Déjà vu: A data-centric forecasting approach through time series cross-similarity

Yanfei Kang[a], Evangelos Spiliotis[b], Fotios Petropoulos[c], Nikolaos Athiniotis[b], Feng Li[d,*], Vassilios Assimakopoulos[b]

[a]*School of Economics and Management, Beihang University, Beijing, China*
[b]*Forecasting and Strategy Unit, School of Electrical and Computer Engineering, National Technical University of Athens, Greece*
[c]*School of Management, University of Bath, UK*
[d]*School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China*

Abstract

Accurate forecasts are vital for supporting the decisions of modern companies. Forecasters typically select the most appropriate statistical model for each time series. However, statistical models usually presume some data generation process while making strong assumptions about the errors. In this paper, we present a novel data-centric approach — 'forecasting with similarity', which tackles model uncertainty in a model-free manner. Existing similarity-based methods focus on identifying similar patterns within the series, i.e., 'self-similarity'. In contrast, we propose searching for similar patterns from a reference set, i.e., 'cross-similarity'. Instead of extrapolating, the future paths of the similar series are aggregated to obtain the forecasts of the target series. Building on the cross-learning concept, our approach allows the application of similarity-based forecasting on series with limited lengths. We evaluate the approach using a rich collection of real data and show that it yields competitive accuracy in both points forecasts and prediction intervals.

Keywords: Forecasting, Dynamic Time Warping, M Competitions, Time Series Similarity, Empirical Evaluation

*Correspondance: Feng Li, School of Statistics and Mathematics, Central University of Finance and Economics, Shahe Higher Education Park, Changping District, Beijing 102206, China.
Email address: feng.li@cufe.edu.cn (Feng Li)

1. Introduction

Effective forecasting is crucial for various functions of modern companies. Forecasts are used to make decisions concerning business operations, finance, strategy, planning, and scheduling, among others. Despite its importance, forecasting is not a straightforward task. The inherent uncertainty renders the provision of perfect forecasts impossible. Nevertheless, reducing the forecast error as much as possible is expected to bring significant monetary savings.

We identify the search for an "optimal" model as the main challenge to forecasting. Existing statistical forecasting models implicitly assume an underlying data generating process (DGP) coupled with distributional assumptions of the forecast errors that do not essentially hold in practice. Petropoulos et al. (2018a) suggest that three sources of uncertainty exist in forecasting: model, parameter, and data. They found that merely tackling the model uncertainty is sufficient to bring most of the performance benefits. This result reconfirms George Box's famous quote, "all models are wrong, but some are useful." It is not surprising that researchers increasingly avoid using a single model, and opt for combinations of forecasts from multiple models (Jose & Winkler, 2008; Kolassa, 2011; Blanc & Setzer, 2016; Bergmeir et al., 2016; Petropoulos et al., 2018b; Montero-Manso et al., 2020). We argue that there is another way to avoid selecting a single model: to select no models at all.

This study provides a new way to forecasting that does not require the estimation of any forecasting models, while also exploiting the benefits of cross-learning (Makridakis et al., 2020). With our proposed approach, a target series is compared against a set of reference series attempting to identify similar ones (déjà vu). The point forecasts for the target series are the average of the future paths of the most similar reference series. The prediction intervals are based on the distribution of the reference series, calibrated for low sampling variability. Note that no model extrapolations take place in our approach. The proposed approach has several advantages compared to existing methods, namely (*i*) it tackles both model and parameter uncertainties, (*ii*) it does not use time series features or other statistics as a proxy for determining similarity, and (*iii*) no explicit assumptions are made about the DGP as well as the distribution of the forecast errors.

We evaluate the proposed forecasting approach using the M1 and M3 competition data (Makridakis et al., 1982; Makridakis & Hibon, 2000). Our approach results in good point forecast accuracy, which is on par with state-of-the-art statistical benchmarks, while a simple combination of our data-centric approach and exponential smoothing significantly outperforms all other approaches tested. Also, forecasting with cross-similarity offers a better estimation of forecast uncertainty, which would allow achieving higher customer service levels.

The rest of the paper is organized as follows. In the next section, we present an overview of the existing literature and provide our motivation behind "forecasting with cross-similarity", . Section 3 describes the methodology for the proposed forecasting approach, while section 4 presents the experimental design and the results. Section 5 offers our discussions and insights, as well as implications for research and practice. Finally, section 6 provides our concluding remarks.

2. Background research

2.1. Forecast model selection

When forecasting with numerous time series, forecasters typically try to enhance forecasting accuracy by selecting the most appropriate model from a set of alternatives. The solution might involve either aggregate selection, where a single model is used to extrapolate all the series, or individual selection, where the most appropriate model is used per series (Fildes, 1989). The latter approach can provide substantial improvements if forecasters are indeed in a position to select the best model (Fildes, 2001; Fildes & Petropoulos, 2015). Unfortunately, this is far from the reality due to the presence of data, model, and parameter uncertainties (Kourentzes et al., 2014; Petropoulos et al., 2018a).

In this respect, individual selection becomes a complicated problem and forecasters have to balance the potential gains in forecasting accuracy and the additional complexity introduced. Automatic forecasting algorithms test multiple forecasting models and select the 'best' based on some criterion. The criteria include information criteria, e.g., the likelihood of a model penalised by its complexity (Hyndman et al., 2002; Hyndman & Khandakar, 2008), or rules based on forecasting performance on past windows of the data (Tashman, 2000). Other approaches to model selection involve discriminant analysis (Shah, 1997), time-series features (Petropoulos et al., 2014), and expert rules (Adya et al., 2001). An interesting alternative is to apply cross-learning so that the series are clustered based on an array of features and the best model is selected for their extrapolation (Kang et al., 2017; Spiliotis et al., 2020).

In any case, the difference between two models might be small, and the selection of one over the other might be purely due to chance. The small differences between models also result in different models being selected when different criteria or cost functions are used (Billah et al., 2006). Moreover, the features and the rules considered may not be adequate for describing every possible pattern of data. As a result, in most cases, a clear-cut for the 'best' model does not exist because all models simply are rough approximations of the reality.

2.2. *The non-existence of a DGP and forecast model combination*

Time series models that are usually offered by the off-the-shelf forecasting software have over-simplified assumptions (such as the normality of the residuals and stationarity), which do not necessarily hold in practice. As a result, it is impossible for these models to capture the actual DGP of the data perfectly. One could work towards defining a complex multivariate model (Svetunkov, 2016), but this would lead to all kinds of new problems, such as data limitations and the inability to accurately forecast some of the exogenous variables, which are identified as significant.

As a solution to the above problem, forecasting researchers have been combining forecasts from different models (Bates & Granger, 1969; Clemen, 1989; Makridakis & Winkler, 1983; Timmermann, 2006; Claeskens et al., 2016; Blanc & Setzer, 2016). The main advantage of combining forecasts is that it reduces the uncertainty related to model and parameter determination, and decreases the risk of selecting a single and inadequate model. Moreover, combining different models enables capturing multiple patterns. Thus, forecast combinations lead to more accurate and robust forecasts with lower error variances (Hibon & Evgeniou, 2005).

Through the years, the forecast combination puzzle (Watson & Stock, 2004; Claeskens et al., 2016; Blanc & Setzer, 2016), i.e., the fact that optimal weights often perform poorly in applications, has been both theoretically and empirically examined. Many alternatives have been proposed to exploit the benefits of combination, including Akakie's weights (Kolassa, 2011), temporal aggregation levels (Kourentzes et al., 2014, 2017), bagging (Bergmeir et al., 2016; Petropoulos et al., 2018a), and hierarchies (Hyndman et al., 2011; Athanasopoulos et al., 2017), among others. Moreover, simple combinations have been shown to perform well in practice (Petropoulos & Svetunkov, 2020). In spite of the improved performance offered by forecast combination, some primary difficulties, e.g., (*i*) determining the pool of models being averaged, (*ii*) identifying their weights, and (*iii*) estimating multiple models, prevent forecast combination from being widely applied by practitioners.

2.3. *Forecasting with cross-similarity*

An alternative to fitting statistical models to the historical data would be exploring whether similar patterns have appeared in the past. The motivation behind this argument originates from the work on structured analogies by Green & Armstrong (2007). Structured analogies is a framework for eliciting human judgment in forecasting. Given a forecasting challenge, a panel of experts is assembled and asked to independently and anonymously provide a list of analogies that are similar to the target problem together with the degree of similarity and

their outcomes. A facilitator calculates the forecasts for the target situation by averaging the outcomes of the analogous cases weighted by the degree of their likeness.

Given the core framework of structured analogies described above, several modifications have been proposed in the literature. Such an approach is practical in cases that no historical data are available (e.g., Nikolopoulos et al., 2015), which renders the application of statistical algorithms impossible. Forecasting by analogy has also been used in tasks related to new product forecast (Goodwin et al., 2013; Wright & Stern, 2015; Hu et al., 2019), in which the demand and the life-cycle curve parameters are possible to estimate based on the historical demand values and life-cycles of similar products.

Even when historical information is available, sharing information across series has been shown to improve the forecasting performance. A series of studies attempted to estimate the seasonality on a group level instead of a series level (e.g., Mohammadipour & Boylan, 2012; Zhang et al., 2013; Boylan et al., 2014). When series are arranged in hierarchies, it is possible to have similarities in seasonal patterns among products that belong to the same category. This renders their estimation on an aggregate level more accurate, especially for the shorter series where few seasonal cycles are available.

The use of cross-sectional information for time series forecasting tasks is a feature of the two best-performing approaches by Smyl (2020) and Montero-Manso et al. (2020) in the recent M4 forecasting competition (Makridakis et al., 2020). Smyl (2020) propose a hybrid approach that combines exponential smoothing with neural networks. The hierarchical estimation of the parameters utilises learning across series but also focuses on the idiosyncrasies of each series. Montero-Manso et al. (2020) use cross-learning based on the similarity of the features in collections of series to estimate the combination weights assigned to a pool of forecasting methods.

A stream of research has focused on similarity-based approaches to forecasting. Similarity-based is based on an assumption nicely articulated by Dudek (2010): "If the process pattern x_α in a period preceding the forecast moment is similar to the pattern x_b from the history of this process, then the forecast pattern y_α is similar to the forecast pattern y_b." In other words, Dudek (2010) suggested that similar patterns may exist within the same signal (process), i.e., the same time series. He applied similarity-based approaches for short-term load forecasting, and empirically demonstrated the usefulness of this approach (Dudek, 2010, 2015a). Also, Dudek (2015b) discussed that similar patterns may be identified via a variety of methods (such as the kernel and nearest neighbour methods). Finally, he discussed that the advantages of forecasting by similarity include simplicity, ease of estimation and calculation, and ability to

deal with missing data.

Along the same lines, Nikolopoulos et al. (2016) explored the value of identifying similar patterns within a series of intermittent nature (where the demand for some periods is zero). They proposed an approach that uses nearest neighbours to predict incomplete series of consecutive periods with non-zero demand values based on past occurrences of non-zero demands. Martínez et al. (2019b) and Martínez et al. (2019a) also used k-nearest neighbours to find similar patterns in fast-moving series and use them for extrapolation. They also suggested the use of multiple k values through ensembles to tackle the need of selecting a single k parameter in the nearest neighbours method.

Li et al. (2019) focused on fast-moving data in the context of maritime, and suggested that the time series is decomposed in low and high frequency components. Subsequently, they suggested similarity grouping of overlapping segments of the high frequency component towards producing its prediction with neural networks. Li et al. (2019) used dynamic time warping (DTW) to measure similarity. DTW is an algorithm for identifying alternative alignments between the points of two series, so that their total distance is minimized. Indeed, Li et al. (2020) showed that DTW is superior to Euclidean distance in classifying and clustering time series, and it could be further improved by considering adaptive constrain. In any case, similar to the previous studies by Dudek, Nikolopoulos and Martínez, the approach by Li et al. (2019) focused on self-similarities: similarities in the patterns within a time series.

We are proposing a novel approach to forecasting that builds on existing approaches on forecasting with cross-similarity, but also extends them in the sense that we suggest that searching for similar patterns can be expanded from within-series to across-series. While cross-learning information has been used in forecasting previously, to the best of our knowledge, it has not been utilised for directly judging the similarity of different series, without the need to extract and estimate time series features. Directly looking for similar observed patterns in the historical information of other series might be particularly relevant in sets of data where appropriate clusters (subsets) are characterised by homogeneity, which could be the case in the sales or demand patterns of a distinct category of products observed by a retailer. Cross-series similarity is also appealing for the cases of short series, where the limited historical information does not allow for learning through self-similarities. In any case, in searching for similarity, it might be useful to consider a decomposition of low and high frequency components, as suggested by Li et al. (2019).

3. Methodology

Given a set with rich and diverse reference series, the objective of forecasting with cross-similarity is to find the most similar ones to a target series, average their future paths, and use this average as the forecasts for the target series. We assume that the target series, y, has a length of n observations and a forecasting horizon of h. Series in the reference set shorter than $n+h$ are not considered. Series longer than $n+h$ are truncated, keeping the last $n+h$ values. The first n values are used for measuring similarity and the last h values serve as the future paths. We end up with a matrix Q of size $m \times (n+h)$. Each row of Q represents the $n+h$ values of a (truncated) reference series, and m is the number of the reference series. A particular reference series is denoted with $Q(i)$, where $i \in 1,\ldots,m$, $Q(i)_{1,\ldots,n}$ is the historical data, and $Q(i)_{n+1,\ldots,n+h}$ represents the future paths. The proposed approach consists of the following steps.

Step 1 **Removing seasonality**, if a series is identified as seasonal.

Step 2 **Smoothing** by estimating the trend component through time series decomposition.

Step 3 **Scaling** to render the target and possible similar series comparable.

Step 4 **Measuring similarity** by using a set of distance measures.

Step 5 **Forecasting** by aggregating the paths of the most similar series.

Step 6 **Inverse scaling** to bring the forecasts for the target series back to its original scale.

Step 7 **Recovering seasonality**, if the target series is found seasonal in Step 1.

In the following subsections, we describe these steps in details. Section 3.1 describes the preprocessing of the data (Steps 1, 2, 3, 6, and 7), section 3.2 provides the details regarding similarity measurement and forecasting (Steps 4 and 5), while section 3.3 explains how prediction intervals are derived.

3.1. Preprocessing

When dealing with diverse data, preprocessing becomes essential for effectively forecasting with cross-similarity. This is because the process of identifying similar series is complicated when multiple seasonal patterns and randomness are present, and the scales of the series to be compared differ. If the reference series are not representative of the target series or the reference set is lack of diversity, the chances of observing similar patterns are further decreased.

To deal with this problem, we consider three steps which are applied sequentially. The first step removes the seasonality if the series is identified as seasonal. By doing so, the target series is more likely to effectively match with multiple reference series, at least when dissimilarities are present due to different seasonal patterns. In the second step, we smooth the seasonally adjusted series to remove randomness and possible outliers from the data, which further reduces

the risk of identifying too few similar series. Finally, we scale the target and the reference series to the same magnitude, so that their values are directly comparable. The preprocessing step is applied to both the reference and target series.

3.1.1. Seasonal adjustment

Seasonal adjustment is performed by utilizing the "Seasonal and Trend decomposition using Loess" (STL) method presented by Cleveland et al. (1990) and implemented in the *stats* package for R. In brief, STL decomposes a time series x_t into the trend (T), seasonal (S), and remainder (R) components, assuming additive interactions among them: $x_t = T_t + S_t + R_t$. An adjustment is only considered if the series is identified as seasonal, through a seasonality test. The test (Assimakopoulos & Nikolopoulos, 2000; Fiorucci et al., 2016) checks for autocorrelation significance on the s^{th} term of the autocorrelation function (ACF), where s is the frequency of the series (e.g., $s = 12$ for monthly data). Thus, given a series of $\hat{n} \geq 3s$ observations, frequency $s > 1$, and a confidence level of 90%, a seasonal adjustment is considered only if

$$|\text{ACF}_s| > 1.645\sqrt{\frac{1 + 2\sum_{i=1}^{s-1}\text{ACF}_i{}^2}{\hat{n}}},$$

where $\hat{n}$ is equal to n and $n+h$ for the target and the reference series, respectively. Non-seasonal series ($s = 1$) and series where the number of observations is fewer than three seasonal periods are not tested and not assumed as seasonal.

As some series may display multiplicative seasonality, the Box-Cox transformation (Box & Cox, 1964) is applied to each series before the STL (Bergmeir et al., 2016). The Box-Cox transformation is defined as

$$x'_t = \begin{cases} \log(x_t), & \lambda = 0, \\ (x_t^{\lambda} - 1)/\lambda, & \lambda \neq 0, \end{cases}$$

where x_t is a time series and $\lambda \in [0, 1]$ is selected using the method of Guerrero (1993), as implemented in the *forecast* package for R (Hyndman et al., 2019). After Box-Cox transformation, x'_t can be decomposed using STL method: $x'_t = T'_t + S'_t + R'_t$. To perform seasonal adjustment on the series x_t with multiplicative seasonality, we first remove the seasonal component S'_t from the Box-Cox transformed series x'_t, and denote the seasonal adjusted series as $x'_{t,\text{SA}} = T'_t + R'_t$. Then the inverse Box-Cox transformation is applied to $x'_{t,\text{SA}}$:

$$x_{t,\text{SA}} = \begin{cases} \exp(x'_{t,\text{SA}}), & \lambda = 0, \\ (\lambda x'_{t,\text{SA}} + 1)^{(1/\lambda)}, & \lambda \neq 0, \end{cases}$$

where $x_{t,\mathrm{SA}}$ is the final seasonal adjusted series of x_t.

As the forecasts produced by the seasonally adjusted series do not contain seasonal information, we need to reseasonalise them with Step 7. Moreover, since the seasonal component removed is Box-Cox transformed, the forecasts must also be transformed using the same λ calculated earlier. Having recovered the seasonality on the transformed forecasts, a final inverse transformation is applied. As in STL decomposition the seasonal component changes over time, seasonality recovery is based on the latest available seasonal cycle. For instance, if the target series is of monthly frequency, then the last twelve estimated seasonal indices are used to reseasonalise the forecasts. If the forecast horizon is longer than the seasonal cycle, then these last estimated seasonal indices are re-used as many times needed.

3.1.2. Smoothing

Smoothing is performed by utilizing the Loess method, as presented by Cleveland et al. (1992) and implemented in the *stats* package for R. In short, a local model is computed, with the fit at point t being the weighted average of the neighbourhood points and the weights being proportional to the distances observed between the neighbours and point t. Similarly to STL, Loess decomposes the series into the trend and remainder components. Thus, by using the trend component, outliers and noise are effectively removed, and it is easier to find similar series. Moreover, smoothing can help us obtain a more representative forecast origin (last historical value of the series), potentially improving forecasting accuracy (Spiliotis et al., 2019).

While we could directly use the smoothed trend component from STL in the previous step, we opt for a separate smoothing on the seasonally adjusted data (which consists of the trend and remainder components from STL). The reason for this is twofold. First, while a Box-Cox transformation is necessary before deseasonalising the data, as the seasonal pattern may be multiplicative and therefore impossible to be properly handled by STL, using the Box-Cox transformed smoothed trend component from STL would not allow us to correctly identify and match different trend patterns (such as additive versus multiplicative) between the target and the reference series. So, we separately smooth the seasonally-adjusted data, after an inverse Box-Cox transformation is applied to the sum of trend and remainder components from STL. Second, keeping the Loess smoothing separate to the deseasonalisation process allows for consistency across series that are identified as seasonal or not, as well as across frequencies of data.

3.1.3. *Scaling*

Scaling refers to translating the target and the reference series at the same magnitude so that they are comparable to each other. This process can be done in various ways, such as by dividing each value of a time series by a simple summary statistic (max, min, mean, etc.), by restricting the values within a specific range (such as in $[0, 1]$), or by applying a standard score. Since the forecast origin, the last historical value of the series, is the most crucial observation in terms of forecasting, we divide each point by this specific value. A similar approach has been successfully applied by Smyl (2020). A different scaling needs to be considered to avoid divisions by zero if either the target or the reference series contain zero values. Finally, inverse scaling is applied to return to the target series's original level with Step 6 once the forecasts have been produced. This is achieved via multiplying each forecast by the origin.

3.2. *Similarity & forecasting*

One disadvantage of forecasting using a statistical model is that a DGP is explicitly assumed, although it might be difficult or even impossible to capture in practice. Notwithstanding, our proposed methodology searches in a set of reference series to identify similar patterns to those of the target series we need to forecast.

Given the preprocessed target series, $\tilde{y}$, and the m preprocessed reference series, $\tilde{Q}$, we search for similar series as follows: For each series, i, in the reference set, $\tilde{Q}(i)$, we calculate the distance between its historical values, $\tilde{Q}(i)_{1,\dots,n}$, and the ones of the target series using a distance measure. The result of this process is a vector of length m distances that correspond to pairs of the target and the reference series available.

In terms of measuring distances, we consider three alternatives. The first one is the $\mathcal{L}_1$ norm, which is equivalent to the sum of the absolute deviations between $\tilde{y}$ and $\tilde{Q}(i)_{1,\dots,n}$. The second measure is the $\mathcal{L}_2$ norm (Euclidean distance), which is equivalent to the square root of the sum of the squared deviations. The third alternative involves the utilization of the DTW. DTW can match sequences that are similar, but locally out of phase, by "stretching" and "contracting", and thus it allows non-linear mapping between two time series. That is, $\tilde{y}_t$ can be matched either with $\tilde{Q}(i)_t$, as done with $\mathcal{L}_1$ and $\mathcal{L}_2$, or with previous/following points of $\tilde{Q}(i)_t$, even if these points have been already used in other matches. The three distance

measures are formally expressed as

$$
\begin{aligned}
d_{\mathcal{L}_1}(\tilde{y}, \tilde{Q}(i)_{1,\dots,n}) &= \left\| \tilde{y}_t - \tilde{Q}(i)_t \right\|_1, \\
d_{\mathcal{L}_2}(\tilde{y}, \tilde{Q}(i)_{1,\dots,n}) &= \left\| \tilde{y}_t - \tilde{Q}(i)_t \right\|_2, \\
d_{\mathrm{DTW}}(\tilde{y}, \tilde{Q}(i)_{1,\dots,n}) &= D(n,n),
\end{aligned}
$$

where $D(n,n)$ is computed recursively as

$$
D(v,w) = |\tilde{y}_v - \tilde{Q}(i)_w| + \min \left\{ \begin{array}{l} D(\tilde{y}_v, \tilde{Q}(i)_{w-1}) \\ D(\tilde{y}_{v-1}, \tilde{Q}(i)_{w-1}) \\ D(\tilde{y}_{v-1}, \tilde{Q}(i)_w) \end{array} \right\}. \tag{1}
$$

Equation (1) returns the total variation of two vectors, $\tilde{y}_{1,\dots,v}$ and $\tilde{Q}(i)_{1,\dots,w}$. Note that DTW assumes a mapping path from $(1,1)$ to (n,n) with an initial condition of $D(1,1) = |\tilde{y}_1 - \tilde{Q}(i)_1|$.

The main differences among the three distance measures are: (1) DTW allows distortion in the time axis, while $\mathcal{L}_1$ and $\mathcal{L}_2$ distances are more sensitive to time distortion. Therefore, DTW introduces more flexibility to the process, allowing the identification of similar series even when they display signal transformations such as shifting and scaling, (2) allowing many-to-one point comparisons, DTW is more robust to outliers or noise, (3) DTW can compare time series with different lengths, while the other two measures are only applicable to time series with the same length, and (4) although DTW is frequently chosen as the distance measure for time series related tasks such as clustering and classification for its aforementioned merits, when dealing with large datasets, DTW does not scale very well due to its quadratic time complexity. In contrast, $\mathcal{L}_1$ and $\mathcal{L}_2$ distance measures are much easier to implement with higher computational efficiency, making them also frequently used in a vast of time series applications. We present the empirical differences among the three measures in the proposed forecasting with cross-similarity approach in Section 4.2.

Having computed the distances between $\tilde{y}$ and $\tilde{Q}$, a subset of reference series is chosen for aggregating their future paths and, therefore, forecasting the target series. This is done by selecting the k most similar series, i.e., the series that display the smaller distances, as determined by the selected measure. In our experiment, we consider different k values to investigate the effect of pool size on forecasting accuracy but demonstrate that any value higher than 100 is a suggested choice.

Essentially, we propose that the future paths from the most similar series can form the basis for calculating the forecasts for the target series. Indeed, we do so by considering the

statistical aggregation of these future paths. The average is calculated for each planning horizon. This is an appealing approach in the sense that it does not involve statistical forecasting in the traditional way: fitting statistical models and extrapolating patterns. Instead, the real outcomes of a set of similar series are used to derive the forecasts. We tested three averaging operators: the arithmetic mean, the median, and the weighted mean[1]. The median operator gave slightly better results than the two other operators, possibly due to its robustness and resistance to outliers. So, our empirical evaluation in section 4 focuses on this operator.

The proposed forecasting approach is demonstrated via a toy example, visualized in Figure 1. The top panel presents the original target series, as well as the seasonally adjusted and smoothed one. The middle panel shows the preprocessed series (scaled values) together with the 100 most similar reference series used for extrapolation. Finally, the bottom panel compares the rescaled and reseasonalised forecasts to the actual future values of the target series.

Note that the above description assumes that Step 5 (forecasting and aggregation) is completed before inverse scaling (Step 6) and recovering of seasonality (Step 7). Equally, one could consider that Steps 6 and 7 are applied to each of the most similar reference series, providing this way k possible paths on the scale of the target series and including the seasonal pattern identified in section 3.1.1. We denote these rescaled and reseasonalised reference series as $\check{Q}_t$. The aggregation of these series would lead to the same point forecasts. Additionally, they can be used as the basis for estimating the forecast uncertainty.

3.3. Similarity & prediction intervals

Time series forecasting uncertainty is usually quantified by prediction intervals, which somehow depend on the forecastability of the target time series. With a model-based forecasting approach, although one could usually obtain a theoretical prediction interval, the performance of such interval depends upon the length of the series, the accuracy of the model, and the variability of model parameters. Alternatively, a straightforward attempt would be bootstrapping the historical time series candidates and calculating the prediction intervals based on their summary statistics (e.g., Thombs & Schucany, 1990; Andrees et al., 2002). Such a procedure is model-dependent, which assumes that a known model provides a promise fit to the data and requires specifying the distribution of the error sequence associated with the model process.

[1] The weighted mean is based on the degree of similarity: the values of the distances of the most similar series to the target series

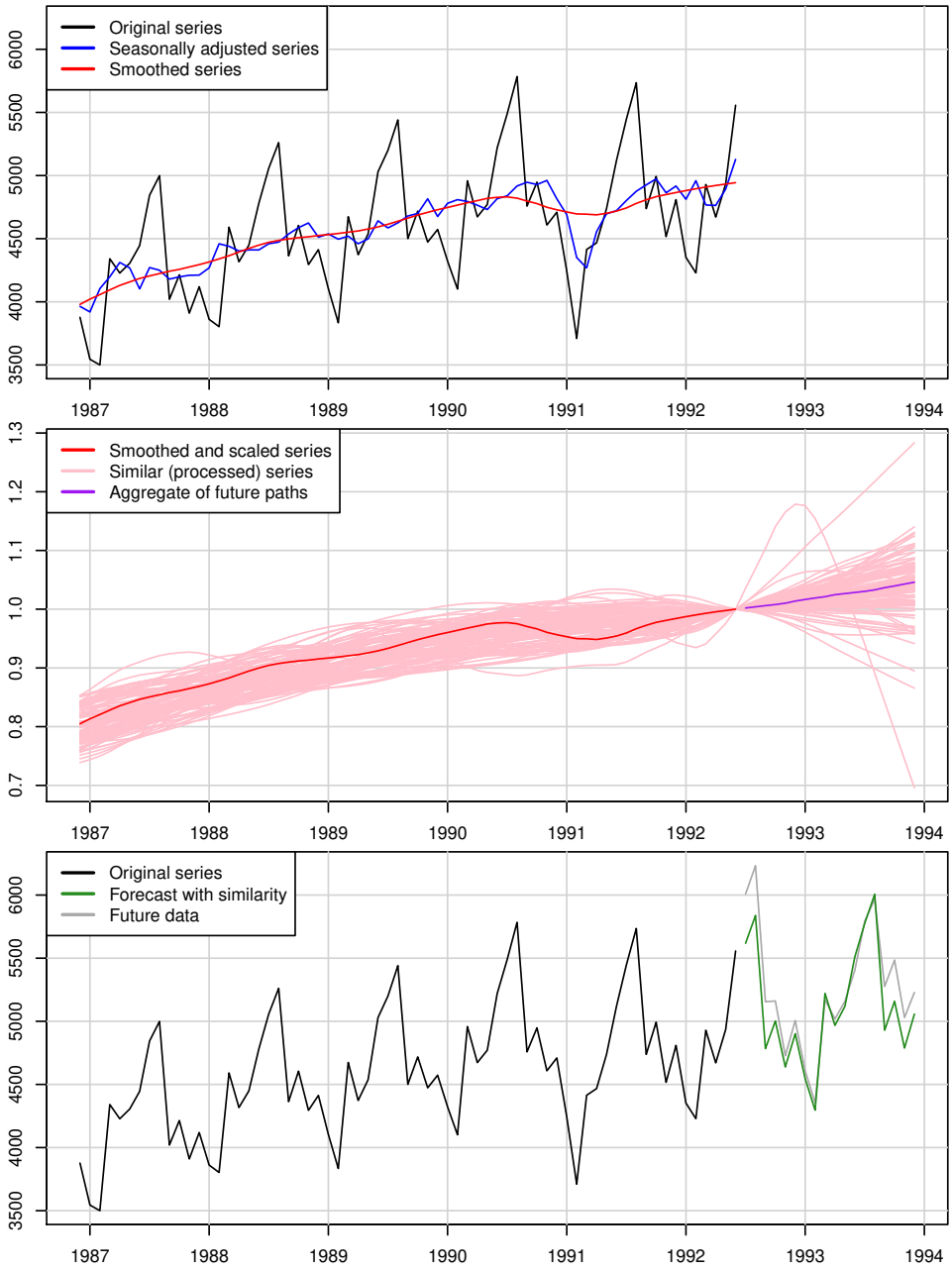

Figure 1: A toy example visualizing the methodology proposed for forecasting with cross-similarity. First, the target series is seasonally adjusted and smoothed (top panel). Then, the series is scaled, and similar reference series are used to determine its future path through aggregation (middle panel). Finally, the computed forecast is rescaled and reseasonalised to obtain the final forecast. The M495 series of the M3 Competition data set is used as the target series. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

Our interest is to find appropriate prediction intervals so that they could quantify the uncertainty of the forecasts based on our similarity approach. We use the variability information from the rescaled and reseasonalised reference series, $\check{Q}_t$, as the source of prediction interval bounds. However, we find that directly using the quantiles or variance of the reference series may lead to lower-than-nominal coverage due to the similarity (or low sampling variability) of reference series. To this end, we propose a straightforward data-driven approach, in which the $(1-\alpha)100\%$ prediction interval for a forecast f_t is based on the a calibrated $\alpha/2$ and $1-\alpha/2$ quantiles of the selected reference series $\check{Q}_t$ for the target y_t. The lower and upper bounds for the prediction interval are defined as

$$L_t = (1-\delta)\, F^{-1}_{\check{Q}_t}(\alpha/2) \text{ and } U_t = (1+\delta)\, F^{-1}_{\check{Q}_t}(1-\alpha/2), \tag{2}$$

respectively, where $F^{-1}_{\check{Q}_t}$ is the quantile based on the selected reference series $\check{Q}_t$, and δ is a calibrating factor.

To evaluate the performance of the generated predictive intervals, we consider a scoring rule, the mean scaled interval score (MSIS), which is defined as

$$\text{MSIS} = \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} (U_t - L_t) + \frac{2}{\alpha}(L_t - y_t)\mathbf{1}\{y_t < L_t\} + \frac{2}{\alpha}(y_t - U_t)\mathbf{1}\{y_t > U_t\}}{\frac{1}{n-s}\sum_{t=s+1}^{n} |y_t - y_{t-s}|}, \tag{3}$$

where n is the historical length of the target time series, s is the length of the seasonal period, and h is the forecasting horizon. We aim to find an optimal calibrating factor $0 \leq \delta \leq 1$, which minimizes the prediction uncertainty score (MSIS). To realize that, the target series y is first split into training and testing period, denoted as $y_{1,\dots,n-h}$ and $y_{n-h+1,\dots,n}$, respectively. We run the proposed forecasting approach to $y_{1,\dots,n-h}$ and apply a grid search algorithm to search from a sequence of values of $\delta \in \{0, 0.01, 0.02, \cdots, 1\}$ and find the optimal calibrating factor δ^* that minimizes the MSIS values of the obtained prediction intervals of $y_{1,\dots,n-h}$. In the end, we get the prediction interval of y by plugging the optimal calibrating factor δ^* into Equation (2).

4. Evaluation

4.1. Design

In this paper, we aim to forecast the yearly, quarterly, and monthly series of the M1 (Makridakis et al., 1982) and M3 (Makridakis & Hibon, 2000) forecasting competitions. These data sets have been widely used in the forecasting literature with the corresponding research paper having been cited more than 1500 and 1600 times, respectively, according to Google Scholar (as of 08/20/2020). The number of the yearly, quarterly, and monthly series is presented in

Table 1: The number of the target series, their lengths, and the forecasting horizon for each data frequency.

Frequency	Number of series	Historical observations					h
		Min	Q1	Q2	Q3	Max	
Yearly	826	9	14	17.5	26	52	6
Quarterly	959	10	36	44	44	106	8
Monthly	2045	30	54	108	116	132	18
Total	3830						

Table 2: The cuts of the target series considered.

Frequency	Up to (in years)							
Yearly	6	10	14	18	22	26	30	34
Quarterly	3	4	5	6	7	8	9	10
Monthly	3	4	5	6	7	8	9	10

Table 1, together with a five-number summary of their lengths and the forecast horizon per frequency.

To assess the impact of the series length, we produce forecasts not only using all the available history for each target series but also considering shorter historical samples by truncating the long series and keeping the last few years of their history. This is of particular interest in forecasting practice as in many enterprise resource planning systems, such as SAP, only a limited number of years is usually available. Table 2 shows the cuts considered per frequency.

For the purpose of forecasting based on similarity described in the previous section, we need a rich and diverse enough set of reference series. For this purpose, we use the yearly, quarterly, and monthly subsets of the M4 competition (Makridakis et al., 2020), which consist of 23000, 24000, and 48000 series, respectively. The lengths of these series are, on average, higher than the lengths of the M1 and M3 competition data. The median lengths are 29, 88, and 202 for the yearly, quarterly, and monthly frequencies in M4, respectively.

The point forecast accuracy is measured in terms of the Mean Absolute Scaled Error (MASE: Hyndman & Koehler, 2006). MASE is a scaled version of the mean absolute error, with the scaling being the mean absolute error of the seasonal naive for the historical data. MASE is widely accepted in the forecasting literature (e.g., Franses, 2016). Makridakis et al. (2020) also use this measure to evaluate the point forecasts of the submitting entries for the M4 forecasting competition. Across all horizons of a single series, the MASE value can be calculated as

$$\text{MASE} = \frac{1}{h} \frac{\sum_{t=n+1}^{n+h} |y_t - f_t|}{\frac{1}{n-s}\sum_{t=s+1}^{n} |y_t - y_{t-s}|},$$

where y_t and f_t are the actual observation and the forecast for period t, n is the sample size, s is the length of the seasonal period, and h is the forecasting horizon. Lower MASE values are better. Because MASE is scale-independent, averaging across series is possible. We also have evaluated our approach using the mean absolute percentage error (MAPE). The results were consistent with the ones by MASE, and as such we do not provide the MAPE results in the manuscript for brevity.

To assess prediction intervals, we set $\alpha = 0.05$ (corresponding to 95% prediction intervals) and consider four measures — MSIS, coverage, upper coverage and spread. MSIS is calculated as in Equation 3. Coverage measures the percentage of times when the true values lie inside the prediction intervals. Upper coverage measures the percentage of times when the true values are not larger than the upper bounds of the prediction intervals: A proxy for achieved service levels. Spread refers to the mean difference of the upper and lower bounds scaled similarly to MSIS: A proxy for holding costs (Svetunkov & Petropoulos, 2018). They are calculated as

$$\text{Coverage} = \frac{1}{h}\sum_{t=n+1}^{n+h} \mathbf{1}\{y_t > L_t \;\&\; y_t < U_t\},$$

$$\text{Upper coverage} = \frac{1}{h}\sum_{t=n+1}^{n+h} \mathbf{1}\{y_t < U_t\},$$

$$\text{Spread} = \frac{\frac{1}{h}\sum_{t=n+1}^{n+h}(U_t - L_t)}{\frac{1}{n-s}\sum_{t=s+1}^{n}|y_t - y_{t-s}|},$$

where y_t, L_t and U_t are the actual observation, the lower and upper bounds of the corresponding prediction interval for period t, n is the sample size, and h is the forecasting horizon. Note that the target values for the Coverage and Upper Coverage are 95% and 97.5%, respectively. Deviation from these values suggest under- or over-coverage. Lower MSIS and Spread values are better.

4.2. Investigating the performance of forecasting with cross-similarity

In this section, we focus on the performance of forecasting with cross-similarity and explore the different settings, such as the choice of the distance measure, the pool size of similar reference series (number of aggregates, k), as well as the effect of preprocessing. Once the optimal settings are identified, in the next subsection, we compare the performance of our proposition against that of four robust benchmarks for different sizes of the historical sample.

Table 3 presents the MASE results of forecasting with cross-similarity for each data frequency separately as well as across all frequencies (Total). The summary across frequencies is a weighted average based on the series counts for each frequency. Moreover, we present

Table 3: The MASE performance of the forecasting with cross-similarity approach for different distance measures and pool sizes of similar reference series (k).

Frequency	Distance Measure	Number of aggregated reference series (k)						
		1	5	10	50	100	500	1000
Yearly	$\mathcal{L}_1$	3.375	2.936	2.884	2.801	2.784	2.785	2.798
	$\mathcal{L}_2$	3.378	2.960	2.876	2.813	2.800	2.794	2.805
	DTW	3.345	2.948	2.846	2.781	2.777	2.783	2.805
Quarterly	$\mathcal{L}_1$	1.468	1.345	1.316	1.279	1.273	1.262	1.260
	$\mathcal{L}_2$	1.488	1.335	1.305	1.278	1.273	1.261	1.261
	DTW	1.440	1.316	1.297	1.257	1.254	1.250	1.250
Monthly	$\mathcal{L}_1$	1.082	0.992	0.964	0.948	0.946	0.943	0.943
	$\mathcal{L}_2$	1.088	0.993	0.970	0.948	0.945	0.942	0.943
	DTW	1.080	0.971	0.950	0.935	0.936	0.932	0.932
Total	$\mathcal{L}_1$	1.673	1.500	1.466	1.431	1.424	1.420	1.422
	$\mathcal{L}_2$	1.682	1.503	1.465	1.433	1.427	1.421	1.424
	DTW	1.659	1.484	1.446	1.414	1.413	1.411	1.416

the results for each distance measure ($\mathcal{L}_1$, $\mathcal{L}_2$, and DTW) in rows and various values of k in columns.

A comparison across the different values for the number of reference series, k, suggests that large pools of representative series provide better performance. At the same time, the improvements seem to tapper off when $k > 100$. Based on the reference set we use in this study, we identify a sweet point at $k = 500$. The analysis presented in section 4.3 focuses on this aggregate size. In any case, we find that both the reference series's size and its similarity with the target series affect the selection of the value of k.

Table 3 also shows that $\mathcal{L}_1$ and $\mathcal{L}_2$ perform almost indistinguishable across all frequencies. DTW almost always outperforms the other two distance measures. However, the differences are small, to the degree of 10^{-2} in our study. Given that the DTW is more computationally intensive than $\mathcal{L}_1$ and $\mathcal{L}_2$ (approximately ×6, ×10, and ×27 for yearly, quarterly, and monthly frequencies, respectively), we further investigate the statistical significance of the achieved performance improvements. To this end, we apply the Multiple Comparisons with the Best (MCB) test that compares whether the average (across series) ranking of each distance measure is significantly different than the others (for more details on the MCB, please see Koning et al. (2005)). With MCB, when the confidence intervals of two methods overlap, their ranked performances are not statistically different. The analysis is done for $k = 500$. The results are presented in Figure 2. We observe that DTW results in the best-ranked performance, which is statistically different from that of the other two distance measures only for the monthly frequency. We argue that if the computational cost is a concern, one may choose between $\mathcal{L}_1$ and

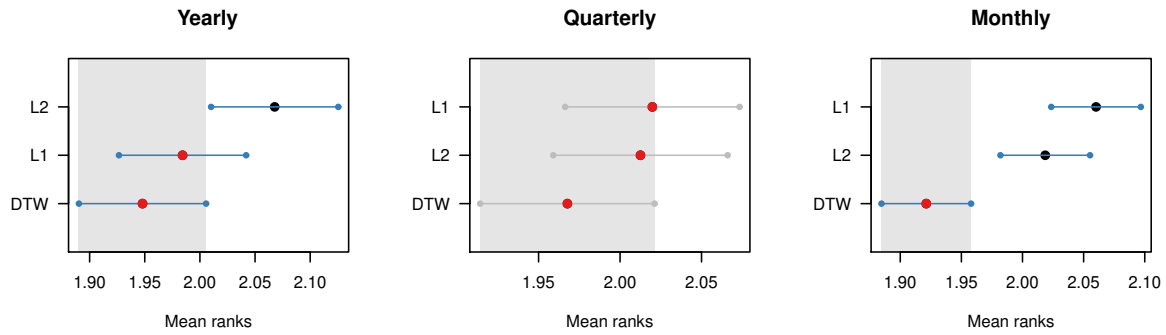

Figure 2: MCB significance tests for the three distance measures for each data frequency.

$\mathcal{L}_2$. Otherwise, DTW is preferable, both in terms of average forecast accuracy and mean ranks. In the analysis below, we focus on the DTW distance measure.

The aforementioned results are based on the application of preprocessing (as described in section 3.1), including seasonal adjustment and smoothing, before searching for similar series. Now we investigate the improvements in seasonal adjustment and smoothing. In the Loess method used for smoothing, the parameter "span" controls the degree of smoothing, which is set to h in Table 3. To investigate how the degree of smoothing in the Loess method influences the accuracies of forecasting with cross-similarity, we consider 30% less and 30% more smoothing. Note that the scaling process (as described in section 3.1.3) is always applied to make the target and reference series comparable. Table 4 presents the MASE results for DTW across different k values with and without the preprocessing described in sections 3.1.1 and 3.1.2, and with different amounts of smoothing. The main findings are: (1) preprocessing always provides better accuracy, so it is recommended with the forecasting with cross-similarity approach, and (2) for yearly and quarterly data, which are usually smooth and relatively short, less smoothing is preferred, while monthly data prefer more smoothing. Therefore, we use 30% less smoothing for yearly and quarterly data, and 30% more smoothing for monthly data in the following sections of the manuscript. Note that the percentage "30%" are arbitrarily selected here to demonstrate that smoothing improves forecasting if properly applied, and that other parameters could be used instead, possibly leading to even better results.

4.3. Similarity versus model-based forecasts

Having identified the optimal settings (DTW, $k = 500$, and preprocessing) for forecasting with cross-similarity, abbreviated from now on simply as 'Similarity', in this subsection we turn our attention to comparing the accuracy of our approach against well-known forecasting benchmarks. We use four benchmark methods. The forecasts with the first method derive from the optimally selected exponential smoothing model when applying selection with the

Table 4: The MASE performance of forecasting with cross-similarity, with and without seasonal adjustment and smoothing. The DTW distance measure is considered. "span" controls the degree of smoothing in the Loess method. A larger value of "span" means more smoothing. h is the forecasting horizon.

Frequency	Seasonal adjustment and smoothing	Number of aggregated reference series (k)						
		1	5	10	50	100	500	1000
Yearly	NO	3.649	2.967	2.898	2.823	2.819	2.828	2.845
	YES, span = $h \times 0.7$	3.474	2.915	2.859	2.778	2.770	2.774	2.796
	YES, span = h	3.345	2.948	2.846	2.781	2.777	2.783	2.805
	YES, span = $h \times 1.3$	3.252	2.901	2.849	2.824	2.808	2.820	2.842
Quarterly	NO	1.734	1.508	1.457	1.471	1.484	1.504	1.507
	YES, span = $h \times 0.7$	1.481	1.319	1.274	1.246	1.247	1.246	1.246
	YES, span = h	1.440	1.316	1.297	1.257	1.254	1.250	1.250
	YES, span = $h \times 1.3$	1.446	1.341	1.313	1.276	1.279	1.277	1.274
Monthly	NO	1.381	1.190	1.130	1.122	1.126	1.153	1.171
	YES, span = $h \times 0.7$	1.159	1.002	0.977	0.959	0.958	0.958	0.959
	YES, span = h	1.080	0.971	0.950	0.935	0.936	0.932	0.932
	YES, span = $h \times 1.3$	1.028	0.955	0.933	0.926	0.926	0.922	0.923

corrected (for small sample sizes) Akakie's Information Criterion (AIC_c). This optimal selection occurs per series individually so that a different optimal model may be selected for different series. We use the implementation available in the *forecast* package for the R statistical software, and in particular the `ets()` function (Hyndman & Khandakar, 2008). The second benchmark is the automatically selected most appropriate autoregressive integrated moving average (ARIMA) model, using the implementation of the `auto.arima()` function (Hyndman & Khandakar, 2008). The third benchmark is the Theta method (Assimakopoulos & Nikolopoulos, 2000), which was the top performing method in the M3 forecasting competition (Makridakis & Hibon, 2000). Finally, the last benchmark is the simple (equally-weighted) combination of three exponential smoothing models: Simple Exponential Smoothing, Holt's linear trend Exponential Smoothing, and Damped trend Exponential Smoothing. This combination is applied to the seasonally adjusted data (multiplicative classical decomposition) if the data have seasonal patterns with the seasonality test described in section 3.1.1. This combination approach has been used as a benchmark in international forecasting competitions (Makridakis & Hibon, 2000; Makridakis et al., 2020) and it is usually abbreviated as SHD.

We have also tested the performance of a self-similarity approach through kNN (k-Nearest-Neighbour) for time series, implemented in the *tsfknn* R package (Martínez et al., 2019a), and found that focusing merely on similar patterns within a series results in very poor forecasting performance. More importantly, *tsfknn* is not applicable when historical information is limited. As such, we decide not to include this approach as a benchmark in our study.

Figure 3 shows the accuracy of our Similarity approach against the four benchmarks, ETS,

ARIMA, Theta and SHD. The comparison is made for various historical sample sizes to examine the effect of data availability. We observe:

- In the yearly frequency, Similarity always outperforms the four benchmarks regardless of the length of the available history. It is worth mentioning that ETS improves when not all available observations are used for model fitting (truncated target series). Using just the last 14 years of the historical samples gives the best accuracy in the yearly frequency for ETS. ARIMA, SHD, and Similarity perform better when more data are available. Theta is not affected by the length of the series for the yearly frequency.
- In the quarterly frequency, similarity performs very competitively against the statistical benchmarks when the length of the series is longer than four years. Only Theta achieves, on average, better performance than similarity. The performance of all methods is improved as the number of observations increases.
- In the monthly frequency, the performance of ETS and Similarity is indistinguishable, outperforming all other statistical benchmarks. Lengthier monthly series generally result in improved performance up to a point: if more than 7 or 8 years of data are available, then the changes in forecasting accuracy are small.

Figure 3 also shows the performance of the simple forecast combination of ETS and Similarity ("ETS-Similarity")[2], which takes their arithmetic mean as the final forecasts. The argument is that these two forecasting approaches are diverse in nature (model-based versus data-centric) but also robust when applied separately. So we expect that their combination will also perform well (Lichtendahl Jr & Winkler, 2020). We observe that this simple combination performs on par to Similarity for the yearly frequency, being much better than any other approach at the seasonal frequencies. Overall, the simple combination of ETS-Similarity is the best approach. This suggests that there are different benefits in terms of forecasting performance improvements with both model-based and data-centric approaches. Solely focusing on one or the other might not be ideal.

Finally, we compare the differences in the ranked performance of the five approaches (ETS, ARIMA, Theta, SHD, and Similarity) and the one combination (ETS-Similarity) in terms of their statistical significance (MCB). The results are presented in the nine panels of Figure 4 for each frequency (in rows) and short, medium, and long historical samples (in columns). We observe:

- Similarity is significantly better than the statistical benchmarks for the short yearly series. At the same time, similarity performs statistically similar to the best of the statistical

[2]Other simple combinations of ARIMA, Theta, and SHD with Similarity were also tested, having on average same or worse performance to the ETS-Similarity simple combination.

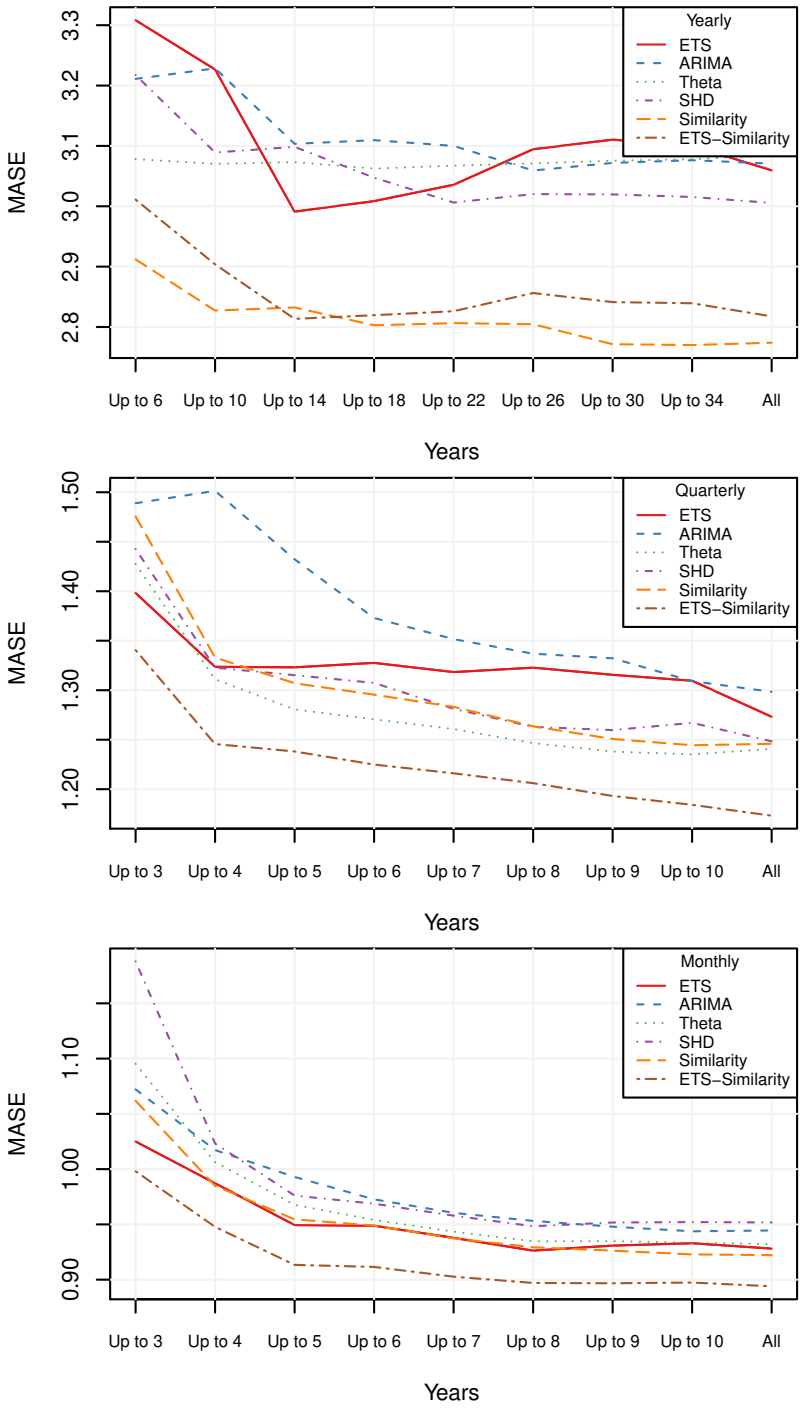

Figure 3: Benchmarking the performance of Similarity against ETS, ARIMA, Theta, and SHD for various historical sample sizes. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

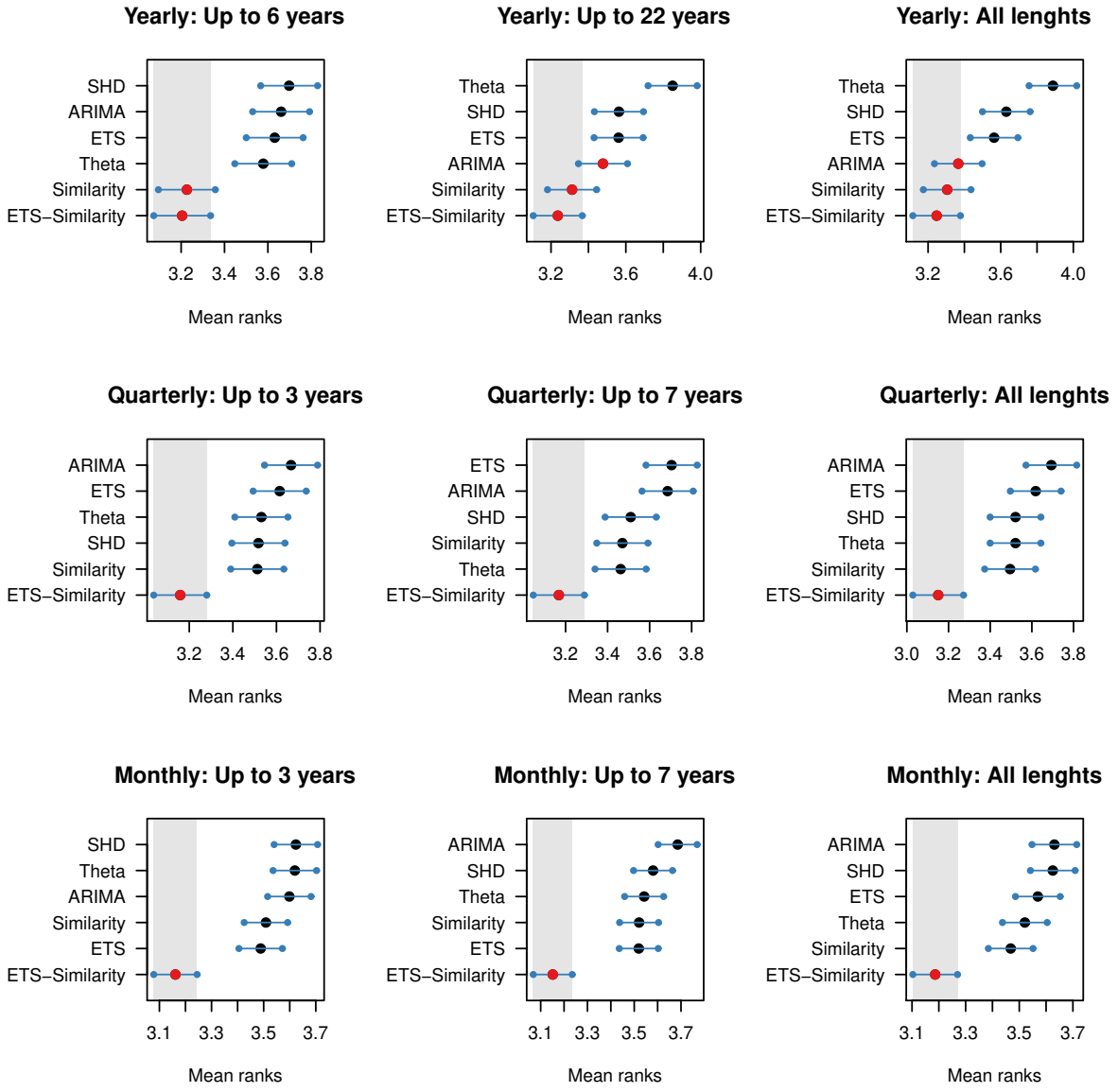

Figure 4: MCB significance tests for ETS, ARIMA, Theta, SHD, Similarity, and ETS-Similarity for each data frequency and various sample sizes.

benchmarks for other lengths and frequencies.

- A simple combination of ETS and Similarity is always ranked 1st. Moreover, its performance is significantly better compared to ETS, Theta, and SHD for all frequencies and historical sample sizes (their intervals do not overlap). ARIMA, Similarity, and ETS-Similarity are not statistically different at the yearly frequency, but the combination approach is better at the seasonal data.

4.4. Evaluating uncertainty estimation

We firstly investigate the importance of the calibrating procedure of prediction intervals by exploring the relationship between the forecastability of the target series and the selected

calibrating factor δ^*. We follow Kang et al. (2017) and use the spectral entropy to measure the "forecastability" of a time series as

$$\text{Forecastability} = 1 + \int_{-\pi}^{\pi} \hat{f}_y(\gamma) \log \hat{f}_y(\gamma) \mathrm{d}\gamma,$$

where $\hat{f}_x(\gamma)$ is an estimate of the spectrum of the time series that describes the importance of frequency γ within the period domain of a given time series y. A larger value of Forecastability suggests that the time series contains more signal and is easier to forecast. On the other hand, a smaller value of forecastability indicates more uncertainty about the future, which suggests that the time series is harder to forecast.

Figure 5 depicts the relationship between forecastability and δ^* for the studied time series by showing the scatter plots of the aforementioned variables for yearly, quarterly, and monthly data, as well as the complete dataset. The corresponding nonparametric loess regression curves are also shown. Along the top and right margins of each scatter plot, we show the histograms of forecastability and δ^* to present their distributions. From Figure 5, we find that time series with lower forecastability values yield higher calibrating factors δ^*. That is, to obtain a more appropriate prediction interval, we need to calibrate more for time series that are harder to forecast. The forecastability of a large proportion of the monthly data is weak when compared to that of the yearly and quarterly data, which makes the overall dataset hard to forecast. The nonparametric loess regression curves indicate that there is a strong dependence between forecastability and the calibration factor, which is strong evidence of elaborating a calibrating factor in the prediction intervals for hard-to-forecast time series.

We proceed by comparing the forecasting performances based on the calibrated prediction intervals of Similarity and other benchmarks. Table 5 shows the performance of Similarity against the four benchmarks, ETS, ARIMA, Theta, and SHD, regarding prediction intervals. The performance of the forecast combination of ETS and Similarity (ETS-Similarity) is also shown. Our findings are as follows:

- For yearly data, similarity significantly outperforms the four benchmarks according to MSIS, while also providing higher coverage and upper coverage. The simple combination of ETS and Similarity achieves similar performance with similarity, with higher coverage and tighter prediction intervals. Overall, we conclude that similarity significantly outperforms ETS, ARIMA, Theta, and SHD for yearly data.

- For quarterly and monthly data, similarity displays similar performance to that of ETS. However, it yields significantly higher upper coverage and at the same time loses some

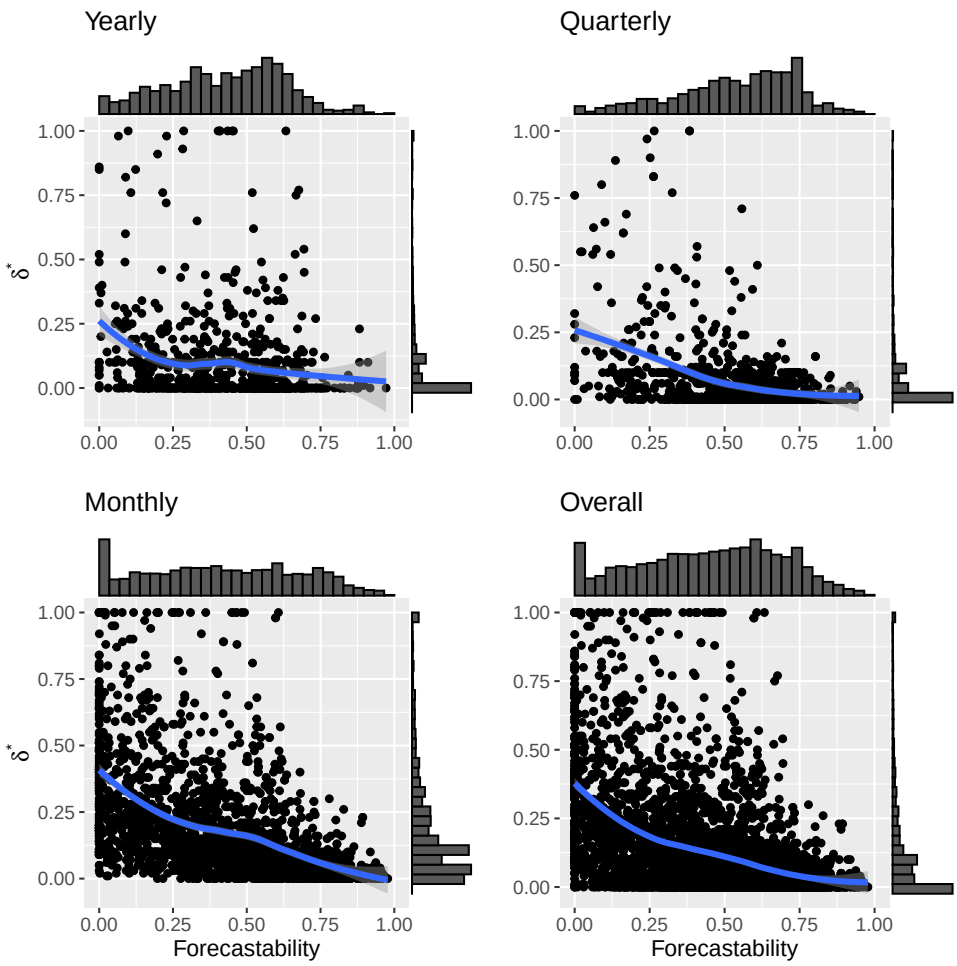

Figure 5: Relationship between forecastability and the optimal calibrating factor (δ) using a nonparametric Loess regression curve (blue line) for yearly (top left), quarterly (top right), monthly (bottom left) and overall (bottom right) data. The top and right margins of each subplot are the histograms of forecastability and the optimal calibrating factor δ^*, respectively.

Table 5: Benchmarking the performance of Similarity against ETS, ARIMA, Theta, SHD, and ETS-Similarity with regard to MSIS, coverage, upper coverage and spread of prediction intervals.

	MSIS	Coverage (%) Target: 95%	Upper coverage (%) Target: 97.5%	Spread
		Yearly		
ETS	37.008	81.578	86.844	11.967
ARIMA	45.590	77.260	86.077	**8.364**
Theta	39.568	80.851	84.705	8.871
SHD	42.424	77.220	83.051	8.506
Similarity	**26.432**	88.680	**94.592**	13.567
ETS-Similarity	26.809	**89.588**	93.119	12.767
		Quarterly		
ETS	12.961	85.076	91.489	4.805
ARIMA	14.982	80.214	91.919	**4.173**
Theta	13.785	84.541	90.667	4.309
SHD	13.409	84.333	90.302	4.391
Similarity	12.823	86.861	94.121	5.778
ETS-Similarity	**11.245**	**89.937**	**94.799**	5.292
		Monthly		
ETS	7.333	90.685	94.224	4.300
ARIMA	8.348	89.343	94.659	4.087
Theta	7.984	88.840	93.371	**4.072**
SHD	7.895	89.343	93.461	4.276
Similarity	7.643	90.497	95.873	4.853
ETS-Similarity	**6.591**	**93.146**	**96.438**	4.576

spread. The simple combination of ETS and Similarity achieves the best performances regarding MSIS and (upper) coverage levels compared with the four benchmarks.

5. Discussions

Statistical time series forecasting typically involves selecting or combining the most accurate forecasting model(s) per series, which is a complicated task significantly affected by data, model, and parameter uncertainties. On the other hand, nowadays, big data allows forecasters to improve forecasting accuracy through cross-learning, i.e., by extracting information from multiple series of similar characteristics. This practice has been proved highly promising, primarily through the exploitation of advanced machine learning algorithms and fast computers (Makridakis et al., 2020). Our results confirm that data-centric solutions offer a handful of advantages over traditional model-based ones, relaxing the assumptions made by the models, while also allowing for more flexibility. Thus, we believe that extending forecasting from within series to across series, is a promising direction to forecasting.

An important advancement of our forecasting approach over other cross-learning ones, is that similarity derives directly from the data, not depending on the extraction of a feature vector that indirectly summarizes the characteristics of the series (Petropoulos et al., 2014; Kang et al., 2017, 2020). To this end, the uncertainty related to the choice and definition of the features used for matching the target to the reference series is effectively mitigated. Moreover, no explicit rules are required for determining what kind of statistical forecasting model(s) should be used per case (Montero-Manso et al., 2020). Instead of specifying a pool of forecasting models and an algorithm for assigning these models to the series, a distance measure is defined and exploited for evaluating similarity. Finally, forecasting models are replaced by the actual future paths of the similar reference series.

Our results are significant for the practice of business research with more accurate forecasts translating into better business decisions. Forecasting is an important driver for reducing inventory associated costs and waste in supply chains (for a comprehensive review on supply chain forecasting, see Syntetos et al., 2016). Small improvements in forecast accuracy are usually amplified in terms of the inventory utility, namely inventory holding and achieved target service levels (Syntetos et al., 2010, 2015). At the same time, forecast accuracy is also essential to other areas of business research, such as humanitarian operations and logistics (Rodríguez-Espíndola et al., 2018), marketing (Qian & Soopramanien, 2014), and finance (Yu & Huarng, 2019).

While the point forecasts are oftentimes directly used in inventory settings, we show that forecasting with cross-similarity allows for better estimation of the forecast uncertainty compared to statistical benchmarks. The upper coverage rates of our approach are superior to that of statistical approaches, directly pointing to higher achieved customer service levels. This is achieved by a minimal increase of the average spread of the prediction intervals, suggesting a small difference in the corresponding holding cost.

Our study also has implications for software providers of forecasting support systems. We offer our code as an open-source solution together with a web interface[3] (developed in R and Shiny) where a target series can be forecasted through similarity, as described in section 3, using the large M4 competition data set as the reference set. We argue that our approach is straightforward to implement based on existing solutions, offering a competitive alternative to traditional statistical modelling. Forecasting with cross-similarity can expand the existing toolboxes of forecasting software. Given that none approach is the best for all cases, a selection framework (such as time series cross-validation) can optimally pick between statistical models

[3] Available here: https://fotpetr.shinyapps.io/similarity/

or forecasting with cross-similarity based on past forecasting performance.

However, the computational time is a critical factor that should be carefully taken into consideration, especially when forecasting massive data collections. This is particularly true in supply chain management, where millions of item-level forecasts must be produced on a daily basis (Seaman, 2018). An advantage of our approach is that the computational tasks in forecasting with cross-similarity can be easily programmed in parallel compared to multivariate models. Moreover, since the DTW distance measure is more computationally intensive than the two other measures presented in this study, an option would be to select between them based on the results of an ABC-XYZ analysis (Ramanathan, 2006). This analysis is based on the Pareto principle (the 80/20 rule), i.e., the expectation that the minority of cases has a disproportional impact on the whole. In this respect, the target series could be first classified as A, B, or C, according to their importance/cost, and as X, Y, or Z, based on how difficult it is to be accurately forecasted. Then, series in the AZ class (important but difficult to forecast) could be predicted using DTW, while the rest using another, less computationally intensive distance measure.

Forecasting with cross-similarity is based on the availability of a rich collection of reference series. In order to have appealing forecasting performance, such a reference dataset should be as representative (see Kang et al. (2020) for a more rigorous definition) as possible to the target series, which is easy to achieve in business cycles because of data accumulation. To illustrate and empirically demonstrate the effectiveness of the approach, we used the M4 competition data set as a reference. This data set is considered to represent the reality appropriately (Spiliotis et al., 2020). However, if our approach is to be applied to the data of a specific company or sector, then it would make sense that the reference set is derived from data of that company/sector so as to be as representative as possible. In the case that it is challenging to identify appropriate reference series for the target series, then generating series with the desirable characteristics (Kang et al., 2020) is an option.

We have empirically tested our approach on three representative data frequencies: yearly, quarterly, and monthly. We have no reason to believe that our approach would not perform well for higher frequency data, such as weekly, daily, or hourly. If multiple seasonal patterns appear, as it could be the case for the hourly frequency with periodicity within a day (every 24 hours) and within a week (every 168 hours), then a multiple seasonal decomposition needs to be applied instead of the standard STL (the *forecast* package for R offers the `mstl()` function for this purpose). On the other hand, our approach is not suitable as-is for intermittent demand data, where the demand values for several periods are equal to zero. In this case, one

could try forecasting with cross-similarity without applying data preprocessing. A similar approach was proposed by Nikolopoulos et al. (2016) who focused on identifying patterns within intermittent demand series rather than across series.

6. Concluding remarks

In this paper, we introduce a new forecasting approach that uses the future paths of similar reference series to forecast a target series. The advantages of our proposition are that it is model-free, in the sense that it does not rely on statistical forecasting models, and, as a result, it does not assume an explicit DGP. Instead, we argue that history repeats itself (déjà vu) and that the current data patterns will resemble the patterns of other already observed series. The proposed approach is data-centric and relies on the availability of a rich, representative reference set of series – a not so unreasonable requirement in the era of big data.

We examined the performance of the new approach on a widely-used data set and benchmarked it against four robust forecasting methods, namely the automatic selection of the best model from the Exponential Smoothing family (ETS), as well as the ARIMA family, the Theta method, and the equal-weighted combination of Simple, Holt, and Damped exponential smoothing (SHD). We find that in most frequencies, the new approach is more accurate than the benchmarks. Moreover, forecasting with cross-similarity is able to better estimate the uncertainty of the forecasts, resulting in better upper coverage levels, which are crucial for fulfilling customer demand. Finally, we propose a simple combination of model-based and model-free forecasts, which results in an accuracy that is always significantly better than the one or the other separately.

The innovative proposition of forecasting with cross-similarity and without models points towards several future research paths. For example, in this study we do not differentiate the reference series to match the industry/field of the target series. It would be interesting to explore if such matching would further improve the accuracy of forecasting with cross-similarity.

Acknowledgements

Yanfei Kang is supported by the National Natural Science Foundation of China (No. 11701022) and the National Key Research and Development Program (No. 2019YFB1404600). Feng Li is supported by the National Natural Science Foundation of China (No. 11501587) and the Beijing Universities Advanced Disciplines Initiative (No. GJJ2019163).

References

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

Andrees, M. A., Pena, D., & Romo, J. (2002). Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference*, *100*, 1–11.

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

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

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Society*, *20*, 451–468.

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

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

Blanc, S. M., & Setzer, T. (2016). When to choose the simple average in forecast combination. *Journal of Business Research*, *69*, 3951–3962.

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

Boylan, J. E., Chen, H., Mohammadipour, M., & Syntetos, A. (2014). Formation of seasonal groups and application of seasonal indices. *The Journal of the Operational Research Society*, *65*, 227–241.

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

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*, 559–583.

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

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. In *Statistical Models in S* chapter 8. (p. 68). Taylor & Francis Group.

Dudek, G. (2010). Similarity-based approaches to short-term load forecasting. *gdudek.el.pcz.pl*, .

Dudek, G. (2015a). Pattern similarity-based methods for short-term load forecasting–part 1: Principles. *Applied Soft Computing*, .

Dudek, G. (2015b). Pattern similarity-based methods for short-term load forecasting–part 2: Models. *Applied Soft Computing*, .

Fildes, R. (1989). Evaluation of aggregate and individual forecast method selection rules. *Management Science*, *35*, 1056–1065.

Fildes, R. (2001). Beyond forecasting competitions. *International Journal of Forecasting*, *17*, 556–560.

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

Fiorucci, J. A., Pellegrini, T. R., Louzada, F., Petropoulos, F., & Koehler, A. B. (2016). Models for optimising the theta method and their relationship to state space models. *International Journal of Forecasting*, *32*, 1151–1161.

Franses, P. H. (2016). A note on the mean absolute scaled error. *International Journal of Forecasting*, *32*, 20–22.

Goodwin, P., Dyussekeneva, K., & Meeran, S. (2013). The use of analogies in forecasting the annual sales of new electronics products. *IMA Journal of Management Mathematics*, *24*, 407–422.

Green, K. C., & Armstrong, J. S. (2007). Structured analogies for forecasting. *International Journal of Forecasting*, *23*, 365–376.

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

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

Hu, K., Acimovic, J., Erize, F., Thomas, D. J., & Van Mieghem, J. A. (2019). Forecasting new product life cycle curves: Practical approach and empirical analysis. *Manufacturing & Service Operations Management*, *21*, 66–85.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2019). *forecast: Forecasting functions for time series and linear models*. R package version 8.7.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*, 2579–2589.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*, 1–22.

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

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

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

Kang, Y., Hyndman, R. J., & Li, F. (2020). GRATIS: GeneRAting TIme series with diverse and controllable characteristics. *Statistical Analysis and Data Mining*, *13*, 354–376.

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

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

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting*, *21*, 397–409.

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

Kourentzes, N., Rostamitabar, B., & Barrow, D. K. (2017). Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research*, *78*, 1–9.

Li, H., Liu, J., Yang, Z., Liu, R. W., Wu, K., & Wan, Y. (2020). Adaptively constrained dynamic time warping for time series classification and clustering. *Information Sciences*, *534*, 97–116.

Li, Y., Liu, R. W., Liu, Z., & Liu, J. (2019). Similarity Grouping-Guided neural network modeling for maritime time series prediction. *IEEE Access*, *7*, 72647–72659.

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

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition.

Journal of Forecasting, *1*, 111–153.

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

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

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science*, *29*, 987–996.

Martínez, F., Frías, M. P., Charte, F., & Rivera, A. J. (2019a). Time series forecasting with KNN in r: the tsfknn package. *The R Journal*, *11*, 229–242.

Martínez, F., Frías, M. P., Pérez, M. D., & Rivera, A. J. (2019b). A methodology for applying k-nearest neighbor to time series forecasting. *Artificial Intelligence Review*, *52*, 2019–2037.

Mohammadipour, M., & Boylan, J. E. (2012). Forecast horizon aggregation in integer autoregressive moving average (INARMA) models. *Omega*, *40*, 703–712.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*, 86 – 92.

Nikolopoulos, K., Litsa, A., Petropoulos, F., Bougioukos, V., & Khammash, M. (2015). Relative performance of methods for forecasting special events. *Journal of Business Research*, *68*, 1785–1791.

Nikolopoulos, K. I., Babai, M. Z., & Bozos, K. (2016). Forecasting supply chain sporadic demand with nearest neighbor approaches. *International Journal of Production Economics*, *177*, 139–148.

Petropoulos, F., Hyndman, R. J., & Bergmeir, C. (2018a). Exploring the sources of uncertainty: Why does bagging for time series forecasting work? *European Journal of Operational Research*, *268*, 545–554.

Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018b). Judgmental selection of forecasting models. *Journal of Operations Management*, *60*, 34–46.

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

Petropoulos, F., & Svetunkov, I. (2020). A simple combination of univariate models. *International journal of forecasting*, *36*, 110–115.

Qian, L., & Soopramanien, D. (2014). Using diffusion models to forecast market size in emerging markets with applications to the chinese car market. *Journal of Business Research*, *67*, 1226–1232.

Ramanathan, R. (2006). Abc inventory classification with multiple-criteria using weighted linear optimization. *Computers & Operations Research*, *33*, 695 – 700.

Rodríguez-Espíndola, O., Albores, P., & Brewster, C. (2018). Disaster preparedness in humanitarian logistics: A collaborative approach for resource management in floods. *European Journal of Operational Research*, *264*, 978–993.

Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, *34*, 822 – 829.

Shah, C. (1997). Model selection in univariate time series forecasting using discriminant analysis. *International Journal of Forecasting*, *13*, 489–500.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*, 75 – 85. M4 Competition.

Spiliotis, E., Assimakopoulos, V., & Nikolopoulos, K. (2019). Forecasting with a hybrid method utilizing data smoothing, a variation of the theta method and shrinkage of seasonal factors. *International Journal of Production*

Economics, *209*, 92–102.

Spiliotis, E., Kouloumos, A., Assimakopoulos, V., & Makridakis, S. (2020). Are forecasting competitions data representative of the reality? *International Journal of Forecasting*, *36*, 37–53.

Svetunkov, I. (2016). True model. https://forecasting.svetunkov.ru/en/2016/06/25/true-model/. Accessed: 2019-5-30.

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

Syntetos, A., Babai, M. Z., & Gardner, E. S. (2015). Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping. *Journal of Business Research*, *68*, 1746–1752.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, *252*, 1–26.

Syntetos, A. A., Nikolopoulos, K., & Boylan, J. E. (2010). Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *International Journal of Forecasting*, *26*, 134–143.

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

Thombs, L. A., & Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, *85*, 486–492.

Timmermann, A. (2006). Forecast combinations. In C. G. G. Elliott, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (pp. 135–196). Elsevier volume 1.

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

Wright, M. J., & Stern, P. (2015). Forecasting new product trial with analogous series. *Journal of Business Research*, *68*, 1732–1738.

Yu, T. H., & Huarng, K. (2019). A new event study method to forecast stock returns: The case of facebook. *Journal of Business Research*, *in press*.

Zhang, K., Chen, H., Boylan, J., & Scarf, P. (2013). Generalised estimators for seasonal forecasting by combining grouping with shrinkage approaches. *Journal of Forecasting*, *32*, 137–150.