Forecasting for lead-time period by temporal aggregation: whether to combine and how

Computers in Industry 145 (2023) 103803

Contents lists available at ScienceDirect

Computers in Industry

journal homepage: www.sciencedirect.com/journal/computers-in-industry

Forecasting for lead-time period by temporal aggregation: Whether to combine and how

Bahman Rostami-Tabar [a,*], Thanos E. Goltsos [a,b], Shixuan Wang [c]

[a] Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff CF10 3EU, UK
[b] PARC Institute of Manufacturing, Logistics and Inventory, Cardiff Business School, Cardiff University, Aberconway Building, Colum Drive, Cardiff CF10 3EU, UK
[c] Department of Economics, University of Reading, Reading, RG6 6AA, UK

ARTICLE INFO

Keywords:
Temporal aggregation
Overlapping
Non-overlapping
Time series forecasting
Combination
M4 dataset
Aggregate forecast

ABSTRACT

Temporal aggregation (TA) refers to transforming a time series from higher to lower frequencies (e.g. monthly to quarterly). There are two different types of aggregation: overlapping and non-overlapping; which, when added to the option of using original time series, present the forecaster with three approaches to produce required forecasts over the lead-time period: (i) non-overlapping aggregation (NOA) (ii) overlapping aggregation (OA); and (iii) bottom-up to aggregate forecast (BU). Forecasters may then need to decide which approach to use or alternatively combine the forecasts generated by the three approaches, instead. In this study, we design and execute an experiment using the M4 competition dataset, to explore the effect of different initial frequencies (i.e. daily, monthly, and quarterly), data aggregation levels and combination methods on forecast accuracy. We are surprised to find that neither temporal aggregation strategies have an overall gain on forecasting accuracy. Equally concerning is the fact that straight (average) combinations of these forecasts are similarly of no benefit to the accuracy. To extract the benefits of both well-supported individual forecasting practices of temporal aggregation and combination, we propose a framework that aims to combine temporal aggregation and forecasting combinations using a polynomially weighted average with multiple learning rates. We find considerable overall improvement in forecasting accuracy by using the proposed combination, especially for longer lead-times. We discuss areas where the framework is expected to perform best in the future and conclude that further research is required in this area. We note that our method can work in parallel of others and close with an agenda for further research on forecasting by temporal aggregation.

1. Introduction

Forecasting is a crucial part of the decision-making process in any organisation (Morariu et al., 2020). Time series forecasting refers to forecasts extrapolated on the basis of observations made sequentially through time (Petropoulos et al., 2022; Singh and Huang, 2019). In practice, such data are collected in a single level of granularity, which does not necessarily need to match the level of granularity of forecasts driving decision-making. The forecast granularity can be more aggregated or disaggregated than the original time series granularity. Generally, it is related to the decision that the forecast aims to inform. In our analysis, we assume that we start with a time series that has a higher frequency (e.g. monthly) than the required forecasting output (e.g. annual = 12 months).

Increasing computing power and improvements in database architectures allow data to be stored in increasingly finer granularity (Kolassa, 2016). At the same time, practitioners are often interested in lower frequency forecasts (e.g. annual) of data that are stored at higher frequencies (e.g. monthly). For instance, a forecast over the lead-time would be required to determine the right level of replenishment. In these situations, it is recommended to convert time series from the higher frequency to the lower frequency that matches the forecast requirement, and then to model the low frequency data to generate forecasts (Goodwin, 2018; Boylan and Babai, 2016). Converting higher frequency data into lower frequencies is generally performed using temporal aggregation. In this case, temporal aggregations should be evaluated, using the lead-time as the level of aggregation. Thus, a forecast of the total value over several time periods ahead (lead-time) is required, which is refereed to lead-time forecast. Therefore, there is no disaggregation mechanism involved in the computation and the lead-time period matches the aggregation level used to create temporally aggregated series. Aggregation over the necessary forecast horizon such as lead-time is often a necessity and not an option (Mohammadipour

* Corresponding author.
E-mail addresses: rostami-tabarb@cardiff.ac.uk (B. Rostami-Tabar), GoltsosA@cardiff.ac.uk (T.E. Goltsos), shixuan.wang@reading.ac.uk (S. Wang).

https://doi.org/10.1016/j.compind.2022.103803
Received 14 February 2022; Received in revised form 13 October 2022; Accepted 31 October 2022
Available online 5 December 2022

and Boylan, 2012; Nikolopoulos et al., 2011; Zotteri and Kalchschmidt, 2007; Rostami-Tabar et al., 2014).

Temporal aggregation (TA) is an intuitively appealing approach that transforms a time series from higher (e.g., daily) into lower frequencies (e.g., monthly), strengthening or attenuating different time series features. There are two different types of temporal aggregation: non-overlapping, where the time series is divided into buckets of the lower frequency's size; and overlapping, where values are replaced by a moving average of the lower frequency's length. Therefore, when interested in forecasting for a given lead-time period, forecasters have three options: (i) using original series, then forecast for ℓ periods ahead, followed by aggregating forecasts (BU) (ii) non-overlapping aggregated series with an aggregation level equal to ℓ, followed by one step-ahead forecast (NOA) and (iii) overlapping aggregated series with an aggregation level equal to ℓ, following one step-ahead forecast (OA). We should note that using the term bottom-up (BU) is more relevant if we consider the temporal aggregation as an hierarchical structure (Athanasopoulos et al., 2017) with higher frequency series (e.g. monthly) at the bottom and lower frequencies (e.g. annual) at the higher levels of the hierarchy.

Recent studies show that three approaches may have their own merit depending on the presence of the autocorrelation in the original series, aggregation level, forecast horizon and the employed forecasting method (see, e.g., Boylan and Babai, 2016; Rostami-Tabar et al., 2022).

At the same time, research shows that forecast combinations improve on the forecast accuracy of their constituent parts. Forecasting competitions, including M4, further cemented this long-established notion (for the M4 competition, all the methods that ranked 2–6 were employing some sort of forecast combinations; Makridakis et al., 2018). The literature proposes various ways to combine forecasts, ranging from simple averages to complex methodologies to derive optimal weights (see, e.g., Kolassa, 2011; Makridakis et al., 2018; Jaganathan and Prakash, 2020).

Further, forecast combinations have been proposed in the temporal aggregation field to combine forecasts created across different frequencies. These approaches have been successfully employed in both intermittent demand[1] (Nikolopoulos et al., 2011; Petropoulos and Kourentzes, 2014) and fast-moving demand contexts (Athanasopoulos et al., 2017).

Temporal aggregation is logically routed on the observation that the same time series (or some of their elements) might prove easier to forecast at a lower frequency than the initial level, often coupled with the need for forecasts covering multiple periods (of the initial granularity). Different time series features can be masked or brought forth by intelligently selecting the correct level for the forecasting purpose. At the same time, literature has explored the forecast accuracy of combined forecasts whose constituents are either originating from different methods or different frequencies, which are produced by temporal aggregation. However, and to the best of our knowledge, no prior research has attempted to explore the effects of combining forecasts generated by temporal aggregation approaches, i.e., non-overlapping and overlapping temporal aggregation, as well as the BU approach.

When it comes to using temporal aggregation and forecast combination, two approaches proposed by Kourentzes et al. (2014) and Athanasopoulos et al. (2017) exploit multiple levels of information generated by the non-overlapping temporal aggregation. In this study, we exploit three levels of information (i.e. the original series, the non-overlapping and the overlapping temporal aggregation series), where each series has a different feature. We propose a way to combine forecasts generated from these three levels. Furthermore, this is a way of combining that can be used in parallel (on top) of the other existing ways (i.e., combinations coming either from model components at different aggregation level or reconciling forecasts in the temporal hierarchy). It should be noted that our work is also in conjunction with an emerging research area of forecast reconciliation which typically deals with series with a hierarchical structure, but can be applicable beyond the hierarchical setting, to improve forecast accuracy under multivariate settings with linear constraints. Basically, our work shares the same nature with forecasting reconciliation to aggregate individual forecasts, but we focus on the temporal dimension rather than cross-sectional dimension. Hollyman et al. (2021) provide an extensive review on the forecast reconciliation and elaborate the connection between hierarchical forecasting and forecasts combinations. We refer the reader to Hollyman et al. (2021) for more details about such a connection.

In this paper, we focus on forecasting for a lead-time period. We generate forecasts originating from three distinct time series: (*i*) the original time series; (*ii*) non-overlapping TA; (*iii*) overlapping TA. In generating non-overlapping and overlapping aggregated series, we assume that aggregation level equals to the lead-time. Although there are studies that investigate the performance of these approaches separately, no study has addressed the potential gain in the forecast accuracy by combining forecasts generated from individual approaches. We attempt to address this gap by use of the vast database of time series from the M4 competition database.

We provide a framework for improving lower frequency forecasts by using a combination of forecasts generated from the original, the overlapping and the non-overlapping temporally aggregated series. We test these approaches independently, using a simple average weight combination rule and a proposed polynomial aggregation rule. Employing a full-factorial empirical experiment, we examine the forecast accuracy of these approaches using M4 competition data with quarterly, monthly, and daily frequencies. Exponential smoothing state space (ETS) and AutoRegressive Integrated Moving Average (ARIMA) are employed as forecasting methods. In using these automatic forecasting models (i.e. ETS() and ARIMA() in the R package "fable"), our intention is to separate the effect of the forecasting method choice from the temporal aggregation and its combinations. However, the experiment could benefit from using some lighter forecasting method such as Theta (Assimakopoulos and Nikolopoulos, 2000) to reduce the running time or a combination approach (e.g. average of ETS and ARIMA) to investigate the effect of the forecasting method choice.

We find that the proposed approach has, on average, promising forecasting performance, with major improvements for long-term forecasts, and that the improvements are irrespective of the forecasting method, data frequency, and lead time. This is the key advantage of this approach since it has the potential to provide improvements on top of any other already tested or implemented approach. Moreover, we examine the claim that lower frequency forecasts tend to be more accurate than higher frequency forecasts when generating lead-time forecasts (Goodwin, 2018; Boylan and Babai, 2016). Our findings suggest that that is not always the case.

The rest of the paper is organised as follows: In Section 2, we briefly discuss the literature of non-overlapping and overlapping temporal aggregation and combinatorial forecasting. In Section 3, we detail the methods, aggregation and combination approaches we employ, our dataset and simulation design. In Section 4, we report and discuss our results. Finally, we conclude and provide directions for practitioners and future research ideas in Section 5.

2. Research background

In this section we review research on aggregating forecast, temporal aggregation and combination of either forecasting methods or forecasts created from temporally aggregated series at various granularities. These reviews will cover both situations where a forecast over lead-time or at the original higher frequency level, which requires disaggregation, is required.

[1] Intermittent demand describes time series where positive instances of demand are dispersed around periods of no demand, see Boylan and Syntetos (2021) for an exposition.

2.1. Aggregate forecast (Bottom-UP)

The first approach to generate the forecast at lower frequency from higher frequency series is to apply the forecasting method to the original time series and then aggregate forecasts, rather than transforming the time series. We may refer to this approach as Aggregate Forecast (AF) or Bottom-Up (BU) (see, e.g., Orcutt et al., 1968; Dunn et al., 1976; Shlifer and Wolff, 1979). This can be considered as a natural benchmark to compare the performance of non-overlapping and overlapping TA against. Athanasopoulos et al. (2011) have conducted an empirical investigation using 366 monthly series and considering some forecasting models including state space models for exponential smoothing (ETS) and ARIMA methodology, and the Theta method. They found that aggregated forecast from either monthly or quarterly to yearly to be more accurate than the forecasts generated from the non-overlapping TA yearly data. In the context of intermittent time series, Willemain et al. (1994) have empirically explored the accuracy of BU approach. They showed that aggregating forecasts does not lead to more accurate forecasts, which later was confirmed by Nikolopoulos et al. (2011). Rostami-Tabar et al. (2013, 2014) analytically demonstrated that aggregating forecasts is more accurate for high values of positive autocorrelation in the original series, when the forecasting method is the Single Exponential Smoothing and the series is an ARMA (1,1) process.

2.2. Temporal aggregation and forecasting

There are two types of temporal aggregation: (*i*) non-overlapping aggregation (NOA) and (*ii*) overlapping aggregation (OA).

As shown in Fig. 1, the non-overlapping aggregated series is created by adding up the values inside a consecutive non-overlapping buckets. The size of the bucket equals the aggregation level, m. It is recommended to start creating the time buckets from the most recent observation (in this sense, when any remainder periods are to be dropped off, they would be dropped off from the beginning of the time series and therefore be the oldest and therefore least relevant data).

Overlapping temporal aggregation creates overlapping buckets of time starting from the last observation, where the bucket's size equals the aggregation level. At each period, the window is moved one step backward, so the newest observation is dropped, and an older one is included. The number of aggregated periods in OA is much higher than NOA.

The frequency of the time series for seasonality (and corresponding seasonal indexes) are derived from the initial frequency of the original series as follows. Assume monthly series which have a frequency of 12, and an aggregation level m = 3. For NOA we move from 12 to 4 'seasons', each representing a quarter of any given year. For OA, we retain the initial frequency of 12, however now the seasons each reflect a three-month period, each ending in a distinct month.

2.2.1. Non-overlapping temporal aggregation

The non-overlapping temporal aggregation approach has been the focus of the literature. The potential value of NOA in the context of intermittent demand forecasting was initially acknowledged by Willemain et al. (1994). The underlying logic is that as we reduce the frequency of the data we end up with less intermittency (fewer periods of no demand), and therefore more well-behaved time series. Nikolopoulos et al. (2011) have exploited this and have shown that an aggregation approach may offer considerable improvements in forecasting accuracy and stock control performance in intermittent demand requirements. They indicate that non-overlapping TA can offer forecast accuracy improvement compared to aggregating the forecast. Similar findings have been reported by Babai et al. (2012), Kourentzes et al. (2014), and Petropoulos and Kourentzes (2015).

Rostami-Tabar et al. (2013, 2014) analytically studied the effect of non-overlapping temporal aggregation on time series forecasting. Assuming an ARMA (1,1) time series process and Single Exponential Smoothing (SES) forecasting method, they reveal that non-overlapping temporal aggregation forecast accuracy improvement depends on three factors: (*i*) value of AR and MA parameters, (*ii*) the value of aggregation level and (*iii*) the smoothing constant of SES method. Additionally, they show that the performance gain generally increases with the aggregation level. Kourentzes et al. (2017) compared the performance of forecasting with multiple aggregation levels with one of using a single optimal aggregation level for real and simulated time series. They show that using non-overlapping TA can improve forecast accuracy compared to aggregating forecast.

Some studies exploit the idea of combining available signals at multiple levels of aggregation when using non-overlapping TA, instead of using only one single optimal temporal aggregation level. Combining forecasts generated from the multiple level of aggregation is intuitively appealing, aiming at capturing different patterns of the time series. Andrawis et al. (2011) used monthly and yearly time series to examine the benefits of combining short-term and long-term forecasts and concluded that the combination can lead to forecast accuracy improvement. Kourentzes et al. (2014) recommended using multiple levels of TA and combining the separate forecasts (MAPA). This approach not only benefits from managing the modelling risk, but also utilises the established gains of forecast combination (Barrow and Kourentzes, 2016; Blanc and Setzer, 2016). Kourentzes et al. (2014) provided empirical evidence to demonstrate gains over conventional forecasting. Since modelling with multiple TA levels has been used successfully to intermittent demand, promotional modelling and inventory management (Petropoulos and Kourentzes, 2014; Kourentzes and Petropoulos, 2016; Barrow and Kourentzes, 2016). Athanasopoulos et al. (2017) implemented a number of weighted least square regimes to combine forecasts created from different non-overlapping temporally aggregated series and find significant improvements compared to base forecasts and bottom-up approaches. The current study can be extended in the future to include the proposed approaches in this stream of research.

2.2.2. Overlapping temporal aggregation

Most of the literature has been focusing on the non-overlapping aggregation and little attention is given to the overlapping aggregation. Porras and Dekker (2008) was the first study that compared the overlapping blocks method with the approach advocated by Willemain et al. (2004), based on an empirical analysis of spare parts from a Dutch petrochemical complex. They examined the inventory cost implications of the two methods, finding that the overlapping blocks method produced lower costs, with both methods attaining a 90% fill rate. Porras and Dekker (2008) used the overlapping temporal aggregation to estimate the distribution of lead-time demand. They compared the inventory performance of overlapping temporal aggregation against the resampling approach proposed by Willemain et al. (2004). The result showed that the overlapping temporal aggregation leads to lower cost savings, but also achieved service levels, compared to the resampling approach. Boylan and Babai (2016) conducted a theoretical analysis of the accuracy of the overlapping to estimate the cumulative distribution function (CDF) under the assumption that demand is independent and identically distributed (i.i.d.). They indicated that the overlapping approach results in an unbiased estimates and showed that it often leads to a better estimate than the non-overlapping one. It also leads to a reduction in backorders by increasing the aggregation level when the target cycle service level is high. Rostami-Tabar et al. (2022) used the overlapping and non-overlapping temporal aggregation approaches, numerically and empirically, to forecast the cumulative demand a finite auto-correlated demand. They showed that the overlapping aggregation approach could be more accurate than the non-overlapping for shorter series. For the longer demand history, the performance becomes similar.

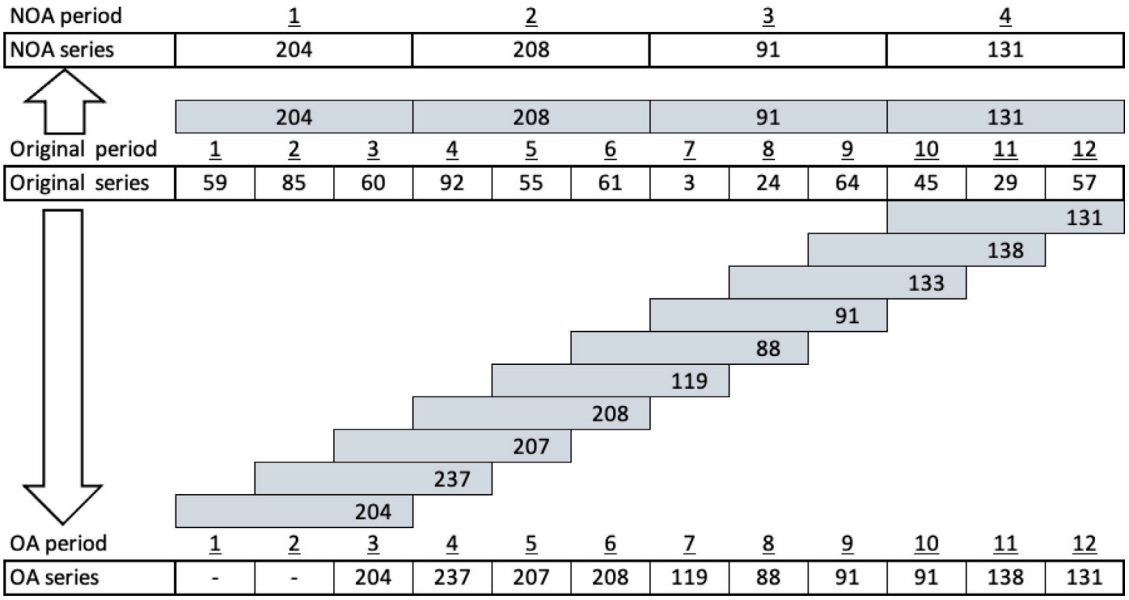


Fig. 1. An example to illustrate how an original series is transformed into overlapping and non-overlapping temporally aggregated series. The middle part shows the original time series with 12 months period. The upper part is the non-overlapping TA series that is created by dividing the original series into consecutive non-overlapping buckets of time where the length of the time bucket equals the aggregation level (m = 3). The lower part is the overlapping TA series that is created by a moving window where the window's size equals the aggregation level (m = 3).

2.3. Forecast combinations

Combining has long been widely considered to be beneficial for forecasting in various fields (Clemen, 1989). For the purposes of this work, we make the distinction between combination of different forecasting methods (statistically or otherwise derived) and between forecasts calculated from different temporally aggregated frequencies discussed in the previous section. Thus, in this part, we discuss briefly combination of forecasts from different forecasting methods. Averaging forecasts from different methods may lead to improvements in accuracy and a lower level of uncertainty (Hibon and Evgeniou, 2005). In the M4 forecasting competition, 5 of 6 top performing methods employed forecast combinations (Makridakis et al., 2018). A number of studies have evaluated sophisticated weighting processes to combine forecasts. Based on theories and methods of self-organising data mining, He and Xu (2005) proposed a self-organising forecast combination method and showed it outperformed linear and a neural network combination approaches. Kolassa (2011) proposed the use of Aikake weights on exponential smoothing forecasts and show that it consistently outperformed the use of single "best" forecasts when those were selected by information criteria (see, e.g. He and Xu, 2005; Kolassa, 2011); however, simple combination approaches seem to perform reasonably well compared to more complex ones (Clemen, 1989; Hibon and Evgeniou, 2005; Jose and Winkler, 2008). This is something that we also explore in our work.

We observe in the literature that there are no comprehensive rules for the outperformance of bottom-up or temporal aggregation approaches. This has been also highlighted by Babai et al. (2021) in a review article on time series aggregation. All BU, NOA and OA approaches may have their own merits when generating forecasts required over the lead-time period, and combining forecasts generated from these approaches may improve the forecast accuracy. To the best of our knowledge, no study has considered the potential benefits of combining overlapping, non-overlapping, and bottom-up approaches for forecast accuracy. To that end, we propose a framework to combine forecasts produced by these approaches, which is discussed in the next section. We compare its forecast accuracy with individual approaches (e.g. BU, NOA, OA) and a simple average forecast combination.

3. Methodology

In this section, we first introduce the forecasting methods, aggregation and combination approaches, followed by a description of the dataset and simulation design.

Assume we are given a time series $y_t, t = 1, 2, \ldots, T$, and we are interested in generating the forecast over lead-time m, $\hat{y}_{T,m}$ for a given forecasting aggregation method. For simplicity, we assume that the lead-time equals the aggregation level, discussed in non-overlapping and over lapping temporal aggregation. Fig. 2 shows the steps in generating the forecast over lead-time for a given time series, a forecasting method i.

3.1. Forecasting methods

The first method is the exponential smoothing state space family of models which can be abbreviated as ETS (Error, Trend, Seasonality). Readers can refer to Hyndman and Athanasopoulos (2021) or Hyndman et al. (2008) for a detailed description of ETS taxonomy. ETS models are capable of capturing trend and seasonality in time series, plus error component. The trend and seasonality components can be none ("N"), Additive ("A") or multiplicative ("M"), while the trend can additionally be damped or not. The error term can also be additive ("A") or multiplicative ("M"). These components can be combined in various forms, creating different possible exponential smoothing models. We use the implementation of ETS models in the "fable" package (O'Hara-Wild et al., 2020) in R using $ETS()$ function. The ETS function in fable uses corrected Akaike's Information Criterion (AICc) to identify the most appropriate model for a given time series. ETS is among the most widely used forecasting models with reliable performance in different applications (Gardner, 2006), in particular with monthly and quarterly series.

Another family of forecasting models that is widely used is AutoRegressive Integrated Moving Average (ARIMA) models. ARIMA may take various forms depending on whether the time series is stationary or not and which values the AutoRegressive (AR) and moving average (MA) orders will take. ARIMA models capture autocorrelation features of time series. In this study, an automatic ARIMA algorithm implemented in $fable$ package using $ARIMA()$ function. This algorithm employs

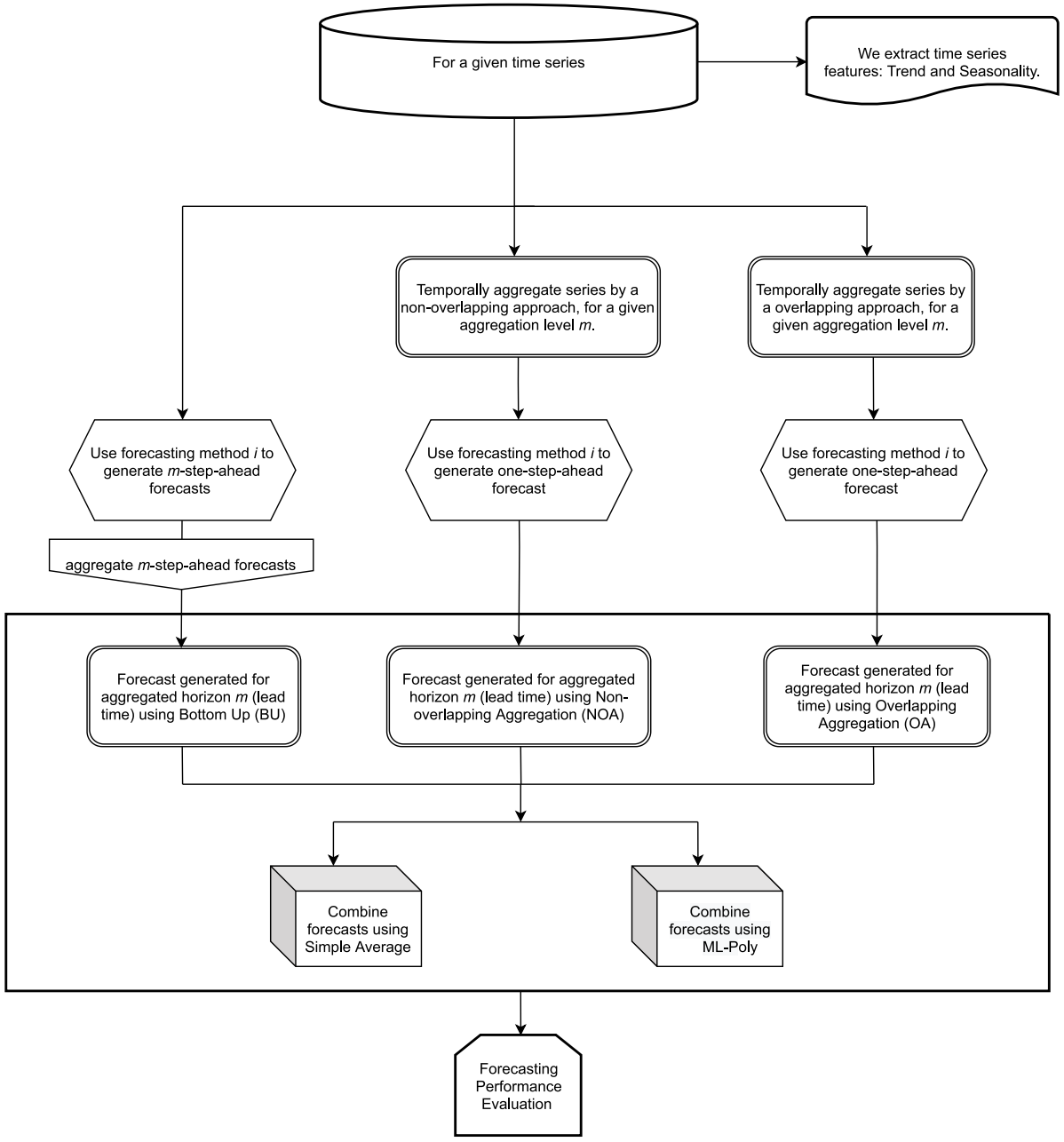


Fig. 2. Flow chart.

unit root tests, minimisation of the AICc, and Maximum Likelihood Estimation (MLE) to select the most appropriate model for a given time series.

3.2. Temporal aggregation and combination approaches

We employ five approaches to generate the required forecasts over lead-time m, $\hat{y}_{T,m}$, per forecasting method.

3.2.1. Bottom-Up (BU)

To generate the forecast over lead-time using Bottom-Up approach, we first create the forecast for m periods ahead using the given forecasting method, i.e. $T+1,\dots T+m$. These forecasts are then added up to obtain the forecast over the lead-time m, where $\hat{y}^{BU}_{T,m} = \sum_{i=1}^{m} \hat{y}_{T+i}$ is the forecast of the time series over the lead-time M.

3.2.2. Non-overlapping

To generate the forecast over lead-time using Non-Overlapping temporal Aggregation (NOA), we first create buckets of aggregated time series based on the aggregation level, m which equals to the lead-time. This process results in a new aggregated time series $Y_n, n = 1,2,\dots,[T/m]$; then the forecasting method is applied to these aggregated series to generate the forecast for one step ahead, which directly corresponds to the forecast over lead-time m where $\hat{y}^{NOA}_{T,m} = \hat{Y}_{[T/m]}$.

3.2.3. Overlapping

To generate the forecast over lead-time using Overlapping temporal Aggregation (OA) approach, we first aggregate series using overlapping buckets of m equals to the lead-time, that results in a new time series $Y_n, n = 1,2,\dots,T-m+1$; and then forecast one step ahead to obtain the forecast over lead-time m, $\hat{y}^{OA}_{T,m} = \hat{Y}_{T-m+1}$.

3.2.4. Simple average combination

The simple average approach is a combination of BU, NOA and OA approaches. To generate the forecast over lead-time m, we take the simple average of these three approaches, where

$$\hat{y}_{T,m}^{Average} = \frac{\hat{y}_{T,m}^{BU} + \hat{y}_{T,m}^{NOA} + \hat{y}_{T,m}^{OA}}{3}$$

3.2.5. Online updating combination

In addition to the simple average combination, we consider an online updating combination scheme, namely the polynomially weighted average with multiple learning rates (ML-Poly). Cesa-Bianchi and Lugosi (2003) show that polynomial potentials are useful in designing combination rule by online updating learning rates. In a follow-up study, Gaillard et al. (2014) develop an algorithm based on polynomial potentials with polynomial order of two and provide the upper bound for the loss function.

We briefly summarise the ML-Poly aggregation rule. In the setting of combination forecasting with the three individual forecasting approaches indexed by $k \in \{BU, NOA, OA\}$, the combiner construct a prediction $\hat{y}_t$ by choosing a vector $\boldsymbol{p}_t = \left(p_t^{BU}, p_t^{NOA}, p_t^{OA}\right)$ of non-negative weights summing to one for each period (aggregate horizon) in the out-of-sample. The prediction of the combiner is $\hat{y}_{T,m}^{ML\text{-}Poly} = p_t^{BU}\hat{y}_t^{BU} + p_t^{NOA}\hat{y}_t^{NOA} + p_t^{OA}\hat{y}_t^{OA}$. We will use a loss function to calculate the loss value of forecasts. The loss function can generally be any form. Following (a reference here), we set it as the squared difference between the true value and predicted value in this paper, i.e. $\ell_t^k = (y_t - \hat{y}_t^k)^2$. Then, the dynamic weights vector is determined by the loss of three forecasting approaches, $\boldsymbol{\ell}_t = \left(\ell_t^{BU}, \ell_t^{NOA}, \ell_t^{OA}\right)$, and the weighted loss of the combiner, $\hat{\ell}_t = p_t^{BU}\ell_t^{BU} + p_t^{NOA}\ell_t^{NOA} + p_t^{OA}\ell_t^{OA}$.

Define an quantity named *regret* which is the cumulative loss of kth the individual method, $R_t^k = \sum_{s=1}^{t}(\hat{\ell}_t - \ell_t^k)$. The objective of the combiner is to control the regret by sequentially updating the learning rates η_t^k. Gaillard et al. (2014) provide the learning rates for ML-Poly:

$$\eta_t^k = \frac{1}{1 + \sum_{s=1}^{t-1}(\hat{\ell}_s - \ell_s^k)^2}. \quad (1)$$

The detailed algorithm is summarised in the Algorithm 1. It is important to point out that ML-Poly is just one of the many ways that can be used to combine forecasts generated by BU, non-overlapping and overlapping temporal aggregation approaches. One might use any other methods proposed in the forecast combination literature, instead. However, the forecast accuracy improvement will depend on the method used for optimally combining the forecasts, the feature of series and the application examined.

3.3. Accuracy measurement

We report the forecasting performance of each approach using Mean Absolute Scaled Error (MASE) to measure the forecast accuracy. We also measured Mean Absolute Percentage Error (MAPE). However, given the similarities in the conclusion, here we only present the results of MASE.

$$\text{MASE} = \text{mean}(|q_j|),$$

$$q_j = \frac{y_j - \hat{y}_j}{\frac{1}{n-1}\sum_{t=2}^{n}|y_t - y_{t-1}|}.$$

where y_j and $\hat{y}_j$ are the actual and forecasted value for out-of-sample observations. The denominator is the mean absolute error of the naive method in the fitting sample of n observations and is used to scale the error.

To summarise the results across the time series of each dataset, the mean and median MASE across all series are computed.

Algorithm 1 ML-Poly combining rule

Initialisation:

- set the vector of learning rates of individual approaches $(\eta_0^{BU}, \eta_0^{NOA}, \eta_0^{OA})$
- set the vector of regrets of individual approaches $(R_0^{BU}, R_0^{NOA}, R_0^{OA}) = (0, 0, 0)$

repeat

At each time (aggregate horizon) in the out-of-sample

1. compute the learning rates η_{t-1}^k according to Eq. (1)
2. calculate the combining weights of each individual method by
$$p_t^k = \frac{\eta_{t-1}^k \max(0, R_{t-1}^k)}{\sum_{k=1}^{K}\eta_{t-1}^k \max(0, R_{t-1}^k)}$$
3. obtain the loss vector $\boldsymbol{\ell}_t = (\ell_t^{BU}, \ell_t^{NOA}, \ell_t^{OA})$ and the weighted loss $\hat{\ell}_t = p_t^{BU}\ell_t^{BU} + p_t^{NOA}\ell_t^{NOA} + p_t^{OA}\ell_t^{OA}$
4. update the regret $R_t^k = R_{t-1}^k + (\hat{\ell}_t - \ell_t^k)$

until *End of the out-of-sample*;

Table 1
The number of time series in M4 competition data.

M4 series	Total	Component			
		(N,N)	(N,S)	(T,N)	(T,S)
Quarterly	24,000	5869	2959	9325	5847
Monthly	48,000	11,412	8752	11,841	15,995
Daily	4227	3335	43	819	30

3.4. Dataset

We use the quarterly, monthly and daily subsets of M4 forecasting competition (Makridakis et al., 2018) dataset to evaluate empirically the forecast accuracy of five approaches for a given forecasting method.

M4 dataset includes time series from various sectors such as demographic, industry, finance, economics, and others, which make it an appropriate choice for this study. We use the R package "M4comp2018" (Montero-Manso et al., 2018) to access the dataset. We use the $ETS()$ function in the "fable" package in R to identify the existing components of each time series. Table 1 shows the total number of time series for each granularity and also the number of time series for each component: (i) (N, N): no trend, no seasonality, (ii) (N, S): no trend, seasonality, (iii) (T, N): trend, no seasonality and (iv) (T, S): trend, seasonality. For quarterly and monthly time series, we observe all type of patterns, however for most of the time series in daily dataset, $ETS()$ did not find any pattern.

3.5. Simulation setup

Our analysis focuses on the lead-time forecast (aggregate forecast horizon); while generating forecasts at disaggregate level (original higher frequency) might be relevant, this is not covered in this paper, however this study could be easily extended to evaluate the forecast accuracy at that level by introducing disaggregation mechanisms. Moreover, we choose the aggregation level to match the lead-time horizon, and this makes sense from a practical point of view.

For the quarterly time series, we use aggregation level, $m = 2, 4$, which corresponds to annual and semi-annual lead-time. For the monthly time series, we consider $m = 2, 3, 4, 6, 12$ corresponding to bi-monthly to Annual lead-times. For the daily series, we consider a lead-time of 2 days to 1 week (7 days) corresponding to aggregation level $m = 2, 3, 4, 5, 6, 7$.

We use a rolling origin forecast evaluation to determine the forecast accuracy of each approach, for a given forecasting method and aggregation level. We use the training set to generate the forecast for the

first given lead-time in the out-of-sample, followed by computing the error metric. Then, we include one new observation in the training set and continue the process until the number of observations left in out-of-sample equals the aggregation level. This will be the last generated forecast.

We should note that supercomputing facilities, with access to 40 cores and 150 GB of memory, were used to run the experiment, including fitting models, generating forecasts, and computing error metrics. The computational time to run the entire experiment with quarterly, monthly, and daily datasets was 9 weeks.

4. Results and discussions

In this section, we present the empirical findings of this study. We first look at the performance of BU, NOA and OA approaches and show for what percentage of time series, each approach provides more accurate forecast, on average. Next, we compare the overall performance of five approaches to forecast the lead-time, as well as the performance for each types of pattern identified in the time series.

4.1. Percentage best for BU, NOA and OA

We here show the percentage of occurrences each approach (i.e. BU, NOA and OA) wins, i.e. it provides more accurate lead-time forecast for a given lead-time using a given forecast method (i.e. ETS, ARIMA) across all time series in quarterly, monthly, and daily M4 competition datasets. We call this percentage best. To calculate the percentage best, we follow the following steps: (*i*) for each time series, we first compute the average of the error metric for each approach, forecasting method, trend & seasonality component and the given lead-time, across all rolling origin samples; (*ii*) next, we determine the winner approach, the one with the smallest error metric; (*iii*) we continue the process for all time series; (*iv*) we count the number of time each approach wins across total series and calculate the percentage.

Figs. 3(a)–3(b), 4(a)–4(b) and 5(a)–5(b) show the percentage best for M4 quarterly, monthly and daily time series, respectively. We have summarised the percentage best with and without trend and seasonality components. Our results indicate that BU approach is more accurate for almost 50% of time series, regardless of the forecasting method employed, the existing pattern and the required lead-time. For the daily time series this percentage is higher and may achieve up to 75% for time series with trend & no seasonality and no trend & no seasonality. We also observe that the percentage of temporal aggregation approaches for time series with seasonality & no trend and trend & seasonality is slightly higher.

The results show that temporal aggregation approaches might not always improve forecast accuracy and BU, overlapping and Non-overlapping temporal aggregation may have their own merit.

This shows that when forecast over lead-time is required, TA approaches might not always provide more accurate forecasts. BU is a reliable competitor regardless of whether there is any pattern such a trend or seasonality in the time series or not. These figures show that BU, NOA, and OA may have their own merit. Additionally, we can argue that the time series available at the original level, the one created by the non-overlapping and the overlapping temporal aggregation contain different information resulting in time series with different features. Exploiting these multiple levels of information and combining them could be beneficial for forecasting. This encourages us to investigate the possibility of combining forecasts generated from these three approaches, instead of using them individually.

4.2. Performance of temporal aggregation forecast combination

In this section, we report the forecasting performance of combining BU, NOA and OA using two forecast combination approaches: (*i*) simple average and (*ii*) an online updating combination scheme, namely ML-Poly. The performance is compared against each approach separately.

Fig. 6 demonstrates the performance of each approach using MASE for M4 quarterly time series. Each figure includes the mean and median of MASE across all time series. Additionally, we added two bars to include 5% (left bar) and 95% (right bar) quantiles for each approach to show the variation in the performance. The overall result shows that ML-Poly approach can beat all approaches when forecasting for the semi-annual lead-time, while it is the second best to forecast annual lead-time, regardless of the forecasting method. BU approaches is a competitive approach in both cases. However, using the simple average combination does not improve forecast accuracy. Both BU and ML-Poly show less variation compared to other approaches.

Tables 2 and A.5 (in the Appendix) present the forecasting performance of all approaches for the M4 quarterly time series using ETS and ARIMA method, respectively. Numbers in brackets refer to median MASE, while the rest to mean MASE. The results for each forecasting method and each pattern are presented separately to assess the accuracy of each approach given the forecasting method and pattern. In each row, the best performing method according to mean and median MASE is highlighted in boldface. These results confirm the observations shown in Fig. 6, however it seems that the performance is not affected by different time series patterns.

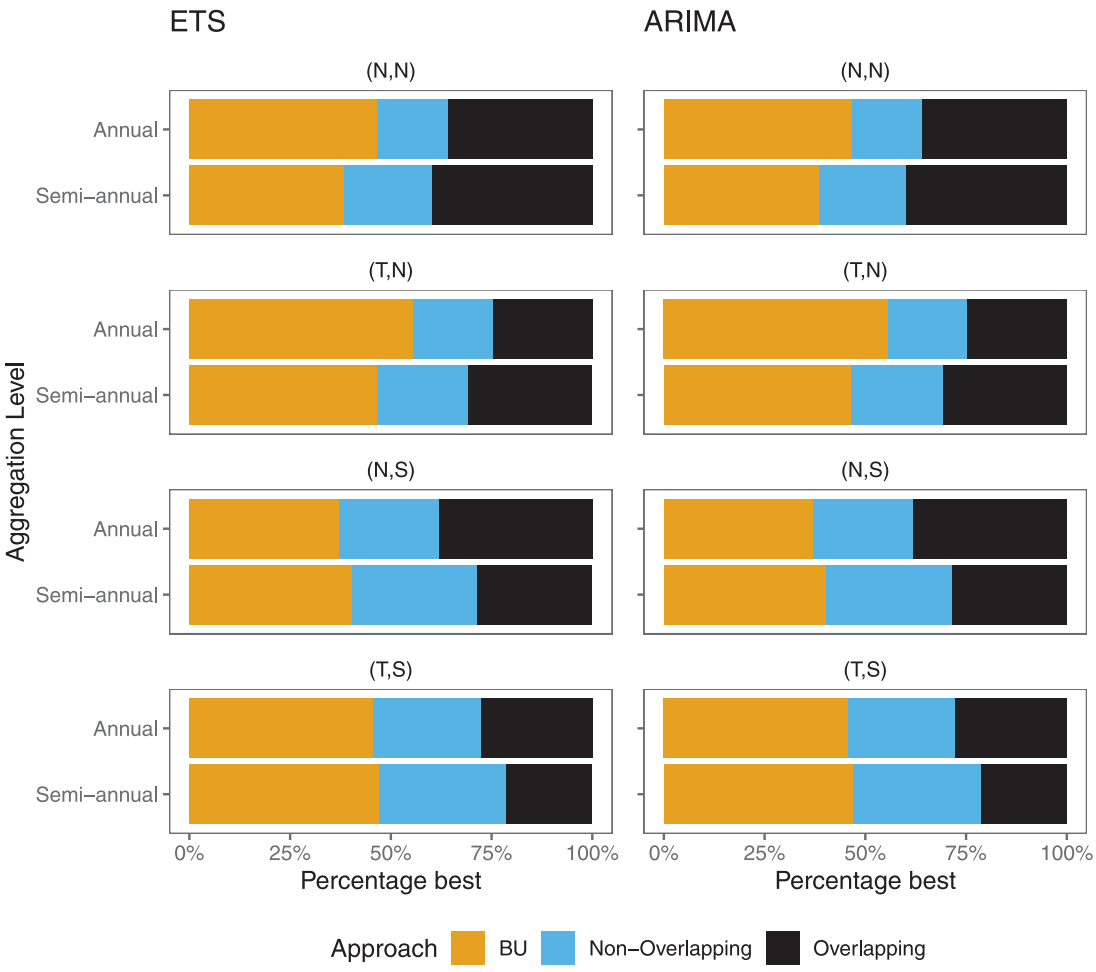
Fig. 7 shows the forecasting performance for the M4 monthly time series. Our overall results indicate that ML-Poly approach outperforms all other approaches, regardless of the forecasting method employed. The gain in forecast accuracy using ML-Poly approach compared with others increases with the lead-time. The highest gain for ML-Poly is achieved when forecasting annual lead-time. BU approach is the second-best approach, followed by simple average, overlapping and non-overlapping temporal aggregation. It is also important to note that ML-Poly approach shows less variation in the performance, followed by BU.

Tables 3 and A.6 (in the Appendix) present the forecast accuracy for each component of the M4 monthly time series, when ETS and ARIMA method is employed. We observe that regardless of the existing time series pattern and the forecasting method, ML-Poly approach is always the most accurate approach. However, the detailed results show that ML-Poly approach is less powerful with shorter lead-time and BU becomes more competitive when forecasting bi-monthly lead-time.

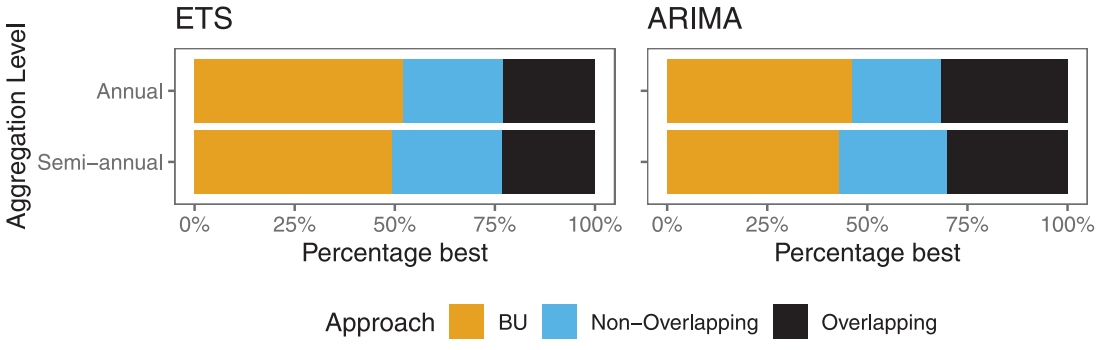
Fig. 8 shows the performance of each approach for M4 daily time series. Both ML-Poly and BU approaches provide accurate result. While ML-Poly becomes more accurate for longer lead-time, BU is more accurate for shorter lead-time. The result is very similar for both ETS and ARIMA.

Tables 4 and A.7 (in the Appendix) show the forecast accuracy for each component of the M4 daily time series. We should note that the majority of time series are identified with no trend and seasonality, therefore the number of time series with Trend, Seasonality and Trend & seasonality are very limited. Hence, results presented for these categories might not be reliable as it is computed based on a very small sample. For the time series with none components, we observe that BU approach is always more accurate for both ETS and ARIMA methods, while for time series with trend component and no seasonality, ML-Poly is more accurate.

It is also important to note that, daily time series may contain multiple seasonal cycles. In that case, an appropriate forecasting method should be used. Both ARIMA and ETS are not suitable for time series with, multiple seasonality.



(a) With trend and seasonality components



(b) Regardless the trend and seasonality components

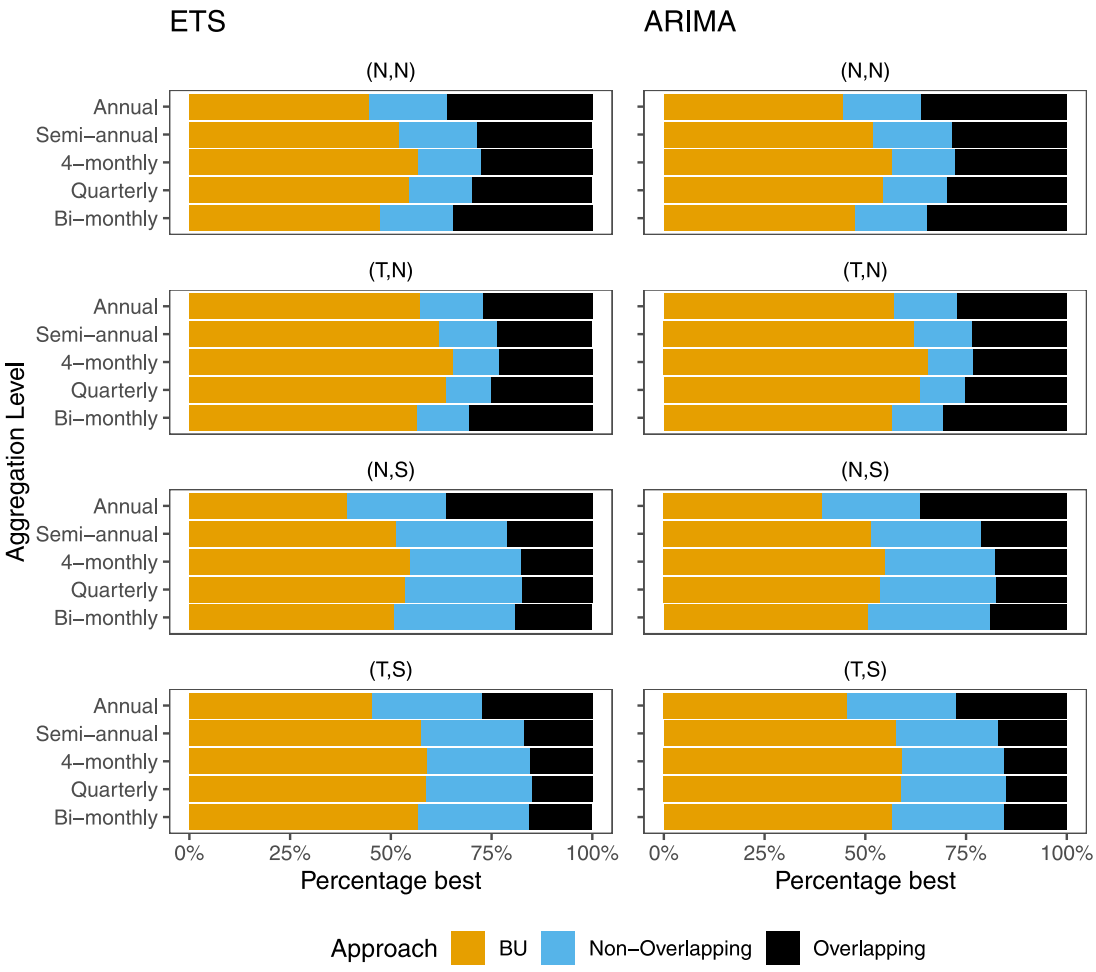
Fig. 3. Percentage best for quarterly time series.

4.3. Significant test

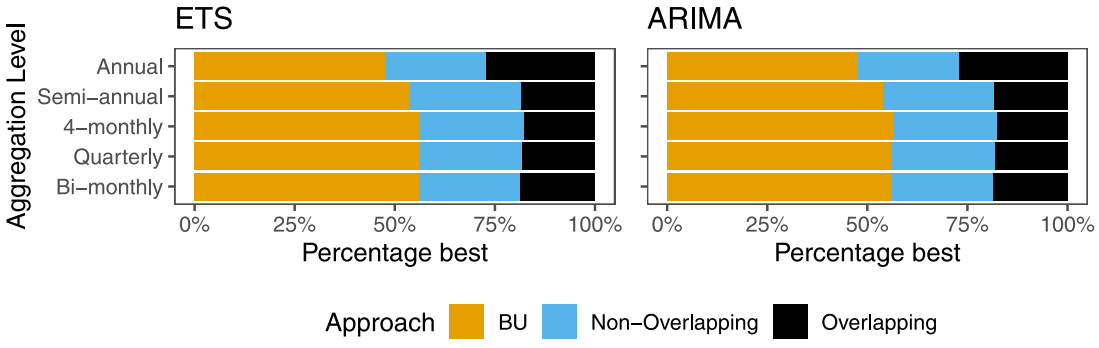
We also conducted the Multiple Comparison with the Best (MCB) method (Demšar, 2006) to investigate the statistical significance in performance of different approaches at various aggregation levels for all series. The test is implemented using the function of rmcb() in the R package "greybox".

The results of the MCB test are plotted in Figs. 9–11 and Figs. B.12–B.14 (in the Appendix), for ETS and ARIMA, respectively. We observe that the difference in the performance of all approaches is statistically significant for quarterly, monthly, and daily series. For almost all cases, either ML-Poly or BU approaches have lower ranks, which also been observed in Tables 2–A.5. It is also clear that using Overlapping and Non-overlapping temporal aggregation approaches are significantly worse than the three others in terms of median MASE.

The intuition behind the superior performance of ML-Poly is mainly due to the online adjustment of the combination weights. In principle, ML-Poly tracks the forecasting loss of individual forecasting approaches



(a) With trend and seasonality components



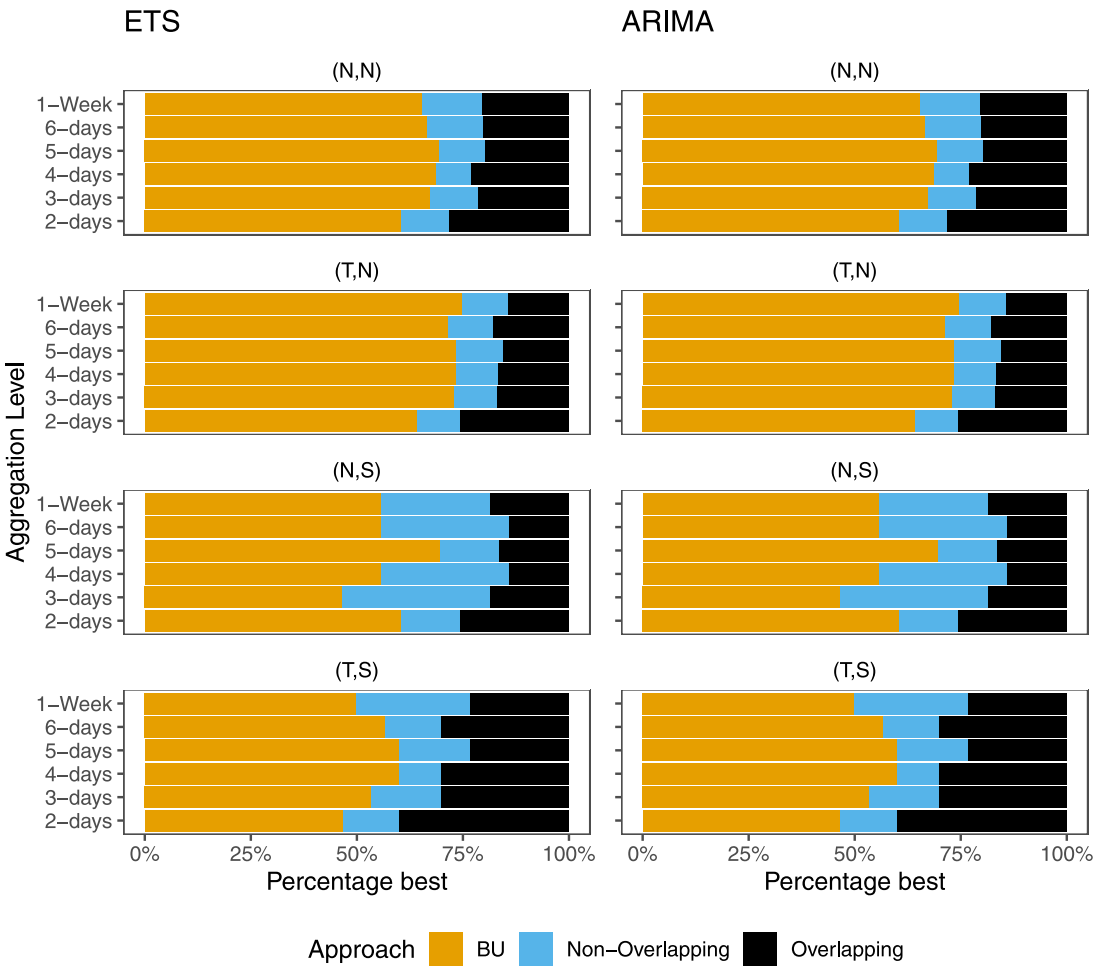
(b) Regardless the trend and seasonality components

Fig. 4. Percentage best for monthly time series.

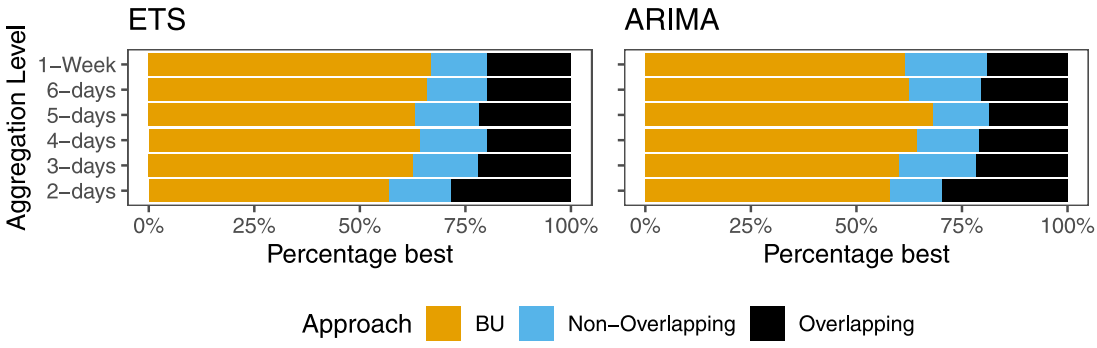
in an online manner, and dynamically reduces the weights on poor approaches and increases the weights on accurate ones. Therefore, the accurate approaches will contribute more to the combiner, and such mechanism helps to improve the forecasting performance, as shown by our comprehensive results based on the M4 dataset. In other words, ML-Poly takes the past forecasting loss into account and penalise on those poor approaches. In this way, ML-Poly can largely leave out the poor individual forecasters and mainly combine the accurate ones.

5. Conclusion and implications

With advances in IT and data collection tools and techniques, data can be collected in the finest time granularity, which can be converted to a time series with equal space intervals such as hourly, daily, or monthly time series. While data is recorded in higher frequency (e.g. monthly), forecasts might in practice be required at lower frequencies (e.g. annual). Often, it is required for a forecast of a parameter over



(a) With trend and seasonality components



(b) Regardless the trend and seasonality components

Fig. 5. Percentage best for daily time series.

several time periods ahead (lead-time) rather than individual periods. For instance, in stock control a forecast is required over the lead-time to determine the safety stock and stock replenishment.

In order to generate the forecast over lead-time for a given time series, there are two different possibilities: (*i*) generate a forecast using the given time series for a forecast horizon equal to the lead-time and then add them up, or (*ii*) aggregate the given time series using time buckets equal to the number of periods required over the lead-time and then generate the forecast for one period ahead. For the second possibility, we can use non-overlapping or overlapping temporal aggregation approaches to create the aggregated time series. Therefore, in total there are three ways to generate lead-time forecast using any given forecasting method named Bottom-Up (BU), Non-overlapping (NOA) and Overlapping (OA) temporal aggregation.

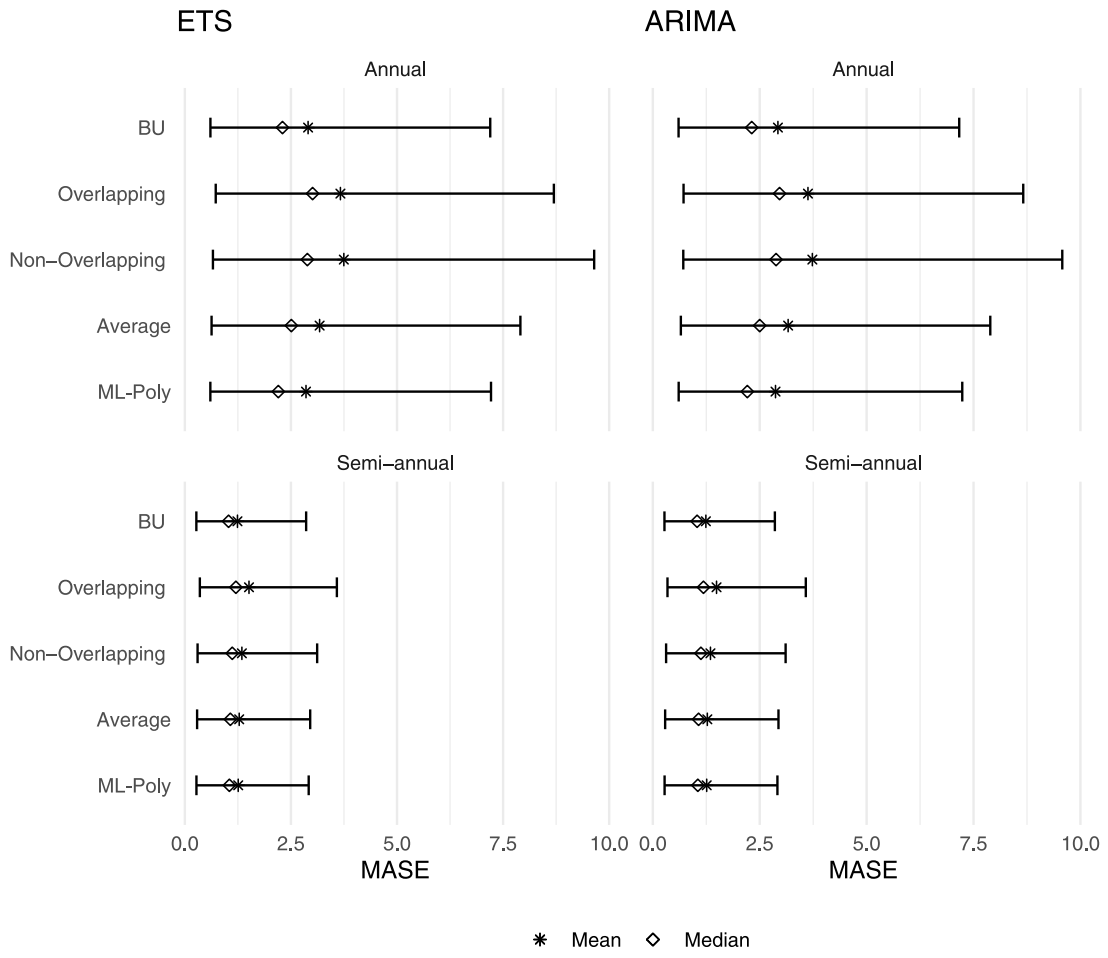


Fig. 6. Mean and median, 5% (left bar) and 95% (right bar) quantiles of MASE for the quarterly time series.

Table 2
Mean (median) MASE for M4 quarterly series with ETS forecasting method. The best approach for each aggregation level and each forecasting method is highlighted in bold.

Aggregation level	Pattern	Approach				
		ML-Poly	Average	Overlapping	Non-overlapping	BU
Annual	(N, N)	2.880 (**2.298**)	3.048 (2.439)	3.088 (2.458)	3.610 (2.843)	**2.845** (2.307)
	(T, N)	**2.820** (**2.096**)	3.294 (2.575)	4.085 (3.423)	4.008 (3.083)	2.868 (2.194)
	(N, S)	**3.086** (**2.473**)	3.178 (2.544)	3.244 (2.586)	3.573 (2.760)	3.159 (2.567)
	(T, S)	**2.769** (**2.126**)	3.115 (2.467)	3.791 (3.184)	3.555 (2.679)	2.894 (2.303)
Semi-annual	(N, N)	1.225 (1.044)	1.226 (1.048)	1.260 (1.068)	1.315 (1.117)	**1.206** (**1.026**)
	(T, N)	1.115 (0.877)	1.132 (0.893)	1.232 (1.007)	1.251 (0.991)	**1.104** (**0.872**)
	(N, S)	1.587 (1.392)	1.614 (1.396)	2.006 (1.647)	1.617 (1.395)	**1.566** (**1.345**)
	(T, S)	1.351 (1.170)	1.406 (1.214)	1.959 (1.558)	1.384 (1.172)	**1.324** (**1.124**)

Critically, and to the best of our knowledge, this is the first time that different temporal aggregation approaches are compared and combined. Using quarterly, monthly, and daily M4 competition series, we design and execute an expansive experiment exploring the performance of these approaches on lead-time forecasting.

When looking at the comparative performance of the three approaches, our findings indicate that neither of the individual approaches have an overall win on forecasting accuracy when accuracy is reported at the series level. We were surprised by the power of aggregating the forecast generated using the original series (BU) rather than forecasting by temporally aggregated time series. This may highlight the fact that if a forecasting method capable of capturing systematic information at the original level is employed, BU approach might be always preferable, which is the case with Monthly and Quarterly M4 data, when ETS and ARIMA is employed.

When it comes to the overall performance of approaches or the performance summarised for each class of existing patterns in the original series, it is surprising to observe that Non-Overlapping and Overlapping temporal aggregation approaches are performing as poorly. Moreover, it seems that the existence of trend and/or seasonality does not affect the overall performance. This is an area which requires more in-depth analysis to connect the features of time series to the performance of each approach. We believe this to be an important finding for practitioners as the prevalent thinking has been to use temporal aggregation when forecasting over a lead-time is required. We provided evidence that this might not be the right choice.

Equally concerning is the fact that simple average combinations of these forecasts are similarly of no benefit to the accuracy either. To extract the benefits of these individual forecasts we propose an on-line forecast combination approach. It combines forecasts generated by BU, NOA and OA to produce the final required lead-time forecast. Overall,

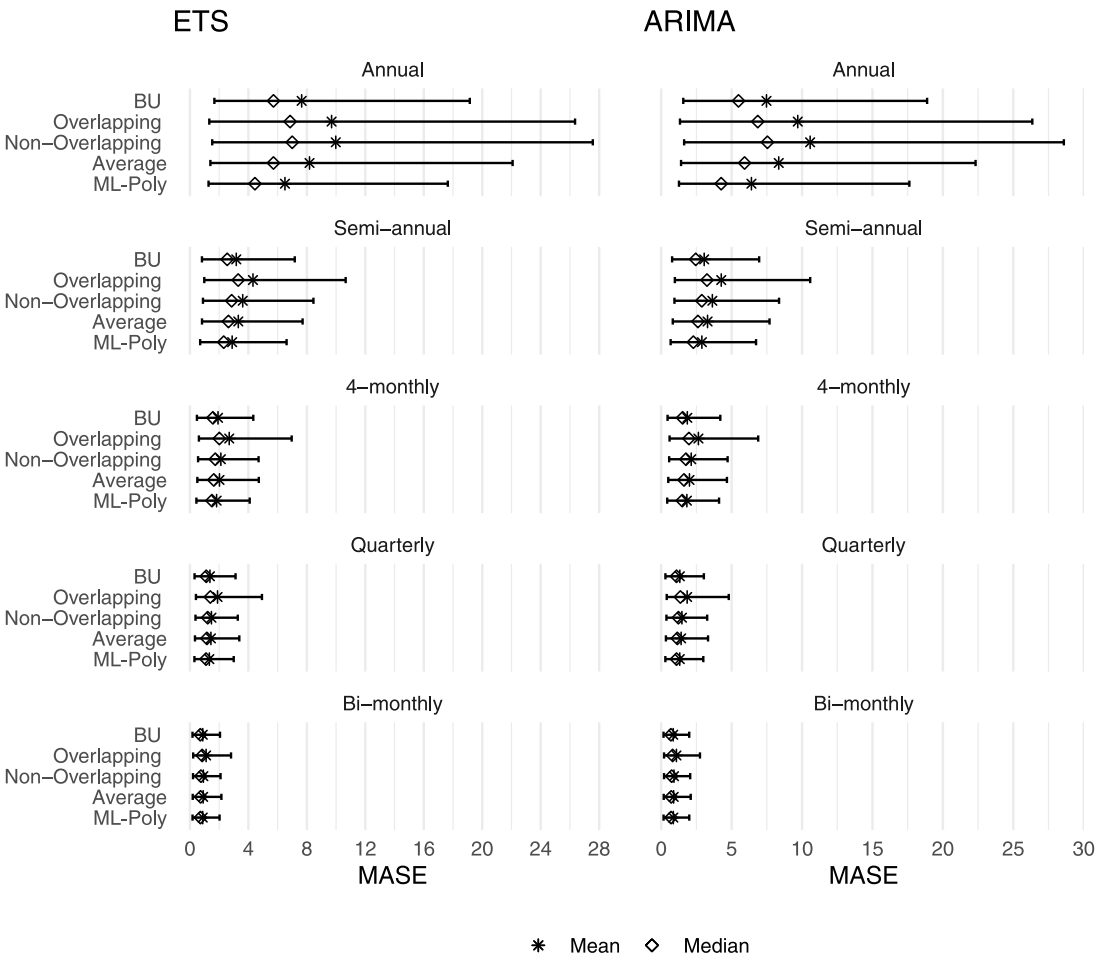


Fig. 7. Mean and median, 5% (left bar) and 95% (right bar) quantiles of MASE for the monthly time series.

Table 3
Mean (median) MASE for M4 monthly series with ETS forecasting method. The best approach for each aggregation level and each forecasting method is highlighted in bold.

Aggregation level	Pattern	Approach				
		ML-Poly	Average	Overlapping	Non-overlapping	BU
Annual	(N, N)	**6.675 (4.456)**	7.958 (5.38)	8.491 (5.723)	9.822 (6.939)	7.673 (5.590)
	(T, N)	**6.483 (4.299)**	8.785 (5.834)	11.161 (7.643)	11.472 (7.66)	7.603 (5.495)
	(N, S)	**6.826 (4.760)**	7.883 (5.579)	8.288 (5.857)	9.191 (6.585)	8.119 (6.252)
	(T, S)	**6.215 (4.394)**	8.064 (5.957)	10.23 (7.96)	9.417 (6.820)	7.396 (5.671)
Semi-annual	(N, N)	**3.023 (2.500)**	3.236 (2.635)	3.398 (2.800)	3.642 (2.871)	3.170 (2.637)
	(T, N)	**2.440 (1.809)**	2.893 (2.153)	3.502 (2.640)	3.511 (2.607)	2.714 (2.045)
	(N, S)	**3.421 (2.803)**	3.875 (3.167)	5.161 (3.992)	4.025 (3.315)	3.82 (3.156)
	(T, S)	**2.835 (2.290)**	3.366 (2.695)	5.105 (3.833)	3.442 (2.768)	3.156 (2.548)
4-monthly	(N, N)	**1.886 (1.620)**	1.951 (1.668)	2.041 (1.751)	2.155 (1.819)	1.916 (1.635)
	(T, N)	**1.430 (1.063)**	1.574 (1.181)	1.845 (1.421)	1.837 (1.420)	1.499 (1.115))
	(N, S)	**2.288 (1.937)**	2.538 (2.115)	3.501 (2.752)	2.551 (2.146)	2.457 (2.056)
	(T, S)	**1.824 (1.497)**	2.121 (1.693)	3.330 (2.354)	2.057 (1.691)	1.948 (1.578)
Quarterly	(N, N)	**1.358** (1.179)	1.382 (1.198)	1.446 (1.259)	1.496 (1.292)	1.366 (**1.167**)
	(T, N)	**0.998 (0.731)**	1.054 (0.772)	1.198 (0.909)	1.195 (0.917)	1.012 (0.733)
	(N, S)	**1.722 (1.471)**	1.862 (1.567)	2.515 (1.986)	1.854 (1.553)	1.807 (1.529)
	(T, S)	**1.338 (1.093)**	1.524 (1.198)	2.372 (1.644)	1.448 (1.181)	1.394 (1.121)
Bi-monthly	(N, N)	0.881 (0.758)	0.886 (0.764)	0.931 (0.799)	0.926 (0.799)	**0.877 (0.745)**
	(T, N)	0.615 (0.423)	0.627 (0.429)	0.684 (0.470)	0.679 (0.493)	**0.608 (0.412)**
	(N, S)	**1.191 (1.030)**	1.218 (1.050)	1.504 (1.241)	1.223 (1.043)	1.208 (1.031)
	(T, S)	**0.895** (0.717)	0.951 (0.745)	1.329 (0.951)	0.919 (0.737)	0.900 (**0.713**)

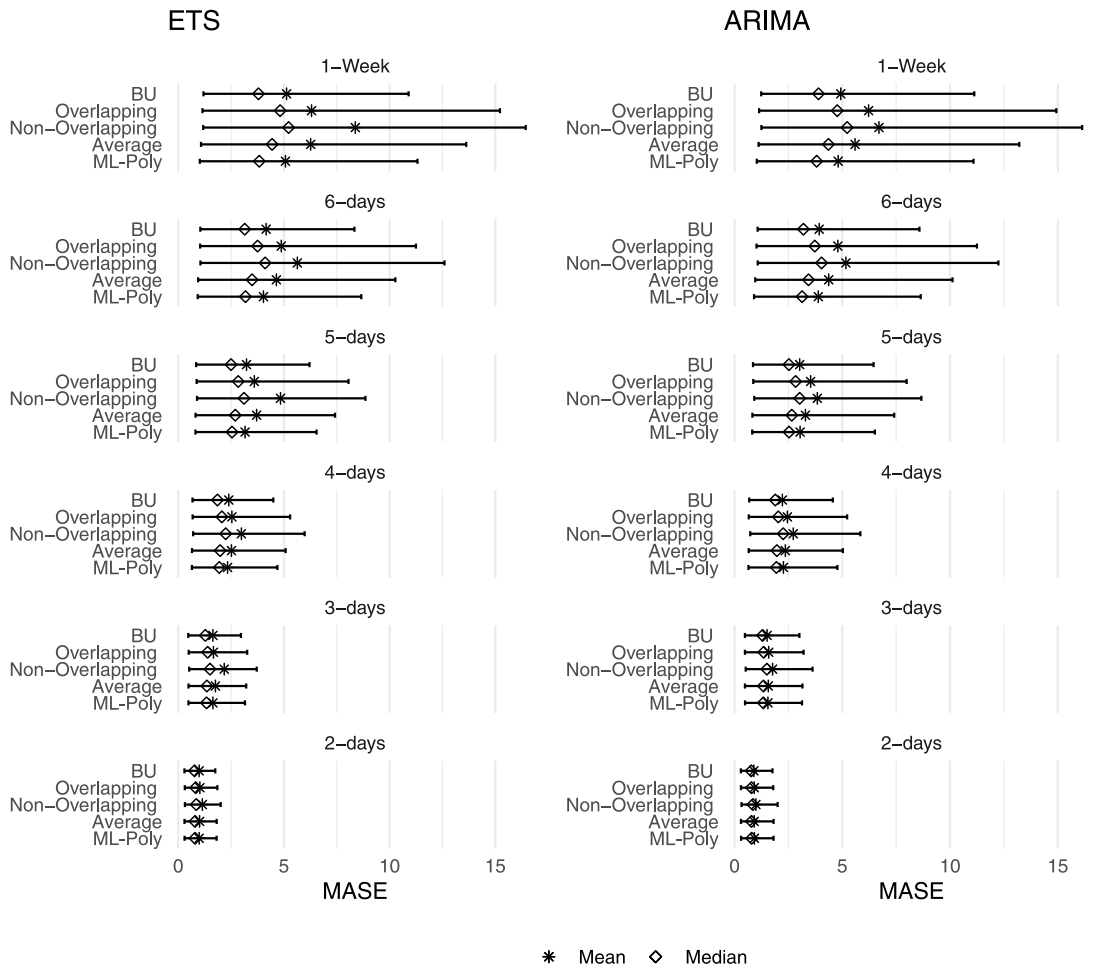


Fig. 8. Mean and median, 5% (left bar) and 95% (right bar) quantiles of MASE for the daily time series.

Table 4
Mean (median) MASE for M4 daily series with ETS forecasting method. The best approach for each aggregation level and each forecasting method is highlighted in bold.

Aggregation level	Pattern	Approach				
		ML-Poly	Average	Overlapping	Non-overlapping	BU
1-week	(N, N)	4.692 (3.749)	5.615 (4.231)	5.819 (4.59)	7.212 (4.969)	**4.604** (**3.718**)
	(T, N)	**6.303** (**2.504**)	6.032 (3.064)	4.707 (3.354)	11.095 (3.505)	8.253 (2.535)
	(N, S)	**5.619** (4.300)	7.05 (5.409)	8.237 (6.22)	8.45 (6.587)	5.622 (**4.212**)
	(T, S)	30.772 (4.732)	57.594 (5.726)	**10.305** (5.435)	132.226 (6.179)	45.217 (**4.694**)
6-days	(N, N)	3.796 (3.133)	4.292 (3.371)	4.516 (3.606)	5.218 (3.962)	**3.685** (**3.085**)
	(T, N)	4.882 (**1.94**)	4.712 (2.268)	**3.752** (2.448)	7.154 (2.997)	6.831 (2.444)
	(N, S)	4.467 (3.454)	5.368 (4.14)	6.145 (4.626)	6.421 (4.97)	**4.458** (**3.355**)
	(T, S)	17.211 (**3.913**)	23.904 (4.231)	**11.85** (4.213)	28.59 (4.968)	44.671 (3.947)
5-days	(N, N)	2.984 (2.512)	3.328 (2.616)	3.363 (2.749)	4.153 (3.012)	**2.851** (**2.455**)
	(T, N)	4.130 (**1.582**)	3.834 (1.593)	**2.838** (1.847)	5.637 (2.195)	6.205 (2.058)
	(N, S)	3.445 (2.739)	3.94 (3.082)	4.398 (3.345)	4.65 (3.631)	**3.405** (**2.642**)
	(T, S)	14.066 (**2.849**)	39.462 (3.321)	**9.237** (3.669)	85.127 (3.694)	36.409 (2.973)
4-days	(N, N)	2.233 (1.92)	2.336 (1.944)	2.375 (2.024)	2.81 (2.198)	**2.104** (**1.836**)
	(T, N)	2.290 (**1.405**)	3.6 (1.411)	**2.09** (1.531)	4.811 (1.589)	5.783 (1.574)
	(N, S)	2.566 (2.053)	2.772 (2.147)	3.003 (2.312)	3.194 (2.49)	**2.504** (**1.98**)
	(T, S)	**7.403** (2.069)	14.271 (2.484)	8.762 (2.35)	14.524 (2.915)	26.583 (**1.991**)
3-days	(N, N)	1.516 (1.323)	1.622 (1.334)	1.565 (1.369)	2.009 (1.47)	**1.436** (**1.266**)
	(T, N)	1.747 (1.126)	2.672 (1.151)	**1.661** (**0.998**)	3.659 (1.202)	4.556 (1.140)
	(N, S)	1.764 (1.432)	1.823 (1.468)	1.895 (1.557)	2.054 (1.631)	**1.709** (**1.384**)
	(T, S)	**11.671** (1.470)	14.084 (1.563)	6.718 (1.341)	22.089 (1.861)	18.328 (**1.418**)
2-days	(N, N)	0.916 (0.780)	0.919 (0.781)	0.96 (0.815)	1.027 (0.839)	**0.861** (**0.757**)
	(T, N)	1.448 (0.691)	2.033 (0.669)	2.146 (**0.664**)	**1.343** (0.72)	3.346 (0.729)
	(N, S)	**1.041** (**0.867**)	1.042 (0.863)	1.061 (0.878)	1.138 (0.935)	1.010 (0.830)
	(T, S)	**7.303** (0.900)	9.112 (0.915)	5.709 (0.837)	13.732 (0.885)	11.575 (**0.872**)

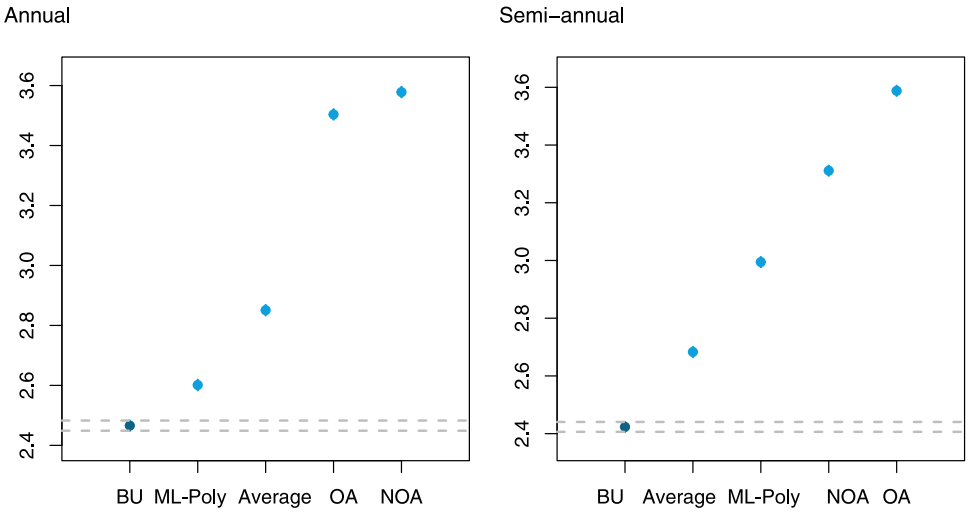


Fig. 9. MCB test for ETS, quarterly series.

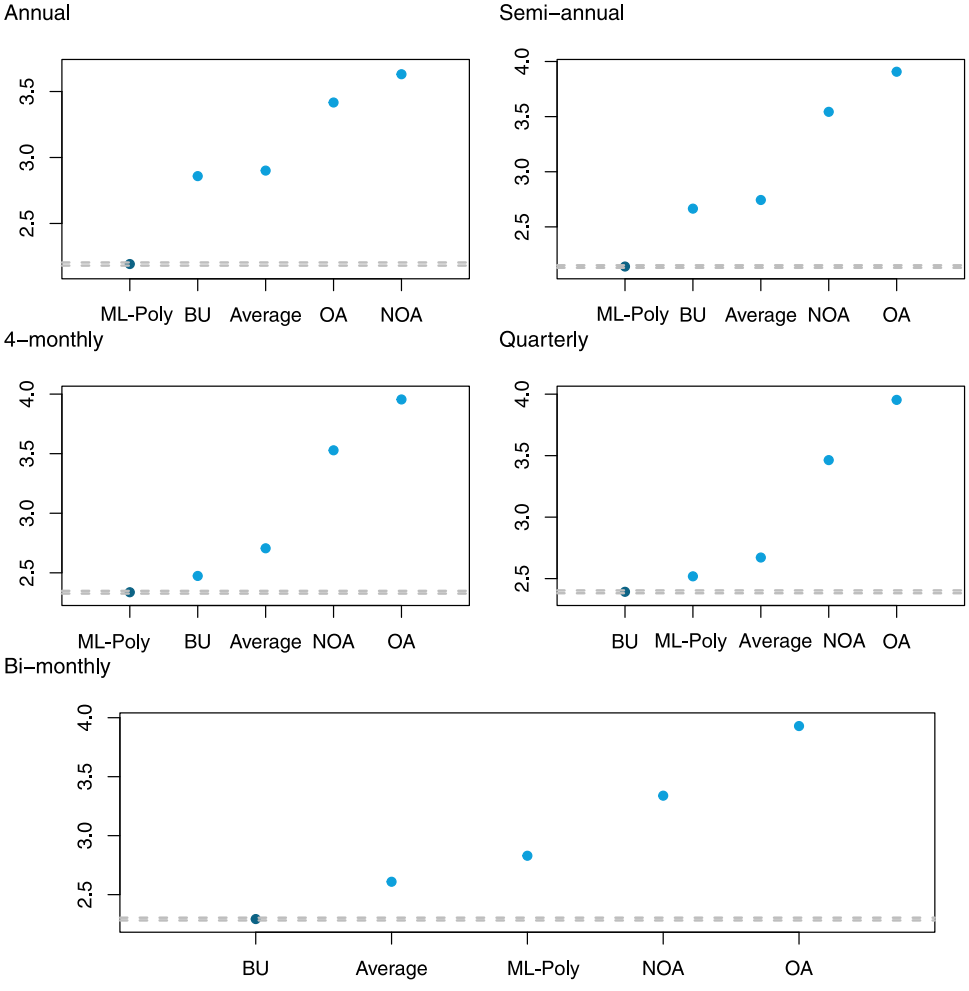


Fig. 10. MCB test for ETS, monthly series.

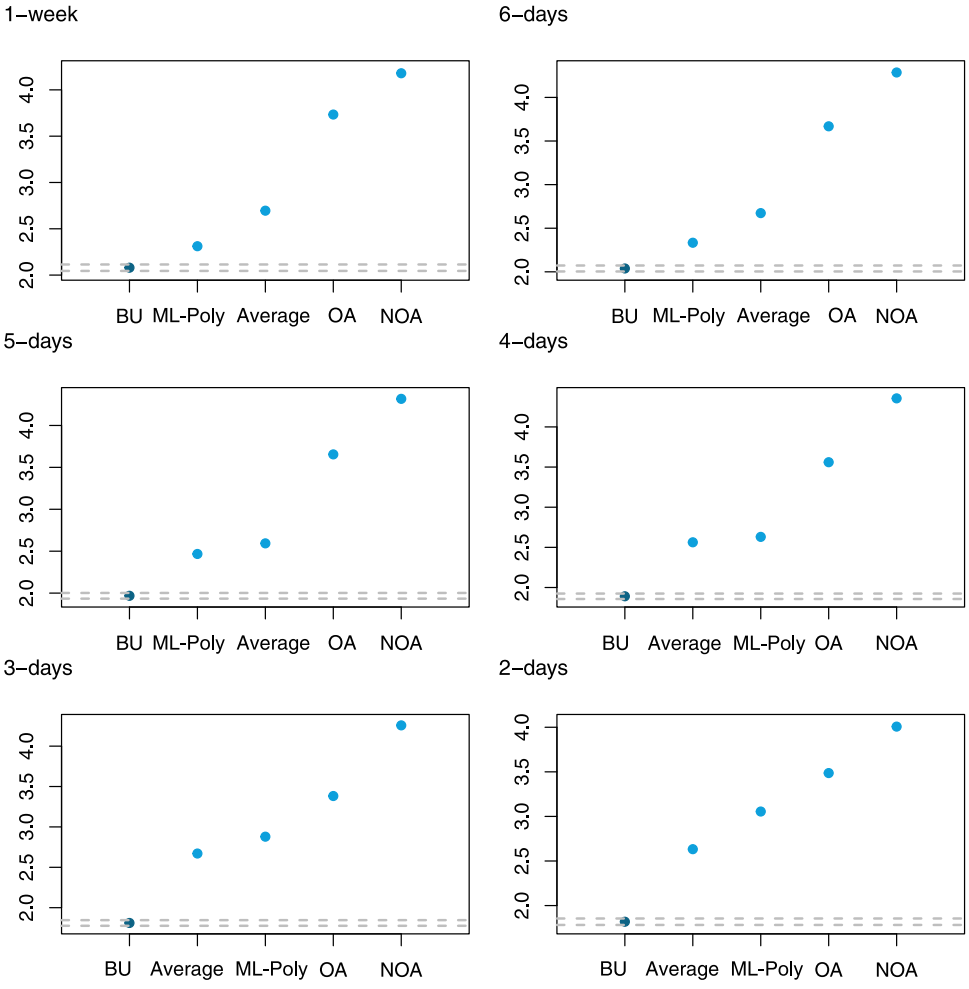


Fig. 11. MCB test for ETS, daily series.

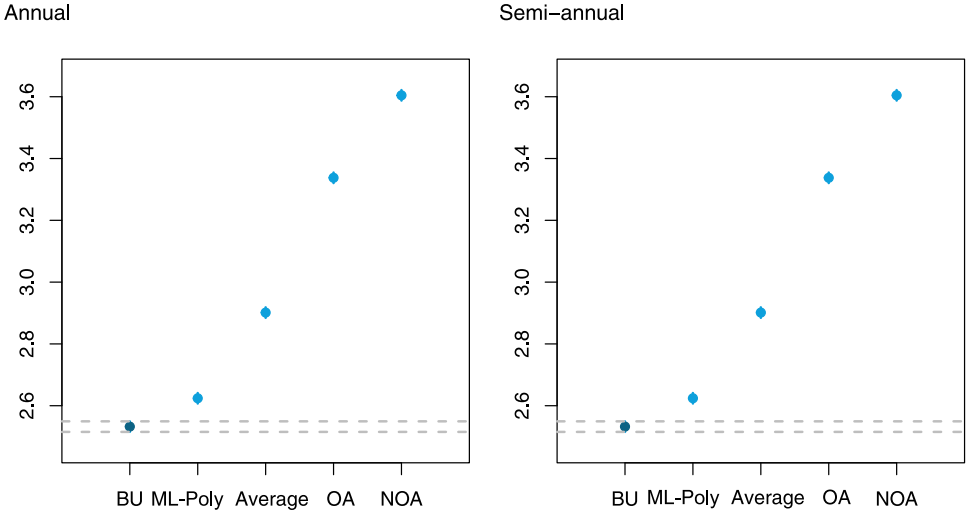


Fig. B.12. MCB test with ARIMA, quarterly series.

Table A.5
Mean (median) MASE for M4 quarterly series with ARIMA forecasting method. The best approach for each aggregation level and each forecasting method is highlighted in bold.

Aggregation level	Pattern	Approach				
		ML-Poly	Average	Overlapping	Non-overlapping	BU
Annual	(N, N)	**2.930 (2.365)**	3.132 (2.538)	3.052 (2.445)	3.917 (3.163)	2.959 (2.420)
	(T, N)	**2.830 (2.111)**	3.230 (2.499)	4.050 (3.395)	3.805 (2.880)	**2.830** (2.161)
	(N, S)	**3.089 (2.444)**	3.226 (2.579)	3.197 (2.584)	3.734 (2.971)	3.258 (2.640)
	(T, S)	**2.759 (2.112)**	3.061 (2.424)	3.750 (3.102)	3.423 (2.591)	2.876 (2.266)
Semi-annual	(N, N)	**1.235 (1.060)**	1.228 (1.055)	1.228 (1.057)	1.355 (1.157)	1.237 (1.065)
	(T, N)	1.099 (0.866)	1.105 (0.872)	1.193 (0.982)	1.220 (0.963)	**1.087 (0.852)**
	(N, S)	1.610 (1.406)	1.624 (1.411)	2.006 (1.624)	1.657 (1.431)	**1.589 (1.391)**
	(T, S)	1.364 (1.169)	1.415 (1.231)	1.965 (1.549)	1.387 (1.178)	**1.313 (1.116)**

Table A.6
Mean (median) MASE for M4 monthly series with ARIMA forecasting method. The best approach for each aggregation level and each forecasting method is highlighted in bold.

Aggregation level	Pattern	Approach				
		ML-Poly	Average	Overlapping	Non-overlapping	BU
Annual	(N, N)	**6.655 (4.490)**	8.277 (5.830)	8.504 (5.734)	11.008 (8.073)	7.661 (5.681)
	(T, N)	**6.549 (4.185)**	9.193 (6.306)	11.180 (7.640)	12.464 (8.809)	7.250(5.128)
	(N, S)	**6.692 (4.559)**	8.069 (5.703)	8.275 (5.811)	9.908 (7.002)	8.034 (6.045)
	(T, S)	**5.965 (4.023)**	7.918 (5.816)	10.222 (7.959)	9.228 (6.612)	7.200 (5.348)
Semi-annual	(N, N)	**3.014 (2.480)**	3.225 (2.618)	3.375 (2.769)	3.778 (3.026)	3.153 (2.643)
	(T, N)	**2.413 (1.767)**	2.852 (2.126)	3.475 (2.607)	3.412 (2.570)	2.561 (1.906)
	(N, S)	**3.470 (2.812)**	3.874 (3.129)	5.089 (3.923)	4.113 (3.36)	3.700 (3.014)
	(T, S)	**2.832 (2.240)**	3.342 (2.645)	5.026 (3.764)	3.427 (2.752)	2.985 (2.391)
4-monthly	(N, N)	**1.882 (1.612)**	1.941 (1.660)	2.008 (1.726)	2.207 (1.874)	1.901 (1.633)
	(T, N)	**1.403 (1.039)**	1.551 (1.162)	1.807 (1.385)	1.812 (1.403)	1.432 (1.047)
	(N, S)	**2.328 (1.957)**	2.540 (2.126)	3.439(2.698)	2.620 (2.194)	2.384 (2.005)
	(T, S)	**1.816 (1.465)**	2.101 (1.65)	3.276 (2.295)	2.046 (1.676)	1.839 (1.484)
Quarterly	(N, N)	**1.353** (1.179)	1.370 (1.188)	1.407 (1.228)	1.529 (1.322)	1.358 (**1.172**)
	(T, N)	**0.971** (0.699)	1.029 (0.749)	1.154 (0.870)	1.177 (0.906)	0.977 (**0.692**)
	(N, S)	**1.735 (1.482)**	1.847 (1.570)	2.456 (1.941)	1.883 (1.596)	1.756 (1.497)
	(T, S)	1.319 (1.062)	1.494 (1.163)	2.323 (1.600)	1.423 (1.156)	**1.315 (1.055)**
Bi-monthly	(N, N)	0.878 (0.753)	0.873 (0.749)	0.903 (0.766)	0.937 (0.811)	**0.874 (0.745)**
	(T, N)	0.593(0.395)	0.603 (0.404)	0.655 (0.436)	0.663 (0.476)	**0.592 (0.391)**
	(N, S)	1.180 (1.024)	1.200 (1.038)	1.477 (1.235)	1.222 (1.054)	**1.174 (1.019)**
	(T, S)	0.868 (0.685)	0.924 (0.718)	1.305 (0.925)	0.892 (0.711)	**0.852 (0.672)**

Table A.7
Mean (median) MASE for M4 daily series with ARIMA forecasting method. The best approach for each aggregation level and each forecasting method is highlighted in bold.

Aggregation level	Pattern	Approach				
		ML-Poly	Average	Overlapping	Non-overlapping	BU
1-week	(N, N)	**4.630 (3.729)**	5.278 (4.167)	5.764 (4.543)	6.311 (5.02)	4.743 (3.819)
	(T, N)	**4.108 (2.564)**	4.731 (3.026)	4.711 (3.284)	5.376 (3.16)	4.638 (3.298)
	(N, S)	**5.563** (4.270)	6.911 (5.39)	8.151 (6.077)	8.364 (6.337)	5.687 (**4.151**)
	(T, S)	**5.307** (4.864)	5.899 (5.928)	6.39 (5.422)	6.773 (6.829)	5.514 (**4.764**)
6-days	(N, N)	**3.756 (3.077)**	4.151 (3.329)	4.485 (3.592)	4.904 (3.874)	3.785 (3.130)
	(T, N)	**3.333 (2.076)**	3.772 (2.225)	3.768 (2.418)	4.337 (2.681)	3.671 (2.846)
	(N, S)	**4.424** (3.397)	5.279 (4.001)	6.095 (4.577)	6.276 (4.775)	4.513 (**3.346**)
	(T, S)	4.528 (**3.852**)	4.821 (4.305)	5.088 (4.131)	5.583 (4.631)	**4.508 (3.852)**
5-days	(N, N)	2.950 (**2.480**)	3.151 (2.571)	3.333 (2.73)	3.676 (2.942)	**2.919** (2.496)
	(T, N)	**2.600 (1.501)**	2.836 (1.575)	2.845 (1.797)	3.145 (1.974)	2.872 (2.649)
	(N, S)	**3.438** (2.689)	3.867 (2.958)	4.371 (3.328)	4.535 (3.433)	3.445 (**2.630**)
	(T, S)	**3.490 (2.85)**	3.769 (3.336)	3.885 (3.524)	4.365 (3.529)	3.549 (2.971)
4-days	(N, N)	2.196 (1.923)	2.263 (1.922)	2.333 (1.985)	2.613 (2.203)	**2.143 (1.865)**
	(T, N)	**2.011 (1.497)**	2.11 (1.686)	2.087 (1.512)	2.308 (1.914)	2.157 (1.854)
	(N, S)	2.573 (2.046)	2.748 (2.149)	2.96 (2.283)	3.185 (2.485)	**2.523 (1.994)**
	(T, S)	**2.697 (1.989)**	2.785 (2.259)	2.79 (2.316)	3.315 (2.644)	**2.697** (2.11)
3-days	(N, N)	1.500 (1.32)	1.508 (1.324)	1.521 (1.338)	1.695 (1.474)	**1.459 (1.286)**
	(T, N)	**1.357 (1.009)**	1.425 (1.335)	1.406 (1.124)	1.548 (1.064)	1.462 (1.255)
	(N, S)	1.754 (1.423)	1.803 (1.483)	1.861 (1.504)	2.051 (1.66)	**1.716 (1.392)**
	(T, S)	**1.846 (1.22)**	1.940 (1.498)	1.925 (1.328)	2.187 (1.646)	1.912 (**1.479**)
2-days	(N, N)	0.894 (0.790)	0.883 (0.777)	0.885 (0.780)	0.961 (0.84)	**0.869 (0.766)**
	(T, N)	**0.842 (0.686)**	0.863 (0.716)	0.843 (0.673)	0.915 (**0.729**)	0.884 (0.739)
	(N, S)	1.037 (0.863)	1.025 (0.841)	1.035 (0.848)	1.119 (0.933)	**1.009** (0.828)
	(T, S)	**1.127 (0.699)**	1.174 (0.893)	1.156 (0.677)	1.342 (0.986)	1.162 (0.887)

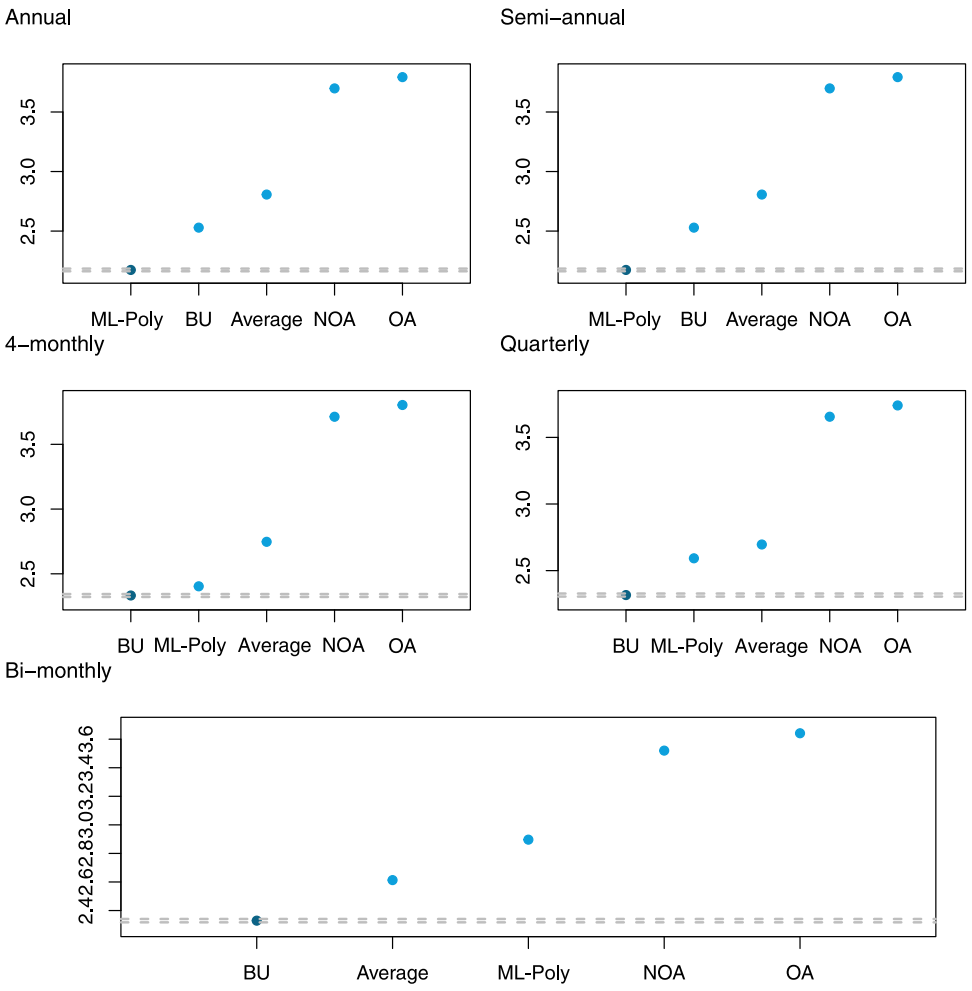


Fig. B.13. MCB test with ARIMA, monthly series.

our proposed combination approach provides better forecasting performance than benchmarks. The gain in forecast accuracy improvement for the combination approach increases for longer lead-times. For shorter lead-times, the performance of BU and combination approach is very similar. The results for daily series seem to be inconclusive, which might be due to the number of time series with systematic information and the fact that ARIMA and ETS might not be suitable for daily series. We note that although the proposed combination approach can improve forecast accuracy, on average, across multiple series, but the BU is still the best alternative for most series individually. This is especially true for daily and quarterly time series.

What is very important about the adopted approach is that it can be applied in parallel with other forecasting combinations. It can also be used to create aggregate lead time forecasts for hierarchical forecasting applications. Further research is needed to quantify such benefits when our method is applied in parallel with others, and importantly, whether forecasts other than that of the lead-time (in effect, whether the disaggregate forecasts also benefit).

Given the lack of comprehensive rules for areas of comparative outperformance of BU, NOA and OA when forecasts are required over a lead-time period, we recommend simulation-experiment comparisons considering both forecast accuracy and utility performance metrics (cost and/or customer service level).

Given the findings of this study, research into any of the following areas of temporal aggregation would prove to be useful:

- The proposed framework could be replicated with intermittent time series. We believe that the combination approach can result in better performance in this context as well;
- As demonstrated in this research and based on the literature, aggregating time series and the corresponding forecasts through temporal aggregation may lead to forecast improvements, but the conditions for this improvement remain unclear. Any research that can shed light on the association between time series features and the performance of these approaches is welcomed;
- More empirical investigations are required to examine the performance of temporal aggregation with daily and sub-daily time series;
- The idea of using available information at various levels of temporal aggregation to improve forecasting performance is promising, and there have been multiple theoretical developments in this area. The current work can be extended to cover approaches such as MAPA and temporal hierarchies (can be applied in parallel to any other approaches), and potentially provide further accuracy improvements to those methods;
- While this study focused on lead-time forecasting, a further investigation is needed to evaluate the performance of the proposed forecast combination approach when producing forecasts at the original higher frequency.

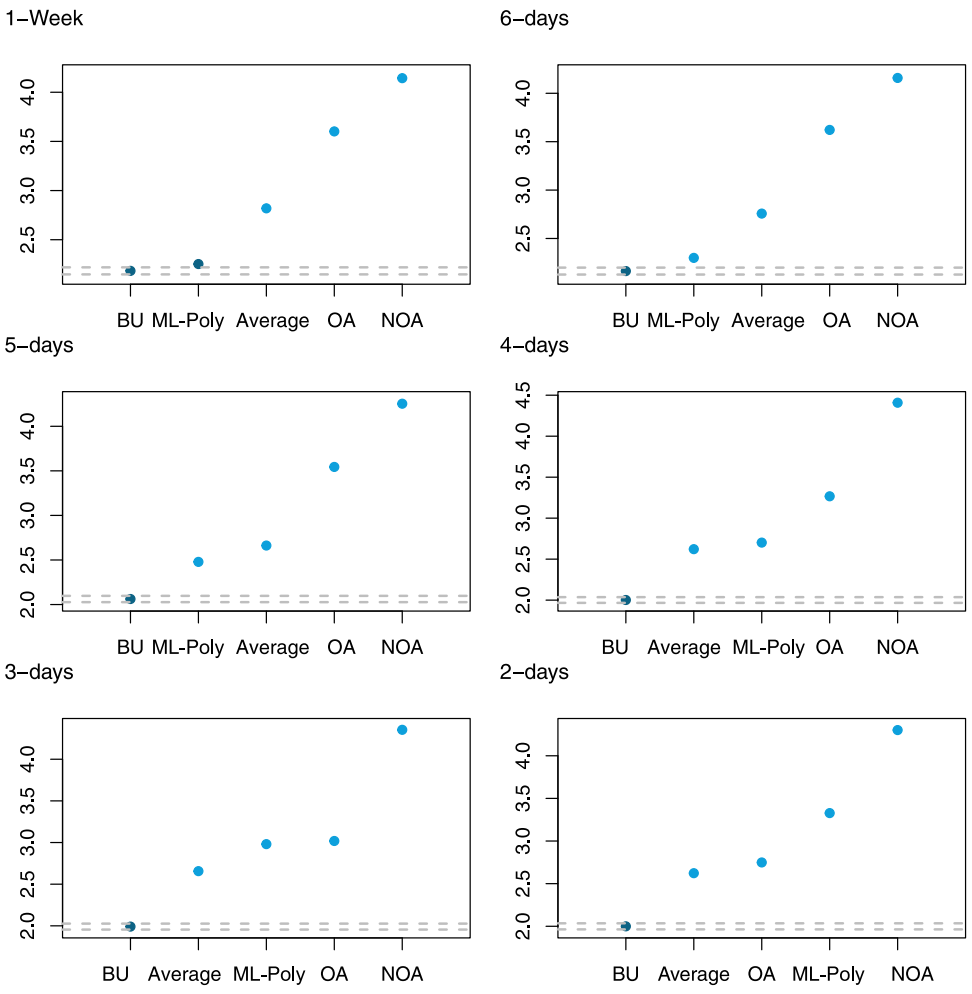


Fig. B.14. MCB test with ARIMA, daily series.

CRediT authorship contribution statement

Bahman Rostami-Tabar: Conceptualization, Programming, Formal analysis, Model development, Literature review, Writing – original draft. **Thanos E. Goltsos:** Conceptualization, Model development, Literature review, Writing – original draft. **Shixuan Wang:** Conceptualization, Programming, Model development, Writing – original draft.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was undertaken using the supercomputing facilities at Cardiff University operated by Advanced Research Computing at Cardiff (ARCCA) on behalf of the Cardiff Supercomputing Facility and the HPC Wales and Supercomputing Wales (SCW) projects. We acknowledge the support of the latter, which is part-funded by the European Regional Development Fund (ERDF) via the Welsh Government.

Appendix A. Performance evaluation for arima method

See Tables A.5–A.7.

Appendix B. Significant test results for ARIMA method

See Figs. B.12–B.14.

References

Andrawis, R.R., Atiya, A.F., El-Shishiny, H., 2011. Combination of long term and short term forecasts, with application to tourism demand forecasting. Int. J. Forecast. 27 (3), 870–886.
Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: a decomposition approach to forecasting. Int. J. Forecast. 16 (4), 521–530.
Athanasopoulos, G., Hyndman, R.J., Kourentzes, N., Petropoulos, F., 2017. Forecasting with temporal hierarchies. European J. Oper. Res. 262 (1), 60–74.
Athanasopoulos, G., Hyndman, R.J., Song, H., Wu, D.C., 2011. The tourism forecasting competition. Int. J. Forecast. 27 (3), 822–844.
Babai, M.Z., Ali, M.M., Nikolopoulos, K., 2012. Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis. Omega 40 (6), 713–721.
Babai, M.Z., Boylan, J.E., Rostami-Tabar, B., 2021. Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. Int. J. Prod. Res. 1–25.
Barrow, D.K., Kourentzes, N., 2016. Distributions of forecasting errors of forecast combinations: implications for inventory management. Int. J. Prod. Econ. 177, 24–33.
Blanc, S.M., Setzer, T., 2016. When to choose the simple average in forecast combination. J. Bus. Res. 69 (10), 3951–3962.
Boylan, J.E., Babai, M.Z., 2016. On the performance of overlapping and non-overlapping temporal demand aggregation approaches. Int. J. Prod. Econ. 181, 136–144.
Boylan, J.E., Syntetos, A.A., 2021. Intermittent Demand Forecasting: Context, Methods and Applications. John Wiley & Sons.
Cesa-Bianchi, N., Lugosi, G., 2003. Potential-based algorithms in on-line prediction and game theory. Mach. Learn. 51 (3), 239–261.
Clemen, R.T., 1989. Combining forecasts: A review and annotated bibliography. Int. J. Forecast. 5 (4), 559–583.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.
Dunn, D.M., Williams, W.H., DeChaine, T., 1976. Aggregate versus subaggregate models in local area forecasting. J. Amer. Statist. Assoc. 71 (353), 68–71.
Gaillard, P., Stoltz, G., Van Erven, T., 2014. A second-order bound with excess losses. In: Conference on Learning Theory. pp. 176–196.
Gardner, Jr., E.S., 2006. Exponential smoothing: The state of the art—Part II. Int. J. Forecast. 22 (4), 637–666.
Goodwin, P., 2018. Profit from Your Forecasting Software: A Best Practice Guide for Sales Forecasters. John Wiley & Sons.
He, C., Xu, X., 2005. Combination of forecasts using self-organizing algorithms. J. Forecast. 24 (4), 269–278.
Hibon, M., Evgeniou, T., 2005. To combine or not to combine: selecting among forecasts and their combinations. Int. J. Forecast. 21 (1), 15–24.
Hollyman, R., Petropoulos, F., Tipping, M.E., 2021. Understanding forecast reconciliation. European J. Oper. Res. 294 (1), 149–160.
Hyndman, R.J., Athanasopoulos, G., 2021. Forecasting: Principles and Practice. OTexts, URL https://otexts.com/fpp3.
Hyndman, R., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media.
Jaganathan, S., Prakash, P., 2020. A combination-based forecasting method for the M4-competition. Int. J. Forecast. 36 (1), 98–104.
Jose, V.R.R., Winkler, R.L., 2008. Simple robust averages of forecasts: Some empirical results. Int. J. Forecast. 24 (1), 163–169.
Kolassa, S., 2011. Combining exponential smoothing forecasts using Akaike weights. Int. J. Forecast. 27 (2), 238–251.
Kolassa, S., 2016. Evaluating predictive count data distributions in retail sales forecasting. Int. J. Forecast. 32 (3), 788–803.
Kourentzes, N., Petropoulos, F., 2016. Forecasting with multivariate temporal aggregation: The case of promotional modelling. Int. J. Prod. Econ. 181, 145–153.
Kourentzes, N., Petropoulos, F., Trapero, J.R., 2014. Improving forecasting by estimating time series structural components across multiple frequencies. Int. J. Forecast. 30 (2), 291–302.
Kourentzes, N., Rostami-Tabar, B., Barrow, D.K., 2017. Demand forecasting by temporal aggregation: using optimal or multiple aggregation levels? J. Bus. Res. 78, 1–9.
Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. The M4 competition: Results, findings, conclusion and way forward. Int. J. Forecast. 34 (4), 802–808.
Mohammadipour, M., Boylan, J.E., 2012. Forecast horizon aggregation in integer autoregressive moving average (INARMA) models. Omega 40 (6), 703–712.
Montero-Manso, P., Netto, C., Talagala, T., 2018. M4comp2018: Data from the M4-competition. R Package Version 0.1. 0.
Morariu, C., Morariu, O., Răileanu, S., Borangiu, T., 2020. Machine learning for predictive scheduling and resource allocation in large scale manufacturing systems. Comput. Ind. 120, 103244.
Nikolopoulos, K., Syntetos, A.A., Boylan, J.E., Petropoulos, F., Assimakopoulos, V., 2011. An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: an empirical proposition and analysis. J. Oper. Res. Soc. 62 (3), 544–554.
O'Hara-Wild, M., Hyndman, R., Wang, E., Caceres, G., 2020. Fable: Forecasting models for tidy time series. R package version 0.2.1. URL https://CRAN.R-project.org/package=fable.
Orcutt, G.H., Watts, H.W., Edwards, J.B., 1968. Data aggregation and information loss. Amer. Econ. Rev. 58 (4), 773–787.
Petropoulos, F., Apiletti, D., Assimakopoulos, V., Babai, M.Z., Barrow, D.K., Taieb, S.B., Bergmeir, C., Bessa, R.J., Bijak, J., Boylan, J.E., et al., 2022. Forecasting: theory and practice. Int. J. Forecast..
Petropoulos, F., Kourentzes, N., 2014. Forecast combinations for intermittent demand. J. Oper. Res. Soc. 66 (6), 914–924.
Petropoulos, F., Kourentzes, N., 2015. Forecast combinations for intermittent demand. J. Oper. Res. Soc. 66 (6), 914–924.
Porras, E., Dekker, R., 2008. An inventory control system for spare parts at a refinery: An empirical comparison of different re-order point methods. European J. Oper. Res. 184 (1), 101–132.
Rostami-Tabar, B., Babai, M.Z., Syntetos, A., 2022. To aggregate or not to aggregate: Forecasting of finite autocorrelated demand. J. Oper. Res. Soc. 1–20.
Rostami-Tabar, B., Babai, M.Z., Syntetos, A., Ducq, Y., 2013. Demand forecasting by temporal aggregation. Nav. Res. Logist. 60 (6), 479–498.
Rostami-Tabar, B., Babai, M.Z., Syntetos, A., Ducq, Y., 2014. A note on the forecast performance of temporal aggregation. Nav. Res. Logist. 61 (7), 489–500.
Shlifer, E., Wolff, R.W., 1979. Aggregation and proration in forecasting. Manage. Sci. 25 (6), 594–603.
Singh, P., Huang, Y.-P., 2019. A new hybrid time series forecasting model based on the neutrosophic set and quantum optimization algorithm. Comput. Ind. 111, 121–139.
Willemain, T., Smart, C., Schwarz, H., 2004. A new approach of forecasting intermittent demand for service parts inventories. Int. J. Forecast. 20, 375–387.
Willemain, T.R., Smart, C.N., Shockor, J.H., DeSautels, P.A., 1994. Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. Int. J. Forecast. 10 (4), 529–538.
Zotteri, G., Kalchschmidt, M., 2007. A model for selecting the appropriate level of aggregation in forecasting processes. Int. J. Prod. Econ. 108 (1–2), 74–83.