Point and interval forecasting of electricity supply via pruned ensembles

Erick Meira*

Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro
Rua Marquês de São Vicente, 225, Ed. Cardeal Leme, 9º andar, Rio de Janeiro 22451-900, Brazil

Energy, Information Technology and Services Division, Brazilian Agency for Research and Innovation (Finep)
Avenida República do Chile, 330, Torre Oeste, 15º andar, Rio de Janeiro 20031-170, Brazil

Fernando Luiz Cyrino Oliveira

Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro
Rua Marquês de São Vicente, 225, Ed. Cardeal Leme, 9º andar, Rio de Janeiro 22451-900, Brazil

Lilian M. de Menezes

Cass Business School, City, University of London
106 Bunhill Row, London EC1Y 8TZ, United Kingdom

Abstract

This paper develops a new ensemble-based approach to point and interval forecasting, and focus on total electricity supply. The proposed approach combines Bootstrap Aggregation (Bagging), time series methods and a novel pruning routine that performs feature selection before the aggregation of forecasts. Monthly time series of the total electricity supplied between January 2000 and September 2020 in 16 countries are considered. Forecasting performance in different horizons is examined. As the data includes the COVID-19 pandemic that affected countries in different ways, with some visible changes in electricity demand, the likely impact of unusual observations on this proposal is also examined. A comparative, multi-step-ahead forecasting with out-of-sample evaluation is conducted using several forecasting accuracy metrics and detailed robustness checks. The results endorse the strength and resilience of the proposed approach in delivering not only accurate point forecasts, but also reliable prediction intervals under different economic settings. Moreover, the methodology presented herein is flexible, in the sense that it can be used to generate reliable point and interval forecasts for any time series in short and medium horizons.

Keywords: Forecasting, Prediction intervals, Ensembles, Electricity supply, Energy planning

*Corresponding author: erickmeira89@gmail.com; research@erickmeira.com; emeira@finep.gov.br
Email addresses: erickmeira89@gmail.com, cyrino@puc-rio.br, l.demenezes@city.ac.uk

First review returned on *May 7, 2021*

1. Introduction

The importance of reliable energy planning cannot be overstated. Every day, corporate leaders, grid operators, regulators, and policymakers are faced with the challenge of making decisions based on the most up-to-date information and their expectations about energy systems. However, developing an energy plan is not straightforward. A major concern, which is inherent to most decision-making, is the uncertainty of future outcomes. In this context, predictions of electricity to be supplied are essential for minimizing energy costs, securing capacity and providing quality services.

Unsurprisingly, extensive research has been conducted to forecast future electricity supply or aggregate consumption[1], with varying degrees of success. Methodological proposals range from classic forecasting approaches, such as exponential smoothing (Wu et al., 2013) and ARIMA (Elamin & Fukushige, 2018), hybrid methods that combine traditional forecasting with machine learning (De Oliveira & Cyrino Oliveira, 2018; Jiang et al., 2020), wavelet transforms and adaptive models (Bashir & El-Hawary, 2009; Bahrami et al., 2014), grey-based models (Xie et al., 2020; Zhu et al., 2020), hierarchical linear models (da Silva et al., 2019), among others. Indeed, a review by Kuster et al. (2017) highlights a growing effort towards forecasting electricity supply and demand. Nevertheless, and despite improvements in forecasting accuracy within the electricity sector, the literature has been focused on point forecasts, and as such, it has emphasized a best-guess strategy.

Having reliable prediction intervals (PIs) for future electricity demand is perhaps more important than precisely balancing supply and demand, as prediction intervals reduce the random variation in classic single-valued load time series forecasts (Petropoulos et al., 2020), and can minimize the risk inherent to decision-making within the management of power systems. Hence, the present study extends a class of ensemble-based approaches, namely *Bootstrap aggregation (Bagging)* algorithms, which have traditionally been used for point forecast generation, to deliver accurate prediction intervals. Specifically, *Bagging* algorithms are combined with time series methods and novel pruning routines, which are capable of feature selection, to forecast and generate prediction intervals of total electricity supply. This approach contrasts with previous proposals for prediction interval forecasting in three ways: first, it delivers prediction intervals without the need to generate point forecasts. This is important, as generally a two-step process is adopted, whereby the point forecast is estimated, and then a prediction interval is constructed. In the present study, a hybrid ensemble-based method is developed, drawing on knowledge from statistics, machine learning and forecasting, and prediction intervals are directly generated. Moreover, the results demonstrate that this new approach is also

[1]Total national electricity supply is computed by considering the sum of a country's indigenous production and adding and subtracting, respectively, the imports from and exports to other countries, when applicable. The total supply is a close proxy to total electricity consumption, given that the latter can be obtained by subtracting from the former the country's Transmission and Distribution (T&D) Losses and the amount of energy used for pumped storage (IEA, 2021).

capable of improving the quality of point forecasts, given the feature selection property of the pruning strategy. Another important point that distinguishes our proposal from existing methods concerns generalization and robustness. Given its hybrid ensemble properties, different stylized facts in the time series (nonlinearities, stochastic components, heteroscedasticity, unusual periods) can be addressed. Finally, the prediction intervals generated via the proposed approach are unlikely to increase significantly over the forecasting horizon, which is characteristic of most prediction intervals. This is particularly important for decision-making in the energy industry, where companies are subject to varying and significant degrees of uncertainty in future electricity supply, but can only address a certain limit, given the high costs involved in energy storage systems and capacity management.

The paper unfolds as follows. The next section describes how *Bootstrap aggregation (Bagging)* algorithms can be combined with time series methods and tackle different sources of uncertainty in building predictive models from data, namely: measurement (data) uncertainty, model uncertainty, and parameter uncertainty (Petropoulos et al., 2018). Section 3 summarizes the most recent methods for prediction interval generation. Together these two sections set the context of the proposed approach, which is described in Section 4. This new approach involves extending previous *Bagging* algorithms, so that prediction intervals are generated, and the development of pruning routines for improving forecasting accuracy. Section 5 describes the data that are used to evaluate forecasting performance, and Section 6 summarizes the results and their implications and highlights directions for future research. Finally, Section 7 concludes.

2. Bagging for point forecasts

The underlying idea of *Bagging* for time series forecasting is to use predictors that are built on bootstrapped versions of the original data. This method is summarized in Figure 1, where its four stages are highlighted.

The first stage concerns pretreatment and decomposition. An initial transformation is conducted to stabilize the variance of the original time series and reduce the skewness of its distribution, when necessary. Subsequently, the transformed series is decomposed into three key components (trend, seasonal and remainder), using a selected decomposition method.

The second stage concerns the generation of replicas of the original time series. A bootstrap method is applied to the remainder of the decomposition. The new version of the remainder, which shares the same properties of the original component but present slightly different values, is added to the trend and seasonal components, and the transformation is inverted. This procedure is repeated $J-1$ times, so that a total of J series is included in the ensemble (the original series and its $J-1$ replicas). The replicas have the same unit of measurement and share the same stylized facts of the

actual, observed time series. Their values, however, are slightly different from the latter. Replicas constitute an ensemble of alternative outputs of the underlying stochastic process of the time series, since the observed time series is one of the many possible outcomes from the true data generation process. In short, by considering multiple replicas, measurement uncertainty is addressed.

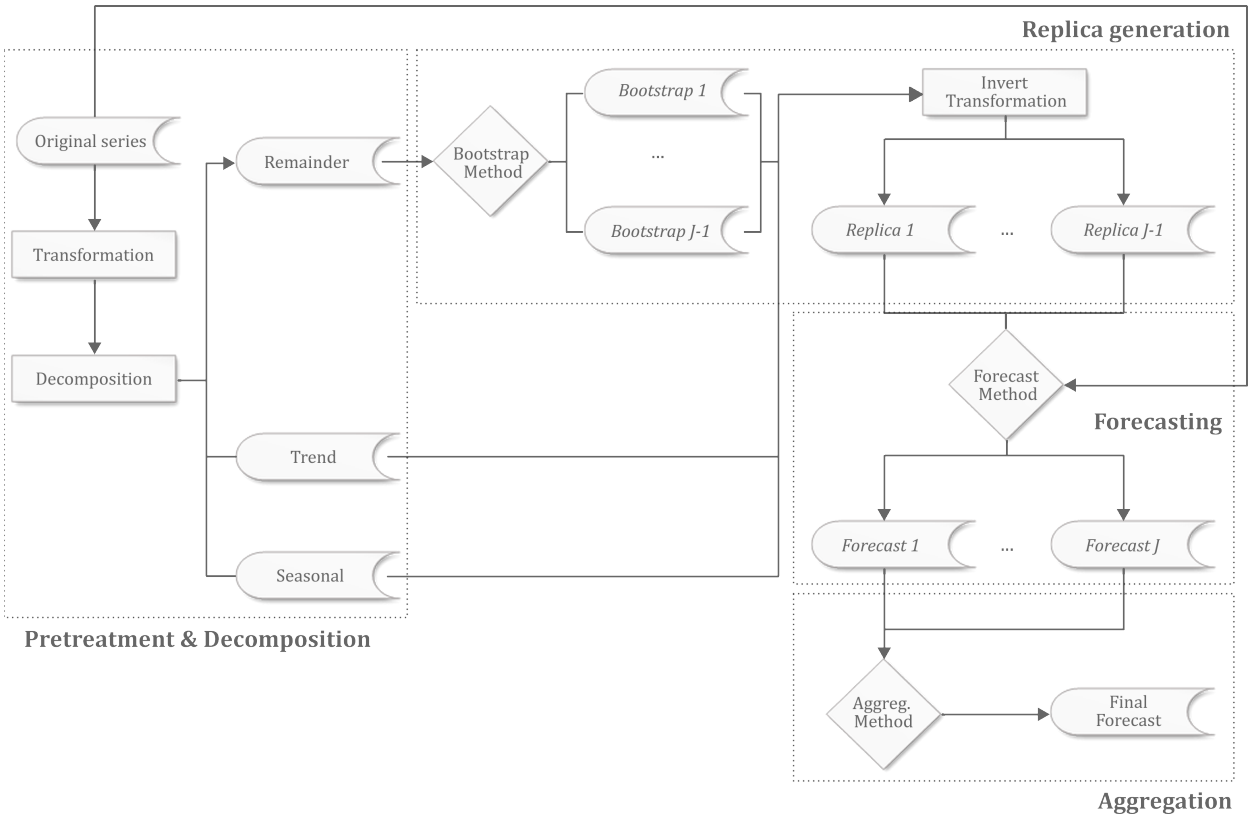

Figure 1: The main stages involved in *Bagging* for point forecast generation.

In the third stage, forecasts are generated for the original series and for each replica in the ensemble, using a selected forecasting method. Models are estimated for the original series and for each replica, so that model parameters vary according to the values in each series. Based on these models, forecasts are independently generated for each series throughout the forecast horizon. Finally, all forecasts are combined using an aggregation strategy, such as median aggregation or pruning followed by median aggregation (our proposed approach). Therefore, as Petropoulos et al. (2018) argued, *Bagging* algorithms can tackle the measurement uncertainty as well as two other sources of uncertainty that are derived from former: model uncertainty, which refers to the uncertainty linked with the selection of the 'optimal' model form; and parameter uncertainty, which is due to the selection of the best set of parameters. Yet, *Bagging* approaches can differ in many aspects/steps of the methodology. In the following subsections, each stage in our proposal is described, and any differences from existing *Bagging* routines for univariate time series forecasting are also highlighted.

2.1. Pretreatment and decomposition

In most forecasting ensembles, the time series is initially filtered or smoothed. A common procedure for treating time series data is the Box–Cox (BC) transformation (Box & Cox, 1964), which is attractive because it can simultaneously stabilize the variance, reduce the skewness of the distribution, and ensure that the components of the time series are additive (Petropoulos et al., 2018). Unsurprisingly, several recent studies in forecasting have applied this transformation (Bergmeir et al., 2016; De Oliveira & Cyrino Oliveira, 2018; Petropoulos et al., 2018; Meira et al., 2021).

The BC transformation is defined as follows:

$$\omega_t = \begin{cases} \log\left(y_t\right), & \lambda = 0 \\ \left(y_t^{\lambda} - 1\right)/\lambda, & \lambda \neq 0 \end{cases} \tag{1}$$

where y_t represents the original time series, ω_t its transformed version, and λ is the transformation parameter. In all *Bagging* approaches considered, we restrict λ to lie in the interval $[0, 1]$ and use the method of Guerrero (1993) to choose its value, following recent studies on *Bagging* for forecasting, such as Bergmeir et al. (2016), De Oliveira & Cyrino Oliveira (2018), Dantas & Cyrino Oliveira (2018) and Meira et al. (2021). In short, the chosen method partitions the original data into subseries of length equal to the seasonality (or length two, if the series is non-seasonal). Subsequently, the sample mean m and the standard deviation s are calculated for each of the subseries, and λ is chosen in such a way that the coefficient of variation of $s/(m(1-\lambda))$ across the subseries is minimized.

Following the initial treatment, time series decomposition can be applied, since the estimation error obtained from further aggregating the extrapolated sub-series is expected to be lower than the estimation error for the whole series. Two types of decomposition have become widespread in the literature: Seasonal-Trend decomposition using Loess (STL) (Cleveland et al., 1990), and Empirical Mode Decomposition (EMD) (Huang et al., 1998). The former consists of six sequential smoothing operations employing Locally-Weighted Regression (*Loess*) that decompose the series into three additive components: trend, seasonal and remainder. When compared to other decomposition methods, STL is robust to outliers, can deal with any type of seasonality regardless of the data-frequency, and allows for controlling trend-cycle smoothness (Hyndman & Athanasopoulos, 2021). By contrast, EMD decomposes the time series into a sum of oscillatory Intrinsic Mode Functions (IMFs) that are symmetric with respect to their local zero-mean. The number of extrema and zero-crossings for each IMF are, by definition, equal or allowed to differ at most by one in the whole data. IMFs are more regular and thus easier to forecast.

Most ensembles adopt STL, prior to generating replicas of the time series. STL has also been integrated to hybrid forecasting methods, such as the Bagged.BLD.MBB.ETS by Bergmeir et al. (2016) and the Bootstrap Model Combination of Petropoulos et al. (2018). However, EMD has

also shown encouraging performance, e.g., EMD-Holt-Winters *Bootstrap aggregation (Bagging)* of Awajan et al. (2018) and the Interval Decomposition Ensemble (IDE) of Sun et al. (2018), which combined bivariate Empirical Mode Decomposition, Interval Multilayer Perceptron and Interval Holt's exponential smoothing method.

2.2. Replica generation

The second stage of *Bagging* approaches concerns the generation of replicas, which are alternative series that share common properties of the original data. This can be achieved in different ways (e.g. Monte Carlo simulation, or by resampling key components of the time series). Our proposal considers replica generation via bootstrapping the remainder component from an STL decomposition. Two points are key when using bootstrap for time series. The first is the prerequisite of stationarity, and this is often fulfilled by the remainder. Secondly, the method must ensure that every value from the original series can be placed anywhere in the bootstrapped series. In this proposal, we deal with this second issue in three different ways, as follows.

Moving Blocks Bootstrap

The Moving Blocks Bootstrap (MBB) of Künsch (1989) draws data blocks of equal size from the time series until the desired length is achieved. Hence, for a series of length n, with a block size of l, $n - l + 1$ (overlapping) possible blocks exist. However, in order to address any non-stationarity and/or autocorrelation in the data, Bergmeir et al. (2016) proposes drawing $[n/l] + 2$ blocks from the remainder series of a STL Decomposition, and discarding a random number of values, between zero and $l - 1$, from the beginning of the bootstrapped series. Subsequently, to obtain a series with the same length as the (original) remainder series, they further discard as many values as necessary. This process ensures that the bootstrapped series do not necessarily begin or end on a block boundary. Finally, the trend and seasonality are combined with the bootstrapped remainder. Their method requires, however, that a block size parameter is set. In this work, we adopt the MBB approach for resampling and use a block size $l = 24$, following previous empirical studies on monthly time series (Bergmeir et al., 2016; Petropoulos et al., 2018; Meira et al., 2021).

Circular Blocks Bootstrap

The Circular Blocks Bootstrap (CBB) by Politis & Romano (1991) is akin to MBB in that it suggests sampling data blocks of equal size until the desired series length is achieved. According to CBB, however, the time series are 'wrapped' in a circle before resampling takes place. That is, the start of the 'construction' of the blocks can occur in any instance, since the time series is 'enveloped'. This procedure aims to ensure that the first and last $l - 1$ observations are not subsampled, which

theoretically makes the CBB superior to MBB. Our choice of block size (l) for CBB follows the same empirical guidelines for monthly series (24 observations).

Linear Process Bootstrap

The Linear Process Bootstrap (LPB) is a five-step algorithm devised by McMurry & Politis (2010) that allows resampling time series with drops in the autocovariance structure without the need to explicitly estimate coefficients. Operating similarly to the Moving Average (MA) sieve bootstrap counterpart, the LPB estimates the autocovariance matrix of the selected series by fitting a MA-type autocovariance function. Subsequently, the algorithm pre-whitens the noise with the estimated autocovariance matrix, and generates bootstraps from the pre-whitened noise. Finally, it post-colors the bootstrap noise with the estimated autocovariance matrix.

2.3. Forecasting

Following replica generation, forecasting models are applied to the original data and each of its replicas separately. Exponential smoothing methods are frequently considered at this stage, given their simplicity and ability to adapt to many different situations (Goodwin, 2010). For example, exponential smoothing models have been recently applied in combination with bootstrap methods to forecast electricity consumption (De Oliveira & Cyrino Oliveira, 2018).

In addition to their simplicity and ease of adaptation, exponential smoothing formulations have a theoretical foundation in state space modelling (Ord et al., 1997), which has allowed for straightforward implementations in statistical packages (Hyndman et al., 2002, 2008; Hyndman & Athanasopoulos, 2021). Exponential smoothing models, when defined according to this framework, are known as ETS, an acronym for 'ExponenTial Smoothing' or 'Error, Trend and Seasonal', thus reflecting the components of the time series that are allowed to vary across formulations. The possible combinations for the trend and seasonal components are depicted in Table 1. In addition, since the error term can be either additive or multiplicative, a total of 30 different formulations can be achieved.

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

Table 1: Possible combinations of seasonal and trend components under the ETS state-space framework.

The ETS algorithm selects one combination for each series in the ensemble (the original series and its replicas). The best model for each time series is then used to generate forecasts for the desired forecast horizon (number of steps ahead).

Other families of models can be considered to generate the forecasts in *Bagging* approaches. ARIMA (Autoregressive, Integrated, Moving Average) formulations, first devised by Box & Jenkins (1970), are a straightforward alternative. They are similar to exponential smoothing as they can model trends and seasonal patterns, but are based on autocorrelation and partial autocorrelation functions of the time series (or transformed stationary series) rather than a structural view of the time series (level, trend and seasonality). ARIMA models may also be automated, and have been recently applied in combination with bootstrap methods to forecast electricity consumption (De Oliveira & Cyrino Oliveira, 2018).

Ensembles of Neural Networks (NNs) have also been used for over 30 years, especially within the artificial intelligence community, and may include a variety of methods (Barrow & Crone, 2016; Li et al., 2016; Szafranek, 2019). They are generally seen as a means to make the most of computing power to address the uncertainty in individual point forecasts. As Rendon-Sanchez & de Menezes (2019) noted in their review of the literature, ensembles of NNs have been particularly successful in forecasting short-term electricity demand, and were inspirational in the development of hybrid approaches that combine forecasts.

Finally, Misiorek et al. (2006) review other linear and non-linear alternatives that were previously used to forecast electricity spot prices. Most of these methods could be extended to electricity supply forecasting in the context of *Bagging* algorithms, as in our case.

2.4. Aggregation

The final stage in *Bagging* consists of aggregating (combining) forecasts. The median is usually favored as it may counter the effects of occasional poor forecasts in the generated (bagged) ensemble – see, for instance, Bergmeir et al. (2016); Dantas et al. (2017); De Oliveira & Cyrino Oliveira (2018). In this study, we use the median forecast as a benchmark. For reference purposes, we will refer to the traditional median aggregation strategy in *Bagging* ensembles as 'BaggedETS', as proposed by Bergmeir et al. (2016).

Petropoulos et al. (2018) proposed a more sophisticated combination strategy known as Bootstrap Model Combination (BMC), which at first sight is similar to the approach of Bergmeir et al. (2016) since replicas are obtained by resampling the remainder from a STL decomposition and are independently predicted using exponential smoothing. However, replicas are used to drive the selection of the best-fit model, and forecasts are combined using weights reflecting the frequency that the selected model specifications were identified as best-fit on the pool of replicas. Considering

all series from two forecast competitions, M (Makridakis et al., 1982) and M3 (Makridakis & Hibon, 2000), the BMC outperformed the approach of Bergmeir et al. (2016).

Pruning, as presented in Section 4.2, is an alternative way to combine the forecasts in a bagged ensemble, because it uses the information retrieved by the prediction intervals to conduct feature selection on the ensemble. In doing so, unwanted forecasts (and corresponding intervals) are removed from the pool of forecasts prior to aggregation. It should be noted that pruning is one of many possibilities to conduct feature selection in ensemble forecasting. Other recent approaches include the hierarchical group-lasso regularization of Lim & Hastie (2015), the Information-Theoretic Criteria devised by Abedinia et al. (2017), the Max-Relevance and Min-Redundancy feature selection filter of Duan et al. (2018), and a recent Bayesian Bootstrap aggregation algorithm proposed by Song et al. (2021).

3. Generating prediction intervals

This section summarizes traditional methods for prediction interval generation, and highlights their main differences. In general, prediction intervals for point forecasts are either based on neural network architectures or follow a two-step process, whereby a forecast is estimated and then, based on an estimate of uncertainty, an interval is constructed. From a neural network perspective, prediction intervals include Delta (Hwang & Ding, 1997), Bayesian (Bishop, 1995) and bootstrap techniques (Heskes, 1996). Khosravi et al. (2010), for instance, construct prediction intervals for outputs of Neural Networks (NNs) via Delta and Bayesian techniques. The downside of these methods is that they make special assumptions about the data distribution. The delta and Bayesian techniques require that the variance of forecasts should be constant, while the bootstrap for NN-based predictions requires a smooth variance. These assumptions can make intervals less reliable when time series are volatile with changing variance.

Two-step prediction interval estimation based on residuals, in turn, has been a common strategy since the seminal paper of Chatfield (1993). More recently, Khosravi et al. (2013) use Moving Blocks Bootstrapped Neural Networks and Generalized Auto Regressive Conditional Heteroscedasticity (GARCH) models to construct prediction intervals for hourly electricity prices in the Australian and New York energy markets. Vilar et al. (2018), by contrast, use residual-based bootstrap procedures to construct prediction intervals associated with the functional nonparametric autoregressive model and the semi-functional partial linear model. Sulandari et al. (2020), in turn, bootstrap the residuals of a hybrid SSA-LRF-NN (Singular Spectrum Analysis, Linear Recurrent Formula and Neural Networks) algorithm to generate prediction intervals. Moreover, Du et al. (2020) develop an interval forecasting approach using the predictions from a hybrid forecasting model combining variational mode decomposition and an optimized outlier-robust extreme learning machine, and the results

of five distribution functions tailored to mining the traits of metal prices. The downside of these proposals is that the quality of prediction intervals depends on both the quality of the point forecasts and the proxies for uncertainty adopted (e.g. standard deviation of the residuals or quantiles of the bootstrapped residuals distribution).

4. Proposed methodology

In this section, we extend *Bagging* strategies presented earlier for point forecast generation to generate prediction intervals. In addition, a variant of the pruning routine developed in Meira et al. (2021) is proposed, which focuses on improving the quality of prediction intervals, but can also increase the accuracy of the point forecasts.

4.1. Extending previous Bagging approaches for interval generation

In this study, prediction intervals are created through *Bagging* replicating the process for point forecasts, as outlined in the Section 2. That is, besides aggregating the point forecasts, their corresponding prediction intervals are combined using the median. This is possible because, regardless of the resampling approach (MBB, CBB or LPB), the point forecasts in each ensemble are generated via ETS, with corresponding prediction intervals, given a predefined coverage level. For instance, if a 95% coverage is aimed, a prediction interval is generated using the 2.5% quantile as lower limit and the 97.5% quantile for the upper limit (Hyndman et al., 2008). In other words, let J be the number of forecasts involved in the *Bagging* ensemble (forecasts of the original data and the $J-1$ replicas generated); the upper and lower limits of the Bagged prediction interval can be obtained as follows:

$$\begin{aligned} U_{t,Bagging} &= \text{median}\left[U_{t,1}, ..., U_{t,J}\right] \\ L_{t,Bagging} &= \text{median}\left[L_{t,1}, ..., L_{t,J}\right] \end{aligned} \tag{2}$$

where $U_{t,1}, ..., U_{t,J}$ and $L_{t,1}, ..., L_{t,J}$ are the upper and lower limits of the J point forecasts in the ensemble, respectively. Equation 2 is applied for each step in the forecast lead time, i.e., $t = 1, ..., h$. Figure 2 summarizes the generation of Bagged Point Forecasts (PFs) and Prediction Intervals (PIs). BaggedETS aggregates the J Point Forecasts (PFs) and their J corresponding Prediction Intervals (PIs) using their medians.

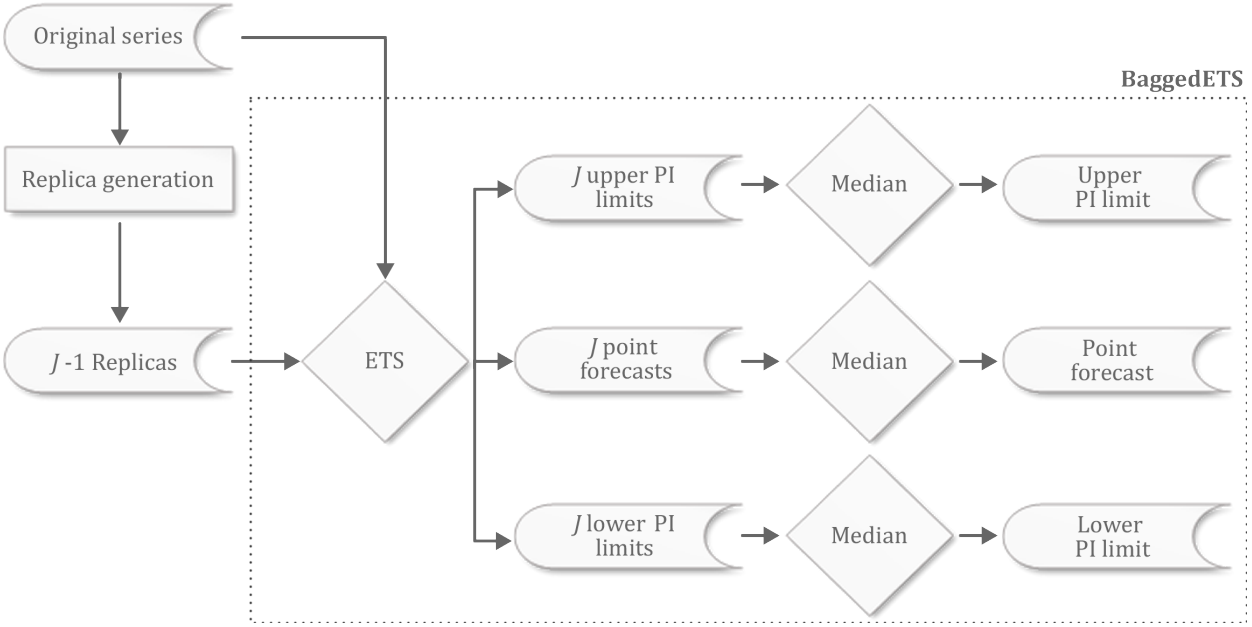

Figure 2: Generation of point forecasts and corresponding prediction intervals via BaggedETS.

4.2. Pruning

Following Meira et al. (2021), the rationale behind pruning for *Bagging* in forecasting algorithms is to compare the J prediction intervals corresponding to each of the J forecasts in the bagged ensemble, and discard those showing unusual behaviors, before final aggregation. To that end, pruning collects the upper and lower limits of the prediction intervals and conducts an outlier detection procedure in the two sets separately. In other words, it searches for outliers among the J upper prediction interval limits and among the J lower prediction interval limits. Outliers are defined as any values lying outside the range of $\pm 1.5 \times IQR$, where $IQR = Q_3 - Q_1$ is the interquartile range.

The outlier detection procedure in pruning is conducted for every step in the forecast horizon and considers all competing upper (or lower) limits, regardless of whether one or several of these limits have already been identified as outlier in a previous step. Finally, every point forecast (and prediction interval) associated with an upper or lower limit that was identified as outlier (even if just once) throughout the forecast horizon is discarded from the ensemble before aggregation. Hence, at the end of the outlier detection procedure, two sets of forecasts (and corresponding prediction intervals) are recommended to be discarded: j_1, corresponding to the forecasts whose upper prediction interval limits were judged outliers among the other upper limits; and j_2, corresponding to the forecasts whose lower interval limits were judged outliers among the other lower limits. Given that j_1 and j_2 are subsets of the same number of forecasts in the original *Bagging* ensemble, the final set of forecasts to be removed via pruning is given by $k = j_1 + j_2 - (j_1 \cap j_2)$. After removing outliers during pruning, the BaggedETS routine proceeds as usual: by aggregating the remaining $J - k$ point forecasts via the median in order to generate the final point forecast, as well as their corresponding

upper and lower prediction limits so as to generate the final prediction interval. Figure 3 illustrates the main steps of pruning.

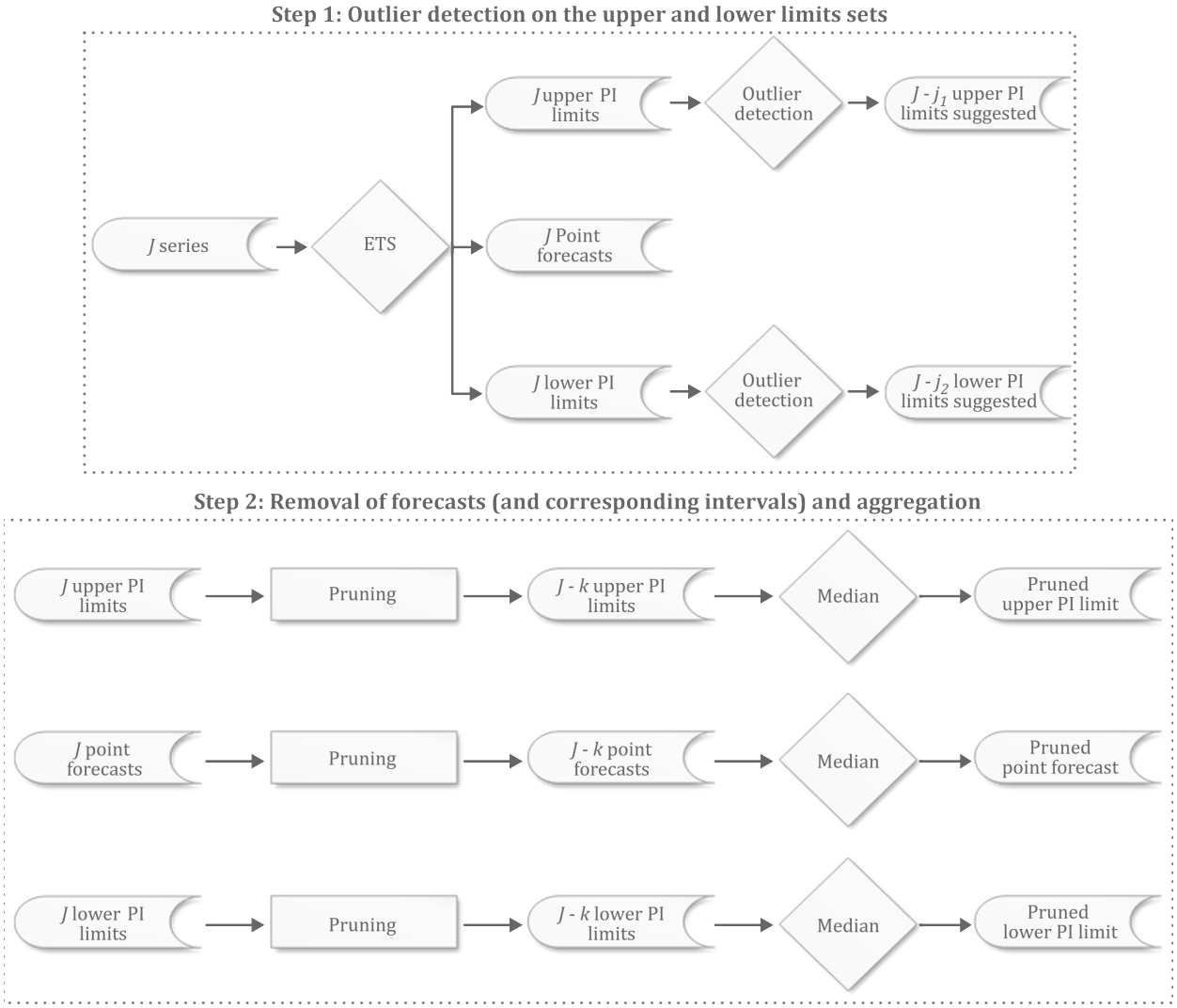

Figure 3: Pruning for BaggedETS routines, where $k = j_1 + j_2 - (j_1 \cap j_2)$ is the set of point forecasts (and prediction intervals) to be removed from the ensemble of J forecasts, prior to median aggregation.

5. Data and evaluation setup

The dataset consists of monthly series of total electricity supplied (in gigawatt-hours, GWh), which were collected from the Statistical Office of the International Energy Agency (IEA, 2021), and span 16 countries from January 2000 to September 2020. Observations from January 2000 to September 2018 are considered as training set. The test set comprises the last 24 observations: October 2018 – September 2020. Figures 4 and 5 depict the original time series. As can be noted, total electricity supply differs considerably between countries. In addition, some series appear to have been significantly affected by the economic distress brought by the COVID-19 pandemic, particularly during the months of March and April 2020, when the negative impacts of social distancing and

lockdown measures were present. Although prediction intervals for electricity supply are desired to be sharp, in the sense that they should not be too large in amplitude, they should also be wide enough to allow for possible downturns in periods of economic stress.

Figure 4: Total electricity supplied in gigawatt-hours (GWh) per country. Source: IEA (2021).

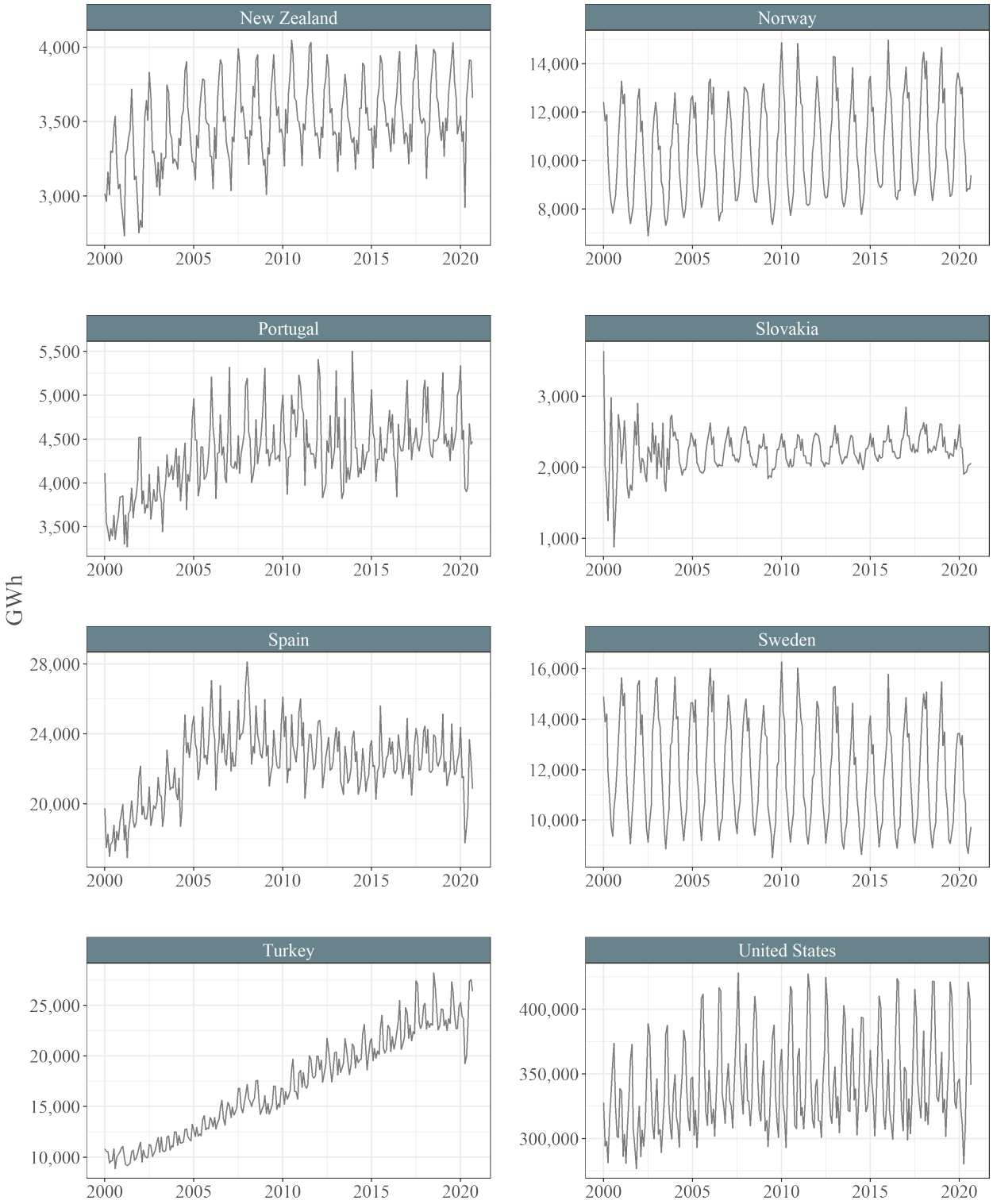

Figure 5: Total electricity supplied in gigawatt-hours (GWh) per country. Source: IEA (2021).

We compare our approach with several forecasting methods, which are summarized in Table 2. We note that implementation is conducted using the R programming language (R Core Team, 2021) and related packages. More specifically, R version 4.0.2 (2020-06-22) and *forecast* version 8.12 for ETS and ARIMA modelling are adopted. Furthermore, a parallel implementation was utilized,

which relied on the following packages: *doSNOW* (1.0.18), *foreach* (1.5.0) and *snow* (0.4–3). For this particular study, 99 replicas were generated per ensemble. All resampling procedures were conducted using the same random seed, which was set to 123 using the `set.seed()` function in R before bootstrapping.

Method	Implementation / Source	Short description
Traditional benchmarks		
ETS	R *forecast* package `ets()` function	Automatic Error, Trend and Seasonality specification
ARIMA	R *forecast* package `auto.arima()` function	Automatically-selected (S)ARIMA model
Holt-Winters	R *forecast* package `hw()`	Three parameter additive Holt-Winters method
Competing Bagging approaches		
BaggedETS	Bergmeir et al. (2016)	see Section 2.4 for details
BMC	Petropoulos et al. (2018)	see Section 2.4 for details

Table 2: Selected methods for comparison with the proposed approaches. Notes: `ets()` and `auto.arima()` are used for model selection. The `forecast()` function is then applied to generate the forecasts. 'BaggedETS' is a shortening for Bagged.BLD.MBB.ETS, proposed by Bergmeir et al. (2016) which considered the Moving Blocks Bootstrap (MBB) as the resampling algorithm. We consider BaggedETS with MBB, and two other alternatives for resampling: Circular Blocks Bootstrap (CBB) and Linear Process Bootstrap (LPB) . We name these strategies CBB BaggedETS and LPB BaggedETS. BMC stands for Bootstrap Model Combination.

To gauge the accuracy of the forecasts, we summarize the results according to the average and average rank across all time series based on several metrics. For point forecasts, we use the Mean Absolute Scaled Errors (MASEs) and the symmetric Mean Absolute Percentage Errors (sMAPEs), which are defined as follows:

$$MASE = \frac{1}{h} \frac{\sum_{t=1}^{h} \left| Y_t - \hat{Y}_t \right|}{\frac{1}{n-m} \sum_{t=m+1}^{n} |Y_t - Y_{t-m}|} \tag{3}$$

$$sMAPE = \frac{1}{h} \sum_{t=1}^{h} \frac{2 \left| Y_t - \hat{Y}_t \right|}{|Y_t| + \left| \hat{Y}_t \right|} \times 100\% \tag{4}$$

where Y_t and $\hat{Y}_t$ are the actual and forecasted values of the series, respectively; t is the forecast lead time from 1 to h steps ahead; n is the number of train set observations; and m is the seasonal period.

For prediction intervals, Mean Scaled Interval Scores (MSISs) are adopted, i.e.:

$$MSIS = \frac{1}{h} \frac{\sum_{t=1}^{h} (U_t - L_t) + \frac{2}{\alpha} (L_t - Y_t) \mathbf{1} \{Y_t < L_t\} + \frac{2}{\alpha} (Y_t - U_t) \mathbf{1} \{Y_t > U_t\}}{\frac{1}{n-m} \sum_{t=m+1}^{n} |Y_t - Y_{t-m}|} \tag{5}$$

where U_t and L_t are the upper and lower limits of the prediction interval produced using the selected method and $1 - \alpha$ is the desired (theoretical) coverage level.

For all metrics herein considered, the lower the values, the more accurate the point forecasts (or prediction intervals, depending on the case) are. Given their advantage of being scale independent, both MASE and sMAPE rank among the most used metrics in point forecast evaluation. As a result, they are usually considered as the official evaluation metrics for point forecasts in forecasting competitions – see, for instance, the recent M4 Competition (Makridakis et al., 2018). The MSIS, in turn, introduces penalties for the width $(U_t - L_t)$ of the prediction interval and for the instances where the actual values lie outside the specified bounds of the interval, thus offering a good balance between spread and coverage (hit rates).

For robustness assessments, besides the average and average rank across all time series of the above metrics, we also compare the boxplots of the values obtained for each metric, when the methods were individually applied to each country involved in the analysis.

6. Results and discussion

6.1. On forecasting performance

Forecasting performance for point forecasts and prediction intervals evaluation is summarized in Tables 3 and 4, respectively. Averages and average ranks of the accuracy metrics (sMAPE, MASE and MSIS) across all time series are provided. For each metric, the best performance is highlighted in **bold**, whilst the second best appears in *italics*. Overall, the best point forecasts are based on combining the MBB algorithm for resampling with the Pruning routine as an additional, intermediary step before aggregation. The same holds for prediction intervals, since the MBB PrunedBaggedETS provides the lowest values for average MSIS, regardless of the desired hit rate (theoretical coverage for the prediction interval). In terms of average rank MSIS, MBB PrunedBaggedETS also provides the most competitive results, as it is only outperformed by the MBB BaggedETS approach on a single occasion, when the desired coverage rate is 95%.

Resampling Algorithm	Combining Method	Average sMAPE	Avg. Rank sMAPE	Average MASE	Avg. Rank MASE
Bagging approaches					
MBB	Pruned BaggedETS	**4.096**	**5.438**	**1.036**	**5.063**
MBB	BaggedETS	*4.109*	*5.438*	*1.039*	*5.188*
MBB	BMC	4.206	6.875	1.103	7.219
CBB	Pruned BaggedETS	4.117	6.094	1.043	5.969
CBB	BaggedETS	4.120	5.906	1.044	5.781
CBB	BMC	4.236	6.938	1.108	7.094
LPB	Pruned BaggedETS	4.132	5.938	1.050	5.938
LPB	BaggedETS	4.153	5.875	1.054	5.875
LPB	BMC	4.289	7.438	1.120	7.875
Traditional benchmarks					
None	ETS	4.284	6.125	1.128	6.125
None	ARIMA	4.350	6.750	1.114	6.750
None	Holt-Winters	4.868	9.188	1.260	9.125

Table 3: Electricity supplied – 24 steps (October 2018 – September 2020) – Point forecasts evaluation. Average and average rank of the evaluation metrics across all countries considering 24 steps ahead forecasts (best in **bold**, second best in *italics*). MBB, CBB and LPB stand for Moving Blocks Bootstrap, Circular Blocks Bootstrap and Linear Process Bootstrap, respectively. Block size for the MBB and CBB algorithms comprises 24 observations. BMC stands for Bootstrap Model Combination.

The combined use of CBB for resampling and Pruning before aggregation is also competitive. It outperforms the traditional benchmarks and the other *Bagging* approaches. Among the latter, the ensembles that consider the LPB algorithm for resampling provide the least competitive forecasts and prediction intervals. Yet, their performances, in most cases, are still superior to the traditional benchmarks.

For each resampling algorithm considered, the Pruned BaggedETS ensembles outperformed the BMC approaches not only in terms of prediction intervals, but also for point forecasts. This is an important contribution of this paper, as it demonstrates that strategies initially developed to sharpen prediction intervals are also capable of improving point forecasts. Given this finding, it should be noted that the BMC has been claimed to be better than the original BaggedETS routine of Bergmeir et al. (2016) for point forecasts, following its performance on forecasting time series from two large competitions (M and M3). This highlights the potential of the extensions to *Bagging* that are introduced in the present study.

Resampling Algorithm	Combining Method	80% coverage	85% coverage	90% coverage	95% coverage
Average MSIS					
MBB	Pruned BaggedETS	**4.810**	**5.297**	**6.010**	**7.246**
MBB	BaggedETS	*4.852*	*5.345*	*6.052*	*7.265*
MBB	BMC	5.244	5.746	6.403	7.472
CBB	Pruned BaggedETS	4.862	5.349	6.058	7.333
CBB	BaggedETS	4.898	5.388	6.093	7.352
CBB	BMC	5.256	5.759	6.409	7.463
LPB	Pruned BaggedETS	4.927	5.436	6.178	7.519
LPB	BaggedETS	4.971	5.472	6.206	7.530
LPB	BMC	5.278	5.788	6.442	7.521
None	ETS	5.287	5.780	6.460	7.684
None	ARIMA	5.448	6.044	6.917	8.606
None	Holt-Winters	5.952	6.466	7.173	8.401
Average Rank MSIS					
MBB	Pruned BaggedETS	**3.813**	**4.438**	**4.688**	*5.188*
MBB	BaggedETS	*5.125*	*5.125*	5.125	**5.063**
MBB	BMC	6.188	6.500	6.250	6.063
CBB	Pruned BaggedETS	5.125	5.375	*5.000*	5.438
CBB	BaggedETS	6.188	6.375	5.875	5.438
CBB	BMC	6.750	6.625	6.500	6.625
LPB	Pruned BaggedETS	6.063	5.563	5.813	6.188
LPB	BaggedETS	7.188	6.688	6.563	6.250
LPB	BMC	6.875	7.188	7.000	6.875
None	ETS	6.438	6.125	7.313	7.063
None	ARIMA	7.375	7.750	7.813	8.500
None	Holt-Winters	10.875	10.250	10.063	9.313

Table 4: Electricity supplied – 24 steps (October 2018 – September 2020) – Prediction intervals evaluation considering four different theoretical coverages. Average MSISs and average rank MSISs across all countries considering 24 steps ahead forecasts (best in **bold**, second best in *italics*). Other notes: please refer to Table 3.

6.2. Distribution of forecast error metrics

We compare the distributions of the error metrics when the methods are individually applied to each time series via boxplots. These are depicted in Figures 6 and 7 for sMAPE and MASE values, respectively, and in Figures 8 and 9 for MSIS values, where for simplicity, boxplots of MSIS computed at the 90% and 95% theoretical hit rates are presented.

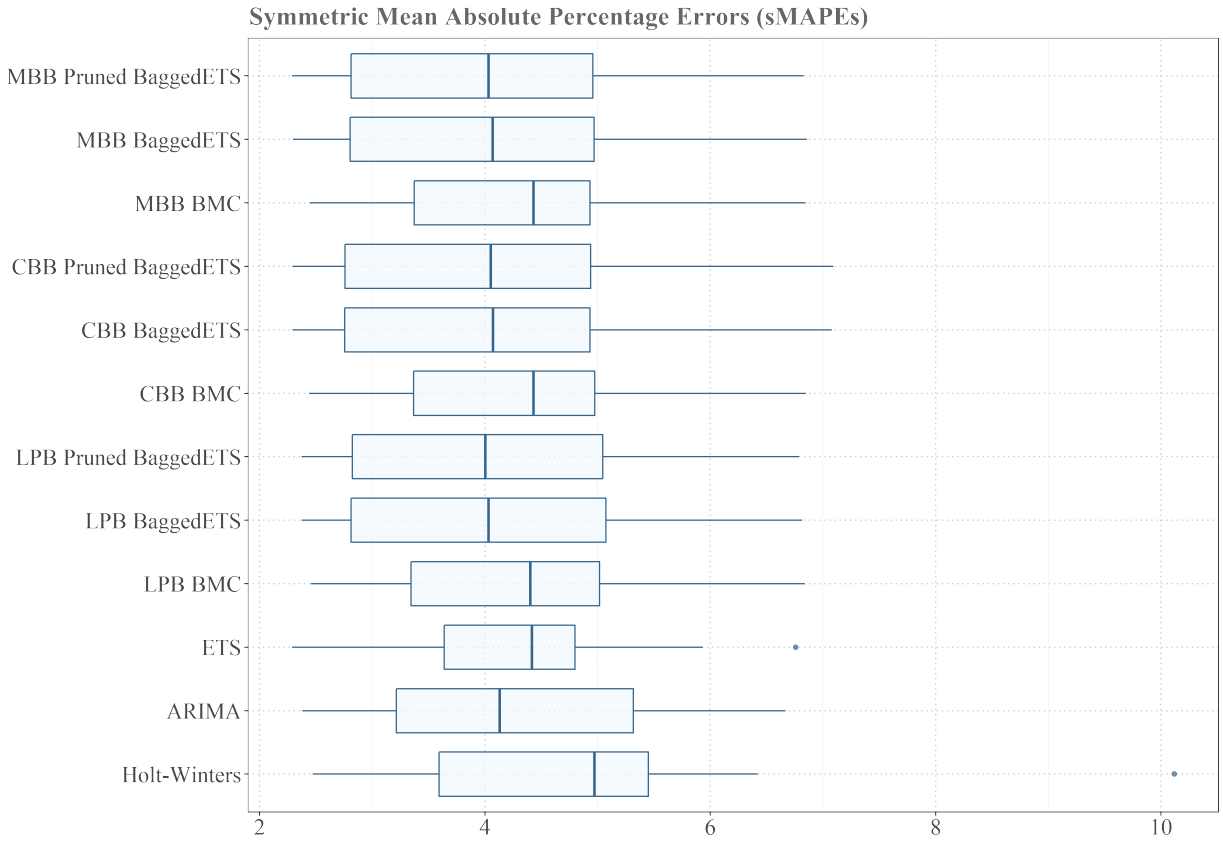

Figure 6: Boxplots – sMAPE values for each forecasting method considered.

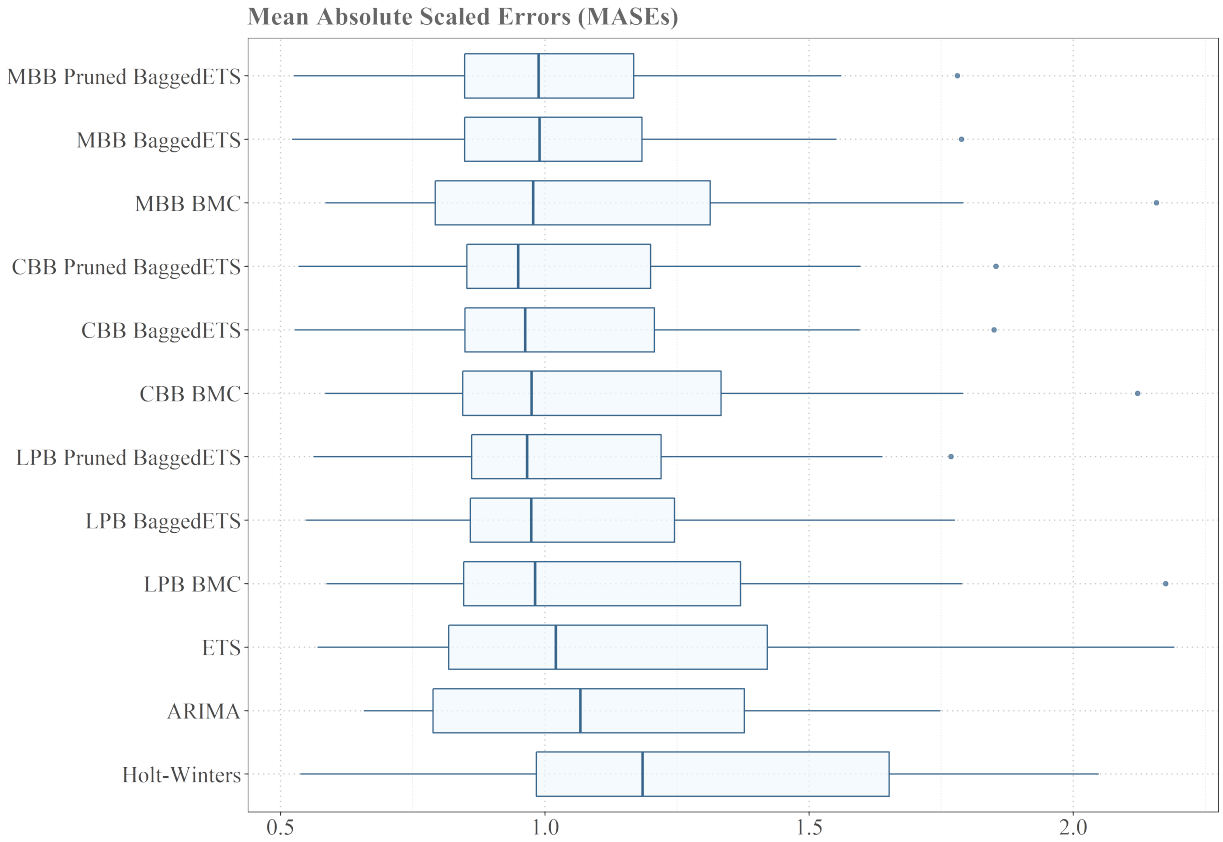

Figure 7: Boxplots – MASE values for each forecasting method considered.

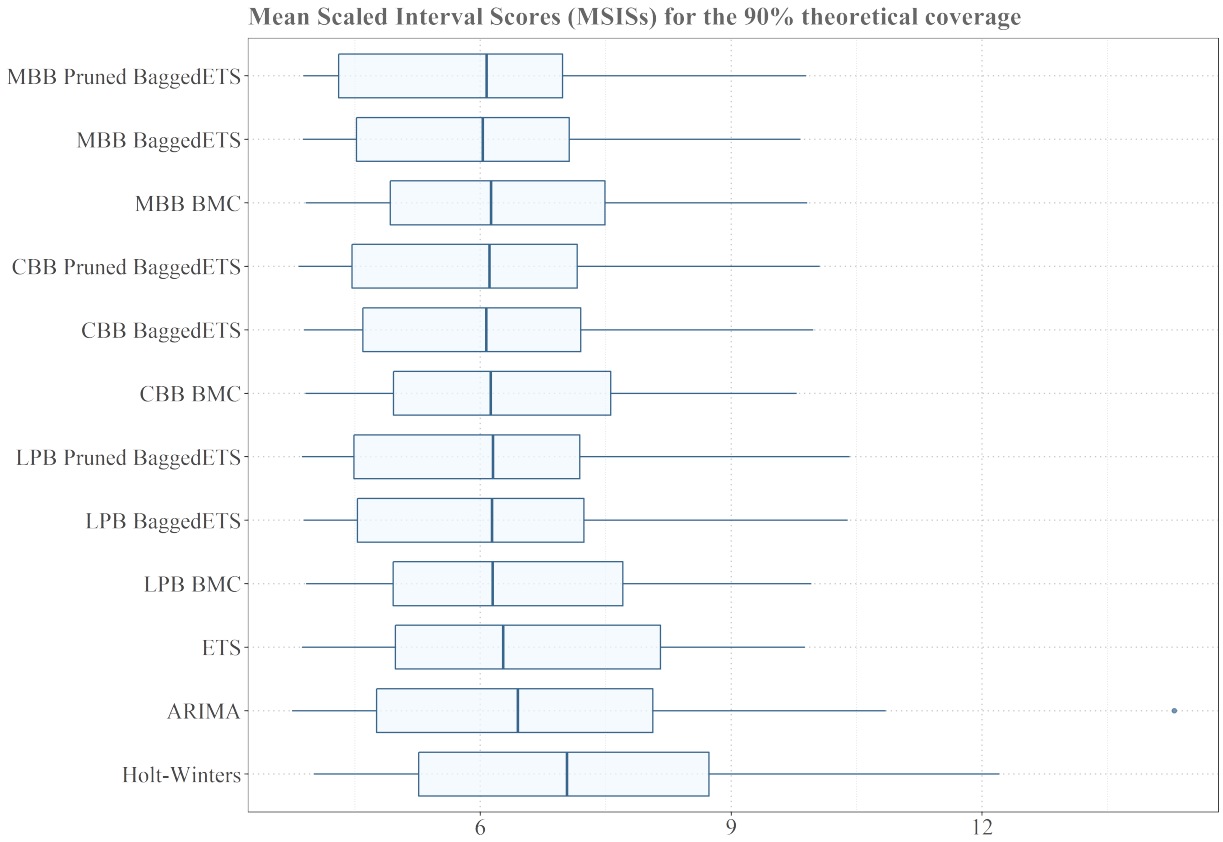

Figure 8: Boxplots – 90% theoretical coverage MSIS values for each forecasting method considered.

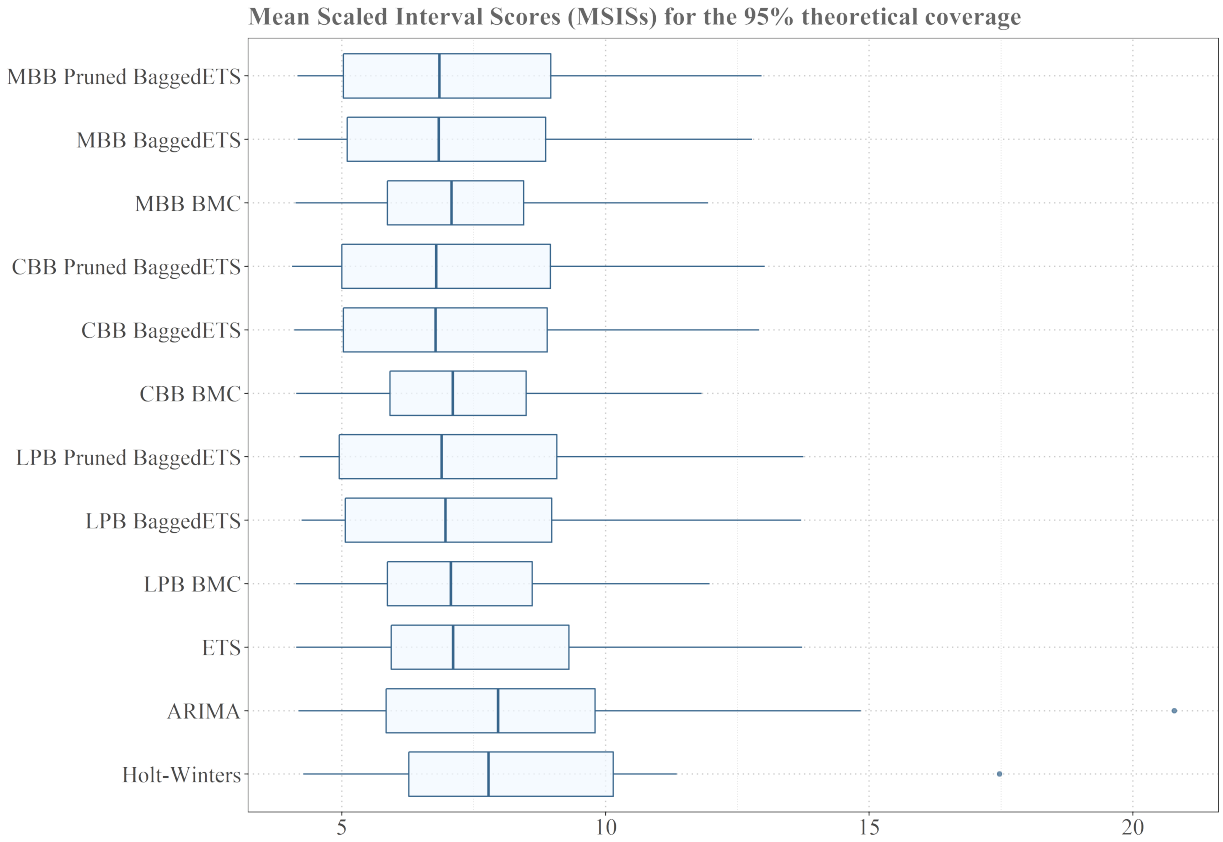

Figure 9: Boxplots – 95% theoretical coverage MSIS values for each forecasting method considered.

The boxplots for point and interval forecast metrics are in line with the average results described in the previous section, in the sense that they suggest better performances from the ensemble methods. Among the latter, the only difference from Table 3 is that the CBB and LPB Pruned BaggedETS have slightly lower median MASEs than the MBB Pruned BaggedETS, which ranked first in terms of lowest average and average rank values for sMAPE and MASE. The differences in median sMAPE among pruned approaches are practically nonexistent. Overall, we can infer that not only pruned bagging approaches result in lower averages, but they also appear to be less sensitive to outliers.

6.3. Does performance vary with country?

Turning the attention to the individual (per country) performance of the best identified method in Tables 3 and 4 (method in bold), Figures 10 and 11 illustrate the differences between the forecasts obtained via the MBB Pruned BaggedETS approach and the observed values throughout the test set period for each country. The figures also depict the prediction intervals generated using the same strategy. In half of the cases considered, the actual values remain within the boundaries of the generated prediction intervals, despite the large forecasting horizon (24 months ahead) and, particularly, the COVID-19 pandemic. For the countries in which total electricity supplied registered values outside the boundaries of the prediction intervals, the actual levels remained below the lower limit of the prediction intervals for two or three months (maximum). This situation occurred mostly during March and April 2020, when the negative impacts of social distancing and lockdown measures due to the widespread dissemination of the new coronavirus led to a significant fall in business and consumer demand, factory closures, and supply chain disruptions. The downturn in total electricity supply, however, did not last long, given the rapid resumption of electricity consumption in the summer (with the exception of New Zealand, all countries herein considered experience summer between June and August). This endorses the importance of having reliable prediction intervals, to support capacity management.

Figure 10: 24 steps ahead forecasting: forecasts and prediction intervals in blue, actual values in red.

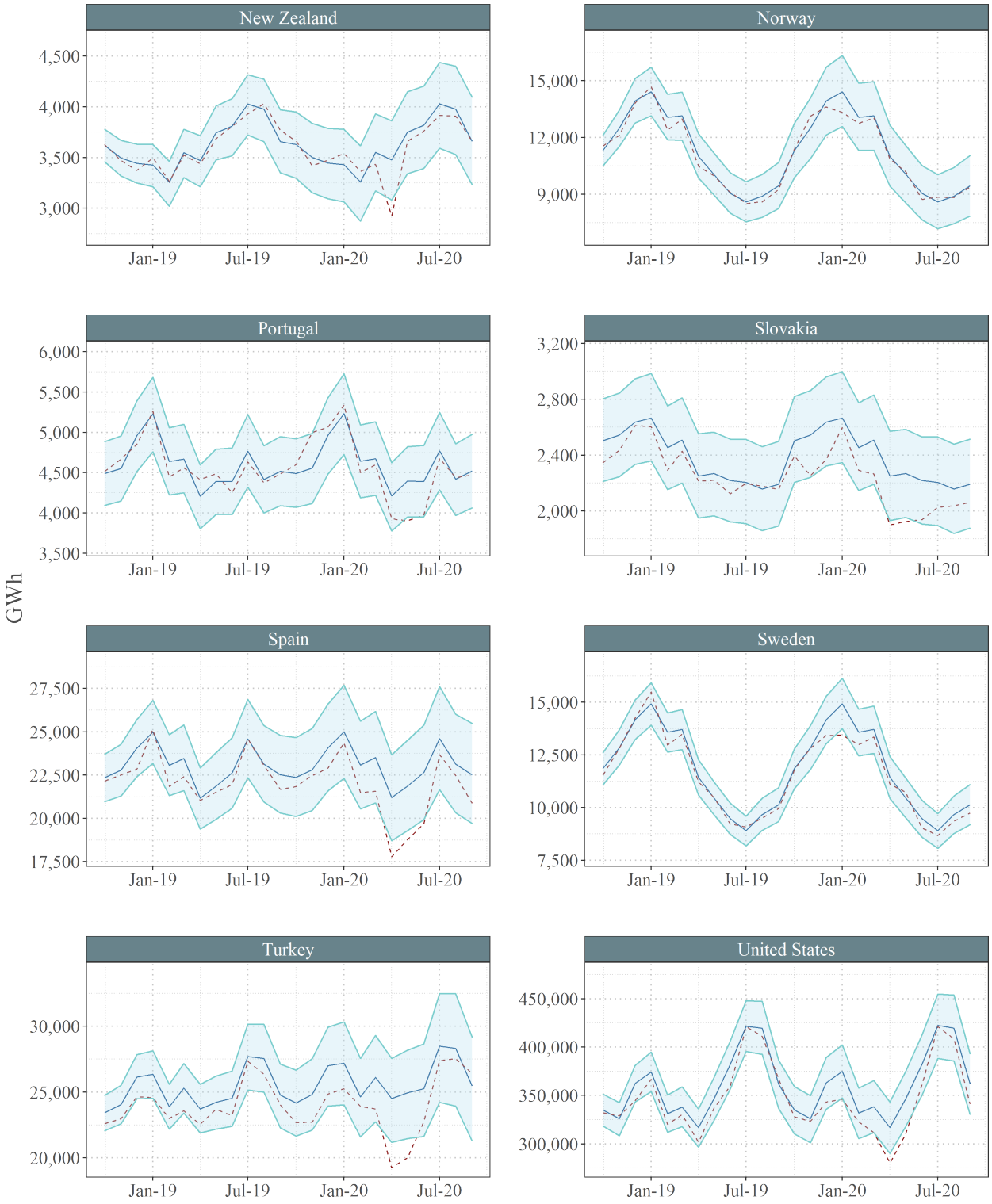

Figure 11: 24 steps ahead forecasting: forecasts and prediction intervals in blue, actual values in red.

6.4. Robustness checks for prediction intervals

In this section, we evaluate prediction intervals under alternative forecasting horizons and test set sample sizes. We consider two different periods which were not affected by the COVID-19 pandemic: the first comprised the months between October 2018 and May 2019; and the second comprised the

eight subsequent months: June 2019 to January 2020. The average accuracy, based on MSIS, for both periods is summarized in Table 5[2].

Resampling Algorithm	Combining Method	Avg MSIS (80% cov)	Avg MSIS (85% cov)	Avg MSIS (90% cov)	Avg MSIS (95% cov)
October 2018 – May 2019					
MBB	Pruned BaggedETS	**3.120**	**3.397**	**3.795**	**4.469**
MBB	BaggedETS	*3.163*	*3.450*	*3.855*	4.525
MBB	BMC	3.790	4.125	4.541	5.169
CBB	Pruned BaggedETS	3.231	3.510	3.898	4.568
CBB	BaggedETS	3.260	3.552	3.941	4.614
CBB	BMC	3.792	4.126	4.535	5.159
LPB	Pruned BaggedETS	3.218	3.506	3.906	4.649
LPB	BaggedETS	3.250	3.533	3.925	4.656
LPB	BMC	3.804	4.137	4.558	5.188
None	ETS	3.728	4.027	4.461	5.077
None	ARIMA	3.390	3.659	3.987	*4.498*
None	Holt-Winters	3.850	4.140	4.559	5.155
June 2019 – January 2020					
MBB	Pruned BaggedETS	**3.978**	**4.340**	**4.843**	**5.512**
MBB	BaggedETS	4.028	4.399	*4.898*	*5.527*
MBB	BMC	4.576	5.035	5.606	6.576
CBB	Pruned BaggedETS	*4.018*	*4.399*	4.925	5.604
CBB	BaggedETS	4.094	4.480	5.015	5.665
CBB	BMC	4.586	5.045	5.621	6.599
LPB	Pruned BaggedETS	4.065	4.463	5.013	5.713
LPB	BaggedETS	4.116	4.505	5.063	5.742
LPB	BMC	4.571	5.026	5.599	6.561
None	ETS	4.504	4.887	5.427	6.363
None	ARIMA	4.543	4.938	5.407	6.065
None	Holt-Winters	5.280	5.756	6.291	7.004

Table 5: Prediction interval evaluation under alternative forecasting horizons (est in **bold**, second best in *italics*).

The relative performance across methods is consistent, with MBB Pruned BaggedETS outperforming the other methods for all theoretical coverages considered. Pruned BaggedETS approaches perform better than the their original counterparts (BaggedETS with no pruning conducted) and the BMC method, regardless of the resampling algorithm considered (MBB, CBB or LPB). This is a substantial improvement over previous methods.

[2]Average rank MSIS results are not shown to conserve space, but can be made available upon request.

6.5. Sensitivity analysis

We consider point and interval forecasting performance under alternative settings. First, the accuracy of the prediction intervals generated by several methods for a wider range of theoretical coverage levels is assessed. Secondly, the average hit rates (actual coverages) and spreads (widths of the prediction intervals) of the same methods included in the first analysis are assessed, with respect to sharpness and calibration. The third analysis concentrates on the average values of the point forecasts and prediction interval error metrics for the Pruned BaggedETS approaches when different measures of outlier detection in pruning are taken into consideration. Finally, the performance of *Bagging* approaches when a larger number of replicas is considered (999 replicas, in lieu of 99, as in previous sections) is addressed.

6.5.1. Interval accuracy for multiple theoretical coverage levels

Figure 12 illustrates the average MSIS obtained from four different forecasting methods for the 80% to 99% range of desired (theoretical) coverage levels. The four methods considered are: the Automatic Error, Trend and Seasonality specification (ETS algorithm), applied to the original series; the Bootstrap Model Combination (BMC); the BaggedETS; and the Pruned BaggedETS. All *Bagging* approaches used the Moving Blocks Bootstrap (MBB) algorithm as resampling method, for comparison purposes[3].

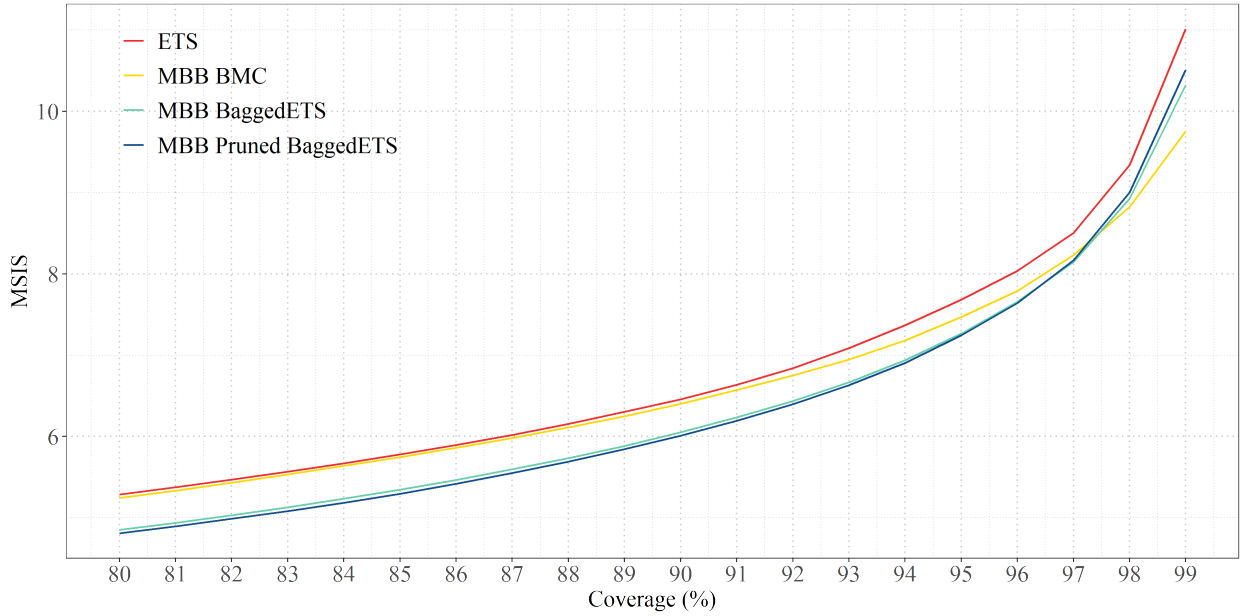

Figure 12: Sensitivity analysis – Average MSIS at different theoretical (desired) coverage levels (80% to 99%).

As previously observed, all ensemble approaches perform considerably better than the single ETS method, regardless of the desired coverage level. Among *Bagging* algorithms, the Pruned BaggedETS

[3]The comparisons among *Bagging* approaches that used alternative resampling methods (CBB or LPB) are not depicted here to conserve space but can be made available upon request.

approach is the most competitive for almost all range of theoretical coverage levels, outperforming the other methods up to the 97% level. When a 98 or 99% coverage is desired, the BMC delivers better results, since its prediction intervals are considerably wider than those generated using BaggedETS or Pruned BaggedETS.

6.5.2. Sharpness and calibration

In order to assess the sharpness and calibration of the generated prediction intervals, we plot the average hit rates per method – the average of their actual, observed coverage levels for each time series –, against their average spreads – their average interval widths divided by the in-sample mean of each time series. A prediction interval is sharp when it is capable of achieving the desired (theoretical) coverage level using a small width, i.e., when the difference between its upper and lower limit is not large, when compared with competing methods. Calibration concerns the ability of the forecasting method to deliver a prediction interval with approximately the same coverage level as the desired coverage level. In other words, the closer the hit rate is to the theoretical coverage, the more calibrated the prediction interval is.

The results of the analysis are depicted in Figure 13, where the average hit rates of the same methods considered in the previous subsection are shown against their average spreads (average standardized widths of their prediction intervals). The figure indicates that, for the same hit rate, the Pruned BaggedETS and the BaggedETS approaches generate narrower prediction intervals than the BMC and ETS methods. Hence, the first pair is sharper. In addition, Figure 13 highlights that the prediction intervals originated via the Pruned BaggedETS and the BaggedETS approaches are better calibrated than the others, since they usually deliver hit rates that are close to the expected hit rates. For instance, the first points of every curve in the figure indicate the hit rates and spreads of the involved methods when an 80% coverage level is desired. At this level, the hit rates of the Pruned BaggedETS and the BaggedETS approaches are close to 80% (the desired coverage), while the ETS prediction interval covered almost 85% of all observed values.

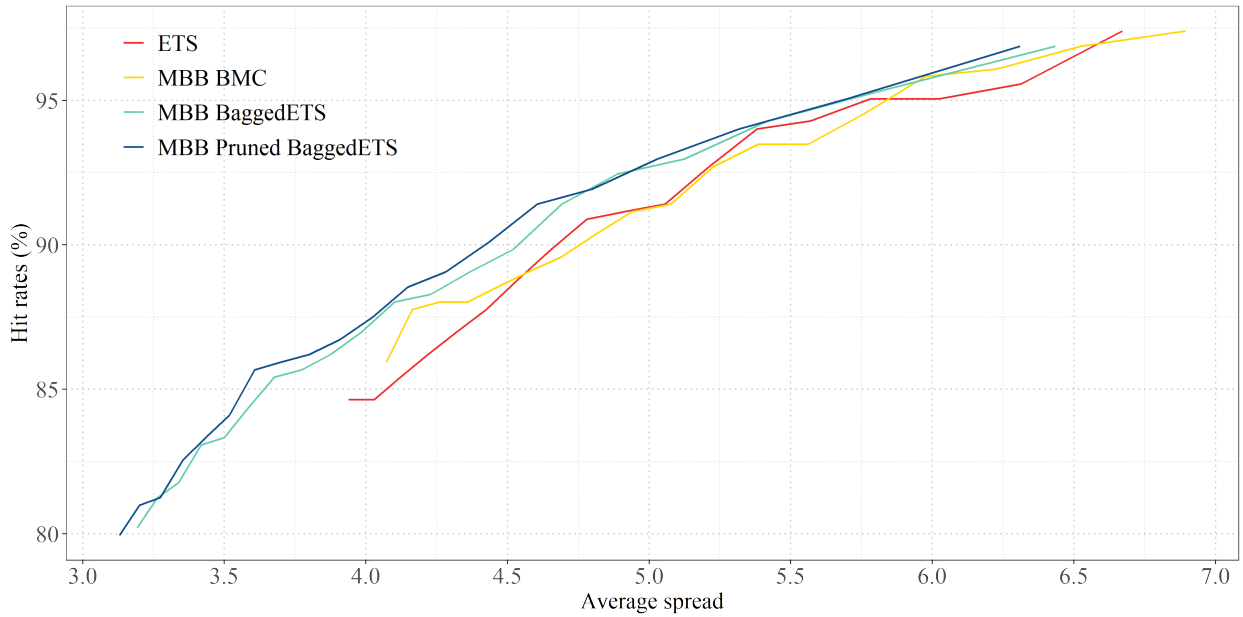

Figure 13: Sensitivity analysis – Average hit rates versus average spread (width) of the prediction intervals.

6.5.3. Different measures of outlier detection

We now focus on the possible differences arising in Pruned Bagging approaches when different measures of outlier detection are considered. The results are summarized in Table 6, for point forecasts and prediction intervals error metrics, and in Figure 14, where the MSIS values across a wide range of theoretical coverage levels are plotted.

Overall, the analysis favors less rigid metrics for prediction interval outlier detection in pruning. For instance, the values of the average error metrics for pruning approaches that considered only the interquartile range (with no multipliers) for outlier detection were slightly lower than the error metrics of the pruning strategies that used larger interquartile ranges to the same end.

Resampling Algorithm	Combining Method	Average sMAPE	Average MASE	Avg MSIS (90% cov)	Avg MSIS (95% cov)
Pruning using IQR for outlier detection					
MBB	Pruned BaggedETS	4.101	*1.036*	**5.996**	**7.236**
CBB	Pruned BaggedETS	**4.090**	1.037	6.028	7.290
LPB	Pruned BaggedETS	*4.090*	1.042	6.190	7.542
Pruning using $1.5 \times IQR$ *for outlier detection*					
MBB	Pruned BaggedETS	4.096	**1.036**	*6.010*	*7.246*
CBB	Pruned BaggedETS	4.117	1.043	6.058	7.333
LPB	Pruned BaggedETS	4.132	1.050	6.178	7.519
Pruning using $3 \times IQR$ *for outlier detection*					
MBB	Pruned BaggedETS	4.108	1.039	6.051	7.268
CBB	Pruned BaggedETS	4.118	1.043	6.096	7.356
LPB	Pruned BaggedETS	4.154	1.055	6.215	7.551
Traditional benchmarks					
None	ETS	4.284	1.128	6.460	7.684
None	ARIMA	4.350	1.114	6.917	8.606
None	Holt-Winters	4.868	1.260	7.173	8.401

Table 6: Sensitivity analysis – Point and interval forecasting accuracy evaluation for pruned approaches using different measures for outlier detection in pruning. Notes: The forecasting horizon is 24 steps (October 2018 – September 2020). Best results are highlighted in **bold** and the second best in *italics*.

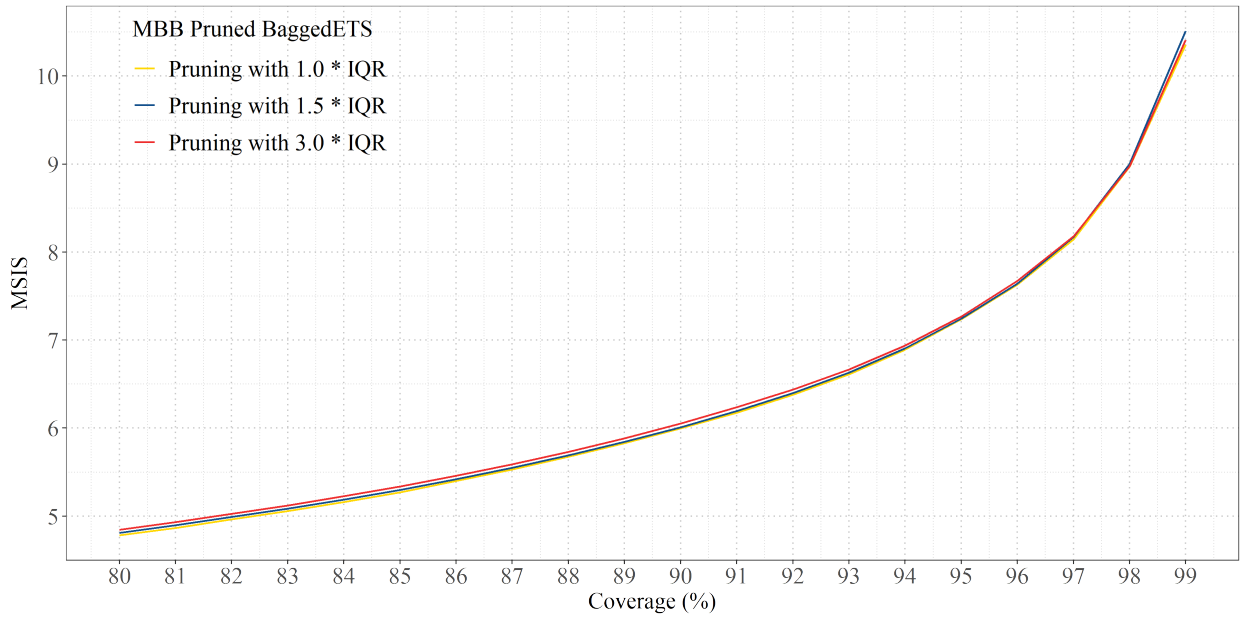

Figure 14: Sensitivity analysis – Average MSIS – MBB Pruned BaggedETS.

6.5.4. Sensitivity to the number of replicas

In order to assess the sensitivity of our proposal to the number of replicas that are generated in its second stage, 24 steps-ahead forecasting performance is compared when 99 and 999 replicas are used. The results are compiled in Table 7. In brief, differences in forecasting performance are not substantial, and perhaps surprising, slightly better results appear to be achieved with 99 replicas. Nonetheless, it can be observed that Bagging and, particularly, with the addition of pruning, can generate reliable forecasts and prediction intervals for different ensemble sizes.

Resampling Algorithm	Combining Method	Average sMAPE	Average MASE	Avg MSIS (90% cov)	Avg MSIS (95% cov)
Bagging approaches using 99 replicas					
MBB	Pruned BaggedETS	**4.096**	**1.036**	**6.010**	**7.246**
MBB	BaggedETS	*4.109*	*1.039*	6.052	7.265
MBB	BMC	4.206	1.103	6.403	7.472
CBB	Pruned BaggedETS	4.117	1.043	6.058	7.333
CBB	BaggedETS	4.120	1.044	6.093	7.352
CBB	BMC	4.236	1.108	6.409	7.463
LPB	Pruned BaggedETS	4.132	1.050	6.178	7.519
LPB	BaggedETS	4.153	1.054	6.206	7.530
LPB	BMC	4.289	1.120	6.442	7.521
Bagging approaches using 999 replicas					
MBB	Pruned BaggedETS	4.116	1.041	6.041	*7.260*
MBB	BaggedETS	4.126	1.043	6.068	7.277
MBB	BMC	4.198	1.103	6.412	7.472
CBB	Pruned BaggedETS	4.103	1.039	*6.023*	7.281
CBB	BaggedETS	4.114	1.042	6.059	7.316
CBB	BMC	4.225	1.106	6.408	7.486
LPB	Pruned BaggedETS	*4.099*	1.040	6.107	7.445
LPB	BaggedETS	4.114	1.043	6.126	7.451
LPB	BMC	4.282	1.116	6.414	7.497
Traditional benchmarks					
None	ETS	4.284	1.128	6.460	7.684
None	ARIMA	4.350	1.114	6.917	8.606
None	Holt-Winters	4.868	1.260	7.173	8.401

Table 7: Sensitivity analysis – Performance of Bagging approaches using different numbers of replicas. Notes: The forecasting horizon is 24 steps (October 2018 – September 2020). Best results are highlighted in **bold** and the second best in *italics*.

6.6. Discussion, implications and suggestions for further research

Overall, the results imply that the proposed methodology is adequate and robust to forecast electricity supply over both short and considerably long time horizons, as well as during periods of

considerable economic stress. Moreover, the prediction intervals generated do not increase much in amplitude over the forecasting horizon, which is rare in traditional forecasting approaches. This is particularly important for decision-making in the energy industry, since the more accurate and sharper the prediction intervals are, the easier it is to plan, thus the method has the potential for real savings in the sector.

The present study demonstrates the value of pruning prior to aggregation in forecasting ensembles. According to our results, the odds of obtaining undesirable outputs are substantially reduced, given that most outliers are removed from the ensemble. It should be noted that, in several countries, electricity demand (and consequently the need for electricity supply) depends on certain macroeconomic indicators, as for example Gross Domestic Product, Gross Fixed Capital Formation and Final Domestic Consumption (Nafidi et al., 2016; Streimikiene & Kasperowicz, 2016). Hence, one may consider a multivariate setting, in order to address the influence of external factors on electricity time series (Maçaira et al., 2018). We emphasize, however, that multivariate formulations usually under perform when forecasting several steps ahead. Hence, the combination of ensemble methods and univariate forecasting techniques is a promising avenue for a wide range of time series in different industries/sectors.

As methodological extensions of this research, investigations of other decomposition schemes and bootstrap algorithms constitute a future research agenda. For country-specific assessments, further studies may benefit from a hierarchical disaggregation approach. For electricity supply forecasting, this would imply using the proposed methods for each class of the domestic electric supply system. Such class-tailored analyses may contribute to a more in-depth understanding of the demand for electricity across countries, thus also potentially improving the quality of the final forecasts for total supply.

7. Summary and conclusions

Accurate prediction intervals of electricity to be supplied plays an increasingly important role in the energy sector, as both over-forecasting and under-forecasting may result in financial losses, particularly in privatized and deregulated markets. Following the philosophy of the 'wisdom of the crowds', this study proposes a novel, ensemble-based approach to generate accurate and precise forecasts and prediction intervals of electricity supply over considerably long periods and for several economies. The methodology combines *Bagging* algorithms, time series methods and new pruning routines capable of feature selection before aggregation.

In all, the results obtained in this study endorse the strength and resilience of the proposed approaches even in periods of economic distress. The performance gains are noteworthy since accurate forecasts and, particularly, prediction intervals for electricity supply, are paramount for profit/cost optimization

and investment strategies in the energy sector. They also provide valuable inputs for policymakers and regulators concerned with the provision of energy infrastructure, affordable high-quality services, and security of supply. Moreover, this new methodology is flexible, for it can be used in different contexts.

Acknowledgements

This work was supported by 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].

References

Abedinia, O., Amjady, N., & Zareipour, H. (2017). A new feature selection technique for load and price forecast of electrical power systems. *IEEE Transactions on Power Systems*, *32*, 62–74. doi:10.1109/tpwrs.2016.2556620.

Awajan, A. M., Ismail, M. T., & Wadi, S. A. (2018). Improving forecasting accuracy for stock market data using EMD-HW bagging. *PLOS ONE*, *13*, e0199582. doi:10.1371/journal.pone.0199582.

Bahrami, S., Hooshmand, R.-A., & Parastegari, M. (2014). Short term electric load forecasting by wavelet transform and grey model improved by PSO (particle swarm optimization) algorithm. *Energy*, *72*, 434–442. doi:10.1016/j.energy.2014.05.065.

Barrow, D. K., & Crone, S. F. (2016). Cross-validation aggregation for combining autoregressive neural network forecasts. *International Journal of Forecasting*, *32*, 1120–1137. doi:10.1016/j.ijforecast.2015.12.011.

Bashir, Z., & El-Hawary, M. (2009). Applying wavelets to short-term load forecasting using PSO-based neural networks. *IEEE Transactions on Power Systems*, *24*, 20–27. doi:10.1109/tpwrs.2008.2008606.

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*, 303–312. doi:10.1016/j.ijforecast.2015.07.002.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford, UK: Oxford University Press.

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

Box, G. E. P., & Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco, CA, USA: Holden-Day, Inc.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business & Economic Statistics*, *11*, 121–135. doi:10.1080/07350015.1993.10509938.

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.

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*, 748–761. doi: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. doi:10.1016/j.jairtraman.2016.12.006.

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. doi:10.1016/j.energy.2017.12.049.

Du, P., Wang, J., Yang, W., & Niu, T. (2020). Point and interval forecasting for metal prices based on variational mode decomposition and an optimized outlier-robust extreme learning machine. *Resources Policy*, *69*, 101881. doi:10.1016/j.resourpol.2020.101881.

Duan, M., Darvishan, A., Mohammaditab, R., Wakil, K., & Abedinia, O. (2018). A novel hybrid prediction model for aggregated loads of buildings by considering the electric vehicles. *Sustainable Cities and Society*, *41*, 205–219. doi:10.1016/j.scs.2018.05.009.

Elamin, N., & Fukushige, M. (2018). Modeling and forecasting hourly electricity demand by SARIMAX with interactions. *Energy*, *165*, 257–268. doi:10.1016/j.energy.2018.09.157.

Goodwin, P. (2010). The Holt-Winters Approach to Exponential Smoothing: 50 Years Old and Going Strong. *Foresight: The International Journal of Applied Forecasting*, (pp. 30–33).

Guerrero, V. M. (1993). Time-series analysis supported by power transformations. *Journal of Forecasting*, *12*, 37–48. doi:10.1002/for.3980120104.

Heskes, T. (1996). Practical confidence and prediction intervals. NIPS'96 (p. 176–182). Cambridge, MA, USA: MIT Press.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., & Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *454*, 903–995. doi:10.1098/rspa.1998.0193.

Hwang, J. T. G., & Ding, A. A. (1997). Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, *92*, 748–757.

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

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice*. (3rd ed.). OTexts: Melbourne, Australia. URL: OTexts.com/fpp3.

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*, 439–454. doi:10.1016/s0169-2070(01)00110-8.

IEA (2021). International Energy Agency Monthly Electricity Statistics. https://www.iea.org/reports/monthly-electricity-statistics. Accessed: 2021-02-20.

Jiang, W., Wu, X., Gong, Y., Yu, W., & Zhong, X. (2020). Holt–winters smoothing enhanced by fruit fly optimization algorithm to forecast monthly electricity consumption. *Energy*, *193*, 116779. doi:10.1016/j.energy.2019.116779.

Khosravi, A., Nahavandi, S., & Creighton, D. (2010). Load forecasting and neural networks: A prediction interval-based perspective. In *Computational Intelligence in Power Engineering* (pp. 131–150). Springer Berlin Heidelberg. URL: https://doi.org/10.1007/978-3-642-14013-6_5. doi:10.1007/978-3-642-14013-6_5.

Khosravi, A., Nahavandi, S., & Creighton, D. (2013). A neural network-GARCH-based method for construction of prediction intervals. *Electric Power Systems Research*, *96*, 185–193. doi:10.1016/j.epsr.2012.11.007.

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

Kuster, C., Rezgui, Y., & Mourshed, M. (2017). Electrical load forecasting models: A critical systematic review. *Sustainable Cities and Society*, *35*, 257–270. doi:10.1016/j.scs.2017.08.009.

Li, S., Goel, L., & Wang, P. (2016). An ensemble approach for short-term load forecasting by extreme learning machine. *Applied Energy*, *170*, 22–29. doi:10.1016/j.apenergy.2016.02.114.

Lim, M., & Hastie, T. (2015). Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, *24*, 627–654. doi:10.1080/10618600.2014.938812.

Maçaira, P. M., Thomé, A. M. T., Oliveira, F. L. C., & Ferrer, A. L. C. (2018). Time series analysis with explanatory variables: A systematic literature review. *Environmental Modelling & Software*, *107*, 199–209. doi:10.1016/j.envsoft.2018.06.004.

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*, 111–153. doi:10.1002/for.3980010202.

Makridakis, S., & Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476. doi: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*, 802–808. doi:10.1016/j.ijforecast.2018.06.001.

McMurry, T. L., & Politis, D. N. (2010). Banded and tapered estimates for autocovariance matrices and the linear process bootstrap. *Journal of Time Series Analysis*, *31*, 471–482. doi:10.1111/j.1467-9892.2010.00679.x.

Meira, E., Oliveira, F. L. C., & Jeon, J. (2021). Treating and pruning: New approaches to forecasting model selection and combination using prediction intervals. *International Journal of Forecasting*, *37*, 547–568. doi:10.1016/j.ijforecast.2020.07.005.

Misiorek, A., Trueck, S., & Weron, R. (2006). Point and interval forecasting of spot electricity prices: Linear vs. non-linear time series models. *Studies in Nonlinear Dynamics & Econometrics*, *10*, 1–34. doi:10.2202/1558-3708.1362.

Nafidi, A., Gutiérrez, R., Gutiérrez-Sánchez, R., Ramos-Ábalos, E., & Hachimi, S. E. (2016). Modelling and predicting electricity consumption in spain using the stochastic gamma diffusion process with exogenous factors. *Energy*, *113*, 309–318. doi:10.1016/j.energy.2016.07.002.

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*, 1621–1629. doi:10.1080/01621459.1997.10473684.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., & et al. (2020). Forecasting: theory and practice. arXiv:2012.03854.

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*, 545–554. doi:10.1016/j.ejor.2018.01.045.

Politis, D. N., & Romano, J. P. (1991). *A circular block-resampling procedure for stationary data*. Technical Report EFS NSF 370 Department of Statistics, Stanford University Stanford, CA, USA.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: https://www.R-project.org/.

Rendon-Sanchez, J. F., & de Menezes, L. M. (2019). Structural combination of seasonal exponential smoothing forecasts applied to load forecasting. *European Journal of Operational Research*, *275*, 916 – 924. doi:10.1016/j.ejor.2018.12.013.

da Silva, F. L., Oliveira, F. L. C., & Souza, R. C. (2019). A bottom-up bayesian extension for long term electricity consumption forecasting. *Energy*, *167*, 198–210. doi:10.1016/j.energy.2018.10.201.

Song, H., Liu, A., Li, G., & Liu, X. (2021). Bayesian bootstrap aggregation for tourism demand forecasting. *International Journal of Tourism Research*, . doi:10.1002/jtr.2453.

Streimikiene, D., & Kasperowicz, R. (2016). Review of economic growth and energy consumption: A panel cointegration

analysis for EU countries. *Renewable and Sustainable Energy Reviews*, *59*, 1545–1549. doi:10.1016/j.rser.2016.01.041.

Sulandari, W., Subanar, Lee, M. H., & Rodrigues, P. C. (2020). Indonesian electricity load forecasting using singular spectrum analysis, fuzzy systems and neural networks. *Energy*, *190*, 116408. doi:10.1016/j.energy.2019.116408.

Sun, S., Sun, Y., Wang, S., & Wei, Y. (2018). Interval decomposition ensemble approach for crude oil price forecasting. *Energy Economics*, *76*, 274–287. doi:10.1016/j.eneco.2018.10.015.

Szafranek, K. (2019). Bagged neural networks for forecasting polish (low) inflation. *International Journal of Forecasting*, *35*, 1042–1059. doi:10.1016/j.ijforecast.2019.04.007.

Vilar, J., Aneiros, G., & Raña, P. (2018). Prediction intervals for electricity demand and price using functional data. *International Journal of Electrical Power & Energy Systems*, *96*, 457–472. doi:10.1016/j.ijepes.2017.10.010.

Wu, J., Wang, J., Lu, H., Dong, Y., & Lu, X. (2013). Short term load forecasting technique based on the seasonal exponential adjustment method and the regression model. *Energy Conversion and Management*, *70*, 1–9. doi:10.1016/j.enconman.2013.02.010.

Xie, W., Wu, W.-Z., Liu, C., & Zhao, J. (2020). Forecasting annual electricity consumption in china by employing a conformable fractional grey model in opposite direction. *Energy*, *202*, 117682. doi:10.1016/j.energy.2020.117682.

Zhu, X., Dang, Y., & Ding, S. (2020). Using a self-adaptive grey fractional weighted model to forecast jiangsu's electricity consumption in china. *Energy*, *190*, 116417. doi:10.1016/j.energy.2019.116417.