International Journal of Forecasting 37 (2021) 547–568

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Treating and Pruning: New approaches to forecasting model selection and combination using prediction intervals

Erick Meira [a,b], Fernando Luiz Cyrino Oliveira [a,*], Jooyoung Jeon [c,d]

[a] Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro, Brazil
[b] Energy, Information Technology and Services Division, Brazilian Agency for Research and Innovation (Finep), Brazil
[c] Korea Advanced Institute of Science and Technology (KAIST), Republic of Korea
[d] School of Management, University of Bath, United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Keywords:
Model selection
Forecast combinations
Prediction intervals
Exponential smoothing
Bagging

ABSTRACT

We propose a new way of selecting among model forms in automated exponential smoothing routines, consequently enhancing their predictive power. The procedure, here addressed as treating, operates by selectively subsetting the ensemble of competing models based on information from their prediction intervals. By the same token, we set forth a pruning strategy to improve the accuracy of both point forecasts and prediction intervals in forecast combination methods. The proposed approaches are respectively applied to automated exponential smoothing routines and Bagging algorithms, to demonstrate their potential. An empirical experiment is conducted on a wide range of series from the M-Competitions. The results attest that the proposed approaches are simple, without requiring much additional computational cost, but capable of substantially improving forecasting accuracy for both point forecasts and prediction intervals, outperforming important benchmarks and recently developed forecast combination methods.

1. Introduction

It has been nearly six decades since the basic structures of exponential smoothing methods were first proposed (Holt, 1957; Winters, 1960). Still, thanks to their ease of use and adaptation to many different situations, exponential smoothing methods are not only widely applied in forecasting but also considered competitive in many cases – see, for instance, the results from the most recent M-Competition (Makridakis et al., 2018), in which automatic selection among exponential smoothing model forms ranked fourth best overall in terms of delivering accurate prediction intervals. In spite of their widespread use, recent literature has demonstrated that it is possible to improve upon exponential smoothing formulations (Hyndman & Athanasopoulos, 2018; Hyndman et al., 2008, 2002; Taylor, 2003).

Concurrently, the literature on forecast combination has now progressed to the point of considering the effect of subsetting the pool of available forecasts before aggregation (Aiolfi & Timmermann, 2006; De Menezes et al., 2000; Diebold & Shin, 2019; Elliott, 2011; Hendry & Clements, 2004; Kourentzes et al., 2019; Matsypura et al., 2018). The rationale behind subsetting has also been recently raised when forecasting using Bootstrap Aggregation (Bagging) routines by Dantas and Cyrino Oliveira (2018), who advocated the use of clustering methods to create a subset with a reduced variance.

In spite of the undeniable achievements on exponential smoothing formulations and on subsetting routines for forecast combination methods, no work has considered looking at the information delivered by Prediction Intervals (PIs) when conducting model selection and/or combination. In fact, it was not until recently that PIs were

* Corresponding author.
E-mail address: cyrino@puc-rio.br (F.L. Cyrino Oliveira).

https://doi.org/10.1016/j.ijforecast.2020.07.005
0169-2070/

considered in most forecasting works. For instance, the M4 Competition (Makridakis et al., 2018) was the first of its kind to explicitly ask participants to deliver PIs for their point forecasts (PFs), and ended with only 20 forecasters providing valid PIs (Makridakis et al., 2019).

We demonstrate that PIs, apart from providing practitioners with a convenient way to estimate the uncertainty of a PF, contain important information that can be used to improve the accuracy of forecasting methods involving model selection and/or combination. Concerning the former, we propose a new way of selecting among competing formulations that involve 'treating' – discarding specific model forms from the set of models – before proceeding to selection via traditional methods, e.g. via information criteria minimization. Regarding the latter, we set forth a 'pruning' strategy that can be used to enhance the accuracy of forecasts arising from any forecast combination method, provided that the models to be combined are able to generate PIs to their PFs. Both treating and pruning are conducted based on the information retrieved from the PIs of the forecasts. We explore the potential gains of these two strategies through an extensive empirical experiment on a wide range of monthly, quarterly and yearly time series from the M, M3 and M4 Competitions (Makridakis et al., 1982; Makridakis & Hibon, 2000; Makridakis et al., 2018). To demonstrate how treating can be used to improve upon model selection approaches, we apply this strategy to the automated exponential smoothing routine implemented in the *forecast* package for the R statistical software (Hyndman et al., 2019; Hyndman & Khandakar, 2008). With regard to pruning, we explore its potential to improve upon forecast combinations by applying it to two recently developed Bagging routines for forecasting, presented in the works of Bergmeir et al. (2016) and Petropoulos et al. (2018). These combining methods were selected in light of their promising results in the M3 Competition. Finally, we also propose different ways that Bagging routines can be extended to deliver PIs for the PF, another important development of this paper.

Foreshadowing our results, we demonstrate that, apart from their simplicity and ease of use, treating and pruning require practically no additional computation cost and can substantially improve the quality of forecasts for a considerable range of forecast approaches that involve model selection or combination. Some examples, as discussed throughout the paper, are model selection algorithms for exponential smoothing model forms, combinations of exponential smoothing model forms and Bootstrap Aggregation (Bagging) approaches.

The paper unfolds as follows. Section 2 provides an overview on the most popular and up-to-date techniques concerning exponential smoothing methods and Bagging approaches to forecasting. The proposed approaches are presented in detail in Section 3. Section 4 introduces the selected data for the empirical analysis and summarizes the results in terms of both PFs and PIs. Finally, Section 5 concludes and suggests directions for future works.

2. Forecasting with exponential smoothing and Bagging – the state of the art

In the following subsections, we provide a brief review on how model selection is typically conducted under most exponential smoothing routines and show the limitations arising from it. We also provide a chronological review of relevant works that use Bagging in time series forecasting contexts.

2.1. Exponential smoothing and current limitations

There are several different approaches to exponential smoothing. Hyndman et al. (2002), building upon the work of Ord et al. (1997), provided a solid theoretical foundation for exponential smoothing in state space modelling, allowing for straightforward implementation in many statistical packages (Hyndman & Athanasopoulos, 2018; Hyndman et al., 2008). Their modelling framework incorporated stochastic models, likelihood calculation, PIs and procedures for model selection among ETS formulations (an acronym standing for ExponenTial Smoothing or, alternatively, Error, Trend and Seasonality, the three components that define a model within the exponential smoothing family of models). They derived analytical results and proposed simulation routines to compute PFs and prediction distributions for 30 standard formulations of ETS, according to the taxonomy proposed by Pegels (1969) and further extended by Gardner Jr. (1985). The possibilities for the trend and seasonal components are depicted in Table 1. In addition, the error term can also vary between additive or multiplicative, resulting in a total of 30 formulations.

Model selection under the framework of Hyndman et al. (2002) is based on the minimization of one or more information criteria. For instance, by default, the `ets()` function from the *forecast* package for the R statistical software (Hyndman et al., 2019; Hyndman & Khandakar, 2008) uses the Akaike's Information Criterion corrected for small sample bias (AICc; Sugiura, 1978) to select an appropriate model. Other information criteria, such as Akaike (1974) or Schwarz (1978) can also be used. A similar procedure is also conducted in the EViews® statistical software (IHS Global Inc., 2015).

Selecting models based on information criteria minimization may seem compelling to practitioners who believe that searching for the 'true' model may not make sense for empirical data, since the optimal model for the real data generating process will not usually be among the candidate models considered in any case (Kolassa, 2011). Nevertheless, selecting a single best model out of a number of competing candidates may be misleading. Multiple models may explain the data almost equally well, and selecting a single model discards the information that could be gauged from alternative models with high explanatory power (Buckland et al., 1997). Another point that is often overlooked is that even if one relies on criteria that partially address the issue of overfitting (such as information criteria), the selected model(s) may still lead to inaccurate and/or unstable forecasts. Conducting some sort of cross-validation routine may circumvent this problem in some cases, but this is not always guaranteed.

In light of the above, we propose looking at the outputs originating from competing ETS formulations as a whole, and let them dictate which models are actually more prone to produce the best forecasts for a given time series

Table 1
Possible variations for the trend and seasonal components under a state space-based approach.

Components	Seasonal		
Trend	None (N)	Additive (A)	Multiplicative (M)
None (N)	N, N	N, A	N, M
Additive (A)	A, N	A, A	A, M
Additive damped (A_d)	A_d, N	A_d, A	A_d, M
Multiplicative (M)	M, N	M, A	M, M
Multiplicative damped (M_d)	M_d, N	M_d, A	M_d, M

and, accordingly, which should be discarded. More specifically, we aim to gather the PIs delivered from competing models and check for deviant behaviors in the ensemble. This 'wisdom of the crowds' approach builds on the same argument of the previous paragraph: since multiple models may explain the data almost equally well, they will usually produce forecasts that are not very distant from one another. However, for models presenting 'hard to estimate' stylized facts such as structural breaks, nonlinear patterns and/or periods with large range of values, a best model may be identified on the basis of traditional criteria, but its forecasts can still be very inaccurate and sometimes display explosive behavior in long forecast lead times. On the other hand, competing models which also delivered low values for most information criteria but were not selected as best during estimation phase may produce better forecasts than the selected model. Under such circumstances, pre-treating the set of candidate models may contribute to reduce the odds of selecting an unstable model and hence improve the accuracy of the forecasting method. We demonstrate the usefulness of 'pre-treatment' in Section 4, using a wide range of time series (98,830 series from the M-Competitions, split into monthly, quarterly and yearly frequencies). We also note that the additional computational cost is negligible, especially when compared with the time ETS routines take to estimate all competing model forms and collect their corresponding information criteria values.

2.2. Bagging in time series forecasting

The Bootstrap Aggregation (Bagging) (Breiman, 1996) is a supervised machine learning technique that generates multiple versions of a predictor via bootstrapping, which are then used to produce an aggregated predictor. In the field of forecasting, the technique can be viewed as a forecast combination method that operates by selectively resampling from the training data to generate derived training sets to which the base learner is applied. Bagging algorithms have demonstrated promising results, particularly when combined with exponential smoothing methods (Bergmeir et al., 2016; Cordeiro & Neves, 2009; Dantas et al., 2017; De Oliveira & Cyrino Oliveira, 2018; Petropoulos et al., 2018). In the following paragraphs, we provide a brief overview of recent studies employing Bagging as a forecast combination method for time series forecasting and indicate what yet remains to be explored in this domain.

Inoue and Kilian (2004) are likely to have pioneered the use of Bagging in time series forecasting. Using a dynamic multiple regression model, they showed that Bagging led to accurate forecasts when the data were stationary. Lee and Yang (2006) observed that Bagging could improve binary predictions in small samples, when asymmetric loss functions were used. Inoue and Kilian (2008) compared variants of Bagging based on untransformed regressors and on orthogonalized regressors and concluded that there were no significant differences in performance among them.

Cordeiro and Neves (2009) are likely to have been the first to combine Bagging and exponential smoothing methods. Their Boot.EXPOS approach, which can be viewed as a variant of the Sieve bootstrap method (Bühlmann, 1998), demonstrated promising results in forecasting time series with marked seasonal and trend components (mainly quarterly and monthly data). Encouraged by these results, Bergmeir et al. (2016) proposed a method where data were pre-treated through a Box–Cox transformation (Box & Cox, 1964) and decomposed via a Seasonal-Trend decomposition using *Loess* (STL decomposition) (Cleveland et al., 1990). Bootstraps for the remainder were then generated via a Moving Blocks Bootstrap (MBB) algorithm (Künsch, 1989). Once the desired number of bootstraps was achieved, the series was reconstructed from its structural components (by adding the trend and seasonal components to the remainder bootstraps) and the Box–Cox transformation inverted. Hence, multiple new series were created, and an exponential smoothing forecasting model was built for the original data and each of the bootstraps separately. Point forecasts were then aggregated using the median. The authors demonstrated that their approach, which became known as Bagged.BLD.MBB.ETS (BLD standing for Box–Cox and Loess-based Decomposition), outperformed Boot.EXPOS and other simple benchmarks, particularly for monthly series from the M3 Competition.

Petropoulos et al. (2018) decomposed the sources of uncertainty arising from Bagging procedures for time series forecasting into the model, data and parameter components, and demonstrated that the benefits of Bagging originate predominantly from model uncertainty. They proceeded to put forth a new strategy (the Bootstrap Model Combination, BMC) that specifically tackled this source of uncertainty. In brief terms, optimal model forms were first identified from the bootstraps and fitted to the original data. Forecasts originating from this strategy were then combined with weights reflecting the frequency that the selected model forms were identified as optimal on the bootstraps. Overall, considering all series from the M and M3 Competitions, the BMC delivered superior results than Bagged.BLD.MBB.ETS and Boot.EXPOS in most cases. Dantas and Cyrino Oliveira (2018), in turn,

aimed at reducing the covariance effect of bagged forecasts by using Partitioning Around Medoids (PAM) to produce clusters of similar forecasts, from which some were selected to create a smaller subset with reduced error-variance to be combined using the median. The methodology was tested on series from the M3 and CIF 2016 competitions and led to more accurate forecasts than several benchmarks from both competitions.

The approach proposed by Dantas and Cyrino Oliveira (2018) was the only one involving subsetting the ensemble of competing forecasts for Bagging routines. The approach, however, is not without issues since the clustering phase is (i) dependent on the number of clusters set by the practitioner (an automatic procedure is offered by the authors, but this is not guaranteed to deliver the best results on every occasion) and (ii) involves a validation set, and is thus also dependent on its size and quality relative to the actual, final values in the out-of-sample period. It should also be noted that the approach is very computationally intensive when compared with previous Bagging routines. Finally, we observe that Bagging has not yet been used to produce PIs (only PFs).

3. Methods

In Section 2.1 we showed that model selection in most automated ETS routines is based on the minimization of one or more information criteria. Despite the benefits arising from this procedure, such as avoiding overfitting, the selected model occasionally leads to unstable forecasts. Therefore, in the following subsections, we introduce the concept of 'treating' and demonstrate how it can be applied to reduce the odds of selecting formulations that will lead to innacurate forecasts, thus enhancing the overall accuracy of automated ETS procedures. We also show that treating can be applied to ETS model combination schemes and provide a brief explanation on how PIs are built in most exponential smoothing routines.

In addition to the above, we outline the rationale behind pruning, a similar concept to treating applied to forecast combination methods, and demonstrate some ways that it can be effectively implemented, using selected Bagging strategies to that end. Finally, we provide some insights on how PIs can be constructed in Bagging routines.

3.1. Treating in ETS model selection

The rationale behind 'treating' is to compare the PIs originating from competing model forms in ETS, and discard the ones showing deviant behaviors from the majority in the ensemble. More specifically, it initially collects the upper limits of the PIs and considers as outliers any values lying outside the range of $\pm 1.5 \times IQR$, where $IQR = Q_3 - Q_1$ is the inter-quartile range (difference between the third and first quartiles). The use of *IQR* for outlier detection is a well-established procedure in descriptive statistics (Vinod, 2014) and has been used in a vast number of applications, including subsetting pools of forecasts (Kourentzes et al., 2019). While it is true that PI symmetry may not hold for all ETS model forms involved, we opted to use the upper limits of the PIs for outlier detection because all series from the M-Competitions involved were strictly positive.

The outlier detection procedure in our 'Treated ETS' approach is conducted for every step in the forecast lead time and considers all competing model forms, regardless of whether a model form has already been identified as an outlier in the first forecasting step, for instance. At the end, every model identified as an outlier (even if just once) throughout the forecast lead time is discarded from the set of competing models. After this treatment, final model selection proceeds as usual: by finding, among the remaining models, the one offering the lowest value for AICc. Albeit unlikely to occur, it may be the case that all competing model forms are identified as outliers at least once during the forecast lead time. Under such circumstances, only the model forms which were the most frequently identified as outliers would be discarded from the set of competing models.

For illustration purposes, Fig. 1 depicts the 15 upper PI limits from 15 ETS model forms estimated for M4 monthly series 16214, considering an 18-steps ahead forecasting experiment – as standard for monthly series in M-Competitions. The shaded area represents the boundaries/range of the $\pm 1.5 \times IQR$. Given the very high values observed in the upper PI limits from three ETS formulations – (M, N, M), (M, A_d, M) and (M, A, M), depicted in red in the figure – these limits lie outside the shaded area on some horizons of the forecasting lead time. Therefore, the ETS formulations which generated these large PIs are discarded during treating and the practitioner is left with 12 competing ETS forms for model selection.

We recall that, for our empirical experiments, we use as benchmark the automated ETS procedure implemented in the `ets()` function from the *forecast* package for the R statistical software (Hyndman et al., 2019; Hyndman & Khandakar, 2008). According to this algorithm, not all the 30 ETS formulations listed in Table 1 are considered by default in model selection. Model forms involving multiplicative trends and combinations of additive errors and multiplicative seasonality are not estimated by default. Thus, at the end, there are 15 competing model forms out of the 30 different possibilities. The number of competing model forms for the yearly data is six. This is rather a small number as this frequency does not involve any seasonality. In addition, for quarterly series with training sets comprising 13 observations or less only, there is insufficient data to estimate models with damped trends. In such cases, the practitioner is left with only 10 out 15 competing model forms. By the same token, for yearly series with training sets comprising 9 observations or less only, the set of candidate models decreases from 6 to 4.

The overall treating procedure for ETS is summarized in the flowchart of Fig. 2. In the figure, *J* stands for the 15 ETS model forms estimated by default in `ets()` and *j* are the model forms discarded during the process of treating. PI stands for prediction intervals.

The reasons behind the use of PIs in lieu of PFs to compare and occasionally discard model forms from the pool of ETS formulations is twofold: first, PIs are quicker, in the sense that they require fewer forecasting steps,

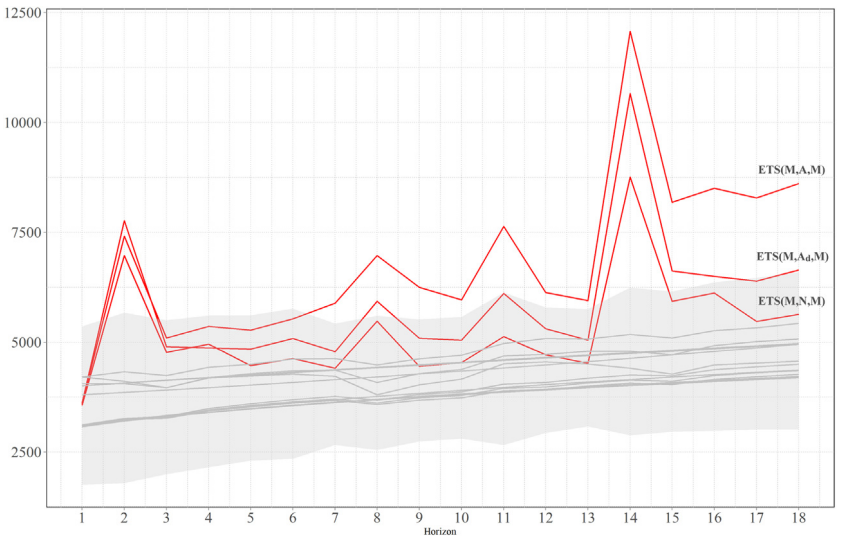

Fig. 1. M4 monthly series 16214 – upper PI limits from 15 ETS model forms.

to indicate explosive patterns in forecasts; second, it can be quite challenging to identify deviant behaviors just by looking at PFs: differences, in relative terms, may not be so big (hampering the task of looking for outliers, for instance); and it is not uncommon to observe forecasts that deviate considerably from the ensemble at specific forecasting steps, usually due to the model from which they were originated, but are still competitive at large forecast lead times.

At this point, a brief explanation on how ETS PIs are constructed is called for. In principle, as pointed out by Hyndman et al. (2008), the most direct method of obtaining prediction distributions for any time series model is to simulate future sample paths from the fitted model and estimate the distributions from the simulated data. However, there are a few drawbacks to this approach. For instance, it does not allow for algebraic analysis of the prediction distributions since PIs are only available numerically rather than algebraically. In addition, computations using simulation routines can be time expensive. To circumvent these issues, Hyndman et al. (2008) derived analytical results on prediction distributions for 15 of the 30 models in their exponential smoothing framework. More specifically, this selected set of models consists of (i) linear models with homoscedastic errors – model forms (A, N, N), (A, N, A), (A, A, N), (A, A, A), (A, A_d, N) and (A, A_d, A); (ii) linear models with heteroscedastic errors – model forms (M, N, N), (M, N, A), (M, A, N), (M, A, A), (M, A_d, N) and (M, A_d, A); and (iii) specific nonlinear model forms which are similar to the seasonal models in (ii) with the exception that the seasonal component is now multiplicative rather than additive – model forms (M, N, M), (M, A, M) and (M, A_d, M). These are also the model forms considered by default in the automated model selection procedure implemented in the `ets()` function from the *forecast* package in R. For these model forms, a Gaussian distribution is assumed for the vector of innovations (errors) and the forecast variance has a closed formula, which depends on the ETS model form. In other words, $100(1-\alpha)\%$ PIs can be calculated in the usual way, i.e.:

$$\mu_{n+h|n} \pm z_{\alpha/2}\sqrt{v_{n+h|n}} \tag{1}$$

where z_q denotes the qth quantile of a standard Gaussian distribution, and $\mu_{n+h|n}$ and $v_{n+h|n}$ are the h steps ahead estimates of the forecast mean and variance, for which closed formulae are available. To conserve space, we refer the interested reader to Hyndman et al. (2008), Chapter 6, for the expressions and derivations on each ETS model form. Finally, we add that prediction distributions for classes (ii) and (iii), outlined in the previous paragraph, are non-Gaussian because of the nonlinearity of the state space equations. However, Hyndman et al. (2008) argue that PIs based on the above (Gaussian) formula will usually give reasonably accurate results for these classes.

3.1.1. Treating in ETS model combination

In this section, we demonstrate how treating can be extended to ETS model combination routines. To that end, we use as benchmark the approach proposed by Kolassa (2011). More specifically, we consider the 'AICc-weighted average' variation. Point forecasts according to this approach are calculated as weighted averages of PFs from all smoothing models investigated, with weights corresponding to the differences from each model AICc to the minimal AICc model. The same rationale is used to construct the PIs. The weights are computed according the following equation, here transcribed in terms of AICc for the benefit of the reader:

$$w_{AICc}(M) = \frac{\exp\left(-\frac{1}{2}\Delta_{AICc}(M)\right)}{\sum_{N\in\eta}\exp\left(-\frac{1}{2}\Delta_{AICc}(N)\right)} \tag{2}$$

where $\Delta_{AICc}(M) = AICc(M) - \min_{N\in\eta}\Delta_{AICc}(N) \geq 0$, η being the number of models (ETS model forms) considered. If no treating is conducted, the 15 ETS model forms estimated by default in the `ets()` function from the

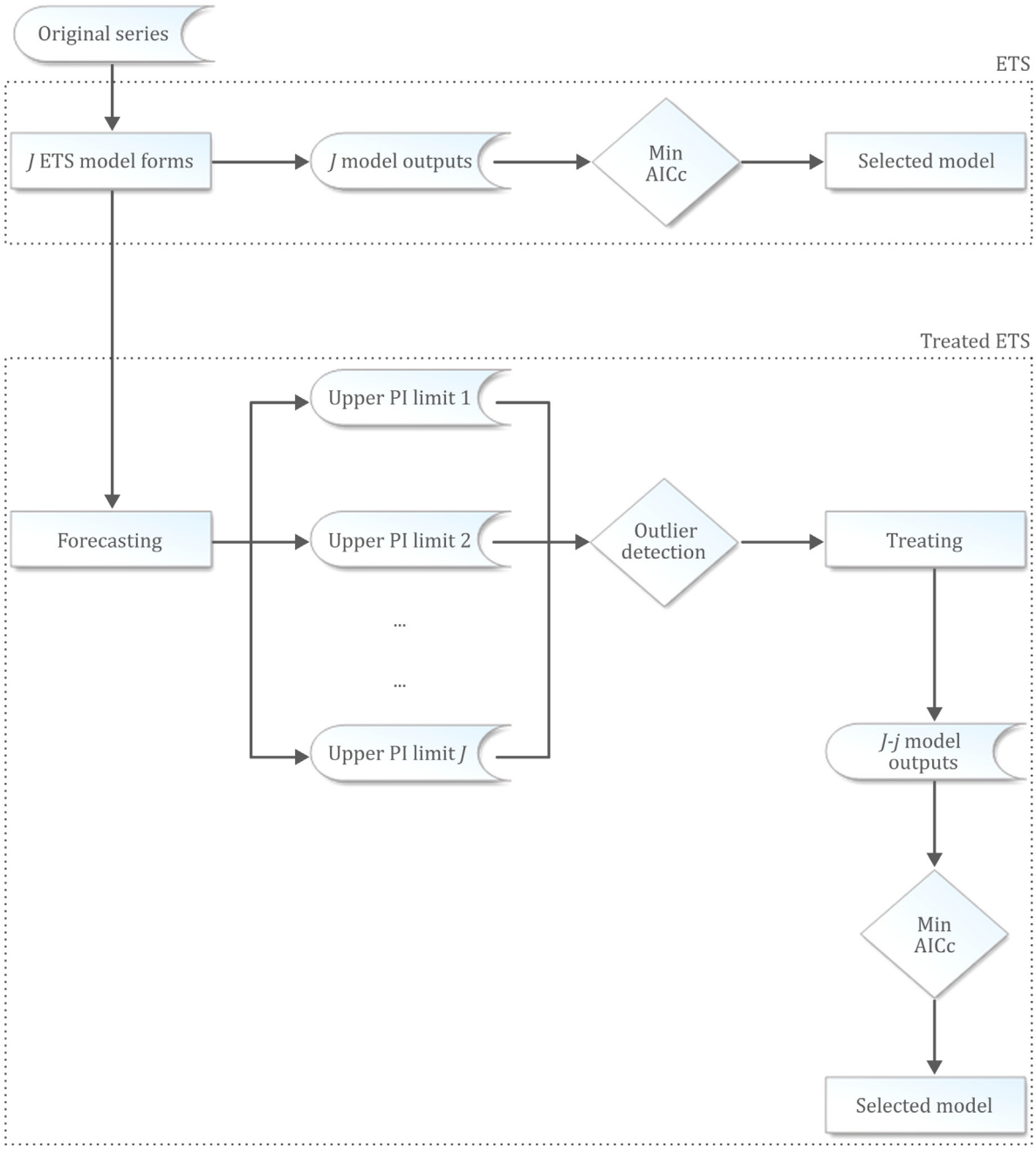

Fig. 2. ETS, as implemented in the *forecast* package in R, and Treated ETS for model selection.

forecast package in R are considered. In the treated version, henceforth addressed as 'Treated AICc weights', in turn, we consider the AICc-weighted average of the remaining ETS models forms after treating, i.e., we combine only the PFs and corresponding PIs from the model forms that were not excluded during the outlier detection phase in treating.

Finally, we highlight that treating could also be applied to the other variations proposed by Kolassa (2011). This could be implemented by changing AICc in Eq. (2) to other model selection measures, such as AIC or BIC.

3.2. Pruning in model combination

As previously outlined, the rationale behind pruning is quite similar to treating. The main difference is that now we aim at subsetting the pool of forecasts to be combined, since some of them may deviate considerably from the rest of the ensemble. Therefore, pruning can be viewed as a feature selection strategy to improve the quality of PIs and PFs of any forecast combination method, as long as the methods to be combined provide PIs along with their PFs. To demonstrate the potential of pruning, in this

paper we opt to apply this procedure to some benchmark Bagging strategies, given their popular use and flexibility to encompass different forecast methods for the ensemble of bootstraps. Particularly, we aim at improving PFs and PIs originating from two different approaches discussed in Section 2.2: the Bagged.BLD.MBB.ETS method proposed by Bergmeir et al. (2016); and the BMC devised by Petropoulos et al. (2018). It should be noted that both approaches were developed with the focus of improving the accuracy of PFs. Therefore, extending their fields of application to generate PIs can also be viewed as a novelty in the paper. In the next subsections, we demonstrate how the two selected Bagging strategies can be used to generate PIs and how pruning can be applied to such cases.

3.2.1. Prediction intervals in bagging

The strategies proposed to generate the PIs in Bagging are built using the same core ideas for the PFs. In the case of Bagged.BLD.MBB.ETS (henceforth 'Bagged ETS' to conserve space), besides aggregating the PFs, we also combine their corresponding PIs using the median. This is possible because, besides the PFs, the `forecast()` function from the *forecast* package, when applied to an ETS model, also generates their corresponding PIs, with a theoretical coverage level set by the practitioner. For instance, if a 95% coverage is aimed for, a PI is generated using the 2.5% quantile as lower limit and the 97.5% quantile for the upper limit. The reader is referred to Hyndman et al. (2008) for details on how the quantiles are computed in ETS formulations.

Let J be the number of forecasts involved in the ensemble (forecasts of the original data and the $J-1$ bootstraps generated). That way, the upper and lower limits in Bagged ETS are obtained as follows:

$$\begin{aligned} U_{t,BaggedETS} &= \text{median}\left[U_{t,1},\ldots,U_{t,J}\right] \\ L_{t,BaggedETS} &= \text{median}\left[L_{t,1},\ldots,L_{t,J}\right] \end{aligned} \tag{3}$$

where $U_{t,1},\ldots,U_{t,J}$ and $L_{t,1},\ldots,L_{t,J}$ are respectively the upper and lower limits of the J PFs in the ensemble. Eq. (3) is applied for every step in the forecast lead time, i.e., $t = 1,\ldots,h$, h being the total number of steps required.

As for BMC, we take a weighted average of the PIs generated from applying the 'unique' ETS model forms on the original data, with weights defined by the frequency that the unique models were identified as optimal. Let K be the number of unique ETS model forms identified among the ensemble of J forecasts. Hence, for each forecast lead time, i.e., $t = 1,\ldots,h$, once again h being the total number of steps required, the upper and lower limits of the BMC PI are obtained according to the following equation:

$$\begin{aligned} U_{t,BMC} &= \sum_{i=1}^{K} w_i\, U_{t,i} \\ L_{t,BMC} &= \sum_{i=1}^{K} w_i\, L_{t,i} \end{aligned} \tag{4}$$

where w_i, $i = 1,\ldots,K$, are the weights of the K unique model forms, and $U_{t,i}$ and $L_{t,i}$ are upper and lower limits of their corresponding PIs.

Fig. 3 illustrates how Bagged ETS and BMC can be used to generate both Bagged PFs and PIs. The figure also foreshadows how pruning can be achieved in each of these strategies (see the next section for details). Bagged ETS aggregates the J PFs and their J corresponding PIs using their medians. BMC, in turn, identifies from the J forecasts the K unique ETS model forms and applies them to the original series. Then, it combines the results from K PFs (and corresponding PIs) using as weights the frequency with which the unique forms were identified as optimal, i.e., the amount of times they were selected divided by J.

3.2.2. Pruning for bagging

Relating once again to Fig. 3, pruning for Bagged ETS and BMC first considers merging the forecasts and PIs from both ensembles, ending up with $J + K$ PFs (and corresponding PIs). This is often recommended in light of the small amount of forecasts comprising the BMC ensemble, which sometimes renders impossible the detection of outliers. By merging the PFs (and corresponding PIs) from both ensembles, it becomes easier to detect and remove unwanted outputs from the BMC ensemble using the $\pm 1.5 \times IQR$ 'rule'. This may also prove beneficial to the Bagged ETS approach since, with the exception of the first forecast from both ensembles (which is the same since it is produced from the original data), the K added forecasts from the BMC ensemble can differ considerably from the J forecasts, bringing more diversity to the merged ensemble and ultimately making outlier detection more effective. Even though pruning is conducted on the merged $(J+K)$ ensemble of forecasts, the final results are separated between Bagged ETS and BMC. In other words, after pruning, the resulting ensembles now encompass $J - j_1$ and $K - k_1$ forecasts (respectively for Bagged ETS and BMC), where j_1 and k_1 are the removed forecasts from each ensemble. Beyond this point, the Bagged ETS and BMC routines proceed as usual.

Pruning can be conducted as many times as desired, until no outliers can be identified in the resulting ensemble. Depending on the case, pruning twice may lead to better results than pruning just once. The gains, however, were not too significant in our empirical tests with the M-Competitions and were usually detected for PIs only, with a slight loss in accuracy for the PFs. We finally note that further pruning (three times or more) frequently led to less accurate results, both in terms of PFs and PIs.

3.2.3. Pruning for alternative combination approaches

In the previous section, we demonstrated some of the possible ways that pruning can be applied to improve the accuracy of Bagging approaches, which can be viewed as a class of forecast combination methods. The two selected approaches were considered in agreement with the recent literature on Bagging for forecasting. Both methods were originally developed considering the auto ETS model form selection algorithm implemented in the `ets()` function from the *forecast* package in R as the forecast method to generate the PFs and PIs for the bootstraps. We argue, though, that Bagging approaches do not require that the bagged forecasts are originated from automated ETS routines. For instance, the ETS forecasts and PIs which

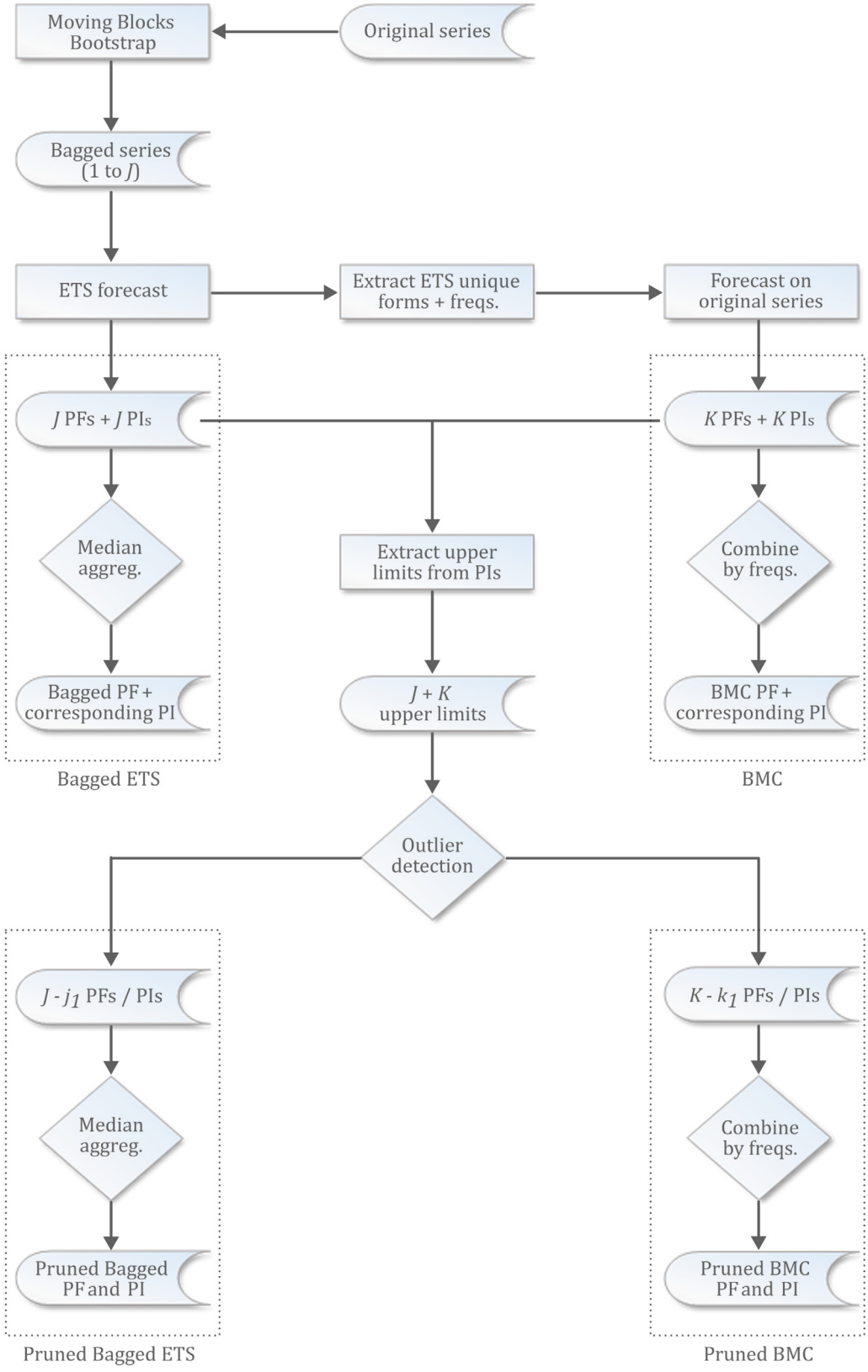

Fig. 3. Bagged ETS and BMC and their pruned versions.

were combined by the median in the last stage of the first Bagging approach (Bagged ETS) could have been generated, for instance, by means of Autoregressive Integrated Moving Average (ARIMA) formulations. In other words, an auto ARIMA model selection routine could be applied to the original series and its bootstraps and the estimated models could be used to produce the forecasts. Since some automatic ARIMA model selection algorithms, such as the `auto.arima()` function from the *forecast* package, generate PIs along with PFs, pruning could be conducted using the same steps proposed in Section 3.2.2.

As for BMC, a similar approach could have been developed for bagged forecasts generated through the auto ARIMA algorithm. In this case, the `auto.arima()` function would be applied to the original data and its bootstraps and the set of unique of ARIMA formulations, selected as best model in each series, would be collected. These formulations would then be applied to the original series and the forecasts generated from the estimated models would be combined using a weighted average, with weights reflecting the number of times each ARIMA formulation was selected as best model among the ensemble of bootstraps. This approach was also proposed by Petropoulos et al. (2018) as a robustness check for the BMC strategy.

Besides Bagging, pruning could also be applied to a wide range of forecast combination methods. For instance, one could consider a simple average among the forecasts originated from selected ETS and ARIMA model forms, or even a combination of ETS, ARIMA and Neural Network forecasts. Provided that each method considered in the combination scheme is able to generate both PFs and PIs, pruning could be implemented. The application of pruning to the above variants of Bagging and to the alternative forecast combination schemes here proposed is beyond the scope of the present work but constitutes a future research direction.

Finally, in spite of the potential benefits brought by pruning, we emphasize that combining models from vastly different natures and formulations may not be the best forecasting approach to take in practice, particularly when several stylized facts of the time series are well known to the forecasting practitioner. It should also be noted that sometimes methods with poor PIs, which will supposedly be filtered more in pruning, may still provide accurate and useful PFs.

4. Empirical investigation

To assess the accuracy of our developed strategies and at the same time provide a common ground for discussion with previous related works, we use databases from three well-known forecasting competitions, the M, M3 and M4 Competitions (Makridakis et al., 1982; Makridakis & Hibon, 2000; Makridakis et al., 2018). We restrict our attention to yearly (181 + 645 + 23,000 = 23,826 series), quarterly (203 + 756 + 24,000 = 24,959 series) and monthly (617 + 1,428 + 48,000 = 50,045 series) data, which are the most used frequencies in practice and also in previous works concerning Bagging approaches (Bergmeir et al., 2016; Dantas & Cyrino Oliveira, 2018; De Oliveira & Cyrino Oliveira, 2018; Petropoulos et al., 2018). The predictive power of the proposed approaches is assessed using the same amount of out-of-sample data suggested in the competitions (6 observations for yearly series, 8 for quarterly and 18 for monthly), to allow comparability with published results. To gauge the accuracy of the developed strategies, we opted to summarize the results according to the following metrics:

- For PFs: the average and median of Mean Absolute Scaled Errors (MASEs) and symmetric Mean Absolute Percentage Errors (sMAPEs);
- For PIs: the average and median of Mean Scaled Interval Scores (MSISs).

The MASE, sMAPE and MSIS are defined as given in Box I, where Y_t and $\hat{Y}_t$ are the actual and forecasted values of the underlying series, respectively; t is the forecast lead time from 1 to h steps ahead; m is the seasonal period; U_t and L_t are the upper and lower limits of the PI produced using the selected method; and $1-\alpha$ is the desired (theoretical) coverage level. By introducing penalties for the width $(U_t - L_t)$ and for the instances where the actual values are outside the specified bounds of the predicted interval, the MSIS offers a good balance between spread and coverage (hit rates).

The choice for the above-mentioned metrics was mainly to allow comparability with published results. It should also be noted that these are the official evaluation metrics for PFs and PIs in the M4 Competition (Makridakis et al., 2018). Apart from depicting the results in terms of the average and median of the above-mentioned metrics, we have also conducted the Multiple Comparisons with the Best (MCB) test (Koning et al., 2005) to assess whether the differences between the error measures were statistically significant.

We summarize the results of the empirical experiments conducted across all competitions in Section 4.1 and consider each M-Competition separately in Section 4.2. The statistical tests of the results using the MCB test are discussed in Section 4.3. We further provide comparison with previously published results in Section 4.4.

Finally, we clarify that we used R version 3.4.3 (2017-11-30) and *forecast* package version 8.6 for ETS modelling. A parallel implementation was adopted in our code, where the following packages were used: *doSNOW* (1.0.16), *foreach* (1.4.4) and *snow* (0.4–3). In addition, we directly accessed data from the M-Competitions through packages *MComp* (version 2.8, for M and M3) and *M4comp 2018* (version 0.1.0, for M4). The `nemenyi()` function from the *tsutils* package (0.9.0) was employed to conduct the Multiple Comparisons with the Best (MCB) test.

4.1. Empirical findings across competitions

Tables 2 and 3 summarize the average and median (across all series) MASE and sMAPEs results, respectively, for PF accuracy evaluation. Table 4, in turn, illustrates the average and median MSIS results for PI accuracy evaluation. Following the M4 Competition guidelines, the PIs

$$MASE = \frac{1}{h}\frac{\sum_{t=1}^{h}\left|Y_t - \hat{Y}_t\right|}{\frac{1}{n-m}\sum_{t=m+1}^{n}|Y_t - Y_{t-m}|} \tag{5}$$

$$sMAPE = \frac{1}{h}\sum_{t=1}^{h}\frac{2\left|Y_t - \hat{Y}_t\right|}{|Y_t| + \left|\hat{Y}_t\right|} \times 100\% \tag{6}$$

$$MSIS = \frac{1}{h}\frac{\sum_{t=1}^{h}(U_t - L_t) + \frac{2}{\alpha}(L_t - Y_t)\,\mathbf{1}\{Y_t < L_t\} + \frac{2}{\alpha}(Y_t - U_t)\,\mathbf{1}\{Y_t > U_t\}}{\frac{1}{n-m}\sum_{t=m+1}^{n}|Y_t - Y_{t-m}|} \tag{7}$$

Box I.

Table 2
All competitions: Average and median MASE of the different forecasting methods (best in **bold**, second best in *italic*).

Method	Average MASE (Monthly)	Median MASE (Monthly)	Average MASE (Quarterly)	Median MASE (Quarterly)	Average MASE (Yearly)	Median MASE (Yearly)
Exponential smoothing						
ETS	0.947	0.736	1.165	0.887	3.431	2.315
Treated ETS	0.939	0.732	1.161	0.887	3.395	2.297
Exponential smoothing forecast combinations						
AICc weights	0.939	0.730	1.150	0.875	3.373	2.290
Treated AICc weights	0.931	0.726	1.147	0.875	3.343	2.273
Bagging strategies						
Bagged ETS	0.955	0.732	1.180	0.899	3.286	2.292
Bagged Treated ETS	0.953	0.731	1.179	0.899	3.284	2.290
BMC ETS	0.925	0.722	1.146	*0.873*	3.323	2.265
BMC Treated ETS	0.922	0.720	1.145	**0.873**	3.318	2.263
Bagging with pruning						
Pruned Bagged ETS	0.955	0.732	1.181	0.898	3.288	2.292
Pruned Bagged Treated ETS	0.953	0.730	1.179	0.899	3.287	2.293
Pruned BMC ETS	*0.920*	*0.719*	*1.144*	0.874	*3.236*	*2.219*
Pruned BMC Treated ETS	**0.917**	**0.718**	**1.143**	0.874	**3.228**	**2.212**

Notes: AICc weights stand for the AICc-weighted average of forecasts from ETS model forms, as proposed by Kolassa (2011). BMC ETS stands for the BMC method devised in Petropoulos et al. (2018). We use the former notation to differentiate it from the BMC Treated ETS, which is the BMC applied to the forecasts generated by using the Treated ETS routine proposed in Section 3.1 on the bootstraps.

were constructed for a desired coverage level (hit rate) of 95%.

For comparison purposes, we contrast the results obtained from the following methods:

(i) The auto state space exponential smoothing (ETS) approach, i.e., the forecasts obtained by selecting the best ETS specification for the original series and subsequently using it for forecasting. In practical terms, this can be achieved by using default options of the `ets()` function from the R *forecast* package on the training set and further forecasting from the selected (estimated) model using the `forecast()` function. Despite its simplicity when compared with other forecasting methods, the ETS provides a sound base for comparison with the proposed Bagging approaches, since they also use ETS models to build the forecasts. In addition, it should be noted that ETS ranked third best overall in terms of closeness to an expected (desired) of 95% hit rate and fourth best in terms of lower MSIS, when all (100,000) series from the M4 Competition were considered (Grushka-Cockayne & Jose, 2019);
(ii) The Treated ETS approach presented in Section 3.1;
(iii) The AICc-weighted average proposed by Kolassa (2011). According to this approach, PFs are calculated as weighted averages of PFs from all smoothing models investigated, with weights corresponding to the differences from each model AICc to the minimal AICc model. The same rationale is used to construct the PIs;
(iv) The treated version of the above method, i.e., the AICc-weighted average of the remaining ETS models forms after treating;

Table 3
All competitions: Average and median sMAPE of the different forecasting methods (best in **bold**, second best in *italic*).

Method	Average sMAPE (Monthly)	Median sMAPE (Monthly)	Average sMAPE (Quarterly)	Median sMAPE (Quarterly)	Average sMAPE (Yearly)	Median sMAPE (Yearly)
Exponential smoothing						
ETS	13.561	7.131	10.331	5.627	15.425	9.059
Treated ETS	13.388	7.117	10.281	5.627	15.244	8.992
Exponential smoothing forecast combinations						
AICc weights	13.441	7.043	10.183	5.545	15.221	8.924
Treated AICc weights	13.274	7.032	10.148	5.547	15.072	8.862
Bagging strategies						
Bagged ETS	13.330	7.392	10.274	5.760	*14.571*	8.709
Bagged Treated ETS	13.305	7.373	10.253	5.755	**14.563**	8.713
BMC ETS	13.107	7.018	10.117	*5.534*	15.014	8.787
BMC Treated ETS	13.059	7.022	**10.096**	**5.525**	15.003	8.771
Bagging with pruning						
Pruned Bagged ETS	13.325	7.367	10.284	5.770	14.603	8.742
Pruned Bagged Treated ETS	13.302	7.353	10.263	5.770	14.593	8.747
Pruned BMC ETS	*13.032*	*7.010*	10.127	5.554	14.631	*8.595*
Pruned BMC Treated ETS	**12.983**	**7.008**	*10.101*	5.555	14.602	**8.568**

Notes: See Table 2.

(v) The Bagged.BLD.MBB.ETS method proposed by Bergmeir et al. (2016) (here abbreviated to 'Bagged ETS');
(vi) The Bootstrap Model Combination (BMC) approach, as proposed in Petropoulos et al. (2018);
(vii) The selected Bagging strategies using, as forecasting method for the original data and the bootstraps, the Treated ETS in lieu of ETS; and,
(viii) The pruning strategy proposed in Section 3.2.2 applied to all Bagging strategies.

Both average and median results of the point and PI forecast evaluation, in Tables 2–4, indicate that Treated ETS provides more accurate results than ETS, with the former outperforming the latter in every case, regardless of the frequency of the time series or the evaluation scenario. This makes a new contribution to the literature, since the automatic ETS routine, as implemented in the `ets()` function from the *forecast* package for R, has been considered the benchmark for automatic model selection among competing ETS model forms and subsequent forecasting. The same rationale applies to forecast combination approaches: The 'Treated AICc weights' approach comfortably outperforms the AICc-weighted average proposed by Kolassa (2011) in all cases, with only one exception: the median sMAPE across quarterly series. The difference between the results in this case, though, is rather insignificant (5.545 versus 5.547). This demonstrates that the proposed treating procedure is capable of improving both PFs and PIs for approaches involving model selection (ETS) and model combination (AICc-weighted averages).

In an attempt to find the main reasons behind the superiority of Treated ETS over ETS (and consequently Treated AICc weights over AICc weights), we analyzed the frequencies with which each ETS model form was selected in the automated ETS routine, before and after treating. The results of this analysis are detailed per frequency and per competition in the online supplement, Section A (Appendix A). A noteworthy pattern common to all series frequencies was the fact that the ETS(M, A, N) model form tends to be selected less often when treating is conducted. The results suggest that this model form is usually replaced by its additive seasonality version – ETS(M, A, A) – in the case of monthly and quarterly series, or by its additive damped trend version – ETS(M, A_d, N) – in yearly series (since no additive seasonal models are considered for this frequency). Another interesting feature in some datasets is the increased preference, after treating, towards certain additive-damped trend model forms, such as ETS(M, A_d, A) and ETS(M, A_d, M) for monthly and quarterly series in M4, and ETS(M, A_d, N) for yearly series. Finally, we note that model forms with multiplicative errors and additive trends are, by far, the ones which are most considered as outliers when treating is applied to ETS in monthly and quarterly series. The opposite can be stated in yearly series, with the simplest model forms, i.e., those involving no trend and seasonal components — ETS(A, N, N) and ETS(M, N, N) — ranking first in terms of times considered as outliers during treating.

Turning once again to Tables 2 and 3, we note that, in addition to the benefits of treating for ETS and AICc-weighted averages, Bagging routines also deliver more accurate PFs when Treated ETS is considered as the forecast method to generated the bagged forecasts, as opposed to the traditional ETS routine implemented in the `ets()` function of R *forecast* package. The benefits arising from pruning the Bagging approaches are also highlighted in terms of PFs, evaluated in Tables 2 and 3, and PIs, evaluated in Table 4.

Turning the attention to the MSIS values only, as shown in Table 4, we note a major issue with using the BMC strategy to generate PIs for monthly time series. When

Table 4
All competitions: Average and Median MSIS, computed at the 95% desired coverage level, of the different forecasting methods (best in **bold**, second best in *italic*).

Method	Average MSIS (Monthly)	Median MSIS (Monthly)	Average MSIS (Quarterly)	Median MSIS (Quarterly)	Average MSIS (Yearly)	Median MSIS (Yearly)
Exponential smoothing						
ETS	8.258	5.027	9.587	5.955	34.970	15.369
Treated ETS	*8.133*	4.963	9.513	5.942	34.466	15.266
Exponential smoothing forecast combinations						
AICc weights	8.167	5.010	9.385	5.941	33.123	14.868
Treated AICc weights	**8.054**	4.931	**9.326**	5.922	32.901	14.790
Bagging strategies						
Bagged ETS	8.699	4.610	9.746	5.528	36.948	14.350
Bagged Treated ETS	8.662	*4.585*	9.724	*5.517*	36.957	14.314
BMC ETS	3.301×10^{11}	4.920	10.355	5.959	32.825	14.392
BMC Treated ETS	6.603×10^{11}	4.898	10.374	5.958	32.633	14.354
Bagging with pruning						
Pruned Bagged ETS	8.727	4.598	9.780	5.523	37.276	14.372
Pruned Bagged Treated ETS	8.693	**4.575**	9.759	**5.513**	37.286	14.368
Pruned BMC ETS	8.342	4.854	*9.345*	5.880	*32.317*	*14.049*
Pruned BMC Treated ETS	8.370	4.826	9.392	5.874	**32.211**	**14.029**

Notes: See Table 2.

no pruning is conducted, regardless of the forecasting method selected for the bootstraps (ETS or Treated ETS), BMC generates very large PIs for some series, resulting in very high overall MSIS values. Upon closer inspection, we note that this issue arises in certain series from the M4 Competition, which display notable structural breaks and/or outliers in the training set. As a consequence, some bootstraps will be generated with extremely large values. ETS model forms for such bootstraps are not optimal for the original series, but they are applied to the latter according to how the BMC algorithm is designed. These model forms will usually generate very large PIs since they contain multiplicative errors and are applied to the original series, which already contains a large range of values. This is illustrated in Fig. 4, which shows the training set ensemble – original series and its corresponding moving blocks bootstraps (MBBs) – of the monthly series 41895 from the M4 Competition, and in Table 5, which reports on the selected model forms in BMC for the same series, along with the upper limits of the PIs generated when such model forms are applied to the original series.

The selected ETS form for the original M4 monthly series 41895 is (A, N, N), a combination of additive errors, no trend and no seasonality. Alternative ETS formulations with additive errors, (A, A, N) and (A, N, A), produce relatively similar values for the upper limits of the PIs, when compared with formulations with multiplicative errors. Irrationally high values for the PIs are noted when the model form involves an additive trend and no seasonality (M, A, N), with the values for the upper limits being more than 10^{16} times higher than the actual (real) values at the last forecasting step. Such irrational behavior is considerably dampened when model forms considers a multiplicative seasonality (M) and/or an additive damped trend (A_d). By conducting pruning following the steps depicted in Section 3.2.2, we were able to discard the forecasts (and corresponding PIs) from the last two ETS formulations – (M, N, M) and (M, A, N) – before proceeding to combination. As a result, the MSIS value decreased substantially, from 1.65×10^{16} with no pruning to 24.10 with pruning. The same pattern is repeated in several cases: the BMC generates very large PIs for at least 15 monthly series from the M4 Competition, and relatively high MSIS values – when compared to ETS, for instance – for more than 100 series. In most cases, a substantial reduction in MSIS is achieved by conducting the proposed pruning strategy.

Turning once again to the overall MSIS results in Table 4, we note that the effect of pruning is substantial for BMC formulations but not for Bagged ETS. This is due to the fact that the latter aggregates the bagged forecasts using the median, which diminishes the effect of the outliers in the ensemble. BMC, in turn, takes a weighted average of the forecasts and will thus always consider the effect of ETS formulations, which generates very large PIs. However, provided that proper pruning is conducted, BMC strategies are superior on average than Bagged ETS, both in terms of PFs and PIs.

The benefits of treating for ETS model selection and pruning for BMC strategies are also shown in Fig. 5, which depicts the average MSIS values computed at alternative (80% to 99%) PI coverage levels for four different methods. By contrasting the results delivered by ETS (in red) and Treated ETS (in yellow) in Fig. 5, we note that the latter outperforms the former in every case scenario, regardless of the time series frequency or desired coverage level. We also compare the results obtained using BMC (in green) and its pruned version (in blue), illustrating the gains one can achieve by considering pruning in forecast combination approaches.

Finally, in terms of calculation speed, we emphasize that the extra time taken to conduct treating on ETS

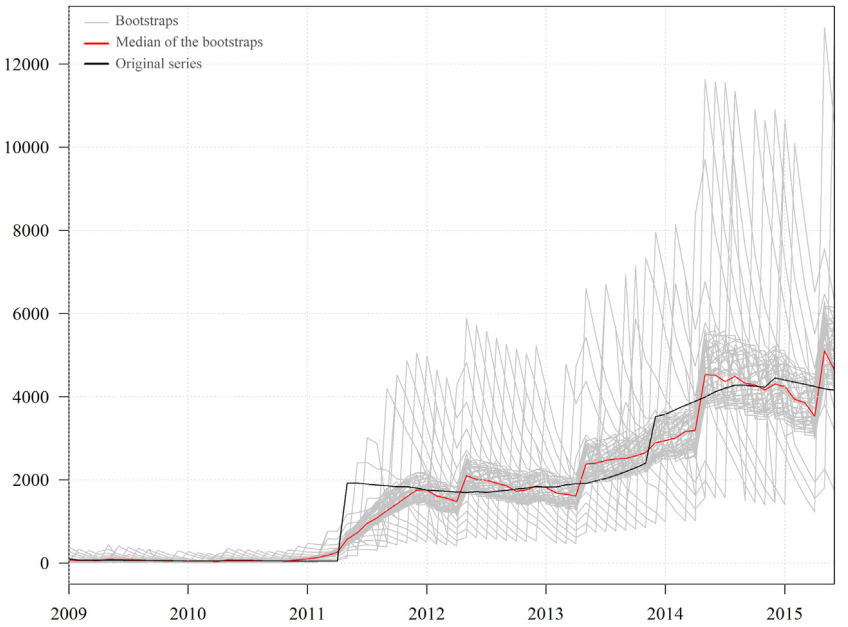

Fig. 4. M4 Competition monthly series 41895, training set: original data in black, bootstraps in gray and median of the bootstraps in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
M4 Competition monthly series 41895, test set: Actual values, selected ETS model forms in the BMC strategy and the corresponding PI upper limits and BMC ETS upper limits before and after pruning.

Lead time	Step 1	Step 2	Step 3	...	Step 16	Step 17	Step 18
Actual values							
out-of-sample	4,119	4,074	4,032	...	3,483	3,443	3,399
ETS forms and upper PI limits							
A, N, N	4,662	4,870	5,029	...	6,164	6,225	6,285
A, N, A	4,607	4,779	4,900	...	6,017	6,034	6,261
A, A, N	4,711	4,969	5,179	...	6,991	7,105	7,217
M, A_d, M	19,094	16,407	16,815	...	16,745	16,054	14,830
M, A_d, N	14,714	14,750	14,784	...	15,131	15,156	15,180
M, A, M	16,476	17,197	15,089	...	14,853	15,388	16,470
M, N, M	13,114	23,203	17,954	...	67,102	60,490	61,591
M, A, N	41,749	363,072	3.38×10^6	...	1.51×10^{19}	1.42×10^{20}	1.33×10^{21}
BMC ETS upper PI limits							
No pruning	15,715	45,060	2.84×10^5	...	1.21×10^{18}	1.13×10^{19}	1.06×10^{20}
With pruning	13,669	13,681	12,662	...	12,876	13,069	13,481

model forms is negligible: while across monthly series from all competitions it took an average of 6.48 s to compute the PFs of the 15 ETS model forms estimated by default in `ets()`, the average time taken to (i) generate the PIs for all possible integer coverage levels (1% to 99%) and (ii) conduct treating on every case – that is, conduct treating 99 times – was 1.29 s. The average extra time to conduct treating on all integer coverage levels for quarterly and yearly series was 0.60 and 0.01 s, respectively. We emphasize that, in most practical applications, the added time to conduct treating will be even less, since the practitioner is usually interested in generating PIs for only one specific coverage level (95%, for instance).

4.2. Empirical findings per competition

To highlight possible differences across competitions and, at the same time, allow straightforward comparison with previously published results, in this section we summarize the PF and PI results per competition considered. The average and median MASEs across M1, M3 and M4 are depicted in Table 6. The average and median sMAPEs are given in Table 7. Finally, the average and median MSISs are presented in Table 8.

Although it is true that the best (most accurate) method may vary across competition, the same patterns observed in Tables 2–4 remain: overall, the treated versions of ETS

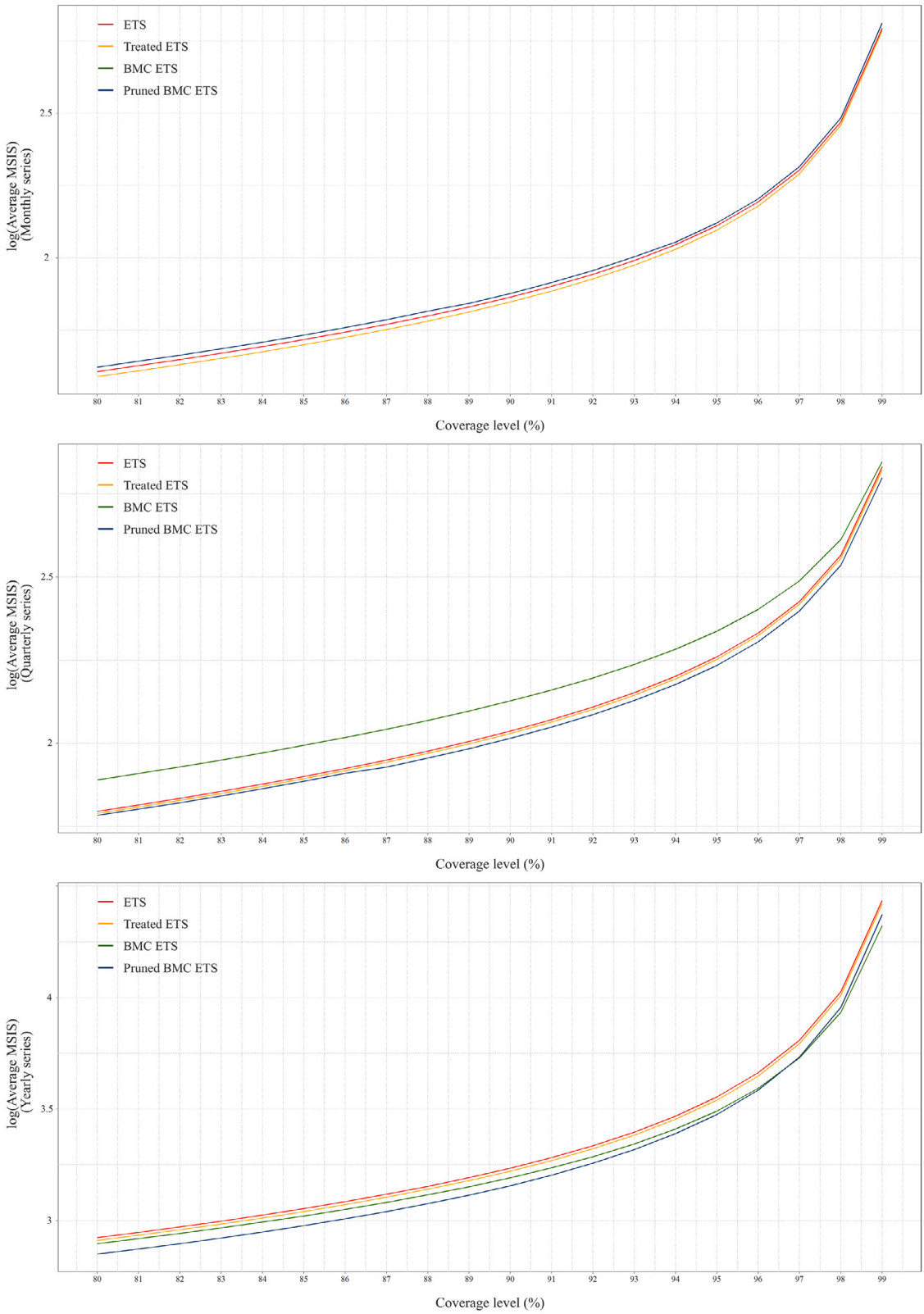

Fig. 5. MSIS per different coverage levels (80% to 99%) for four different methods (y-axis in log scale). MSIS results for BMC ETS (in green) across monthly series are not shown in light of their very high values. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Average and Median MASEs of the different forecasting methods across the M1, M3 and M4 Competitions (best method(s) per competition in **bold**, second best in *italic*).

Method	Average MASE (Monthly)	Median MASE (Monthly)	Average MASE (Quarterly)	Median MASE (Quarterly)	Average MASE (Yearly)	Median MASE (Yearly)
M1						
ETS	1.074	0.851	1.658	1.196	3.771	2.324
Treated ETS	1.066	0.851	*1.623*	1.196	3.757	2.313
AICc weights	1.060	**0.845**	1.625	**1.171**	3.692	2.328
Treated AICc weights	1.054	**0.845**	**1.604**	**1.171**	3.675	2.328
Bagged ETS	1.076	0.852	1.796	1.240	*3.628*	2.388
Bagged Treated ETS	1.075	0.852	1.786	1.240	**3.612**	2.354
BMC ETS	1.055	0.857	1.661	1.201	3.670	**2.264**
BMC Treated ETS	*1.050*	0.854	1.650	1.203	3.660	**2.264**
Pruned Bagged ETS	1.077	0.852	1.799	1.240	3.644	2.375
Pruned Bagged Treated ETS	1.077	*0.850*	1.789	1.240	3.629	2.354
Pruned BMC ETS	1.054	0.853	1.656	*1.173*	3.661	*2.269*
Pruned BMC Treated ETS	**1.049**	0.854	1.645	1.201	3.654	*2.269*
M3						
ETS	0.865	0.712	1.170	0.855	2.860	1.907
Treated ETS	0.859	0.711	1.153	0.851	2.825	1.897
AICc weights	0.849	0.707	*1.137*	0.834	*2.769*	1.887
Treated AICc weights	0.845	0.705	**1.124**	0.819	**2.750**	**1.850**
Bagged ETS	0.838	0.686	1.163	0.853	2.896	1.879
Bagged Treated ETS	**0.834**	*0.681*	1.158	0.850	2.893	1.867
BMC ETS	0.853	0.707	1.146	*0.799*	2.806	1.858
BMC Treated ETS	0.849	0.705	1.141	**0.798**	2.803	*1.857*
Pruned Bagged ETS	0.839	0.690	1.165	0.860	2.894	1.885
Pruned Bagged Treated ETS	*0.835*	**0.681**	1.160	0.858	2.894	1.871
Pruned BMC ETS	0.850	0.695	1.145	0.810	2.815	1.858
Pruned BMC Treated ETS	0.846	0.693	1.140	0.808	2.807	*1.857*
M4						
ETS	0.948	0.736	1.161	0.886	3.444	2.329
Treated ETS	0.940	0.731	1.158	0.886	3.408	2.310
AICc weights	0.940	0.729	1.146	0.874	3.387	2.304
Treated AICc weights	0.932	0.725	1.144	0.875	3.357	2.290
Bagged ETS	0.957	0.732	1.175	0.897	3.294	2.302
Bagged Treated ETS	0.955	0.731	1.174	0.898	3.293	2.302
BMC ETS	0.926	0.722	1.141	*0.873*	3.335	2.280
BMC Treated ETS	0.923	0.719	1.140	**0.872**	3.330	2.276
Pruned Bagged ETS	0.957	0.732	1.176	0.897	3.297	2.303
Pruned Bagged Treated ETS	0.955	0.730	1.175	0.898	3.295	2.305
Pruned BMC ETS	*0.920*	*0.719*	*1.139*	0.874	*3.245*	*2.229*
Pruned BMC Treated ETS	**0.918**	**0.717**	**1.139**	0.874	**3.236**	**2.224**

Notes: See Table 2.

Table 7
Average and Median sMAPEs of the different forecasting methods across the M1, M3 and M4 Competitions (best method(s) per competition in **bold**, second best in *italic*).

Method	Average sMAPE (Monthly)	Median sMAPE (Monthly)	Average sMAPE (Quarterly)	Median sMAPE (Quarterly)	Average sMAPE (Yearly)	Median sMAPE (Yearly)
M1						
ETS	14.971	10.878	17.467	8.404	18.613	13.008
Treated ETS	14.806	10.887	16.535	8.391	18.531	12.592
AICc weights	14.775	10.979	17.090	8.372	18.095	**11.764**
Treated AICc weights	14.642	11.032	**16.310**	8.372	18.005	**11.764**
Bagged ETS	15.042	*10.673*	17.558	9.472	*17.733*	12.694
Bagged Treated ETS	14.994	10.677	17.412	9.472	**17.643**	12.049
BMC ETS	14.579	10.960	16.725	*8.147*	17.970	11.807
BMC Treated ETS	**14.474**	10.768	16.618	*8.147*	17.919	*11.801*
Pruned Bagged ETS	15.041	10.693	17.592	9.283	17.825	12.695
Pruned Bagged Treated ETS	15.002	**10.620**	17.422	9.163	17.741	12.694

(*continued on next page*)

Table 7 (*continued*).

Method	Average sMAPE (Monthly)	Median sMAPE (Monthly)	Average sMAPE (Quarterly)	Median sMAPE (Quarterly)	Average sMAPE (Yearly)	Median sMAPE (Yearly)
Pruned BMC ETS	14.585	10.920	16.637	*8.147*	17.854	12.357
Pruned BMC Treated ETS	*14.486*	10.759	*16.509*	**7.989**	17.831	12.337
M3						
ETS	14.139	9.133	9.684	5.528	17.003	11.518
Treated ETS	14.083	9.094	9.454	5.465	16.744	11.437
AICc weights	13.832	8.933	*9.311*	5.392	*16.436*	*11.075*
Treated AICc weights	13.828	8.903	**9.184**	5.353	**16.307**	**10.934**
Bagged ETS	13.686	**8.788**	9.780	5.726	17.324	11.597
Bagged Treated ETS	**13.597**	*8.790*	9.723	5.769	17.282	11.533
BMC ETS	13.873	8.854	9.528	**5.344**	16.871	11.666
BMC Treated ETS	13.798	8.828	9.476	**5.344**	16.790	11.378
Pruned Bagged ETS	13.697	8.883	9.808	5.713	17.380	11.577
Pruned Bagged Treated ETS	*13.616*	8.858	9.744	5.682	17.362	11.614
Pruned BMC ETS	13.803	8.914	9.590	*5.349*	16.835	11.250
Pruned BMC Treated ETS	13.726	8.915	9.538	*5.349*	16.771	11.186
M4						
ETS	13.525	6.995	10.291	5.608	15.356	8.966
Treated ETS	13.349	6.979	10.254	5.605	15.176	8.900
AICc weights	13.413	6.915	10.152	5.529	15.164	8.848
Treated AICc weights	13.240	6.907	10.127	5.526	15.014	8.797
Bagged ETS	13.298	7.270	10.228	5.739	*14.469*	8.642
Bagged Treated ETS	13.274	7.252	10.209	5.733	**14.462**	8.642
BMC ETS	13.065	6.899	10.080	*5.515*	14.939	8.697
BMC Treated ETS	13.019	6.899	**10.060**	**5.511**	14.930	8.691
Pruned Bagged ETS	13.292	7.264	10.238	5.740	14.500	8.661
Pruned Bagged Treated ETS	13.270	7.242	10.218	5.735	14.491	8.666
Pruned BMC ETS	*12.989*	*6.887*	10.089	5.529	14.544	*8.492*
Pruned BMC Treated ETS	**12.942**	**6.887**	*10.064*	5.531	14.516	**8.467**

Notes: See Table 2.

Table 8
Average and Median MSISs, computed at the 95% desired coverage level, of the different forecasting methods across the M1, M3 and M4 Competitions (best method(s) per competition in **bold**, second best in *italic*).

Method	Average MSIS (Monthly)	Median MSIS (Monthly)	Average MSIS (Quarterly)	Median MSIS (Quarterly)	Average MSIS (Yearly)	Median MSIS (Yearly)
M1						
ETS	9.625	5.430	21.289	7.385	59.784	16.505
Treated ETS	9.294	5.407	20.701	7.358	59.509	15.236
AICc weights	9.249	5.384	**19.969**	6.996	57.585	**13.215**
Treated AICc weights	**8.982**	5.283	*19.988*	**6.877**	57.725	*13.348*
Bagged ETS	10.702	*4.996*	23.564	7.324	63.008	15.629
Bagged Treated ETS	10.569	**4.986**	23.416	7.389	63.143	15.629
BMC ETS	9.221	5.405	20.764	*6.896*	*56.014*	14.010
BMC Treated ETS	*9.050*	5.333	20.423	*6.896*	**55.652**	14.121
Pruned Bagged ETS	10.836	5.014	23.819	7.283	63.965	15.762
Pruned Bagged Treated ETS	10.720	5.012	23.657	7.439	63.984	15.742
Pruned BMC ETS	9.574	5.287	21.375	7.239	58.018	14.165
Pruned BMC Treated ETS	9.441	5.213	21.185	7.085	57.612	14.343
M3						
ETS	6.342	4.371	10.717	5.259	30.616	11.289
Treated ETS	6.216	4.357	10.640	5.253	28.830	11.190
AICc weights	*6.132*	4.335	*10.342*	5.208	27.118	10.877
Treated AICc weights	**6.051**	4.298	**10.304**	5.208	**26.369**	10.730
Bagged ETS	6.427	4.033	11.492	4.804	32.015	*10.135*
Bagged Treated ETS	6.360	**3.998**	11.382	**4.784**	32.141	**10.120**
BMC ETS	6.203	4.304	10.427	5.238	27.257	10.964
BMC Treated ETS	6.160	4.299	10.388	5.248	*26.992*	10.908

(*continued on next page*)

Table 8 (*continued*).

Method	Average MSIS (Monthly)	Median MSIS (Monthly)	Average MSIS (Quarterly)	Median MSIS (Quarterly)	Average MSIS (Yearly)	Median MSIS (Yearly)
Pruned Bagged ETS	6.473	4.035	11.527	4.797	32.486	10.206
Pruned Bagged Treated ETS	6.404	*4.009*	11.411	*4.791*	32.547	10.136
Pruned BMC ETS	6.225	4.263	10.551	5.187	28.614	10.892
Pruned BMC Treated ETS	6.179	4.251	10.577	5.196	28.514	10.817
M4						
ETS	8.297	5.040	9.452	5.977	34.897	15.487
Treated ETS	*8.175*	4.975	9.383	5.958	34.427	15.370
AICc weights	8.214	5.025	9.265	5.964	33.099	15.011
Treated AICc weights	**8.102**	4.945	**9.205**	5.945	32.889	14.927
Bagged ETS	8.741	4.622	9.574	5.550	36.881	14.474
Bagged Treated ETS	8.706	*4.595*	9.556	*5.541*	36.885	14.406
BMC ETS	3.442×10^{11}	4.935	10.265	5.980	32.799	14.503
BMC Treated ETS	6.884×10^{11}	4.909	10.289	5.976	32.610	14.467
Pruned Bagged ETS	8.767	4.609	9.606	5.545	37.200	14.486
Pruned Bagged Treated ETS	8.735	**4.588**	9.589	**5.537**	37.208	14.465
Pruned BMC ETS	8.389	4.864	*9.205*	5.905	*32.219*	*14.160*
Pruned BMC Treated ETS	8.422	4.838	9.254	5.897	**32.114**	**14.140**

Notes: See Table 2.

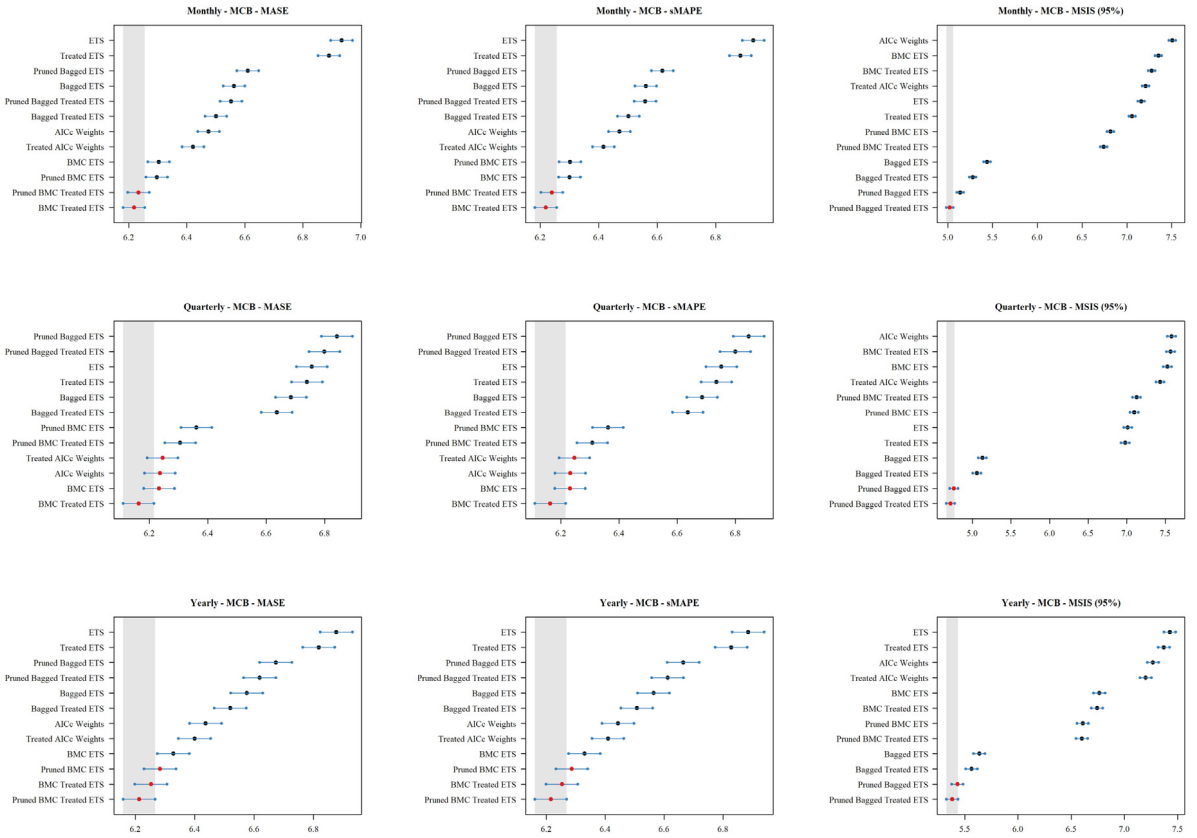

Fig. 6. Multiple comparisons with the best for MASE, sMAPE and MSIS values per frequency.

and AICc weights are superior than the original methods; Bagging approaches benefit from the use of Treated ETS as forecasting method for the bootstraps, in lieu of the traditional ETS; and pruning usually improves the accuracy of Bagging, particularly BMC, approaches.

4.3. Statistical tests of the results

We now turn our attention to the results of the Multiple Comparisons with the Best (MCB) tests, illustrated in Fig. 6. These are also reported in table format in the

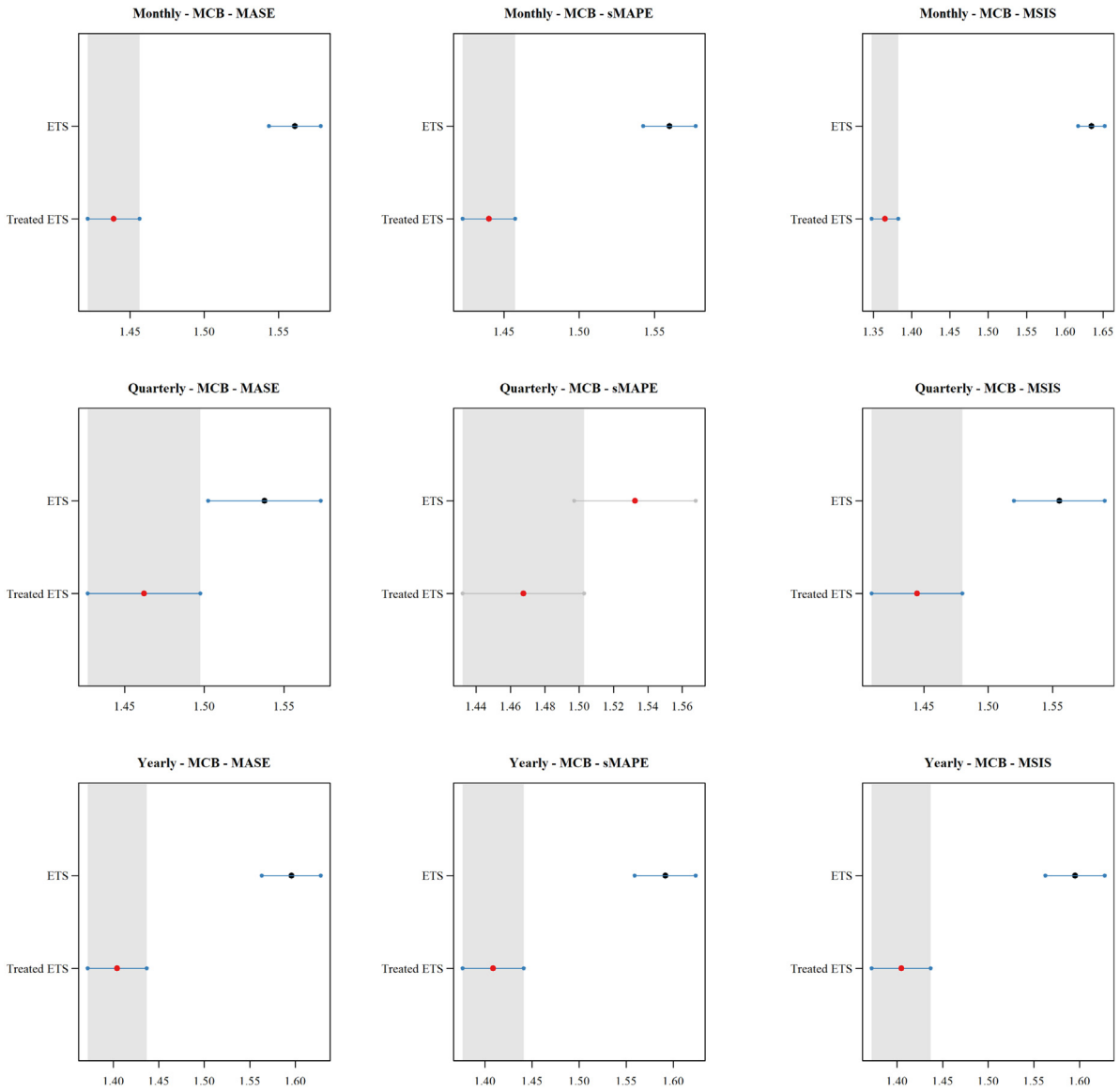

Fig. 7. Multiple Comparisons with the Best (MCB) – ETS vs Treated ETS – only series for which the selected ETS form changed after treating.

online supplement, Section B (Appendix A). The results in terms of average MASE and sMAPE ranks are, to a greater extent, in line with the results from Tables 2 and 3, with BMC Treated ETS and Pruned BMC Treated ETS depicted as the best methods and statistically significant from the others. The only exception was for quarterly series, where Pruned BMC Treated ETS ranked fifth and was considered different from BMC ETS and BMC Treated ETS. Average MSIS ranks, however, tell a slightly different history from the average MSIS values illustrated in Table 4. Pruned Bagged Treated ETS now ranks as the best overall in terms of average MSIS rank, in every case considered (monthly, quarterly or yearly series). An explanation lies in the fact that Pruned Bagged Treated ETS is usually the best method across the series, but when Bagging strategies fail in generating accurate and calibrated PIs for their PFs, Pruned Bagged Treated ETS usually delivers worse results than Pruned BMC Treated ETS, which is usually among the best methods in terms of average MSIS, as depicted in Table 4. Even so, the results from Tables 2–4 and Fig. 6 make it clear that Treated ETS consistently outperforms ETS, both in terms of Point Forecasts and Prediction Intervals, and that pruned Bagging strategies are usually more accurate than their traditional versions.

Finally, in terms of MCB test results in Fig. 6, where the whole set of series from the competitions is compared, it is unlikely to observe any significant statistical difference between the original ETS model selection routine and its treated version, Treated ETS. This is reasonable because, despite the ability of treating in identifying outliers, in most cases the 'best' model form selected in the original ETS routine will still be among the remaining model forms after treating. Therefore, it will still be the 'best' model form in the Treated ETS version. However,

Table 9
M3: Average sMAPEs of the proposed approaches and of the four best methods at the time of the competition.

Method	Average sMAPE (Monthly)	Average sMAPE (Quarterly)	Average sMAPE (Yearly)
Exponential smoothing			
ETS	14.14	9.68	17.00
Treated ETS	14.08	9.45	16.74
Exponential smoothing forecast combinations			
AICc weights	13.83	9.31	16.44
Treated AICc weights	13.83	*9.18*	**16.31**
Bagging strategies			
Bagged ETS	13.69	9.78	17.32
Bagged Treated ETS	**13.60**	9.72	17.28
BMC ETS	13.87	9.53	16.87
BMC Treated ETS	13.80	9.48	16.79
Bagging with pruning			
Pruned Bagged ETS	13.70	9.81	17.38
Pruned Bagged Treated ETS	*13.62*	9.74	17.36
Pruned BMC ETS	13.80	9.59	16.84
Pruned BMC Treated ETS	13.73	9.54	16.77
Best methods from the M3 competition			
1st best method	13.85	**8.96**	*16.42*
2nd best method	13.86	9.22	16.48
3rd best method	14.83	9.33	16.52
4th best method	15.03	9.36	16.90

Notes: Best (most accurate) approach in **bold**, second best in *italic*. In M3, the best forecast methods in terms of average sMAPE for monthly data were, from best to worst: Theta, ForecastPro, ForecastX and Comb S-H-D. For quarterly data, the best methods were: Theta, Comb S-H-D, Dampen and PP-autocast. For yearly data: RBF, ForecastX, Autobox2 and Theta.

if we consider only the cases in which the best model form is changed after treating, which amount to 3144 monthly, 769 quarterly and 914 yearly series across all three competitions considered, the difference in ranking is significant as shown in the MCB test results in Fig. 7.

4.4. Relative performance on the M3 and M4 competitions

As a final experiment, we compare the relative performance of the methods developed here with the best methods from M3 and M4, by the time the competitions took place in 2000 and 2018, respectively. Table 9 depicts, per series frequency, the average sMAPE values of the proposed approaches and of the four best methods in the competition. Table 10 in turn, provides the same sMAPE comparisons, but considering the M4 Competition. Finally, Table 11 depicts the comparisons in M4 in terms of average MSIS values at the 95% coverage level. We hasten to add that the MASE, proposed by Hyndman and Koehler (2006), is posterior to the year the M3 Competition took place. In addition, no official values for the best average MASE results per series frequency were provided in M4. For such reasons, the comparisons with the best PF methods from M3 and M4 were restricted to average sMAPE values. Finally, we also note that M3 did not require PIs for the PFs provided by the competitors, so the average MSIS comparisons are restricted to M4.

Overall, the proposed approaches provided competitive PFs when compared to the best methods in M3 and competitive PIs when compared to the best methods in M4. For the M3 monthly series, for instance, all Bagging strategies that used the Treated ETS as the forecasting method to generate the bagged forecasts provided better PFs than Theta, the best method from the competition. The same story is also held true for the Pruned strategies. For M3 yearly series, in turn, the 'Treated AICc weights' provided better PFs than RBF, the best method from the competition for this frequency, and the methods based on BMC ranked between the third and fourth best methods. Against M4 Competition, our approaches were not as competitive, in terms of PFs, as the best methods in that competition.

In terms of PIs in the M4 Competition, we note that (i) the Treated ETS and Treated AICc weights approaches ranked third among the best methods for monthly series; (ii) for quarterly series, the Treated AICc weights and the Pruned BMC ETS strategies provided better MSIS results than the second best method; and (iii) finally, for yearly series, several strategies considering treating and/or pruning ranked between the third and fourth best methods. One must take into account the fact that these are ex-post results, after the end of the competition. On the other hand, treating and pruning were not designed to outperform benchmarks and/or rank among the best methods in the M3 and M4 Competitions and yet provided very competitive results. It is also worth recalling that the overall accuracy of pruning in the cases considered is restricted to how good Bagging strategies perform in practice.

Table 10
M4: Average sMAPEs of the proposed approaches and of the four best methods at the time of the competition.

Method	Average sMAPE (Monthly)	Average sMAPE (Quarterly)	Average sMAPE (Yearly)
Exponential smoothing			
ETS	13.53	10.29	15.36
Treated ETS	13.35	10.25	15.18
Exponential smoothing forecast combinations			
AICc weights	13.41	10.15	15.16
Treated AICc weights	13.24	10.13	15.01
Bagging strategies			
Bagged ETS	13.30	10.23	14.47
Bagged Treated ETS	13.27	10.21	14.46
BMC ETS	13.07	10.08	14.94
BMC Treated ETS	13.02	10.06	14.93
Bagging with pruning			
Pruned Bagged ETS	13.29	10.24	14.50
Pruned Bagged Treated ETS	13.27	10.22	14.49
Pruned BMC ETS	12.99	10.09	14.54
Pruned BMC Treated ETS	12.94	10.06	14.52
Best methods from the M4 Competition			
1st best method	**12.13**	**9.68**	**13.18**
2nd best method	*12.49*	*9.73*	*13.37*
3rd best method	12.64	9.80	13.53
4th best method	12.74	9.80	13.67

Notes: Best (most accurate) approach in **bold**, second best in *italic*. In M4, the best forecast methods in terms of average sMAPE for monthly data were, from best to worst: submissions IDs (please refer to https://mofc.unic.ac.cy/ for the IDs) 118, 972, 245 and 069. For quarterly data, the best methods were: 118, 245, 237 and 036. For yearly data: 118, 260, 245 and 036.

5. Conclusions and future directions

We introduced in this paper a new way of selecting among model forms in automated ETS forecasting routines, here addressed as treating. The approach operates by subsetting the pool of competing models based on the information delivered by their Prediction Intervals (PIs). An application to exponential smoothing formulations gave rise to alternative forecasting methods, the 'Treated ETS' and the 'Treated AICc weights'. By the same token, we also proposed a pruning strategy that can be used to enhance the accuracy of forecasts arising from any forecast combination method, provided that the models to be combined are able to generate PIs to their Point Forecasts (PFs).

The gains originating from treating and pruning were empirically demonstrated by means of an extensive experiment on a wide range of monthly, quarterly and yearly time series from the M-Competitions. We used as benchmarks for forecast combination two recently developed Bagging routines, which were originally formulated with the focus of improving the accuracy of PFs. To demonstrate how the accuracy of these methods could be improved with the use of pruning, we first extended the fields of application of Bagging to generate PIs, another important development of this paper.

The implications of our study are significant both in terms of theory and practice. First, we demonstrate that model selection via traditional information criteria minimization may lead to inaccurate forecasts and unstable PIs in some cases. Second, we show that PIs, apart from providing practitioners with a convenient way to estimate the uncertainty of a PF, contain important information that can be used to improve the accuracy of forecasting methods without having to resort to procedures which are dependent on the choice of the practitioner, such as the use of a validation set, for instance. Third, based on these two previous findings, we set forth strategies that can be used to improve the accuracy of forecasts for a considerable range of forecast approaches that involve model selection or combination. Finally, we note that, apart from their simplicity and ease of use, these strategies require practically no additional computational cost.

As methodological extensions of this research, future works may benefit from alternative schemes for subsetting the pool of competing model forms, in the case of treating, or the ensemble of forecasts to be combined, in the case of pruning. For the latter, for instance, we restricted our attention to demonstrate how subsetting could be achieved in Bagging routines. It would be interesting to see how the concept could be extended to other forecast combination methods. Some directions in this regard were already discussed in the methodology section. The use of alternative methods for outlier detection in ensembles, such as nonparametric methods, also constitutes a future research agenda.

Table 11
M4: Average MSISs, computed at the 95% desired coverage level, of the proposed approaches and of the four best methods at the time of the competition.

Method	Average MSIS (Monthly)	Average MSIS (Quarterly)	Average MSIS (Yearly)
Exponential smoothing			
ETS	8.30	9.45	34.90
Treated ETS	8.18	9.38	34.43
Exponential smoothing forecast combinations			
AICc weights	8.21	9.27	33.10
Treated AICc weights	8.10	*9.20*	32.89
Bagging strategies			
Bagged ETS	8.74	9.57	36.88
Bagged Treated ETS	8.71	9.56	36.89
BMC ETS	3.44×10^{11}	10.26	32.80
BMC Treated ETS	6.88×10^{11}	10.29	32.61
Bagging with pruning			
Pruned Bagged ETS	8.77	9.61	37.20
Pruned Bagged Treated ETS	8.73	9.59	37.21
Pruned BMC ETS	8.39	9.21	32.22
Pruned BMC Treated ETS	8.42	9.25	32.11
Best methods from the M4 Competition			
1st best method	**7.21**	**8.55**	**23.90**
2nd best method	*8.03*	9.38	*27.48*
3rd best method	8.23	9.42	30.20
4th best method	8.23	9.85	35.84

Notes: Best (most accurate) approach in **bold**, second best in *italic*. In M4, the best forecast methods in terms of average MSIS for monthly data were, from best to worst: Submissions 118, 069, 036 and 132. For quarterly data: Submissions 118, 245, 069 and 238. For yearly data: Submissions 118, 245, 238 and 069.

Acknowledgments

The authors acknowledge the thoughts shared by Dr. Fotios Petropoulos, University of Bath, UK, which contributed to the conceptualization of the present study. The authors also acknowledge support from the Balena High Performance Computing (HPC) Service at the University of Bath, UK.

Funding

This work was supported by the Brazilian Coordination for the Improvement of Higher Level Personnel (CAPES) under Grant [number 001]; the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant [number 307403/2019-0]; and the Carlos Chagas Filho Research Support Foundation of the State of Rio de Janeiro (FAPERJ) under Grants [numbers 202.673/2018 and 211.086/2019].

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2020.07.005.

References

Aiolfi, M., & Timmermann, A. (2006). Persistence in forecasting performance and conditional combination strategies. *Journal of Econometrics*, *135*(1–2), 31–53. http://dx.doi.org/10.1016/j.jeconom.2005.07.015.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. http://dx.doi.org/10.1109/tac.1974.1100705.

Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2016). Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *International Journal of Forecasting*, *32*(2), 303–312. http://dx.doi.org/10.1016/j.ijforecast.2015.07.002.

Box, G., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *26*(2), 211–252.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. http://dx.doi.org/10.1007/bf00058655.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics*, *53*(2), 603. http://dx.doi.org/10.2307/2533961.

Bühlmann, P. (1998). Sieve bootstrap for smoothing in nonstationary time series. *The Annals of Statistics*, *26*(1), 48–83. http://dx.doi.org/10.1214/aos/1030563978.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, *6*, 3–73.

Cordeiro, C., & Neves, M. M. (2009). Forecasting time series with BOOT.EXPOS procedure. *REVSTAT – Statistical Journal*, *7*(2), 135–149.

Dantas, T. M., & Cyrino Oliveira, F. L. (2018). Improving time series forecasting: An approach combining bootstrap aggregation, clusters and exponential smoothing. *International Journal of Forecasting*, *34*(4), 748–761. http://dx.doi.org/10.1016/j.ijforecast.2018.05.006.

Dantas, T. M., Oliveira, F. L. C., & Repolho, H. M. V. (2017). Air transportation demand forecast through bagging holt winters methods. *Journal of Air Transport Management*, *59*, 116–123. http://dx.doi.org/10.1016/j.jairtraman.2016.12.006.

De Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*(1), 190–204. http://dx.doi.org/10.1016/s0377-2217(98)00380-4.

De Oliveira, E. M., & Cyrino Oliveira, F. L. (2018). Forecasting mid-long term electric energy consumption through bagging ARIMA and exponential smoothing methods. *Energy*, *144*, 776–788. http://dx.doi.org/10.1016/j.energy.2017.12.049.

Diebold, F. X., & Shin, M. (2019). Machine learning for regularized survey forecast combination: partially-egalitarian LASSO and its derivatives. *International Journal of Forecasting*, *35*(4), 1679–1691. http://dx.doi.org/10.1016/j.ijforecast.2018.09.006.

Elliott, G. (2011). *Averaging and the optimal combination of forecasts*: *Working Paper*, San Diego: University of California.

Gardner Jr., E. S. (1985). Exponential smoothing: The state of the art. *Journal of Forecasting*, *4*(1), 1–28. http://dx.doi.org/10.1002/for.3980040103.

Grushka-Cockayne, Y., & Jose, V. R. R. (2019). Combining prediction intervals in the M4 competition. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2019.04.015.

Hendry, D. F., & Clements, M. P. (2004). Pooling of forecasts. *The Econometrics Journal*, *7*(1), 1–31.

Holt, C. C. (1957). *ONR Memorandum*: *Vol. 52*, *Forecasting seasonals and trends by exponentially weighted moving averages*. Pittsburgh: PA7 Carnegie Institute of Technology.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). Melbourne, Australia: OTexts, URL:.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2019). Forecast: Forecasting functions for time series and linear models. URL: http://pkg.robjhyndman.com/forecast.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(3), 1–22. http://dx.doi.org/10.18637/jss.v027.i03.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. http://dx.doi.org/10.1016/j.ijforecast.2006.03.001.

Hyndman, R., Koehler, A., Ord, K., & Snyder, R. (2008). *Forecasting with exponential smoothing: The state space approach*. Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-540-71918-2.

Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439–454. http://dx.doi.org/10.1016/s0169-2070(01)00110-8.

IHS Global Inc. (2015). *EViews® illustrated* (9th ed.). IHS Global Inc., URL: https://www.eviews.com/illustrated/EViewsIllustrated.pdf.

Inoue, A., & Kilian, L. (2004). *Bagging time series models*: *CEPR Discussion Paper No. 4333*, Centre for Economic Policy Research (CEPR), URL: https://ssrn.com/abstract=540262.

Inoue, A., & Kilian, L. (2008). How useful is bagging in forecasting economic time series? A case study of U.S. consumer price inflation. *Journal of the American Statistical Association*, *103*(482), 511–522. http://dx.doi.org/10.1198/016214507000000473.

Kolassa, S. (2011). Combining exponential smoothing forecasts using akaike weights. *International Journal of Forecasting*, *27*(2), 238–251. http://dx.doi.org/10.1016/j.ijforecast.2010.04.006.

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. (2005). The m3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*(3), 397–409. http://dx.doi.org/10.1016/j.ijforecast.2004.10.003.

Kourentzes, N., Barrow, D., & Petropoulos, F. (2019). Another look at forecast selection and combination: Evidence from forecast pooling. *International Journal of Production Economics*, *209*, 226–235. http://dx.doi.org/10.1016/j.ijpe.2018.05.019.

Künsch, H. R. (1989). The Jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, *17*(3), 1217–1241. http://dx.doi.org/10.1214/aos/1176347265.

Lee, T.-H., & Yang, Y. (2006). Bagging binary and quantile predictors for time series. *Journal of Econometrics*, *135*(1–2), 465–497. http://dx.doi.org/10.1016/j.jeconom.2005.07.017.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting*, *1*(2), 111–153. http://dx.doi.org/10.1002/for.3980010202.

Makridakis, S., & Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451–476. http://dx.doi.org/10.1016/s0169-2070(00)00057-1.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, *34*(4), 802–808. http://dx.doi.org/10.1016/j.ijforecast.2018.06.001.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2019). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, http://dx.doi.org/10.1016/j.ijforecast.2019.04.014.

Matsypura, D., Thompson, R., & Vasnev, A. L. (2018). Optimal selection of expert forecasts with integer programming. *Omega*, *78*, 165–175. http://dx.doi.org/10.1016/j.omega.2017.06.010.

Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, *92*(440), 1621–1629. http://dx.doi.org/10.1080/01621459.1997.10473684.

Pegels, C. C. (1969). Exponential forecasting: Some new variations. *Management Science*, *15*(5), 311–315.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, *268*(2), 545–554. http://dx.doi.org/10.1016/j.ejor.2018.01.045.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. http://dx.doi.org/10.1214/aos/1176344136.

Sugiura, N. (1978). Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics. Theory and Methods*, *7*(1), 13–26. http://dx.doi.org/10.1080/03610927808827599.

Taylor, J. W. (2003). Exponential smoothing with a damped multiplicative trend. *International Journal of Forecasting*, *19*(4), 715–725. http://dx.doi.org/10.1016/s0169-2070(03)00003-7.

Vinod, H. D. (2014). Matrix algebra topics in statistics and economics using R. In *Handbook of statistics* (pp. 143–176). Elsevier, http://dx.doi.org/10.1016/b978-0-444-63431-3.00004-8.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324–342. http://dx.doi.org/10.1287/mnsc.6.3.324.