combine heterogeneous base learners? More specifically, can we use information on the recent performance of the base models in order to improve future predictions?

Q3: How does the performance of the proposed dynamic heterogeneous ensemble relate to the state-of-the-art methods for time series forecasting tasks and state-of-the-art methods for the adaptive combination of forecasters?

First we use 16 real world time series and several baseline models to prove our hypotheses. Then, to further improve our point we also include a more in-depth analysis where we employ our model in an artificial time series. All these experiments are reproducible. The datasets and code for this work are available[1] as a R package.

A. Experimental Setup

The above-mentioned hypotheses were tested using 16 real world time series briefly described in Table I. Only time series with variance above 1 (normalized by their respective range) were considered. The rationale behind this choice is that highly volatile time series are more prone to comprise non-linear dynamics with different regimes. From Table I, time series with ID 1–8 are related to a residential power load [23]; 9–11 are associated with water demand levels from different delivery points[2] in a city; 12–13 are related to ozone level detection [24]. Finally, time series 14–16 were collected from a solar radiation monitoring system [25].

Algorithm 1: Dynamic and Heterogeneous Ensemble

```
Input: Time series $Y_N$ of size $N$; Set of heterogeneous
       base learners $S$; Embedding Dimension $K$
– Embed $Y_N$ into $Y_{[N,K]}$
– Compute extra predictors $\Delta$ onto $Y_{[N,K]} \rightarrow Y_{[N,\{K,\Delta\}]}$
foreach base learner s in S do
    foreach $n \subseteq N, k \leq K$ do
        train $s$ using $Y_{[n,\{k,\Delta\}]}$
    end
end
Compute $EMASE_s, \quad \forall s \in S$
Initialize $\Lambda \leftarrow Percentile_{1-\lambda}(EMASE)$
Initialize Committee $S' \leftarrow \{s \in S : EMASE_s \geq \Lambda\}$
foreach upcoming new data point d do
    Get predictions $\hat{y}_{s'}$ from models $s' \in S'$
    Compute weights $\omega_{s'}, \forall s' \in S'$
    Compute prediction $\hat{y}_d = Aggregate(\hat{y_{s'}}, \omega_{s'})$
    Compute loss $L(\hat{y}_s, y_d), \quad \forall s \in S$
    Update $EMASE_s \quad \forall s \in S; \quad \Lambda; \quad S'$
end
```

[1] repository: https://github.com/vcerqueira/tsensembler

[2] Downloaded from Águas do Douro e Paiva: http://addp.pt

TABLE I: Datasets and respective summary

ID	Time series	Data source	Data characteristics
1	Wholehouse Power	Residential Loads [23]	Every 30 secs. – May 5, 2016 – from 8h34min to 12h58min (649 values)
2	Wholehouse Reactive Power		
3	Condenser Power		
4	Dryer Power		
5	Range Power		
6	Washer Power		
7	Dishwasher Power		
8	Lights Power		
9	Preciosa Mar	Oporto Water Consumption from different locations	Half-hourly values from Nov. 11, 2015 to Jan. 11, 2016 (2929 values)
10	Ameal		
11	Montes Burgos		
12	Sea Level Pressure	Ozone Level Detection [24]	Daily values from Jan. 2, 1998 to Dec. 31, 2004 (2533 values)
13	K Index		
14	Global Horiz. Radiation	Solar Radiation Monitoring [25]	Hourly values from Apr. 25, 2016 to Aug. 25, 2016 (2950 values)
15	Direct Normal Radiation		
16	Diffuse Horiz. Radiation		

The experiments were performed using the framework provided by the **performanceEstimation** [26] R package.

The methods were evaluated using the Mean Squared Error (MSE) on ten Monte Carlo repetitions. For each repetition, a random point in time is chosen from the full time window available for each series, and the previous window N consisting of 60% of the data set size is used for training the ensemble while the following window of size 25% is used for testing. Moreover, we follow the guidelines provided by Demšar in [27] for the statistical comparison of the different methods.

B. Ensemble Methods Setup

Finding the appropriate embed size is dependent on the data itself. In order to test for robustness, we tried two different values of maximum embed: **20** and **40**. We tested 3 different levels of responsiveness to changes. This is accomplished by having P, the number of periods used to calculate EMASE, take the values **10**, **25** and **50**.

The base models comprising each ensemble are the following: SVM [28], Neural Networks [29], Gaussian Processes [28], MARS [30], Generalized Linear Models [31], Generalized Boosted Models [32], Random Forests [33], Rule-based Regression [34] and PPR [35]. The heterogeneous ensembles include several parameter variants of each of these base models, in a total of 324 models for each ensemble.

The percentile λ is set to **10**, which means that at each prediction time the best 10% base models are combined to make the final prediction.

1) Baselines: We considered the following six different models as baselines:

- **ARIMA**: An ARIMA model, estimated using the function *auto.arima* from [19], which automatically tunes the model for an optimal parameter setting;
- **BAGT**: A static homogeneous ensemble. We include the best performing variant of the ensemble of bagged trees proposed in [11] (BaggingDE±S). This model extends standard bagging by using summary statistics as predictors as well as exploring different embedding dimensions and is specially designed for time series forecasting tasks;
- **S-S**: A static heterogeneous ensemble. This is a variant of our method where the predictions of all base learners are simply averaged using the arithmetic mean. This strategy goes back to [12] where the author proves its robustness for time series forecasting tasks;
- **S-W**: Another static heterogeneous ensemble. In this variant the base models are weighted according to past performance such as in [13], where the weights are linear and determined using all past information;
- **NG-W50**: A dynamic ensemble in which all available models are weighed according to their past performance in the past 50 observations [14]. We tested different values for the window size and 50 provided the best results;
- **R-W**: A dynamic variant of Algorithm 1 where EMASE is replaced by an exponentially weighted average function with theoretical bounds [10, Chapter 2], which minimizes a regret loss function;
- **AEC** [17]: a method for the adaptive combination of forecasting models – check Section II for a description;
- **ERP** [18]: a method for the adaptive combination of forecasting models – check Section II for a description;

Our proposed method is denoted as E-W*P*.

C. Results

Figures 3 to 6 summarise the paired comparisons results from the Wilcoxon Signed Rank test for two different baselines, **BAGT** and **S-W**. Paired comparisons are depicted by back-to-back barplots. Outer bars are wins – left for the respective variant, right for the baseline. Inner bars are statistically significant wins. We picked these two models as

TABLE II: Mean and deviation of rank of the workflows across the 16 experiments. A method with rank position 1 in a given experiment means it is the best performing model in such experiment.

K	20	40
ARIMA	8.1 ± 3.7	7.3 ± 3.9
BAGT	4.6 ± 3.7	4.9 ± 3.6
S-S	7.7 ± 1.9	8.0 ± 2.3
S-W	7.1 ± 4.0	7.0 ± 3.9
NG-W50	5.3 ± 2.8	4.9 ± 2.8
E-W10	6.4 ± 1.6	5.8 ± 2.4
E-W25	4.9 ± 1.7	5.3 ± 2.1
E-W50	$\mathbf{4.0 \pm 2.2}$	$\mathbf{4.3 \pm 2.2}$
R-W	7.1 ± 3.4	7.5 ± 2.9
AEC	5.8 ± 3.8	6.3 ± 3.9
ERP	5.0 ± 2.4	4.5 ± 2.1

baselines for the paired comparisons results to verify two of our hypothesis **Q1** and **Q2**. Table II shows the mean rank position and respective deviation of all the methods across all 16 experiments. A critical difference diagram is presented in Figure 2 using the Bonferroni-Dunn post-hoc test. In the graphic, methods that are not connected with the horizontal line show significantly differences with respect to **E-W50**, the proposed method. K was set to 40. Similar conclusions were drawn using K equal to 20.

By using BAGT as baseline we can evaluate hypothesis **Q1**. From the inspection of Figures 3 and 4, the static heterogeneous ensembles (i.e. S-S and S-W) show a competitive performance relative to BAGT. This is more evident if we use the information about the performance of the base models in the training data (S-W). Although S-W presents the best mean rank position (among static ensembles) these are comparable with each other if we take into account the variability in the rank position. In summary, the results of our experiments show empirical evidence of heterogeneous methods being able to overcome the performance of homogeneous ensembles.

From the perspective of the method S-W as baseline – a static heterogeneous method – we can test the validity of hypothesis **Q2**. This analysis is supported by Figures 5 and 6, which provide a sense of the impact of using a dynamic aggregation function. The proposed dynamic approaches show a clearly superior performance relative to the static methods. Further, the dynamic variants with greater stability (i.e. $P = 50$) perform better than similar but more reactive versions ($P = 10$). The mean rank positions also corroborate the idea that it is in fact worthwhile to dynamically combine heterogeneous base learners. Particularly, the dynamic ensemble E-W50 achieves a remarkable mean rank.

According to the diagram in Figure 2, the proposed method shows a consistent superior average rank with respect to the remaining methods, including other state-of-art approaches for adaptive combination of forecasting models. Moreover, the difference to the methods ARIMA, S-S and R-W is statistically significant. These results answer the research question **Q3** about the comparison of the proposed method to other state-of-art approaches used in time series forecasting tasks.

In order to emphasise our point we tested the methodology in a synthetic environment, whose underlying dynamics are made to change.

D. Synthetic Analysis

In order to ensure the presence of regime shifts we resort to artificial data. We created an artificial time series with a large structural change (c.f. Figure 7) based on those presented in [36]. The time series includes six marked regimes each of which originated from a different process and comprising 250 observations. Our working hypothesis in this scenario is that the dynamic ensemble should better adapt to new regimes relative to a static ensemble due to its responsiveness component. As we mentioned before, we expect this property to improve the combined model by giving more importance to the best recent models. Log transformations were produced using the following equation: $sign(x) \cdot log(|x| + 1)$.

For simplicity we focus this analysis on the variants BAGT, S-W and E-W50, which represent a Static Homogeneous Ensemble, the best ranked Static Heterogeneous Ensemble and the best ranked Dynamic Heterogeneous Ensemble according to Table II, respectively. Moreover, we set K to 40. Similar conclusions were drawn with K equal to 20.

We compared the dynamic ensemble over time to the static ensembles by measuring the percentual difference in squared error, which is depicted in Figure 8. The figure supports the idea that the dynamic ensemble fits the data better than the static ensembles (**Q2**). The dynamic ensemble shows not only a lower error throughout the series, but also shows a better response to regime changes.

E. Base Learner's Analysis

We present an more in-depth analysis of the ensemble E-W50 as well as its base learners. Particularly, we study the impact of our diversity strategy described in Section III-B. Moreover, through a bias-variance decomposition we analyse the behavior of E-W50 with respect to some of its base learners. For simplicity we focus on time series IDs 15 and 9 (c.f. Table I) for those tasks, respectively. We set K to 40 and use an holdout strategy to perform the experiments. The first 70% of the time series is used to train the base learners and the remaining 30% is used for testing.

1) Diversity Analysis: In Section III-B we described sampling strategies used to encourage diversity in the ensemble. Particularly we sample the data with two strategies related to the embedding dimension K and the training window N. For example, a combination of $K - N/2$ means that the data comprises all the embeds up to K (columns in Matrix 1) and the half most recent embedding vectors (rows in Matrix 1). Other combinations have a similar intuition. This resulted in 9 different data combinations available for training each base learner.

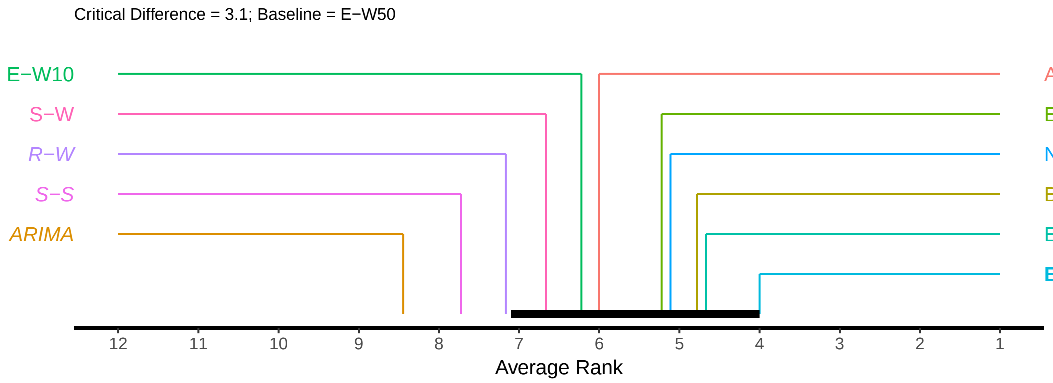

Fig. 2: Bonferroni-Dunn post-hoc test comparing the performance of the methods in the datasets for $K = 20$.

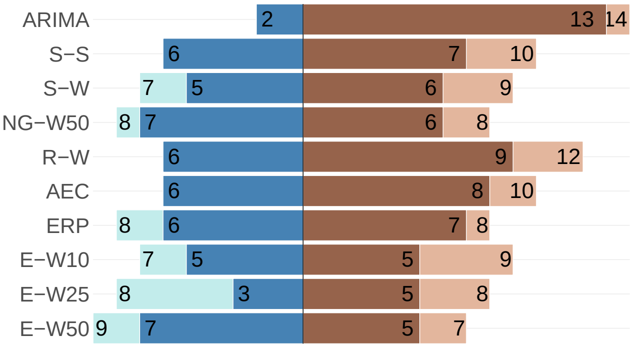

Fig. 3: baseline: BAGT ($K = 20$) – Paired comparisons depicted by back-to-back barplots. Outer bars are wins - left for the respective variant, right for the baseline. Inner bars are statistically significant wins.

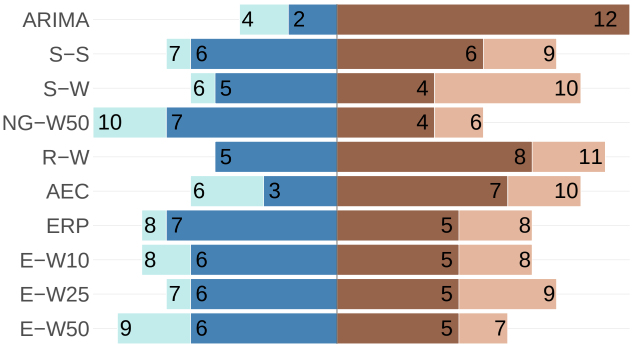

Fig. 4: baseline: BAGT ($K = 40$)

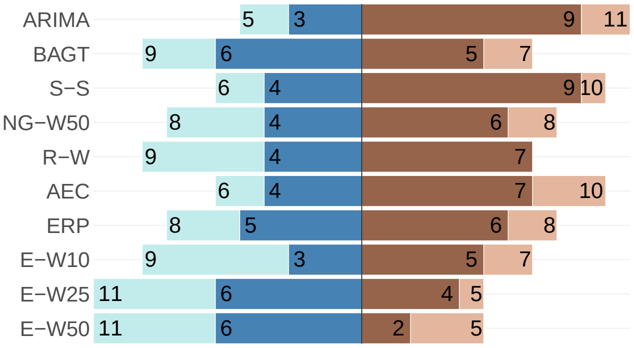

Fig. 5: baseline: S-W ($K = 20$)

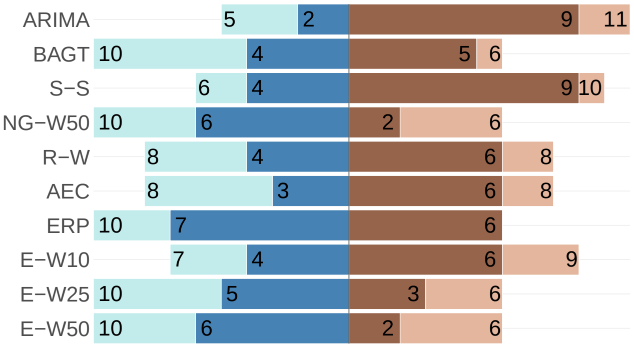

Fig. 6: baseline: S-W ($K = 40$)

In order to understand the impact of this strategy we analyse the performance of models grouped by each data combination. To accomplish this we study the rank of each data combination in terms of squared error. This metric is produced by computing the rank of the average rank of individual models grouped by data combination. For example, a data combination with rank 1 means that it comprises the base learners that on average have the lowest squared error. We report this rank in Figure 11 for time series id 15, where the results are averaged over periods of 20 consecutive observations. Overall the base learners training in the full data (N-K) have a lower mean rank (i.e. better performance) than other combinations which are subsets of the N-K embedded series. However, there are some peaks along the series in which base learners trained in subsets of the full embedded series, particularly those with lower embedding size (K/2 or K/4), have better performance. For example, in the highlighted area in Figure 11 there is a time interval in which the base models trained on subsets of K are performing better. The same, although less clear and less

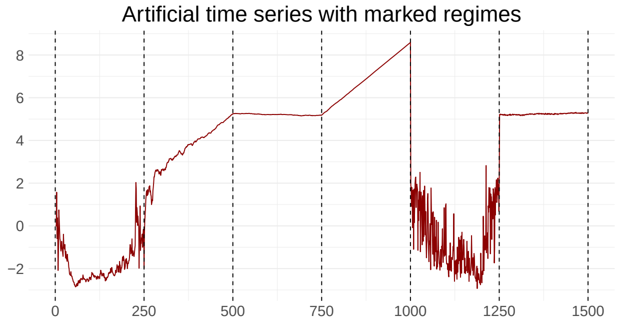

Fig. 7: Synthetic series with large structural change. The series is depicted in a log scale with six different marked regimes.

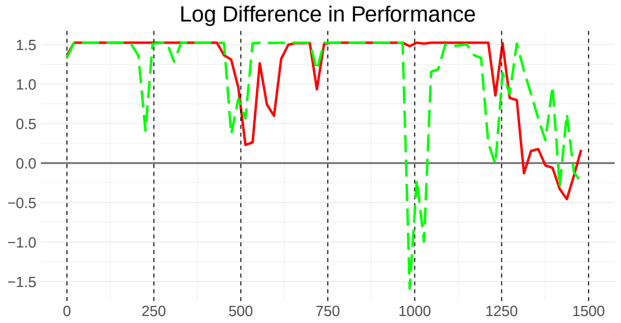

Fig. 8: Percentual difference (log-scaled) in squared error of E-W50 w.r.t. BAGT and S-W. Values above 0 represent periods when E-W50 is performing better than the static models. Regime changes are marked by vertical dotted lines. The red solid and green dashed lines represent the relative performance of E-W50 to S-W and BAGT, respectively.

frequent happens for subsets of the training window N.

Effectively, this analysis shows that the strategy we employed for diversity by using different subsets of data is worthwhile in some time intervals and helpful to cope with different regimes. Even if the subsetting strategies typically underperform (e.g. K/4 subsets) this is handled by the committee, which only considers a fraction of the best recent performing models.

2) Bias-Variance Analysis: We perform a bias-variance decomposition to understand how the performance of E-W50 relates to the performance of its base learners. For illustration purposes we focus on the top 3 base learners with lower mean squared error on the time series.

The bias of a model is the difference between the expected prediction and the actual value. Figure 9 shows the log-scaled bias of each model, computed incrementally at each test observation. The combined model has a better expected performance than the base learners. Additionally, the figure illustrates how the combined model mitigates fluctuations in bias error with respect to its base learners, rendering a stabler algorithm. Initially, the ensemble presents considerable fluctuations because the EMASE score is only fully computed after 50 observations due to the number of periods P to average over the squared error. We also show the log-scaled variance in Figure 10, which is lower for E-W50 and generally stable throughout the series for all models.

In summary, this posterior analysis of the results further strengthens our argument by studying how the proposed ensemble is better able to cope with changing environments, particularly univariate time series. Although this analysis is presented here only for two particular time series, similar conclusions were drawn in other ones.

V. Discussion

Overall, the results from the experiments demonstrate the competitiveness of the proposed method relative to other approaches. These include state-of-the-art methods for dynamically combining forecasting models.

Our starting hypothesis was that the dynamic aggregation renders a combined model that is self-adaptable when a change in regime occurs. This became more clear when we applied our dynamic method in a synthetic environment against two static ensembles (c.f. Figure 8). At each changing point (denoted by vertical dotted lines) the advantage of the dynamic ensemble is diminished, and sometimes lost, to its static competitors.

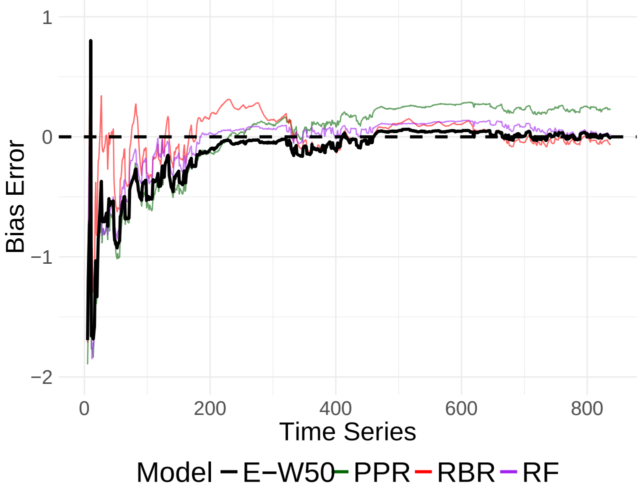

Fig. 9: Log-scaled bias of E-W50 ensemble and some of its base learners in time series id 9

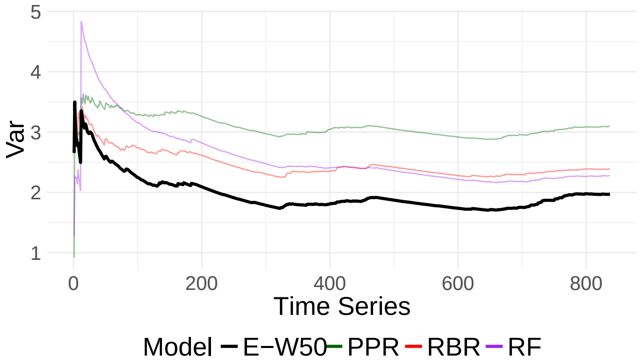

Fig. 10: Log-scaled variance of E-W50 ensemble and some of its base learners in time series id 9

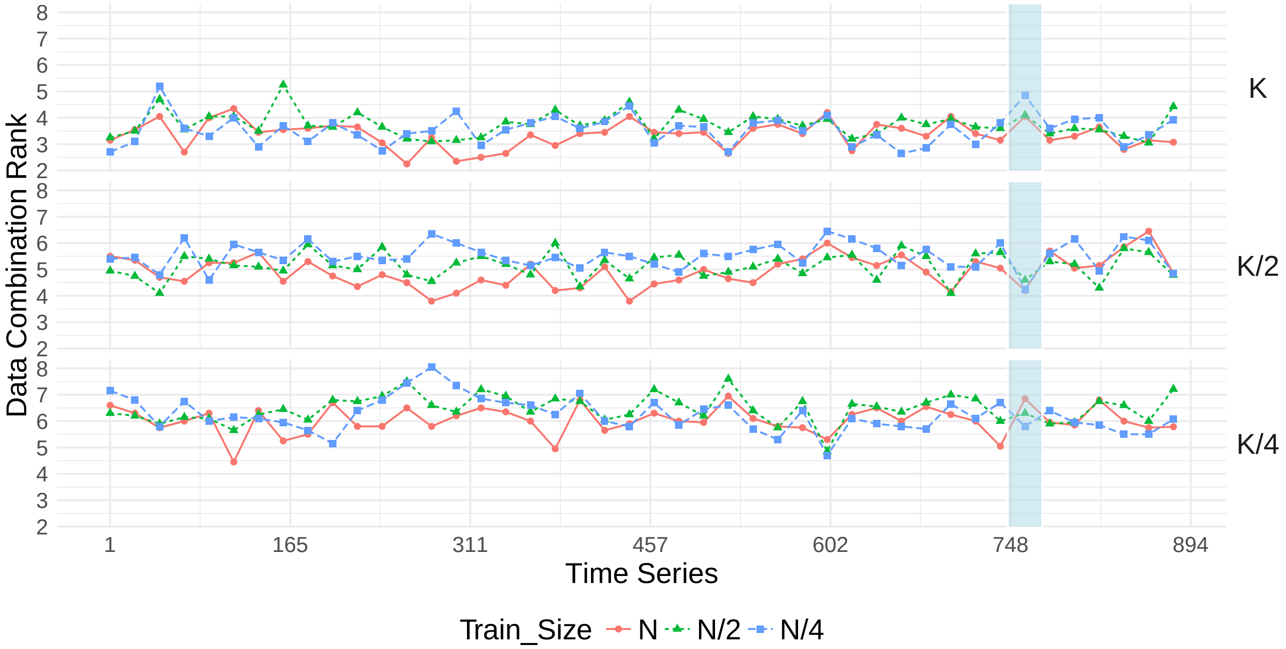

Fig. 11: Diversity analysis for time series id 15 – Squared error rank of data combination strategy. A combination with rank 1 means it comprises the base learners that have the lowest squared error, on average.

However, E-W50 is able to quickly adapt itself to the new regime and regain its previous advantage. The figure also illustrates the limitation of traditional methods in the face of concept drift. The static variants are not adaptable and then consistently under-perform relative to E-W50.

Many time series comprise recurrent patterns due to for example seasonality [37]. In this perspective, our model is capable of adapting between different concepts. When the underlying series generating process changes to a novel concept we expect the heterogeneity and responsiveness of our method to cope with such scenarios. Nonetheless, as future work we plan to include a re-training parameter to the base learners.

Regarding our combination strategy we used a complementary Gaussian error function to weight the base models. This function was picked instead of a linear transformation to further favor well performing ones and hence penalize bad performing ones. Figure 1 provides the intuition behind this choice. The weight of a given base learner decays exponentially as its loss increases.

In Section IV-E we studied the behavior of the base models comprising the proposed dynamic ensemble. First, we showed that training models in subsets of the original available series (c.f. Section III-B) might be worthwhile and helpful to cope with a changing environment. Second, we illustrated how the combined model improves and stabilizes the bias with respect to some of the individual models that comprise it.

VI. Conclusions

In this paper we presented new forms of ensemble methods for time series forecasting tasks, an extensively researched field in Machine Learning. Our main goal was to uncover new techniques for adaptively combining diverse base learners. In this context, we proposed a combination of the following strategies: **(i)** base learner *heterogeneity*, meaning that distinct modelling algorithms are used to create an ensemble, injecting a natural diversity in the combined model; and **(ii)** *responsiveness*, achieved by dynamically picking and aggregating the best recent base models. This dynamic ability is essential for predictive models in changing environments, as is frequently the case in time series forecasting tasks. Similar strategies have already been used before in classification tasks, but we extend them to numerical domains in univariate time series. Besides heterogeneity and dynamic aggregation, we also include varying embedding dimensions, different learner parameters, summary statistics as predictors and a non-linear combination function.

We conducted experimental comparisons of several variations of the proposed methods against other ensemble learning approaches designed to deal with time series forecasting tasks as well as classical auto-regressive methods used for time series forecasting. Our experiments included 16 real-world time series with unknown dynamics as well as an artificially generated time series with clear regime shifts.

Results from Monte Carlo simulations on the real-world series show the competitiveness of our methods and, in particular, the advantages of ensemble heterogeneity and responsiveness. This was validated using hypothesis testing with the Wilcoxon Signed Rank test and by the computation of the methods' average ranks over all time series. Experiments with the artificially generated time series further confirm our claims, as our dynamic method is visibly more adaptable to concept drift than ensembles using static prediction aggregation.

Future work includes: **(i)** implementing the ability to learn additional base models (or re-train existing ones) as the performance of the current pool degrades beyond tolerance; and **(ii)** introducing the ability to deal with multivariate dependencies.

To encourage reproducible research the methodology proposed as well as the time series used in this paper are publicly available as an R package.

Acknowledgment

This work is financed by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project Fotocatgraf - UTAP-ICDT/CTM-NAN/0025/2014. This work is also financed by the ERDF European Regional Development Fund through the Operational Programme for Competitiveness and Internationalisation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT Fundação para a Ciência e a Tecnologia as part of project UID/EEA/50014/2013.

References

[1] G. Brown, *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2010, ch. Ensemble Learning, pp. 312–320.

[2] L. I. Kuncheva, *Multiple Classifier Systems: 5th International Workshop, MCS 2004, Cagliari, Italy, June 9-11, 2004. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, ch. Classifier Ensembles for Changing Environments, pp. 1–15.

[3] J. Z. Kolter and M. A. Maloof, "Dynamic weighted majority: An ensemble method for drifting concepts," *J. Mach. Learn. Res.*, vol. 8, pp. 2755–2790, Dec. 2007.

[4] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, p. 10, 2012.

[5] J. C. Schlimmer and R. H. Granger, Jr., "Incremental learning from noisy data," *Mach. Learn.*, vol. 1, no. 3, pp. 317–354, Mar. 1986.

[6] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[7] H. Wang, W. Fan, P. S. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers," in *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: ACM, 2003, pp. 226–235.

[8] H.-L. Nguyen, Y.-K. Woon, W.-K. Ng, and L. Wan, *Advances in Knowledge Discovery and Data Mining: 16th Pacific-Asia Conference, PAKDD 2012, Kuala Lumpur, Malaysia, May 29 – June 1, 2012, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Heterogeneous Ensemble for Feature Drifts in Data Streams, pp. 1–12.

[9] F. Pinto, C. Soares, and J. Mendes-Moreira, "Chade: Metalearning with classifier chains for dynamic combination of classifiers," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2016.

[10] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge University Press, 2006.

[11] M. Oliveira and L. Torgo, "Ensembles for time series forecasting," in *ACML Proceedings of Asian Conference on Machine Learning. JMLR: Workshop and Conference Proceedings*, 2014.

[12] A. Timmermann, "Forecast combinations," *Handbook of economic forecasting*, vol. 1, pp. 135–196, 2006.

[13] V. R. R. Jose and R. L. Winkler, "Simple robust averages of forecasts: Some empirical results," *International Journal of Forecasting*, vol. 24, no. 1, pp. 163–169, 2008.

[14] P. Newbold and C. W. Granger, "Experience with forecasting univariate time series and the combination of forecasts," *Journal of the Royal Statistical Society. Series A (General)*, pp. 131–165, 1974.

[15] M. Aiolfi and A. Timmermann, "Persistence in forecasting performance and conditional combination strategies," *Journal of Econometrics*, vol. 135, no. 1, pp. 31–53, 2006.

[16] D. W. Bunn, "A bayesian approach to the linear combination of forecasts," *Journal of the Operational Research Society*, vol. 26, no. 2, pp. 325–329, 1975.

[17] I. Sánchez, "Adaptive combination of forecasts with application to wind energy," *International Journal of Forecasting*, vol. 24, no. 4, pp. 679–693, 2008.

[18] A. Timmermann, "Elusive return predictability," *International Journal of Forecasting*, vol. 24, no. 1, pp. 1–18, 2008.

[19] R. J. Hyndman, with contributions from George Athanasopoulos, S. Razbash, D. Schmidt, Z. Zhou, Y. Khan, C. Bergmeir, and E. Wang, *forecast: Forecasting functions for time series and linear models*, 2014, R package version 5.6.

[20] F. Takens, *Dynamical Systems and Turbulence, Warwick 1980: Proceedings of a Symposium Held at the University of Warwick 1979/80*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981, ch. Detecting strange attractors in turbulence, pp. 366–381.

[21] V. Cerqueira, L. Torgo, and C. Soares, "Arbitrated ensemble for solar radiation forecasting," in *International Work-Conference on Artificial Neural Networks*. Springer, Cham, 2017, pp. 720–732.

[22] G. A. Carpenter, S. Grossberg, and J. H. Reynolds, "Artmap: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network," *Neural Netw.*, vol. 4, no. 5, pp. 565–588, Sep. 1991.

[23] B. Sparn. (2016, May) Residential loads, national renewable energy laboratory. [Online]. Available: https://data.nrel.gov/submissions/48

[24] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[25] S. Andreas, A.; Wilcox, "Solar radiation monitoring station (sorms): Humboldt state university, arcata, california (data); nrel report no. da-5500-56515." 2007.

[26] L. Torgo, *An Infra-Structure for Performance Estimation and Experimental Comparison of Predictive Models*, 2013, R package version 0.1.1.

[27] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[28] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "kernlab – an S4 package for kernel methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.

[29] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0.

[30] S. Milborrow, *earth: Multivariate Adaptive Regression Spline Models. Derived from mda:mars by Trevor Hastie and Rob Tibshirani.*, 2012.

[31] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[32] G. Ridgeway, *gbm: Generalized Boosted Regression Models*, 2015, R package version 2.1.1.

[33] M. N. Wright, *ranger: A Fast Implementation of Random Forests*, 2015, R package version 0.3.0.

[34] M. Kuhn, S. Weston, C. Keefer, and N. C. C. code for Cubist by Ross Quinlan, *Cubist: Rule- and Instance-Based Regression Modeling*, 2014, R package version 0.0.18.

[35] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[36] K. Pan Pang and K. M. Ting, "Improving the centered cusums statistic for structural break detection in time series," in *Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence*, ser. AI'04. Berlin, Heidelberg: Springer-Verlag, 2004, pp. 402–413.

[37] J. Gama and P. Kosina, "Tracking recurring concepts with meta-learners," in *Portuguese Conference on Artificial Intelligence*. Springer, 2009, pp. 423–434.