Improving forecasting by subsampling seasonal time series

Xixi Li[a], Fotios Petropoulos[b], Yanfei Kang[c,*]

[a]*Department of Mathematics, The University of Manchester, UK.*
[b]*School of Management, University of Bath, UK.*
[c]*School of Economics and Management, Beihang University, China.*

Abstract

Time series forecasting plays an increasingly important role in modern business decisions. In today's data-rich environment, people often aim to choose the optimal forecasting model for their data. However, identifying the optimal model requires professional knowledge and experience, making accurate forecasting a challenging task. To mitigate the importance of model selection, we propose a simple and reliable algorithm to improve the forecasting performance. Specifically, we construct multiple time series with different sub-seasons from the original time series. These derived series highlight different sub-seasonal patterns of the original series, making it possible for the forecasting methods to capture diverse patterns and components of the data. Subsequently, we produce forecasts for these multiple series separately with classical statistical models (ETS or ARIMA). Finally, the forecasts are combined. We evaluate our approach on widely-used forecasting competition data sets (M1, M3, and M4) in terms of both point forecasts and prediction intervals. We observe performance improvements compared with the benchmarks. Our approach is particularly suitable and robust for the data with higher frequency. To demonstrate the practical value of our proposition, we showcase the performance improvements from our approach on hourly load data that exhibit multiple seasonal patterns.

Keywords: forecasting; combinations; subsampling; sub-seasonal patterns; load forecasting

*Corresponding author

Email addresses: xixi.li@manchester.ac.uk (Xixi Li), f.petropoulos@bath.ac.uk (Fotios Petropoulos), yanfeikang@buaa.edu.cn (Yanfei Kang)

URL: https://orcid.org/0000-0001-5846-3460 (Xixi Li), https://orcid.org/0000-0003-3039-4955 (Fotios Petropoulos), https://orcid.org/0000-0001-8769-6650 (Yanfei Kang)

1. Introduction

Time series forecasting is an indispensable part of modern businesses and a valuable input in the decision process for production, finance, planning, scheduling, and other activities (Petropoulos et al., 2021). We are living in a big data era where large collections of time series constantly emerge. For example, large retailers need to forecast the sales for tens of thousands of products in each forecast cycle. There is a need to have robust solutions to process and forecast these large volume of series in a batch, automatic manner. However, given a set of candidate forecasting models, identifying the most appropriate and reliable model for each data is not always a straightforward task.

One way to deal with modelling time series in an automatic way would be to resort to statistical model selection approaches, like information criteria or cross-validation. Such approaches are able to select the most "appropriate" forecasting model, based on the in-sample fit (penalised for model complexity) or based on past performance on a hold-out set of observations. However, real-life time series data do not abide by particular data generation processes, and established patterns may change over time. In a famous quote, George Box pointed out that "all models are wrong, but some are useful" (Box and Draper, 1987, p. 424). Wolpert (1996) also argued that a single model cannot fit in all situations. Instead, combining the forecasts across multiple models has shown not only to perform well but also to reduce the variance of the forecasts (Bates and Granger, 1969; Hibon and Evgeniou, 2005).

The good performance of forecast combinations is also linked with the three inherent uncertainties that a forecaster needs to tackle (Hyndman et al., 2008; Petropoulos et al., 2018). First, the model uncertainty refers to being able to identify the correct model form. Within the exponential smoothing (ExponenTial Smoothing or Error, Trend, Seasonality; ETS) family of models, this means identifying if trend and/or seasonal components need to be included, but also their interactions (additive or multiplicative). In the autoregressive integrated moving average (ARIMA) family, tackling model uncertainty entails correctly identifying the appropriate (non-seasonal and seasonal) autoregressive and moving average terms. Second, even if the correct model form has been selected, there is uncertainty related to the estimates of the model's parameters, such as the smoothing parameters of the exponential smoothing models. Finally, data uncertainty refers to the variation of the inherent random component of the series.

While fitting many different models on the same data and averaging their forecasts is sometimes sufficient to achieve performance improvements (Kolassa, 2011; Petropoulos and Svetunkov, 2020), a better approach is to manipulate the data and extract additional information from them, if there is any. For example, Assimakopoulos and Nikolopoulos (2000) and Fiorucci et al. (2016), among others, changed the local curvatures of the seasonally-adjusted data to be able to predict short and long-term behaviours. The resulting method, Theta, performed well in a variety of real-life settings, including a first place in the M3-competition (Makridakis and Hibon, 2000). Kourentzes et al. (2014) and Athanasopoulos et al. (2017) worked with multiple temporal aggregation levels of the same signal and performed forecast combinations across different sampling frequencies. Bergmeir et al. (2016) and Petropoulos et al. (2018) performed bootstrapping on the remainder of the time series decomposition to create several instances of the same series, which were then modelled independently and their forecasts were aggregated.

Our work builds on previous studies in an attempt to extract as much information from the data as possible. We particularly focus on the challenge of correctly modelling seasonal patterns. A seasonal time series pattern exists when the data is influenced by seasonal factors (e.g., the quarter or the month of the year, Hyndman, 2011). Seasonality is observed in a fixed and known period, unlike other cyclical patterns. Most of the existing time series models are designed to adapt to simple seasonal patterns with short periodic cycles with respect to the series' time granularity. Examples of such simple models are the Seasonal Naïve method, Holt-Winters' Exponential Smoothing (Winters, 1960), Damped Trend Seasonal Method (Gardner Jr, 2006), and Seasonal ARIMA (SARIMA) models. In the case of multiple, complex seasonal patterns, more complicated forecasting methods need to be considered (De Livera et al., 2011).

For series that may display seasonal patterns, it is reasonable and feasible to use the observations from one season of the historical data to forecast the corresponding season in the future. Intuitively and similarly, several adjacent seasons could be predicted using the respective past observations. Starting with the conjecture that we can use the observations of some adjacent seasons in the past to predict the values of the corresponding seasons in the future, we construct multiple time series which consist of the observations of only one or some adjacent seasons. We refer to these new subsampled series as *sub-seasonal* series, as they contain part of but not all the possible seasonal information.

These derived series highlight different sub-seasonal patterns of the original time series, thus simplifying their modelling and estimation.

After constructing all possible sub-seasonal series, we extrapolate these derived series as well as the original series using standard forecasting workhorses. In this paper, we focus our empirical investigation on two widely-used forecasting families: ETS and ARIMA (Hyndman and Athanasopoulos, 2018). However, our framework is model-agnostic and could be expanded to any other forecasting model suitable to deal with simple seasonal patterns. In modelling each sub-seasonal series, we select the 'optimal' model form and set of parameters separately and independently from other sub-seasonal series (or the original series). The forecasts produced for the different series are then combined, effectively tackling issues surrounding model and parameter uncertainty.

The key innovations of our approach are as follows:

- We zoom in the sub-seasonal patterns of the original series that are simpler to model.
- We do not rely on a single model and its forecasts that are based on the original series. Instead, we mitigate the importance of model selection by combining forecasts across many sub-seasonal series.
- Our proposed approach is simple, transparent, and does not rely on a particular family of forecasting models. In particular to the latter, we do not propose a new forecasting model per se, but a framework that can be plugged into any existing model.

We conduct an extensive empirical evaluation of our proposed framework using 75 thousand real time series from the Makridakis forecasting competitions (Makridakis et al., 1982; Makridakis and Hibon, 2000; Makridakis et al., 2020). Our framework works as a self-improving mechanism for ETS and ARIMA families of models, resulting in better point-forecast accuracy and uncertainty estimation (captured through prediction intervals) than simply applying ETS and ARIMA on the original series. We observe that improvements in performance are greater for longer forecasting horizons, where uncertainty is also higher. When applied to the case of load forecasting, our approach produces reliable forecasts, showcasing its efficacy on time series data with multiple and complex seasonal patterns.

The rest of the article is organised as follows: Section 2 offers a short literature review

on relevant studies. Section 3 describes the methodology for the proposed forecasting approach. Section 4 presents the experimental results and their statistical significance. Section 5 shows the application to electric load forecasting. Section 6 offers our discussions and insights. Finally, Section 7 provides our conclusions.

2. Background research

Given the plethora of available models to choose from, several selection strategies have been developed over the years. Such strategies compare the performance of different candidate models in order to choose the best one, based on some criteria. Some approaches, namely information criteria, select the model with the maximum likelihood by adding a penalty to compensate for the over-fitting of more complex models (Bishop, 2006). Popular information criteria include Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC). However, Bishop (2006) pointed out that information criteria could not properly account for the uncertainty of the models' parameters and tend to prefer simple models to more complex ones.

An alternative to selecting with information criteria is judging the models based on their past performance over multiple lead times. Such approaches are known as validation and cross-validation for time series, and are closely related to the concepts of fixed and rolling origin evaluation (Tashman, 2000). Fildes and Petropoulos (2015) claimed that (cross-)validation approaches are better than information criteria as the former ones take into account longer forecast horizons when evaluating past forecasts while the latter methods are based on the one-step-ahead in-sample forecast errors. Montero-Manso et al. (2020) achieved a good performance on the M4 forecasting competition using an approach that is based on a validation set-up enhanced by time series features. On the other hand, Billah et al. (2006) showed that the performance of validation is similar to that of information criteria. In any case, (cross-)validation approaches require longer series and more computational resources.

Other approaches to model selection take a more data-driven approach, essentially attempting to answer the question: Which is the best forecasting method for my data? Many attempts have been made to tackle the task of model selection using time series features. For example, Reid (1972) pointed out that the nature of the data has an influence on the performance of the forecasting methods. Collopy and Armstrong (1992)

developed a rule-based system containing 99 rules to produce forecasts based on the features of the data. Adya et al. (2001) identified time series features automatically for rule-based forecasting. Petropoulos et al. (2014) explored the factors that affect forecasting accuracy in the field of demand forecasting, and proposed related selection protocols. Kang et al. (2017) used Principal Component Analysis (PCA) to visualise the forecasting algorithm performance in the time series instance spaces and had a better understanding of their relative performance. Finally, Talagala et al. (2018) used a decision tree to select the best forecasting method based on 42 manual features.

Regardless of the available options to perform model selection between forecasting models, the *No Free Lunch* theorem suggests that there does not exist one method that will perform always best across series, nor across time, due to the dynamic nature of the problem. As such, instead of considering selecting a single model (per series or across series), combinations of forecasts are an alternative way forward. Forecast combinations tend to yield better results when one is averaging forecasts from robust models but also when there is significant diversity among the forecasts (Wang and Petropoulos, 2016; Thomson et al., 2019; Lichtendahl and Winkler, 2020; Kang et al., 2021). Combinations of forecasts address two of the sources of uncertainty in forecasting (Petropoulos et al., 2018), model's form and model's parameters uncertainty, as they do not rely on a single model anymore.

Apart from obtaining diverse and robust forecasts, another critical factor in the efficiency of forecast combinations is the estimation of the weights. Recently scholars tried to use advanced machine learning techniques to optimise optimal weights with a focus on the time series features of the target time series. Montero-Manso et al. (2020) trained XGBoost (Chen and Guestrin, 2016) to obtain optimal weights for each method based on the 42 manual time series features, achieving the second-best performance in the M4 Competition. Li et al. (2020) used a similar approach to estimate optimal combination weights on the automatic image features of the series and produced comparable performance with the top performers in M4 Competition. Regardless of the good performance of some optimal-weighting combination strategies, equal-weights combinations also perform remarkably well (Petropoulos and Svetunkov, 2020), and often are hard to beat by other more complex approaches (Jose and Winkler, 2008; Genre et al., 2013). This is referred to as the "forecast combination puzzle" in the literature (Watson and Stock,

2004; Smith and Wallis, 2009; Claeskens et al., 2016).

A special case in forecast combinations refers to approaches that, instead of fitting different models to the same data, manipulate the original data and create other series in an attempt to amplify some particular series characteristics while suppressing others. One such approach can be found in the core of the theta method (Assimakopoulos and Nikolopoulos, 2000). Instead of applying forecasting models on the original data, the theta method works on the seasonally-adjusted data. These are then further decomposed into theta lines, aiming to capture the short and long-term dynamics. The theta method essentially changes the data by manipulating the residual of the regression model. This simple method is a robust benchmark in the forecasting literature and was the top-performing method at the M3 forecasting competition (Makridakis and Hibon, 2000).

Another approach that also uses data manipulation to extract additional information is (non-overlapping) temporal aggregation (Nikolopoulos et al., 2011). This technique down-samples the original series in order to obtain more series of lower frequencies. Higher temporal aggregation levels offer smoother, less intermittent series, allowing for better extrapolation of trend patterns. Still, lower levels of aggregation allow for better estimation of seasonal patterns. Producing forecasts for many levels independently and then combining such forecasts has shown sizable improvements in accuracy both for fast and slow-moving series (Kourentzes et al., 2014; Petropoulos and Kourentzes, 2015). More recently, the combination of forecasts from several temporal aggregation levels is made in a hierarchical fashion (Athanasopoulos et al., 2017; Kourentzes and Athanasopoulos, 2021).

A third approach that follows the concept of extracting more information from the data is bootstrapping (e.g., Bergmeir et al., 2016; Hasni et al., 2019). In particular for time series forecasting, Bergmeir et al. (2016) proposed decomposing the original signal to the trend, seasonal, and remainder components, bootstrapping the latter in order to create several series of remainders, and reconstructing the series by putting together trend, seasonality, and each bootstrapped series of remainders. Following the above process, one can construct several series of the same structure (trend and seasonality) but with different noise. Instead of simply forecasting the original series, one can forecast all reconstructed series and then aggregate (average) the forecasts. 'Bagging', when applied in conjunction with an algorithm that considers a large number of candidate

models and automatically selects the best one (such as automatic exponential smoothing or ARIMA), can tackle the three sources of uncertainty: model form, model parameters, and data (Petropoulos et al., 2018).

3. Forecasting through integrating the forecasts of multiple time series with different sub-seasonal patterns

Suppose we observe a quarterly series of three years, and our target is to forecast the next year (four quarters). The standard approach would be to consider all 12 available data points and forecast using a time series method. In our approach, we construct various sub-seasonal series (e.g., series consisting of only some of the quarters of each year) from the original target series. The final forecasts can be obtained by combining the forecasts produced from the series with different sub-series patterns. Figure 1 graphically illustrates our method. For this particular series, the proposed forecasting procedure is as follows.

1. Create a new series that consists only of the first quarters (Q_1) of each year. Repeat for the second (Q_2), third (Q_3), and fourth (Q_4) quarters of the year. This is the first level of information, where each sub-seasonal series contains exactly one quarter.
2. Create a new series that only consists of two adjacent quarters of each year. Specifically, extract the observations of the first and second quarters (Q_1&Q_2) of each year and construct a new series accordingly. Repeat for Q_2&Q_3, Q_3&Q_4 and Q_4&Q_1 (or Q_1&Q_4). This is the second level of information, where each sub-seasonal series contains exactly two adjacent quarters.
3. Create a new series that only consists of three adjacent quarters of each year. Specifically, extract the observations of the first, second and third quarters (Q_1&Q_2&Q_3) of each year and construct a new series accordingly. Repeat for Q_2&Q_3&Q_4, Q_3&Q_4&Q_1 (or Q_1&Q_3&Q_4) and Q_4&Q_1&Q_2 (or Q_1&Q_2&Q_4). This is the third level of information, where each sub-seasonal series contains exactly three adjacent quarters.
4. Forecast the constructed sub-seasonal series, and the original series (Q_1&Q_2&Q_3&Q_4, which is the fourth level of information), using standard time series forecasting methods (e.g., ETS or ARIMA). That is, we use the historical observations of some adjacent seasons to forecast the corresponding seasons in the future.

5. Combine the obtained forecasts with equal weights. Note that the original series, $Q_1\&Q_2\&Q_3\&Q_4$, is repeated four times so that equal importance is given to all information levels.

More generally, assume we are interested in forecasting a seasonal time series $\{y_t, t = 1, 2, \cdots, T\}$ with a sampling frequency m, the forecasting procedure through integrating the forecasts of multiple time series with different sub-seasonal patterns is as follows.

1. **Constructing** multiple series with sub-seasonal patterns. Create new time series that contain observations of only one or a few adjacent seasons and the corresponding frequency is equal to the number of seasons.
2. **Forecasting** each of the newly derived series using standard time series forecasting methods. Theoretically, the number of the constructed series is m^2 for all the information levels. However, in practice, we only need to forecast the original series once for the last information level, and when the forecasting horizon h is less than the frequency m, we do not need to construct and forecast all the subsampled series. Therefore, the number of subsampled series M that need to be forecast can be written as

$$M = \begin{cases} \frac{(m-h)(m+h-1)}{2} + (h-1)m + 1, & \text{if } h < m, \\ m(m-1) + 1, & \text{otherwise.} \end{cases}$$

3. **Combining** the forecasts produced from the sub-seasonal series with equal weights. Note that when combining, the original series in the last information level is repeated m times so that equal importance is given to all information levels. Since the 'optimal' model form and set of parameters are estimated separately and independently when modelling each sub-seasonal series, forecast combination makes it possible to tackle issues around model and parameter uncertainty (Petropoulos et al., 2018).

In the following sections, we introduce each step of our proposed method in detail.

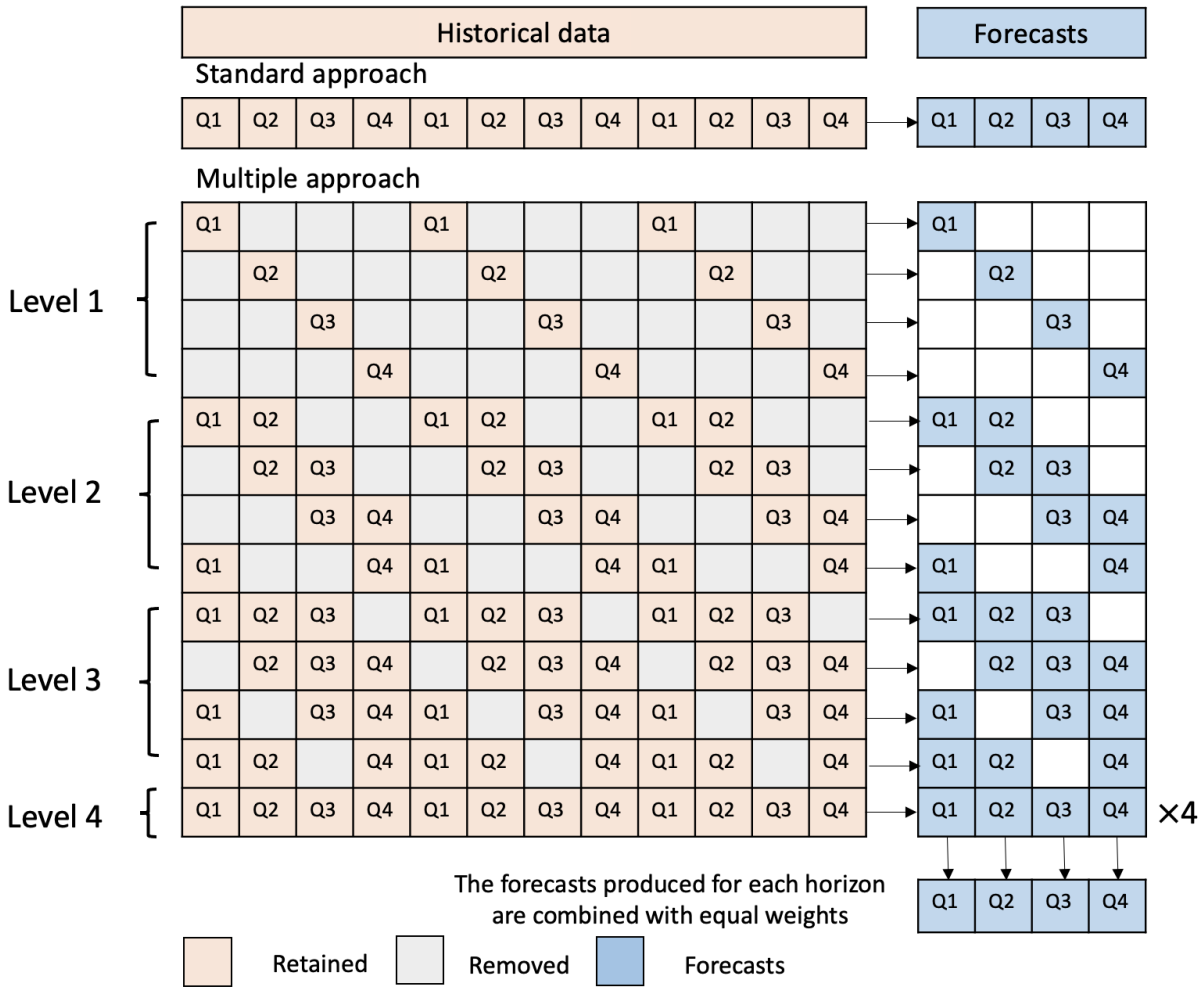

Figure 1. Caption: A graphical illustration of the standard approach and the proposed multiple approach applied to a quarterly series.

Figure 1. Alt text: Sequences of adjacent boxes showing the standard approach and the steps of multiple approach applied to a quarterly series, with each box representing one observation.

Figure 1. Long description: A graphical illustration of the proposed approach applied to a quarterly series with three years of historical observations. The aim is to forecast the next year (four quarters). The light orange, grey and blue boxes represent the retained, the removed observations and forecasts respectively. The forecasting procedure consists of three main steps: 1) create new time series only containing one or more adjacent seasons, i.e., Q_1, Q_2, Q_3, Q_4, $Q_1\&Q_2$, $Q_2\&Q_3$, $Q_3\&Q_4$, $Q_4\&Q_1$, $Q_1\&Q_2\&Q_3$, $Q_2\&Q_3\&Q_4$, $Q_3\&Q_4\&Q_1$, $Q_4\&Q_1\&Q_2$, and the original series $Q_1\&Q_2\&Q_3\&Q_4$. The level value in the figure represents the corresponding several adjacent quarterly observations are retained, 2) forecast separately each sub-seasonal and the original series, and 3) combine the produced forecasts with equal weights. Note that when combining, $Q_1\&Q_2\&Q_3\&Q_4$ is repeated four times so that equal importance is given to all information levels.

3.1. Constructing multiple time series with sub-seasonal patterns

In this section, we use the quarterly time series Q520 from the M3 competition data set (Makridakis and Hibon, 2000) to demonstrate the construction of sub-seasonal series from the original time series and illustrate the importance of forecasting with sub-seasonal patterns. Figure 2 shows the constructed sub-seasonal series. For this particular time series, we can construct $m(m-1)+1 = 13$ series with different sub-seasonal patterns containing the observations of only one or a few adjacent quarters. For each sub-seasonal series, the corresponding frequency is equal to the number of the adjacent quarters. Note that among the thirteen series, the last series shown in the last row of Figure 2 corresponds to the original, target series. The series in each row of Figure 2 contain the same number of quarters, i.e., belong to the same information level. In particular, each series in the first row contains observations of only one quarter from the original series. The series in the second row contain observations of two adjacent quarters, and the third row shows the sub-seasonal patterns from observations of three adjacent quarters.

We observe from Figure 2 that each newly created series highlights different sub-seasonal patterns and components of the original series. The four series in the first row, containing only one quarter, show an upward trend, while the trend in series that contains Q_1 or Q_2 is almost linear. From the second row of Figure 2, we can see that the sub-seasonal series containing $Q_1\&Q_2$, $Q_2\&Q_3$, $Q_3\&Q_4$ and $Q_4\&Q_1$ exhibit different amounts of variability. Also, all four series demonstrate an increasing trend. The first three series in the third row demonstrate larger variability around the upward trend. However, these patterns differ across the series. Moreover, the ranges of the values of the observations from different sub-seasonal patterns differ. Thus, the constructed series highlight different sub-seasonal patterns of the target series, which can be used to improve the forecasting performance.

3.2. Producing forecasts for each sub-seasonal series

After constructing multiple sub-seasonal series, we apply two families of models, namely ETS and ARIMA, on each of the constructed series to forecast the corresponding seasons in the future. Table 1 provides details of the implementations used.

ETS is able to capture the components of a time series: error, trend and seasonality (Hyndman et al., 2002). The model form can be abbreviated as ETS (error, trend, seasonality). The error component can be additive ("A") or multiplicative ("M"). The

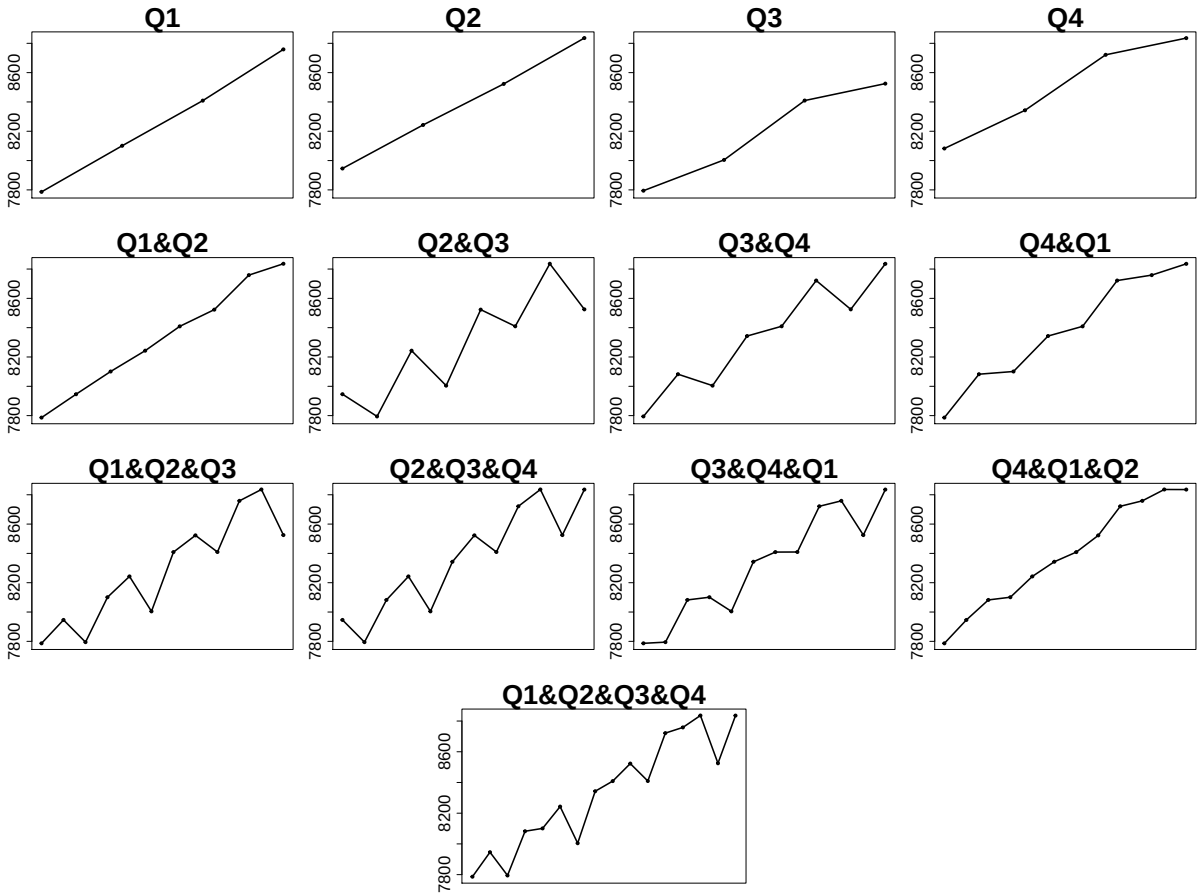

Figure 2. Caption: An example visualising the constructed time series by our proposed method. The quarterly time series Q520 from M3 competition is used to be the target series. Thirteen sub-seasonal series that contain the observations of only one or a few adjacent quarters are constructed.

Figure 2. Alt Text: Thirteen line graphs in four rows showing the constructed sub-seasonal series that contain the observations of only one or a few adjacent quarters of an example series Q520 from M3 competition.

trend and seasonality terms can be none ("N"), "A" or "M", while the trend can additionally be damped or not. The corrected Akaike's Information Criterion (AICc) is used to determine which of the ETS models is most appropriate for a given series. The ETS model selection and the corresponding parameter estimation for each sub-seasonal series are made using the function `ets()` in the R package **forecast** (Hyndman et al., 2020).

We also consider ARIMA models based on an algorithm that searches for the model form with the smallest AICc value and estimates the corresponding parameters (Hyndman et al., 2020). By default, it combines unit root tests, minimisation of the AICc, and Maximum Likelihood Estimation (MLE) to obtain the final ARIMA model among models with various AutoRegressive (AR) orders and moving average (MA) orders, up to a

Table 1. The standard methods used to forecast each sub-seasonal series, which are implemented in the **forecast** package of R.

Forecasting method	Description	R function
ETS	The exponential smoothing state family of models (Hyndman et al., 2002).	`ets()`
ARIMA	The autoregressive integrated moving average family of models (Hyndman and Khandakar, 2008).	`auto.arima()`

maximum of five.

It is worth emphasising that the constructed sub-seasonal series can also be forecast with other standard methods. Moreover, one may choose to apply different families of models on each constructed series or each information level. In any case, a different model form and set of parameters are selected for each sub-seasonal series, and a different set of forecasts is produced accordingly. The produced sets of forecasts for the sub-seasonal series are used in the next step: forecast combination.

3.3. Forecast combination

After extrapolating each series separately, forecasts produced from the multiple sub-seasonal series are averaged with equal weights (see Figure 1). In this section, in order to illustrate how our method works, we use the same example as in Section 3.1 to compare the combined forecasts with the standard forecasts of the target time series, based on the ETS model.

Figure 3 visualises the historical and test data of the quarterly time series Q520 from the M3 competition, and their standard (using the original series only) and multiple (using all sub-seasonal series) forecasts, respectively. The right panel of Figure 3 visualises the forecasts of each sub-seasonal series individually. We can see that our method efficiently forecasts the variation as the information level of the series decreases (fewer adjacent quarters are considered) compared with the standard ETS method. This shows that even if the modelling for the original series "fails" and a seasonal model is not selected, the availability of further sub-seasonal patterns will render the forecasts more robust due to the capacity of capturing the diverse patterns. As a result, no single sub-seasonal series might be enough to offer improvements over the standard approach of modelling the

original series. The benefit of our approach stems from the combination of 'suboptimal' forecasts from all the sub-seasonal series. The combination (or aggregation) of forecasts from different "interpretations" of the original series has been successfully applied before in the contexts of multiple temporal aggregation (Petropoulos and Kourentzes, 2014) and bagging (Bergmeir et al., 2016).

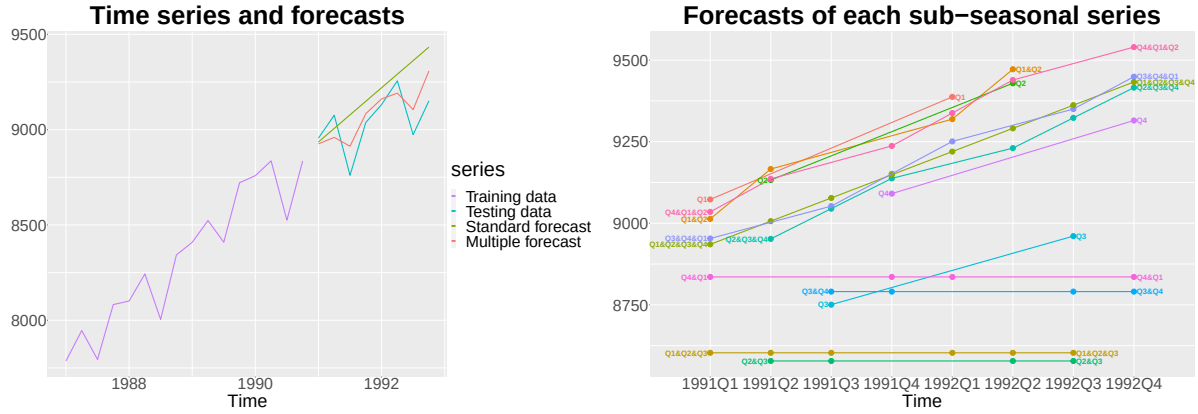

Figure 3. Caption: Left: standard and multiple forecasts of Q520 from M3 competition. Right: individual ETS forecasts for each sub-seasonal series.

Figure 3. Alt Text: Two panels, with the left line graph showing the standard and multiple forecasts of Q520 from M3 competition, and the right line graph showing individual ETS forecasts for each sub-seasonal series.

The diverse patterns captured by the sub-seasonal series are presented in Figure 4. Despite the fact that no sub-seasonal series is identified as seasonal, the combined forecast from our approach displays seasonality due to the proper alignment of the sub-seasonal forecasts to the corresponding periods and the different degrees of trend identified (see, for example, the right panel of Figure 3 and the $Q_1\&Q_2$ series).

We also use the proposed methodology to produce prediction intervals (PIs) by simply averaging the PIs produced from the subsampled series. That is, for each forecasting horizon, we calculate the PI lower bound by combining the lower bounds of the PIs generated from the corresponding sub-seasonal series with equal weights. The same process applies for calculating the upper bounds.

4. Empirical evaluation on M competitions

4.1. Data

Table 2 presents the data used in our empirical experiments. We consider the quarterly, monthly and hourly subsets of M1 (Makridakis et al., 1982), M3 (Makridakis and

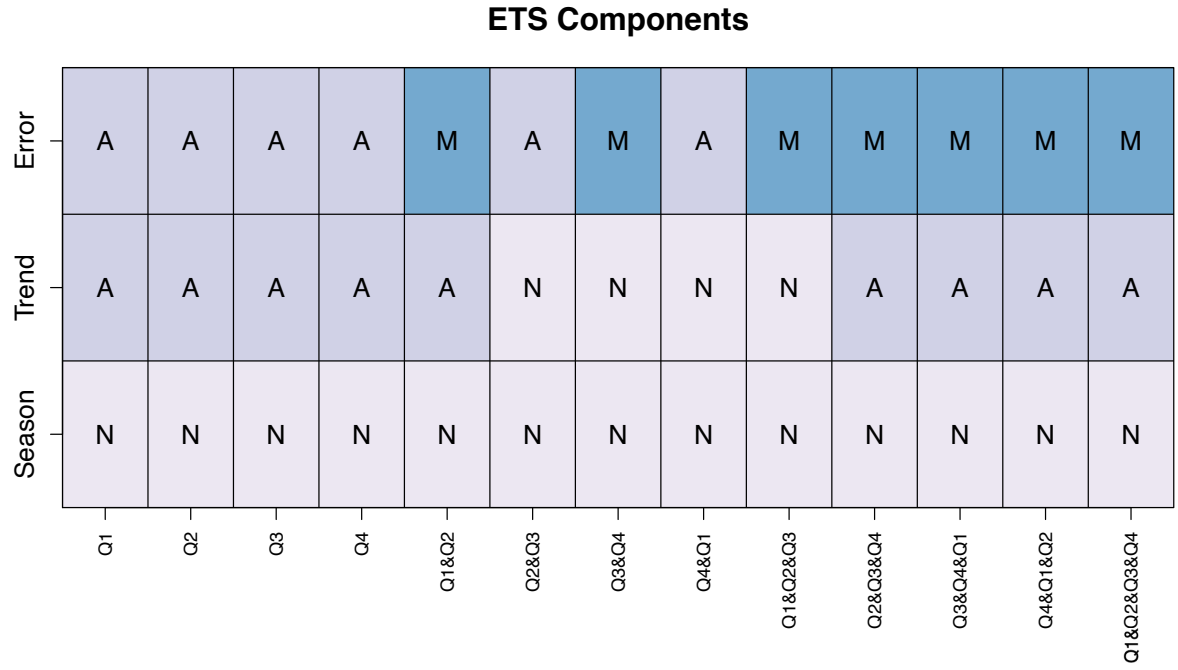

Figure 4. Caption: Optimal ETS components for each sub-seasonal series of Q520 from M3 competition data.

Figure 4. Alt Text: Three rows of adjacent boxes showing the optimal ETS components for each sub-seasonal series of Q520 from M3 competition data, with each box showing "N", "A" or "M" for the corresponding optimal component.

Hibon, 2000), and M4 (Makridakis et al., 2020) forecasting competitions, which refer to various categories such as demographic, industry, finance, economics, and others. The M1 and M3 data sets are publicly available in the **Mcomp** R package (Hyndman, 2018), and the M4 data set in the **M4comp2018** R package (Montero-Manso et al., 2018). The forecasting horizons for quarterly, monthly, and hourly data are 8, 18, and 48, with the total number of series being 24,959, 50,045, and 414, respectively. The inclusion of the hourly data offers an evaluation of our approach on series with higher frequencies.

Table 2. Data used for the empirical evaluation. h denotes the forecasting horizon.

Category	Source	Number of series	Frequency	h
Quarterly	M1	203	4	8
	M3	756	4	8
	M4	24000	4	8
	Total	24959		
Monthly	M1	617	12	18
	M3	1428	12	18
	M4	48000	12	18
	Total	50045		
Hourly	M4	414	24	48

4.2. Evaluation metrics

To evaluate the accuracy of the point forecasts, we use a commonly employed measure, the Mean Absolute Scaled Error (MASE, Hyndman and Koehler, 2006). This accuracy measure was also used in the M4 competition. It is calculated as:

$$\text{MASE} = \frac{1}{h}\frac{\sum_{t=1}^{h} \mid y_{T+t} - \hat{y}_{T+t} \mid}{\frac{1}{T-m}\sum_{t=m+1}^{T} \mid y_t - y_{t-m} \mid},$$

where y_{T+t} is the real value of the time series at time point $T+t$, T is the length of the historical data, $\hat{y}_{T+t}$ is the point forecast, h is the forecasting horizon and m is the frequency of the data (e.g., 4, 12, and 24 for quarterly, monthly, and hourly series, respectively).

In addition to MASE, the Absolute value of the Scaled Mean Error (ASME, Spiliotis et al., 2019) is employed to measure the bias of the forecasts. ASME is calculated as:

$$\text{AMSE} = \left| \frac{1}{h}\frac{\sum_{t=1}^{h}(y_{T+t} - \hat{y}_{T+t})}{\frac{1}{T}\sum_{t=1}^{T} y_t} \right|.$$

In order to quantify the performance on forecasting uncertainty, we use the $(1-\alpha)\times 100\%$ prediction intervals (PIs), and apply the Mean Scaled Interval Score (MSIS, Gneiting and Raftery, 2007) to measure the accuracy of PIs.

$$\text{MSIS} = \frac{1}{h}\frac{\sum_{t=1}^{h}\left(f^u_{T+t} - f^l_{T+t}\right) + \frac{2}{\alpha}\left(f^l_{T+t} - y_{T+t}\right)\mathbb{1}\left\{y_{T+t} < f^l_{T+t}\right\} + \frac{2}{\alpha}\left(y_{T+t} - f^u_{T+t}\right)\mathbb{1}\left\{y_{T+t} > f^u_{T+t}\right\}}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|},$$

where f^l_{T+t} and f^u_{T+t} are the lower and upper bounds of the generated prediction intervals, where we set $\alpha = 0.05$ (corresponding to 95% prediction intervals), and $\mathbb{1}$ is the indicator function, which equals to 1 when the condition is true and 0 otherwise.

4.3. Standard versus multiple forecasts

To evaluate the proposed method, we compare it with the standard benchmarks (i.e., ETS and ARIMA applied on the original series) with regard to the MASE, AMSE, and MSIS values for the 95% confidence level over the quarterly, monthly, and hourly data sets. Table 3 presents the forecasting performance regarding the mean MASE, AMSE, and MSIS values for short-term, medium-term, long-term, and overall horizons over the M1, M3, and M4 quarterly, monthly, and hourly subsets. "Standard" represents the standard benchmarks, while "Multiple" represents the proposed method. We observe that:

- For quarterly data, we find that "Multiple" performs better than the standard benchmarks in most mid and long-term horizons (when $h > 3$) in terms of the forecasting accuracy and bias.
- For monthly data, "Multiple" performs very competitively against "Standard" for almost all the forecasting horizons regarding the mean results. Our approach is also less biased, based on AMSE.
- For hourly data, "Multiple" performs better than "Standard" for all the forecasting horizons. The differences are very large for the case of ETS, where the drops in the values of the mean MASE and mean MSIS are more than 40%. Also, our approach is less biased for all the horizons in terms of AMSE.
- The proposed approach is more suitable for the time series with higher frequencies (higher values of m), which yield a larger number of sub-seasonal series. The diversity in the sub-seasonal patterns is beneficial in improving the forecasting performance by forecast combination.

Table 3. Benchmarking the performance of our proposed method against all the benchmark models with regard to the mean of MASE, AMSE and MSIS values for the 95% confidence level over the M1, M3 and M4 quarterly, monthly and hourly data sets. Entries in **bold** highlight that our method outperforms the corresponding benchmarks.

		Quarterly					Monthly					Hourly				
							MASE									
		h=1	1-3	4-6	7-8	1-8	h=1	1-6	7-12	13-18	1-18	h=1	1-16	17-32	33-48	1-48
ETS	Standard	0.599	0.774	1.261	1.607	1.165	0.454	0.651	0.965	1.225	0.947	0.390	1.410	1.611	2.450	1.824
	Multiple	0.644	0.797	**1.252**	**1.567**	**1.160**	**0.451**	**0.633**	**0.936**	**1.173**	**0.914**	**0.375**	**0.879**	**0.919**	**1.303**	**1.034**
ARIMA	Standard	0.596	0.779	1.273	1.605	1.171	0.441	0.627	0.953	1.213	0.931	0.366	0.708	0.867	1.272	0.949
	Multiple	0.641	0.807	1.274	**1.584**	1.177	0.442	0.628	**0.932**	**1.172**	**0.911**	**0.336**	**0.648**	**0.799**	**1.157**	**0.868**
							AMSE									
		h=1	1-3	4-6	7-8	1-8	h=1	1-6	7-12	13-18	1-18	h=1	1-16	17-32	33-48	1-48
ETS	Standard	0.086	0.091	0.149	0.194	0.126	0.079	0.084	0.128	0.168	0.117	0.038	0.179	0.136	0.239	0.172
	Multiple	0.089	0.092	**0.146**	**0.188**	**0.123**	**0.078**	**0.079**	**0.122**	**0.156**	**0.108**	**0.036**	**0.110**	**0.085**	**0.135**	**0.100**
ARIMA	Standard	0.083	0.091	0.150	0.195	0.127	0.076	0.079	0.125	0.164	0.113	0.043	0.076	0.080	0.122	0.080
	Multiple	0.088	0.094	**0.149**	**0.191**	**0.126**	0.076	**0.078**	**0.121**	**0.155**	**0.108**	**0.035**	**0.072**	**0.074**	**0.114**	**0.075**
							MSIS									
		h=1	1-3	4-6	7-8	1-8	h=1	1-6	7-12	13-18	1-18	h=1	1-16	17-32	33-48	1-48
ETS	Standard	4.729	6.154	10.354	13.585	9.587	3.698	5.137	8.493	11.143	8.258	3.127	11.685	15.107	25.669	17.487
	Multiple	5.046	6.253	**10.004**	**12.908**	**9.323**	**3.645**	**5.023**	**8.109**	**10.409**	**7.847**	**2.923**	**6.534**	**8.937**	**12.999**	**9.490**
ARIMA	Standard	5.543	7.221	12.230	15.787	11.241	4.045	5.470	9.132	11.640	8.747	3.295	5.753	7.463	9.279	7.498
	Multiple	**5.480**	**7.159**	**11.848**	**15.176**	**10.921**	**3.936**	**5.458**	**8.909**	**11.119**	**8.495**	**3.141**	**5.248**	**6.533**	**7.889**	**6.557**

4.4. Significance tests

To investigate the statistical significance of the performance differences between the proposed method and the benchmarks, we carry out Diebold-Mariano (DM) tests (Harvey et al., 1997). In DM tests, the null hypothesis is that the two approaches have the same forecast accuracy. We report the percentage of times that the DM tests statistic falls in the lower or upper 2.5% tail of a standard Normal distribution. The DM tests is implemented using `forecast::dm.test()` in R. The entries in Table 4 present the percentage of times "Multiple" is significantly better or worse than the standard benchmarks for different horizons. We observe that the percentage of times that our approach outperforms the benchmarks significantly increases as the horizon and frequency increase. For instance, the multiple ETS performs significantly better than the standard ETS in 53% of the cases for the longer horizons (33-48 hours ahead) of the hourly data, and only 18% of the cases worse.

Table 4. Diebold-Mariano (DM) tests for comparing the forecasting accuracy of the standard method with the multiple method. The entries show the percentage of times "Multiple" is significantly better or worse than "Standard" for different horizons.

		Quarterly				Monthly				Hourly			
		h=1-3	4-6	7-8	1-8	h=1-6	7-12	13-18	1-18	h=1-16	17-32	33-48	1-48
Multiple ETS	better	2.893	**3.698**	**4.559**	15.313	**16.275**	**19.085**	**29.190**	**33.276**	**53.382**	**44.686**	**53.382**	**61.353**
	worse	6.431	2.821	3.362	18.611	16.040	14.655	23.039	24.464	11.836	10.870	18.116	14.010
Multiple ARIMA	better	2.757	**3.666**	**3.971**	13.566	14.015	**20.054**	**28.045**	**30.884**	**31.643**	**32.367**	**44.444**	**51.691**
	worse	5.862	3.337	3.277	17.749	16.759	15.374	22.476	24.886	10.870	13.285	13.043	13.768

4.5. Linking time series features to the performance of the proposed approach

To gain a better understanding of how the time series features affect the performance of the proposed approach and further investigate when our method works well, we consider dividing our data into four categories: series with no trend nor seasonality, series with trend (but no seasonality), series with seasonality (but no trend), and series with both trend and seasonality. They are denoted as '(N,N)', '(T,N)', '(N,S)' and '(T,S)', respectively. The trend and seasonality identification of a given time series is based on automatically fitting an ETS model using the function `ets()` of the **forecast** package in R.

Figure 5 visualises the forecasting performance of our proposed approach against the benchmarks with regard to the mean of MASE, AMSE, and MSIS values over the quarterly (first row), monthly (second row) and hourly (third row) series with different characteristics. We observe that:

- For quarterly data and the ETS family of models, the proposed approach is consistently better than the benchmark ("Standard") for each of the four categories considered. Using ARIMA, the multiple approach does not improve forecasting in terms of point forecasting but still it provides better interval forecasts and less bias across all four categories. There are no noticeable differences in the scale of improvement or deterioration across the four categories, '(N,N)', '(T,N)', '(N,S)' and '(T,S)'.
- For monthly data, "Multiple" always performs better than "Standard" for the data with different characteristics in terms of MASE, AMSE and MSIS. It appears that

in terms of accuracy (MASE) and bias (AMSE), the improvements of "Multiple" over "Standard" are slightly larger when one or both of the time series patterns (trend and seasonality) are not picked up by ETS on the original series.

- For hourly data, "Multiple" almost always performs better than "Standard" for all the four data categories, except that for the '(N,N)' category where the multiple ARIMA method does not improve forecasting compared with the standard ARIMA. Moreover, focusing on ETS, the forecasts for the '(T,N)' category significantly benefit from using multiple sub-seasonal series.

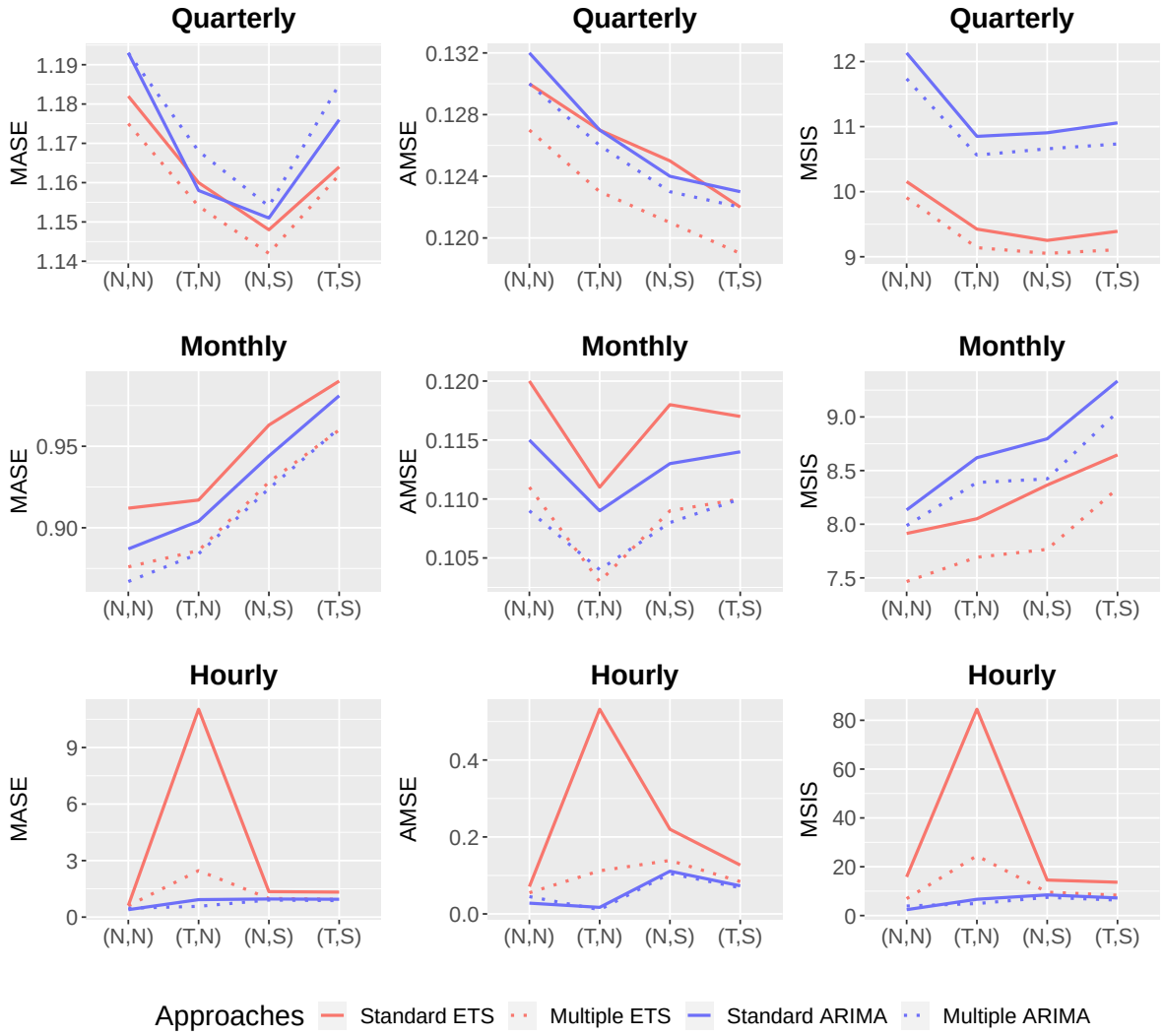

Figure 5. Caption: Benchmarking the performance of our proposed method against the benchmarks for series with different characteristics. '(N,N)', '(T,N)', '(N,S)' and '(T,S)' denote series with no trend nor seasonality, series with trend only, series with seasonality only, and series with both trend and seasonality, respectively.

Figure 5. Alt Text: A 3×3 plot showing nine line graphs to benchmarking the performance of our proposed method against the benchmarks for series with different characteristics.

4.6. Analysis of the results with a focus on management of production systems and logistics

As we mentioned in Section 4.1, the empirical sets of data that we used for our empirical evaluation comprise of different data categories, which include demographic, industry, finance, economics, and others. In this subsection, we focus our attention on the results for the series within the "industry" category. Petropoulos et al. (2019) also used the industry subset of the M3 competition data in a forecasting/inventory application context. After analysing the descriptions of these data, they concluded that the majority of the M3 industry series correspond to sales, shipments, or production. As such, focusing our investigation on the industry subsets of the M competitions, we aim to offer insights on how our proposed solution works towards improving the management of production systems and logistics.

Table 5 offers the summary results when the errors across all planning horizons (in this case, lead times) are averaged. We observe that our proposed solution that is based on multiple subseries and multiple models outperforms the state-of-the-art approaches (standard ETS and standard ARIMA). Using multiple ETS or multiple ARIMA results in better point forecast accuracy (MASE), lower bias (AMSE) and better estimation of the uncertainty (MSIS). The latter measure is relevant to inventory and logistics applications, as the accuracy of the prediction intervals can be associated with the inventory holding costs required but also the achieved service levels (Svetunkov and Petropoulos, 2018).

Table 5. The average forecasting performance of each criterion for each frequency and measure.

		Quarterly			Monthly		
		MASE	AMSE	MSIS	MASE	AMSE	MSIS
ETS	Standard	1.143	0.104	9.236	0.983	0.113	8.260
	Multiple	**1.138**	**0.100**	**8.943**	**0.963**	**0.107**	**7.921**
ARIMA	Standard	1.151	0.102	10.353	0.988	0.111	9.017
	Multiple	1.155	**0.100**	**9.952**	**0.979**	**0.108**	**8.764**

5. Applications to load forecasting

In this section, we aim to further validate the good performance of our approach when complex seasonal patterns exist. As a case study, we focus on two data sets of

electricity load, which were also used in previous studies (Taylor, 2003; Rendon-Sanchez and de Menezes, 2019). The first data set is the hourly electricity demand in England and Wales from Monday 5 June 2000 to Sunday 27 August 2000 (Taylor, 2003), as depicted in the top panel of Figure 6. The second one is the hourly electricity demand in England and Wales from 1 January 2016 to 31 December 2016 (Rendon-Sanchez and de Menezes, 2019), as shown in the bottom panel of Figure 6.

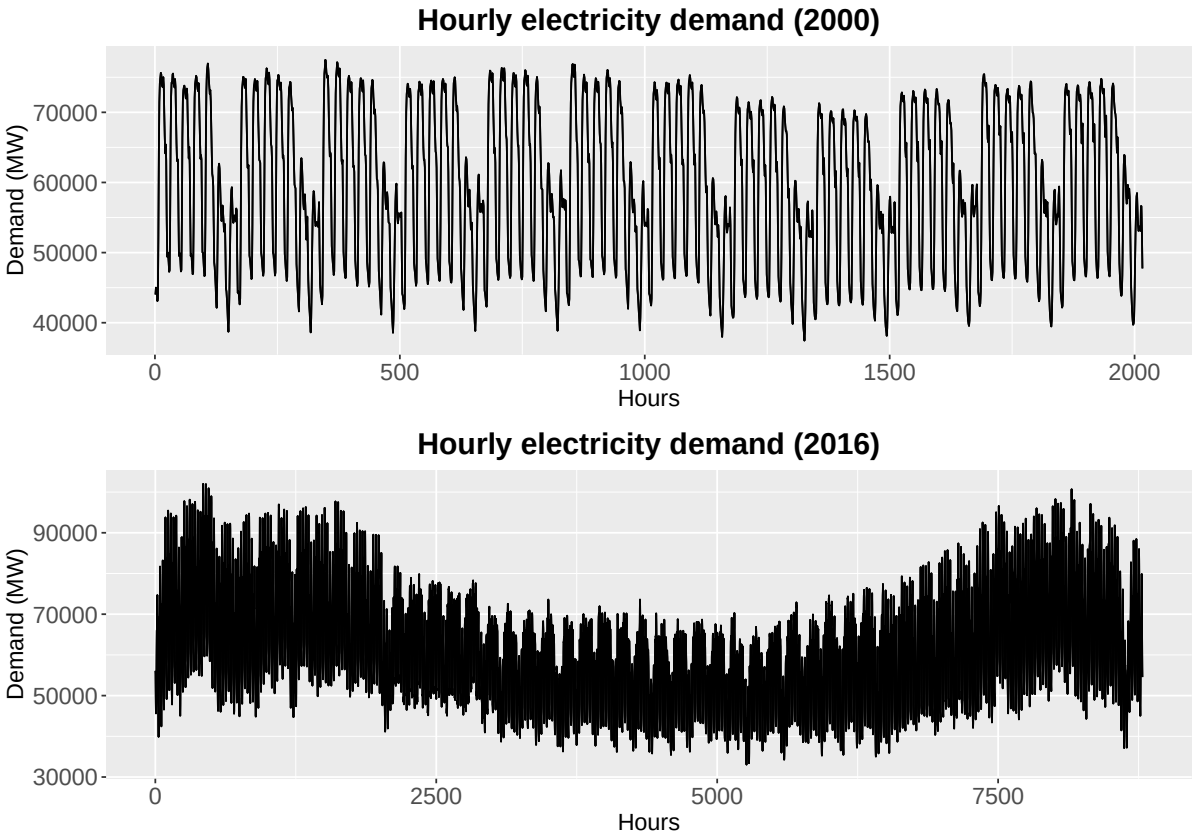

Figure 6. Caption: Top: Hourly electricity demand in England and Wales from 5 June 2000 to 27 August 2000. Bottom: Hourly electricity demand in England and Wales, 1 January 2016 to 31 December 2016. Figure 6. Alt Text: Two panels of line graphs, with the top one showing the hourly electricity demand in England and Wales from 5 June 2000 to 27 August 2000, and the bottom one showing the hourly electricity demand in England and Wales, 1 January 2016 to 31 December 2016.

We focus on the hour within day and day within week patterns in this case study (the day within year pattern is not considered since the lengths of the two load data sets are both less than one year). As shown in Table 6, the standard that we apply is the Double-Seasonal Holt-Winters method (DSHW, Taylor, 2003) which considers two seasonal cycles (24 and 168, with respect to the hour within day and day within week patterns, respectively). DHSW is implemented using the `dshw()` function of the R **forecast**

package. In the multiple approach, we consider 24 information levels. The subsampled series in the first level have single seasonality, while all the other levels exhibit double seasonality. Table 6 presents our subsampling strategy for the electricity data with double seasonal patterns, together with the seasonal periods and the respective functions applied for each information level.

Table 6. Standard and multiple approaches for load data with multiple seasonality implemented using the R package **forecast**.

Approach	Number of levels	Subsamping	Seasonality	Period(s)	R function
Standard	1	None	Double	(24, 168)	`dshw()`
Multiple	24	1h/day	Single	7	`hw()`
		2h/day	Double	(2, 14)	`dshw()`
		3h/day	Double	(3, 21)	`dshw()`
		...			
		24h/day	Double	(24, 168)	`dshw()`

5.1. Forecasting hourly electricity demand in England and Wales from 5 June 2000 to 27 August 2000

The top panel of Figure 6 shows the 12 weeks of hourly electricity demand in England and Wales from Monday 5 June 2000 to Sunday 27 August 2000. In line with Taylor (2003), the first eight weeks of data are used as the initial training data and the remaining four weeks are used to evaluate the post-sample forecasting performance of forecasts up to 24 hours ahead. As such, 1344 hourly observations are used for training and 672 for rolling-origin evaluation.

The left panel of Figure 7 shows the forecasting accuracy for each forecasting horizon in terms of MASE using the standard and multiple approaches. We can observe significant benefits from using information from sub-seasonal patterns. Our approach performs better than the standard method over all horizons. On average, the multiple approach improves the forecasting performance by 22.12% compared to the standard approach. Larger improvements are observed for horizons 10 to 16 hours ahead.

Figure 7. Caption: The forecasting performances of the standard and multiple approaches for all horizons over the hourly electricity demand in England and Wales for the year 2000 (from 5 June 2000 to 27 August 2000, left panel) and the year 2016 (from 1 January 2016 to 31 December 2016, right panel).
Figure 7. Alt Text: Two panels of line graphs showing the forecasting performances of the standard and multiple approaches for all horizons over the hourly electricity demand in England and Wales for the year 2000 (from 5 June 2000 to 27 August 2000, left panel) and the year 2016 (from 1 January 2016 to 31 December 2016, right panel).

5.2. Forecasting hourly electricity demand in England and Wales from 1 January 2016 to 31 December 2016

The bottom panel of Figure 6 visualises the hourly electricity demand in England and Wales for the year 2016 (from 1 January 2016 to 31 December 2016). This is a much longer series compared to the one in the top panel of Figure 6, and it will allow us to validate the usefulness of our approach for longer sequences.

To be consistent with Rendon-Sanchez and de Menezes (2019), the series is split into a training period that consists of 35 weeks and a testing period of 17.3 weeks, with rolling forecasts up to 24 hours ahead being generated and evaluated. The right panel of Figure 7 depicts the MASE values for the standard and multiple approaches. Compared with the standard approach, using information from multiple sub-seasonal levels improves the forecasting accuracy by 28.12% on average. Significant improvements are observed for horizons 8 to 20 hours ahead.

6. Discussion

We are living in a big data era where large collections of time series are constantly generated. Forecasters often aim to choose the best model for their data automatically.

However, accurate model selection requires professional knowledge and rich experience, making forecasting a difficult task. To mitigate the importance of model selection, we propose a novel seasonal time series forecasting method that constructs multiple series with diverse sub-seasonal patterns and subsequently extrapolates each new series separately. Finally, the forecasts of these subsampled series are averaged with equal weights.

In this paper, we propose the use of sub-seasonal patterns to forecast seasonal time series. Our approach makes it possible for forecasting methods to amplify time series patterns and hence improve forecasting performance. To make accurate and automatic extrapolations for these series individually, we use two widely-used statistical time series forecasting benchmarks, ETS and ARIMA, which are available in the R package **forecast**. Automatic and optimal model and parameter selection for these sub-seasonal series can ensure local optimal predictions.

We apply the proposed method on a large number of real-life series (the widely used data sets M1, M3, and M4) for empirical evaluation. We show that our approach performs better than the benchmarks in most horizons, whether in point or interval prediction for the monthly and hourly data sets, which indicates that our approach is more suitable for the time series with higher frequencies. On the other hand, the proposed approach does not improve as much over the standard benchmarks for the quarterly data sets. To compare the statistical significance of the accuracy improvements over the benchmarks, we carry out DM significance tests. The results further verify and strengthen the conclusion that our approach is more suitable for the time series with higher frequencies.

Our approach is also suitable for data with multiple seasonal cycles. We apply the proposed approach to the hourly electricity demand data that exhibit complex seasonal patterns to verify its effectiveness and stability. The empirical results show that our method indeed produces better and more robust point forecasts. Given that when ETS or ARIMA is used, different models might be involved at each information level when applying the multiple approach. Another implication of this case study is that our method also works well when only one model (DSHW) is used, highlighting the effect of subsampling.

Why does the proposed approach work? We construct multiple time series consisting of one or several adjacent seasons, amplifying the sub-seasonal patterns and complex components of the original series. Furthermore, by extrapolating each new series separately

and forecast combination, we effectively mitigate the importance of selecting a single model for (the original) series, making it possible to offer diverse and reliable forecasts. Our approach can significantly improve the performance of ETS and ARIMA for high frequency data through constructing multiple sub-seasonal series. More importantly, due to its simplicity and transparency, our method can be transferred in different contexts and works with other families of models.

Apart from the research implications, our approach also has clear implications for practice, particularly for production and logistics. Understanding the demand patterns and efficiently producing sales forecasts will not only lead to the maximisation of the customer service levels but also the minimisation of costs related to the inventory and logistics (Doganis et al., 2008; Wang and Petropoulos, 2016; Mircetic et al., 2021). Sales patterns of retailers' stock keeping units are often recorded in high frequencies (daily) usually display strong seasonal patterns. Our approach can help in improving the forecasts of such historical sales, allowing for more informed decisions not only at an inventory but also a logistics (transportation from distribution centres to stores) and production level.

Our approach **fo**recasting with **s**ub-seasonal **s**eries (**foss**) is implemented as an R package. Our code is open-source and publicly available at `https://github.com/lixixibj/foss`.

7. Conclusions

To mitigate the importance of model selection, we propose a novel forecasting method that constructs new series containing the observations of only one or a few adjacent seasons of the original time series. Each subsampled time series is used to make forecasts for the corresponding seasons in the future. Finally, the forecasts produced for each horizon are combined with equal weights. The main advantage of our proposed method is that it makes full use of the sub-seasonal patterns that may be available within a time series.

In our empirical experiments, the proposed approach yielded better forecasting performance than the benchmark methods for the widely-used forecasting competition data sets M1, M3, and M4. Our approach is particularly robust and stable for the data sets with higher frequencies, especially those with a trend or seasonal pattern. We also ap-

plied our approach to data with multiple seasonal cycles and showed its effectiveness in improving forecasting accuracy in the context of electricity demand.

In this paper, we focused our attention on creating sub-seasonal series using adjacent periods from the original seasonal series. An alternative would be to consider even more sub-seasonal series that are not necessarily limited to ones with adjacent periods, thus creating a larger pool of series and their accompanying forecasts. Averaging across such series could help us further reduce the variance of the forecasts even, possibly with a positive effect on the forecasting performance. Moreover, in this work, we set equal weights for combining the forecasts from the various sub-seasonal series. A potential avenue for future research would be investigating the performance of optimal (unequal) combination weights.

Data Availability Statement

The M1 and M3 data sets that support the findings of this study are publicly available in the **Mcomp** R package (Hyndman, 2018), and the M4 data set in the **M4comp2018** R package (Montero-Manso et al., 2018). Both the two data sets of electricity load for the application studies in Section 5 are publicly available at National Grid[1].

Acknowledgements

Yanfei Kang is supported by the National Key Research and Development Program (No. 2019YFB1404600) and the National Natural Science Foundation of China (Nos. 72171011 and 72021001). This research was supported by the high-performance computing (HPC) resources at Beihang University. Fotios Petropoulos thanks Ivan Svetunkov for his feedback in the early-stages of the development of the algorithm proposed in this paper.

References

Adya, M., Collopy, F., Armstrong, J. S. and Kennedy, M. (2001), 'Automatic identification of time series features for rule-based forecasting', *International Journal of Forecasting* **17**(2), 143–157.

[1] https://demandforecast.nationalgrid.com/efs_demand_forecast/faces/DataExplorer

Assimakopoulos, V. and Nikolopoulos, K. (2000), 'The theta model: a decomposition approach to forecasting', *International Journal of Forecasting* **16**(4), 521–530.

Athanasopoulos, G., Hyndman, R. J., Kourentzes, N. and Petropoulos, F. (2017), 'Forecasting with temporal hierarchies', *European Journal of Operational Research* **262**(1), 60–74.

Bates, J. M. and Granger, C. W. J. (1969), 'The combination of forecasts', *Journal of the Operational Research Society* **20**(4), 451–468.

Bergmeir, C., Hyndman, R. J. and Benítez, J. M. (2016), 'Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation', *International Journal of Forecasting* **32**(2), 303–312.

Billah, B., King, M. L., Snyder, R. and Koehler, A. B. (2006), 'Exponential smoothing model selection for forecasting', *International Journal of Forecasting* **22**(2), 239–247.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer.
URL: *https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/*

Box, G. E. and Draper, N. R. (1987), *Empirical model-building and response surfaces*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York.

Chen, T. and Guestrin, C. (2016), XGBoost: A scalable tree boosting system, *in* 'ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', pp. 785–794.

Claeskens, G., Magnus, J. R., Vasnev, A. L. and Wang, W. (2016), 'The forecast combination puzzle: A simple theoretical explanation', *International Journal of Forecasting* **32**(3), 754–762.

Collopy, F. and Armstrong, J. S. (1992), 'Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations', *Management Science* **38**(10), 1394–1414.

De Livera, A. M., Hyndman, R. J. and Snyder, R. D. (2011), 'Forecasting time series with complex seasonal patterns using exponential smoothing', *Journal of the American Statistical Association* **106**(496), 1513–1527.

Doganis, P., Aggelogiannaki, E. and Sarimveis, H. (2008), 'A combined model predictive control and time series forecasting framework for production-inventory systems',

International Journal of Production Research **46**(24), 6841–6853.

Fildes, R. and Petropoulos, F. (2015), 'Simple versus complex selection rules for forecasting many time series', *Journal of Business Research* **68**(8), 1692–1701.

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., Petropoulos, F. and Koehler, A. B. (2016), 'Models for optimising the theta method and their relationship to state space models', *International Journal of Forecasting* **32**(4), 1151–1161.

Gardner Jr, E. S. (2006), 'Exponential smoothing: The state of the art—Part II', *International Journal of forecasting* **22**(4), 637–666.

Genre, V., Kenny, G., Meyler, A. and Timmermann, A. (2013), 'Combining expert forecasts: Can anything beat the simple average?', *International Journal of Forecasting* **29**(1), 108–121.

Gneiting, T. and Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Harvey, D., Leybourne, S. and Newbold, P. (1997), 'Testing the equality of prediction mean squared errors', *International Journal of Forecasting* **13**(2), 281–291.

Hasni, M., Aguir, M., Babai, M. and Jemai, Z. (2019), 'Spare parts demand forecasting: a review on bootstrapping methods', *International Journal of Production Research* **57**(15-16), 4791–4804.

Hibon, M. and Evgeniou, T. (2005), 'To combine or not to combine: selecting among forecasts and their combinations', *International Journal of Forecasting* **21**(1), 15–24.

Hyndman, R. (2018), *Mcomp: Data from the M-Competitions*. R package version 2.8.
URL: *https://CRAN.R-project.org/package=Mcomp*

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E. and Yasmeen, F. (2020), *forecast: Forecasting functions for time series and linear models*. R package version 8.13.
URL: *http://pkg.robjhyndman.com/forecast*

Hyndman, R. J. (2011), *Cyclic and seasonal time series*.
URL: *https://robjhyndman.com/hyndsight/cyclicts/*

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

Hyndman, R. J. and Khandakar, Y. (2008), 'Automatic time series forecasting: the forecast package for R', *Journal of Statistical Software* **26**(3), 1–22.

Hyndman, R. J. and Koehler, A. B. (2006), 'Another look at measures of forecast accuracy', *International Journal of Forecasting* **22**(4), 679–688.

Hyndman, R. J., Koehler, A. B., Snyder, R. D. and Grose, S. (2002), 'A state space framework for automatic forecasting using exponential smoothing methods', *International Journal of Forecasting* **18**(3), 439–454.

Hyndman, R., Koehler, A. B., Ord, J. K. and Snyder, R. D. (2008), *Forecasting with exponential smoothing: the state space approach*, Springer Science & Business Media.

Jose, V. R. R. and Winkler, R. L. (2008), 'Simple robust averages of forecasts: Some empirical results', *International Journal of Forecasting* **24**(1), 163–169.

Kang, Y., Cao, W., Petropoulos, F. and Li, F. (2021), 'Forecast with forecasts: Diversity matters', *European Journal of Operational Research* (in press).
URL: *https://www.sciencedirect.com/science/article/pii/S0377221721008730*

Kang, Y., Hyndman, R. J. and Smith-Miles, K. (2017), 'Visualising forecasting algorithm performance using time series instance spaces', *International Journal of Forecasting* **33**(2), 345–358.

Kolassa, S. (2011), 'Combining exponential smoothing forecasts using Akaike weights', *International Journal of Forecasting* **27**(2), 238–251.

Kourentzes, N. and Athanasopoulos, G. (2021), 'Elucidate structure in intermittent demand series', *European Journal of Operational Research* **288**(1), 141–152.

Kourentzes, N., Petropoulos, F. and Trapero, J. R. (2014), 'Improving forecasting by estimating time series structural components across multiple frequencies', *International Journal of Forecasting* **30**(2), 291–302.

Li, X., Kang, Y. and Li, F. (2020), 'Forecasting with time series imaging', *Expert System with Applications* **160**, 113680.

Lichtendahl, K. C. and Winkler, R. L. (2020), 'Why do some combinations perform better than others?', *International Journal of Forecasting* **36**(1), 142–149.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E. and Winkler, R. (1982), 'The accuracy of extrapolation (time series) methods: Results of a forecasting competition', *Journal of Forecasting* **1**(2), 111–153.

Makridakis, S. and Hibon, M. (2000), 'The M3-Competition: results, conclusions and implications', *International Journal of Forecasting* **16**(4), 451–476.

Makridakis, S., Spiliotis, E. and Assimakopoulos, V. (2020), 'The M4 competition: 100,000 time series and 61 forecasting methods', *International Journal of Forecasting* **36**(1), 54–74.

Mircetic, D., Rostami-Tabar, B., Nikolicic, S. and Maslaric, M. (2021), 'Forecasting hierarchical time series in supply chains: an empirical investigation', *International Journal of Production Research* **0**(0), 1–20.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J. and Talagala, T. S. (2020), 'FFORMA: Feature-based forecast model averaging', *International Journal of Forecasting* **36**(1), 86–92.

Montero-Manso, P., Netto, C. and Talagala, T. S. (2018), *M4comp2018: Data from the M4-Competition*. R package version: 0.1.0.

Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., Petropoulos, F. and Assimakopoulos, V. (2011), 'An aggregate-disaggregate intermittent demand approach (ADIDA) to forecasting: An empirical proposition and analysis', *The Journal of the Operational Research Society* **62**(3), 544–554.

Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M. Z., Barrow, D. K., Ben Taieb, S., Bergmeir, C., Bessa, R. J., Bijak, J., Boylan, J. E., Browell, J., Carnevale, C., Castle, J. L., Cirillo, P., Clements, M. P., Cordeiro, C., Cyrino Oliveira, F. L., Baets, S. D., Dokumentov, A., Ellison, J., Fiszeder, P., Franses, P. H., Frazier, D. T., Gilliland, M., Gönül, M. S., Goodwin, P., Grossi, L., Grushka-Cockayne, Y., Guidolin, M., Guidolin, M., Gunter, U., Guo, X., Guseo, R., Harvey, N., Hendry, D. F., Hollyman, R., Januschowski, T., Jeon, J., Jose, V. R. R., Kang, Y., Koehler, A. B., Kolassa, S., Kourentzes, N., Leva, S., Li, F., Litsiou, K., Makridakis, S., Martin, G. M., Martinez, A. B., Meeran, S., Modis, T., Nikolopoulos, K., Önkal, D., Paccagnini, A., Panagiotelis, A., Panapakidis, I., Pavía, J. M., Pedio, M., Pedregal, D. J., Pinson, P., Ramos, P., Rapach, D. E., Reade, J. J., Rostami-Tabar, B., Rubaszek, M., Sermpinis, G., Shang, H. L., Spiliotis, E., Syntetos, A. A., Talagala, P. D., Talagala, T. S., Tashman, L., Thomakos, D., Thorarinsdottir, T., Todini, E., Trapero Arenas, J. R., Wang, X., Winkler, R. L., Yusupova, A. and Ziel, F. (2021), 'Forecasting: theory and practice', *International Journal of Forecasting* (in press).

Petropoulos, F., Hyndman, R. J. and 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.

Petropoulos, F. and Kourentzes, N. (2014), 'Improving forecasting via multiple temporal aggregation', *Foresight: The International Journal of Applied Forecasting* **34**, 12–17.

Petropoulos, F. and Kourentzes, N. (2015), 'Forecast combinations for intermittent demand', *The Journal of the Operational Research Society* **66**(6), 914–924.

Petropoulos, F., Makridakis, S., Assimakopoulos, V. and Nikolopoulos, K. (2014), "Horses for courses' in demand forecasting', *European Journal of Operational Research* **237**(1), 152–163.

Petropoulos, F. and Svetunkov, I. (2020), 'A simple combination of univariate models', *International Journal of Forecasting* **36**(1), 110–115.

Petropoulos, F., Wang, X. and Disney, S. M. (2019), 'The inventory performance of forecasting methods: Evidence from the M3 competition data', *International Journal of Forecasting* **35**(1), 251–265.

Reid, D. (1972), 'A comparison of forecasting techniques on economic time series', *Forecasting in Action. Operational Research Society and the Society for Long Range Planning* .

Rendon-Sanchez, J. F. and de Menezes, L. M. (2019), 'Structural combination of seasonal exponential smoothing forecasts applied to load forecasting', *European Journal of Operational Research* **275**(3), 916–924.

Smith, J. and Wallis, K. F. (2009), 'A simple explanation of the forecast combination puzzle', *Oxford Bulletin of Economics and Statistics* **71**(3), 331–355.

Spiliotis, E., Petropoulos, F. and Assimakopoulos, V. (2019), 'Improving the forecasting performance of temporal hierarchies', *PloS one* **14**(10), e0223422.

Svetunkov, I. and Petropoulos, F. (2018), 'Old dog, new tricks: a modelling view of simple moving averages', *International Journal of Production Research* **56**(18), 6034–6047.

Talagala, T. S., Hyndman, R. J., Athanasopoulos, G. et al. (2018), Meta-learning how to forecast time series, Technical report, Monash University, Department of Econometrics and Business Statistics.

Tashman, L. J. (2000), 'Out-of-sample tests of forecasting accuracy: an analysis and review', *International Journal of Forecasting* **16**(4), 437–450.

Taylor, J. W. (2003), 'Short-term electricity demand forecasting using double seasonal exponential smoothing', *Journal of the Operational Research Society* **54**(8), 799–805.

Thomson, M. E., Pollock, A. C., Önkal, D. and Gönül, M. S. (2019), 'Combining forecasts: Performance and coherence', *International Journal of Forecasting* **35**(2), 474–484.

Wang, X. and Petropoulos, F. (2016), 'To select or to combine? the inventory performance of model and expert forecasts', *International Journal of Production Research* **54**(17), 5271–5282.

Watson, M. W. and Stock, J. H. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**(6), 405–430.

Winters, P. R. (1960), 'Forecasting sales by exponentially weighted moving averages', *Management Science* **6**(3), 324–342.

Wolpert, D. H. (1996), 'The lack of a priori distinctions between learning algorithms', *Neural Computation* **8**(7), 1341–1390.