Forecasting hierarchical time series in supply chains: an empirical investigation

ABSTRACT

Demand forecasting is a fundamental component of efficient supply chain management. An accurate demand forecast is required at several different levels of a supply chain network to support the planning and decision-making process in various departments. In this paper, we investigate the performance of bottom-up, top-down and optimal combination forecasting approaches in a supply chain. We first evaluate their forecast performance by means of a simulation study and an empirical investigation in a multi-echelon distribution network from a major European brewery company. For the latter, the grouped time series forecasting structure is designed to support managers' decisions in manufacturing, marketing, finance and logistics. Then, we examine the forecast accuracy of combining forecasts of these approaches. Results reveal that forecast combinations produce forecasts that are more accurate and less biased than individual approaches. Moreover, we develop a model to analyse the association between time series characteristics and the effectiveness of each approach. Results provide insights into the interaction among time series characteristics and the performance of these approaches at the bottom level of the hierarchy. Valuable insights are offered to practitioners and the paper closes with final remarks and agenda for further research in this area.

Keywords: Supply chain forecasting, Forecast combination, Hierarchical forecasting, Grouped time series forecasting, Time series characteristics.

1. INTRODUCTION

Demand forecasting is the starting point for most planning and control organizational activities (Rostami-Tabar et al. 2015). It is vital to supply chains (SCs), as it provides the basic inputs for the planning and control of all functional areas, including logistics, marketing, production, and finance (Ballou 2004). Demand forecasting performance is subject to the uncertainty underlying the time series demand a SC is dealing with (Rostami-Tabar 2013). Therefore, capturing, managing and characterizing uncertainty in SCs represents one of the main problems confronting managers while planning and synchronizing operations in SCs. The demand uncertainty is among the most important challenges facing modern SCs (Syntetos et al. 2016; Chen and Blue 2010; Mircetic et al. 2017; Teunter et al. 2018; Babai, Ali, and Nikolopoulos 2012; Trapero, Cardos, and Kourentzes 2019; Nikolopoulos, Babai, and Bozos 2016). The challenge that demand uncertainty may bring to supply chains is also highlighted by the recent Covid-19 pandemic causing many disruptions (Nikolopoulos et al., 2020; Singh et al., 2020)

that poses considerable difficulties in terms of the SC planning and control (Syntetos et al. 2016). Hence, the purpose of demand forecasting in SCs is to inform SC planning decisions by providing an accurate estimation of the future demand in a given situation.

Demand forecasting for SC often concerns many items. SC forecasters may extrapolate the time series for each Stock Keeping Unit (SKU) individually. However, most of the SC time series have natural groupings of SKUs; that is, the SKUs may be aggregated to get higher levels of forecasts across different dimensions such as geographical areas, customer types, supplier types and product families (Chen and Boylan 2007). Therefore, forecasts are required at various levels to inform decisions at different parts of SC. The level at which forecast is generated will depend on the desired level of the decision making (Rostami-Tabar et al. 2015). For instance, a retailer may use the point-of-sale data to produce forecasts at the store level. However, a manufacturer may use the forecasts of aggregated demand series for the production planning (Chopra and Meindl 2007). For the transportation manager in charge of the distribution planning, crucial information may include spatial fragmentation of demand, shipments size (replenishment orders) for each distribution channel, the timing of the shipments and the type of product in shipments. Inventory manager might be interested in forecasts related to the type of materials needed at the SKU level, how much will be needed, and when it will arise (Caplice and Sheffi 2006). Accordingly, SC managers may disaggregate the total demand on the dimensions which are important for a particular party in the chain. This is the situation where hierarchical (HF) or grouped forecasting (GF)[a] should be used instead of producing forecasts at each level separately.

A considerable part of the forecasting literature has been dedicated to methods for single time series, but in reality, there are often many related time series that can be organized hierarchically or in groups (Rostami-Tabar et al. 2015). Hierarchical time series can be represented as a hierarchically organized multiple time series that may be aggregated at several different levels in groups based on different features (Hyndman et al. 2011). Grouped time series are hierarchical time series that do not impose a unique hierarchical structure in the sense that the order by which the series can be grouped is not unique (Hyndman and Athanasopoulos 2018).

HF naturally reflects important SC characteristics and offers an ample scope for the introduction of innovative forecasting methodologies (Syntetos et al. 2016), improvement of the forecast accuracy and planning, reduction of the overall forecast burden and delivery of the high service level (Strijbosch, Heuts, and Moors 2008; Caplice and Sheffi 2006; Turbide 2015). Existing approaches for forecasting hierarchical and grouped times series may involve bottom-up (BU), top-down (TD) and optimal

[a] GF can be considered as a special case of HF. Depending on the demand structure of the SC, HF or GF methodology might be used.

combination (OC) approach. In this paper, we may use GF or HF approaches to refer to BU, TD and OC approaches. In the TD approach, the forecast is generated at the top level of the forecasting structure and then disaggregated to the bottom level series. Oppositely, the BU generates multiple forecasts at the bottom level of the forecasting structure and then aggregates these forecasts to the upper levels in the hierarchy. Hyndman et al. (2011) proposed the OC approach as a new methodology for HF. OC uses all the information available in the hierarchy by forecasting all of the series independently and then uses a regression model to reconcile them.

When forecasting demand for a hierarchical/grouped SC network, practitioners need to determine three things: *i)* the forecasting model to use when generating the base forecasts, which are independent forecasts created at different levels of the forecasting structure; *ii)* an appropriate forecasting structure, we refer to forecasting structures as an umbrella term for all different data structures and designs performed to tailor the data structure in line with the needs of the decision-makers (e.g. hierarchical, grouped or temporal structure); and *iii)* an approach which provides the most accurate forecasts. The latter has attracted the attention of many researchers as well as practitioners over the last few decades (Huber, Gossmann, & Stuckenschmidt, 2017; Pennings & van Dalen, 2017; Rostami-Tabar, 2013; Syntetos et al., 2016; Widiarta, Viswanathan, & Piplani, 2009). Recently, there are several studies dealing with the application of HF approaches in real-world data sets (Abolghasemi, Hyndman, Spiliotis, & Bergmeir, 2020; Abolghasemi, Hyndman, Tarr, & Bergmeir, 2019; Punia, Singh, & Madaan, 2020; Spiliotis, Abolghasemi, Hyndman, Petropoulos, & Assimakopoulos, 2020). Although the hierarchical/grouped forecasting has been studied for decades, there is no agreement on which HF approach provides more accurate forecasts. Moreover, there is a lack of studies in the literature linking time series characteristics to the accuracy of HF models especially using real datasets of a SC structure. Abolghasemi et al. (2020) suggest that the improvements in terms of forecasting accuracy could be possibly achieved if forecasters were able to select the most appropriate HF method according to the characteristics of the series that form a hierarchy. To the best of our knowledge, this is the first study that uses a real dataset of a multi-echelon SC to investigates not only the effectiveness of HF approaches but also the forecast performance of HF combinations. Kahn (1998) was the first to suggest that it is time to combine the existing methodologies so that we can enjoy the good features of both methods, but no specific idea was provided in that discussion.

In this paper, we evaluate the performance of different approaches in a SC context. To do so, we conduct a simulation study and an empirical investigation using real data from a SC distribution network of a major European brewery company. Our contribution to the literature is fourfold: *i)* we demonstrate the application of grouped demand forecasting in SC and compare the effectiveness of approaches on a multi-echelon SC from a major European brewery company; *ii)* we examine the

forecast performance of HF combination *iii)* we comprehensively evaluate the performance of BU, TD and OC approaches and *iv)* we develop a model to analyse the association of time series characteristics with the forecast performance of different approaches.

The remainder of the paper is structured as follows: Section 2 introduces the reserach background and a review of the literature. Section 3 provides forecasting approaches for hierarchical and grouped time series. Section 4 and Section 5 present the simulation and empirical evaluation, subsequently. Section 6 presents the association of time series characteristics with the forecasting performance of HF models. We discuss the findings and conclude the paper with future research and final remarks. in Section 7.

2. HIERARCHICAL AND GROUPED TIME SERIES FORECASTING

Compared with traditional forecasting of univariate time series, forecasting hierarchical or group time series is a more challenging and demanding task for forecasters. One of the reasons is because hierarchical or grouped data structures impose an additional aggregation constraint, which needs to be taken into account during the forecasting process. This constraint is related to generating the forecasts which need to be consistent through all levels in the hierarchy or grouped structure. In literature, we refer to this constraint as “aggregate consistency” or “coherent forecasts”. Therefore, the objective is to generate the final forecast that will add up in a way that is consistent with the aggregation structure of the collection of time series (Hyndman and Athanasopoulos 2018). The HF/GF could be seen as the principle on how the base forecasts are aggregated, disaggregated, reconciled or combined during the process of generating the final forecasts for each series in the forecasting structure.

There are three main methodologies which can be used when dealing with forecasting hierarchical and grouped time series: BU, TD and OC. The main criterion for selecting among different methodologies is the forecast accuracy. This is essential as an effective planning and operation logistics system require an accurate, disaggregated demand forecasts (Caplice and Sheffi 2006). Errors in forecasting may cause significant misallocation of the resources in inventory, facilities, transportation, sourcing, pricing, and even in information management (Chopra and Meindl 2007). Forecasting accuracy is directly connected to inventory management, lower errors result in reduced stock-keeping without compromising the service level (Trapero, Kourentzes, and Fildes 2012). Moreover, inaccurate forecasts will inevitably lead to inefficient, high-cost operations and/or poor levels of customer service. Therefore, one of the most important action we may take to improve the efficiency and effectiveness of the logistics process is to improve the quality of the demand forecasts (Caplice and Sheffi 2006).

Starting from the 1950s, there have been extensive discussions in the literature about the merits of TD and BU models. Studies that favour the BU approach are predominantly from the economy field (Dunn, Williams, and Spivey 1971; Dunn, Williams, and DeChaine 1976; Kinney 1971; Edwards and Orcutt 1969; Collins 1976). Others argue that TD can produce more accurate aggregate forecasts at top levels (Grunfeld and Griliches 1960; Aigner and Goldfeld 1973; Barnea and Lakonishok 1980). Generally, the proponents of a TD approach argue that the lower-level data is often more error-prone and more volatile (Vogel 2013) and suggest that the TD approach is superior because of its lower cost and greater accuracy during times of a reasonably stable demand (Weatherford, Kimes, and Scott 2001). On the other hand, when the distinction between individual demand patterns is important, BU is preferable (Dunn, Williams, and DeChaine 1976; Weatherford, Kimes, and Scott 2001). Schwarzkopf, Tersine, and Morris (1988) argue against using the TD approach for forecasting the bottom level series in a hierarchy. They also challenge the premise that aggregating series reduce the variability at the top level by developing equations which demonstrate that the variability will increase in cases of a positive correlation between bottom level series. We found similar conclusions in Gordon, Morris, and Dangerfield (1997); Dangerfield and Morris (1992). While the empirical results tend to point towards the superiority of the BU approach, there is no general consensus on whether a TD or BU approach performs better (Vogel 2013). In an analytical study, Rostami-Tabar et al. (2015) provide the superiority conditions for BU and TD in the one level hierarchy with two nodes in sub-aggregate level, where series follow a non-stationary integrated moving average process of order one.

The application of the HF/GF especially in SCs requires the need for accurate forecasts at all levels and not only in the aggregate top level (Fliedner 1999; Vogel 2013). In some studies concerning the forecasting accuracy across all levels in the hierarchy, BU shows the better overall performance (Hyndman et al. 2011; Athanasopoulos, Ahmed, and Hyndman 2009; Seongmin, Hicks, and Simpson 2012). There is still a “dead heat race” between the accuracy of TD and BU at the top level of the hierarchy; however, when the entire hierarchy is considered, BU significantly outperforms the TD approach. At the same time, the OC represents a new promising methodology, which has shown excellent results and outperformed others in forecasting tourism, mortality, prison population and labour market data (Hyndman et al. 2011; Shang and Hyndman 2017; Hyndman, Lee, and Wang 2016; Hyndman and Athanasopoulos 2018). To the best of our knowledge, there are only few recent studies that investigate with the application of OC on SC data (Abolghasemi et al., 2019; Punia et al., 2020; Spiliotis et al., 2020). Therefore, there is a need to quantify its effectiveness on new larger data sets. Besides choosing among different methodologies, there is also an additional dilemma about choosing the right forecasting model, which comes from the diversity of models in TD and OC methodologies.

There are several variations of models in TD and OC methodologies which all have different forecasting performances.

By considering the fact that there is no consensus which approach provides the most accurate forecasts and the importance of the forecasting process for practitioners in SCs, we fill the gap in the literature by comparing the forecast accuracy of the different approaches in a real SC and using a simulation study. Additionally, we examine the performance of combining forecasts generated from different models against individual approaches. Finally, we develop a model to analyse the association of time series characteristic with HF approaches.

3. FORECASTING APPROACHES FOR HIERARCHICAL AND GROUPED TIME SERIES

Common approaches to forecast hierarchical or grouped time series often include BU, TD and OC models. Each of them has its own unique principle as well as advantages and disadvantages, which will be further explained in the following subsections. In addition to these approaches, we introduce two combination schemes for combining the forecasts of different approaches in subsection 3.4.

3.1. Bottom-up (BU) methodology

BU approach first generates the base forecasts in the bottom level of the forecasting structure, using a forecasting model. All other forecasts in the structure are generated through aggregating of the base forecast to the higher levels, in a manner which is consistent with the observed data structure. Therefore, summing matrix $\boldsymbol{S}$ can be used to represent the matrix that dictates how the aggregation of higher-level series is calculated from the bottom level series. Accordingly, the final forecasts in the BU approach can be expressed as follows:

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S} \cdot \hat{\boldsymbol{y}}_{B,h}. \quad (1)$$

Where $\tilde{\boldsymbol{y}}_h$ represents the vector of all final forecasts in a given structure for *h*-step-ahead periods and $\hat{\boldsymbol{y}}_{B,h}$ represents the vector of all the bottom level forecasts, generated for *h*-step-ahead.

Since the BU creates the base forecasts at the bottom level, it uses a significant amount of information available in the data. This could result in a better capturing of the individual dynamics of the series in the bottom level. On the other hand, series in the bottom level may be noisy and hard to forecast which may lead to inaccurate forecasts, especially in the top level of the forecasting structure.

3.2. Top-down (TD) methodology

TD consists of generating the forecast at the top level of the structure and then disaggregate it to the bottom level in the structure. For disaggregating the top level forecasts, TD methodology uses the disaggregation proportions (p_j). Hence, the forecasting principle of TD can be presented as:

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S} \cdot \hat{y}_h \cdot \boldsymbol{p}. \quad (2)$$

Where $\hat{\boldsymbol{y}}_h$ represents the top level base forecast generated for the *h*-step-ahead periods and $\boldsymbol{p} = [p_j]$ is a vector containing all disaggregation proportions corresponding to the series in the bottom level. Where *j* = 1, …, *n*; and *n* is the number of bottom level series in the forecasting structure.

Generally, there is a lot of criticism in the literature regarding the performance of TD methodology in the lower levels of the forecasting structures (Hyndman et al., 2011; Mircetic, 2018; Schwarzkopf et al., 1988). The poor performance of the TD approach in the lower levels lies in the disaggregation proportions. There are several variations of the TD approach based on how the disaggregating proportions are determined. These variations could be classified into two groups: approaches that use historical proportions and those that use future forecasts to determine disaggregation proportions. TD methodology can only be used for forecasting the hierarchical time series, but not for the grouped time series.

3.2.1. Top-down approaches based on the historical proportions

In the literature, there are three TD approaches based on historical proportions to determine disaggregation weights. Gross and Sohl (1990) examined twenty-one different proportional disaggregation schemes, which include simple averages of the sales proportions, lagged proportions and combined lagged proportions. They suggest two disaggregation proportions as best for disaggregating the top-level forecasts: i) average historical proportions (TD1) and ii) proportions of the historical averages (TD2). The majority of practitioners are still using disaggregation proportions, suggested by Gross and Sohl (1990).

For the TD1, the disaggregation proportions are determined in the following way:

$$p_j = \frac{1}{T}\sum_{t=1}^{T} \frac{y_{j,t}}{y_t}. \quad (3)$$

Disaggregation proportions of TD1 represent the mean value of the proportions between the series in the bottom level ($y_{j,t}$) and the top level series (y_t), observed in the historical period *t* = *1*,..., *T*. Similarly, disaggregation proportions for TD2 reflect the relationship between the average historical values of the same series and they are determined as follows:

$$p_j = \frac{\sum_{t=1}^{T} \frac{y_{j,t}}{T}}{\sum_{t=1}^{T} \frac{y_t}{T}}. \quad (4)$$

Based on the TD1 and TD2 models, Chen, Yang, and Hsia (2008) attempted to improve the accuracy of the TD methodology. They propose an algorithm to minimise the sum of squared errors to determine the disaggregating proportions as a result of that process. In their approach, disaggregating proportions are determined as follows:

$$p_j = \frac{\sum_{t=1}^{T} y_{j,t} \cdot y_t}{\sum_{t=1}^{T} {y_t}^2}. \quad (5)$$

We will refer to this approach as TD3 in the following Sections.

3.2.2. Top-down approaches based on future forecasts

Given that disaggregation proportions can change over time, it could significantly deteriorate the forecast accuracy at the bottom level. Therefore, it is crucial to capture the dynamic nature of disaggregation proportions by using the future forecasts of the series in the forecasting structure.

Fliedner (2001) was among the first to propose such a TD model and suggested using final forecasts of the BU model for that purpose. The author proposed calculating the ratio of the direct child forecast divided by the sum of the direct child forecasts comprising their families. The parent forecast is multiplied by this ratio. For more details refer to Appendix in (Fliedner 2001). We will refer to this approach as TD4 in the following Sections.

Top-down forecasted proportions (TDFP) is another TD approach for generating the disaggregating proportions by using future forecasts (Athanasopoulos, Ahmed, and Hyndman 2009). For that purpose, the TDFP is–using future forecasts of the top and bottom level series, which as a result significantly improved the accuracy of the TD methodology. Boylan (2010) note that although this has not been tested on SC data, the use of forecasted proportions rather than historical proportions appears to be promising. The principle of determining TDFP forecasted proportions is the following:

$$p_j = \prod_{l=0}^{K-1} \frac{\hat{y}_{j,h}^{(l)}}{\hat{S}_{j,h}^{(l+1)}}. \quad (6)$$

Where $\hat{y}_{j,h}^{(l)}$ is the *h*-step-ahead base forecast of the node that is *l* levels above *j,* and $\hat{S}_{j,h}^{(l)}$ refers to the sum of the *h*-step-ahead base forecasts below the node which is *l* levels above the node *j* and directly connected to that node (Hyndman and Athanasopoulos 2018).

3.3. Optimal combination (OC) methodology

The OC approach uses all the information that is available in the series by generating the univariate forecasts for all of the series in the forecasting structure. Since the independent univariate forecasts do not meet the condition of "aggregate consistency", OC is performing the reconciliation of the forecasts. The aim of reconciliation is to produce the final forecasts which are mutually coherent and at the same time close to the initial independent base forecasts. The generic formula for producing all final *h*-step-ahead forecasts ($\tilde{\boldsymbol{y}}_h$) in the OC approach is the following:

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\hat{\boldsymbol{y}}_h. \quad (7)$$

Where $\boldsymbol{W}_h$ represents the variance-covariance matrix of the base forecast errors.

There are four variations of the OC approach, depending on how the estimation of the $\boldsymbol{W}_h$ matrix is performed. These estimators are: *i*) ordinary least square, *ii*) weighted least squares, *iii*) structural scaling and *iv*) the minimum trace. In this paper, we used the minimum trace with shrinkage estimation of the variance-covariance matrix since it provids the most accurate forecasts in the simulation and empirical study. For more details regarding the OC approach and its different estimators, see (Hyndman, Ahmed, and Athanasopoulos 2007; Hyndman, Lee, and Wang 2016; Hyndman and Athanasopoulos 2018).

3.4. Combination approaches

In this paper, we also use two forecast combination approaches based on forecasts generated from existing approaches: *i)* COMB - the combination of models forecasts with no weights which is shown in the Eq. 8 and *ii)* COMBw - the weighted combination of models forecasts, shown in the Eq. 9 (Ballou 2004).

$$\tilde{\boldsymbol{y}}_{COMB,h} = \frac{1}{\mathrm{m}}\sum_{m=1}^{M}\hat{\boldsymbol{y}}_{B,h_m}. \quad (8)$$

$$\tilde{\boldsymbol{y}}_{COMBw,h} = \sum_{m=1}^{M}\tilde{\boldsymbol{y}}_{COMBw,h_m} = \sum_{m=1}^{M}\boldsymbol{W}_m \cdot \hat{\boldsymbol{y}}_{B,h_m}. \quad (9)$$

Where $\tilde{\boldsymbol{y}}_{COMB,h}$ and $\tilde{\boldsymbol{y}}_{COMBw,h}$ represent the vector of *h*-step-ahead bottom level forecasts, created from the combination of *h*-steps-ahead bottom level forecasts of other HF models ($\hat{\boldsymbol{y}}_{B,h_m}$). $\boldsymbol{W}_m$ is the scaling matrix which transforms the bottom level forecasts of different HF models used in combination, which could be represented as:

$$\underbrace{\begin{bmatrix}\tilde{y}_{1,m}\\ \tilde{y}_{2,m}\\ .\\ .\\ .\\ .\\ .\\ \tilde{y}_{n,m}\end{bmatrix}}_{\tilde{\boldsymbol{y}}_{COMBw,h_m}} = \underbrace{\begin{bmatrix}\omega_{1,m} & 0 & 0 & . & . & . & . & 0\\ 0 & \omega_{2,m} & 0 & . & . & . & . & 0\\ 0 & 0 & \omega_{3,m} & . & . & . & . & 0\\ . & . & . & . & . & . & . & .\\ . & . & . & . & . & . & . & .\\ . & . & . & . & . & . & . & .\\ 0 & 0 & 0 & 0 & 0 & 0 & \omega_{n-1,m} & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \omega_{n,m}\end{bmatrix}}_{\boldsymbol{W}_m} \cdot \underbrace{\begin{bmatrix}\hat{y}_{1,m}\\ \hat{y}_{2,m}\\ .\\ .\\ .\\ .\\ .\\ \hat{y}_{n,m}\end{bmatrix}}_{\hat{\boldsymbol{y}}_{B,h_m}}$$

Diagonal elements of $\boldsymbol{W}_m$ correspond to the weighted coefficients $\boldsymbol{\omega}_m$=$[\omega_{1,m}, \omega_{2,m}, \ldots, \omega_{n,m}]^T$. The weight associated with each $\boldsymbol{\omega}_{j,m}$ coefficient represents inverse forecasting error of the model *m*, scaled by the sum of inverse errors of all observed HF models on the particular node *j* (Eq. 10).

$$\omega_{j,m} = \frac{\frac{1}{\varepsilon_{j,m}}}{\sum_{m=1}^{M}\frac{1}{\varepsilon_{j,m}}}; for\ j = 1,2, \ldots n. \tag{10}$$

Where $\varepsilon_{j,m}$ represents the forecasting error of model *m* at the bottom level node *j*, *n* is the number of nodes at the bottom level and *M* is the number of HF models used in combining.

There are three possibilities to combine forecasts of separate HF/GF models that result in reconciled combined forecasts: *i)* at the top level, *ii)* at the bottom level or *iii)* in all levels. In this study, we combine the forecasts at the bottom level where individual bottom level forecasts of the best performing models are combined via two combination schemes presented in Eq. 8 and 9. Therefore, we use the forecasts at the bottom level to generate combination forecasts. Forecasts at higher levels are generated through aggregating the combined base forecast to the higher levels, in a way that is consistent with the observed data structure (***S***). Therefore, the final coherent forecasts of the COMB and COMBw approaches can be expressed as following:

$$\underbrace{\begin{bmatrix}\tilde{y}_h\\ \tilde{y}_{A,h}\\ \tilde{y}_{B,h}\\ \tilde{y}_{C,h}\\ \tilde{y}_{D,h}\\ \tilde{y}_{AA,h}\\ \tilde{y}_{AB,h}\\ \tilde{y}_{BA,h}\\ \tilde{y}_{BB,h}\\ \tilde{y}_{CA,h}\\ \tilde{y}_{CB,h}\\ \tilde{y}_{DA,h}\\ \tilde{y}_{DB,h}\end{bmatrix}}_{\tilde{\boldsymbol{y}}_h} = \underbrace{\begin{bmatrix}1&1&1&1&1&1&1&1\\ 1&1&0&0&0&0&0&0\\ 0&0&1&1&0&0&0&0\\ 0&0&0&0&1&1&0&0\\ 0&0&0&0&0&0&1&1\\ 1&0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0&0\\ 0&0&1&0&0&0&0&0\\ 0&0&0&1&0&0&0&0\\ 0&0&0&0&1&0&0&0\\ 0&0&0&0&0&1&0&0\\ 0&0&0&0&0&0&1&0\\ 0&0&0&0&0&0&0&1\end{bmatrix}}_{\boldsymbol{S}} \cdot \underbrace{\begin{bmatrix}\tilde{y}_{AA,h}\\ \tilde{y}_{AB,h}\\ \tilde{y}_{BA,h}\\ \tilde{y}_{BB,h}\\ \tilde{y}_{CA,h}\\ \tilde{y}_{CB,h}\\ \tilde{y}_{DA,h}\\ \tilde{y}_{DB,h}\end{bmatrix}}_{\tilde{\boldsymbol{y}}_{B,h}}$$

Or in a compact form shown in the Equation 11:

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S} \cdot \tilde{\boldsymbol{y}}_{B,h}. \tag{11}$$

We use the combination approaches in the simulation and the empirical study. Different combination of BU, TD and its varieties and OC approaches are used to generate $\tilde{\boldsymbol{y}}_{B,h}$; depending on their individual performances in the simulation and the empirical study. We also evaluate the performance of the combination approaches via several forecast accuracy and bias measures.

4. NUMERICAL SIMULATION

In Section 4, we perform a simulation study to evaluate: *i)* the relative performance of the TD, BU and the OC approaches; and *ii)* the performance of the forecast combination approaches.

4.1. Experiment design

The simulation hierarchy consists of three levels, where the top aggregated series (Total) is subdivided into four series at level 1 (A, B, C and D) and each of series is further disaggregated into two additional series at the level 2 (AA, AB, BA, BB, CA, CB, DA and DB). Therefore, there are eight-time series in the bottom and 13 series in total (Fig. 1). The seasonal *Autoregressive Integrated Moving Average* (S-ARIMA) process is used for generating the monthly simulated series at the bottom level of Fig. 1. For that purpose, we used the *sarima.Sim* function in R software (Smith 2019). Generally, the ARIMA framework of the analysis has been the most useful for research in the SC forecasting (Syntetos et al. 2016; Rostami-Tabar et al. 2015).

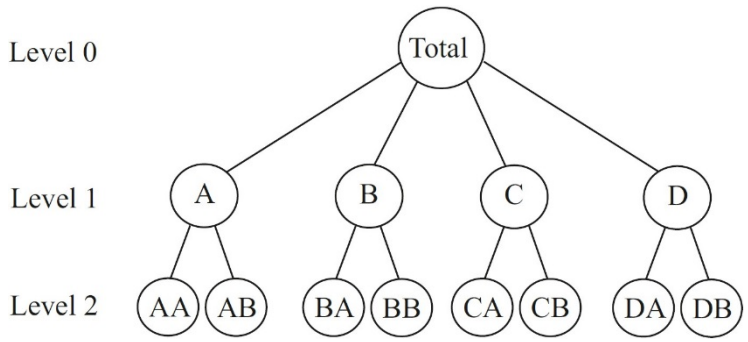

Fig. 1. The hierarchical structure of the simulation study.

During the simulation, orders of the S-ARIMA process (*d, D* - differencing; *p, P* - autoregression and *q, Q* - moving average) were chosen randomly and restricted to values of 0, 1 and 2. Moving average (θ, Θ) and autoregressive parameters (ϕ, Φ) were also chosen randomly from the interval [-0.99, 0.99], with controlling stationarity and invertibility of the simulated series. The error term is normally distributed white noise with mean zero and variance one. Therefore, we generate the bottom level series ($\boldsymbol{y}_{B,t}$) corresponding to eight nodes at level 2 of Fig. 1. We then obtain all other series ($\boldsymbol{y}_t$) by aggregating the bottom level series. The process of obtaining all the series in the hierarchy could be represented as:

$$\underbrace{\begin{bmatrix} y_t \\ y_{A,t} \\ y_{B,t} \\ y_{C,t} \\ y_{D,t} \\ y_{AA,t} \\ y_{AB,t} \\ y_{BA,t} \\ y_{BB,t} \\ y_{CA,t} \\ y_{CB,t} \\ y_{DA,t} \\ y_{DB,t} \end{bmatrix}}_{\boldsymbol{y}_t} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}}_{S} \cdot \underbrace{\begin{bmatrix} y_{AA,t} \\ y_{AB,t} \\ y_{BA,t} \\ y_{BB,t} \\ y_{CA,t} \\ y_{CB,t} \\ y_{DA,t} \\ y_{DB,t} \end{bmatrix}}_{\boldsymbol{y}_{B,t}}$$

Or in a compact form as :

$$\boldsymbol{y}_t = S \cdot \boldsymbol{y}_{B,t}. \tag{12}$$

Each generated series has 100 observations, and all series are restricted to be positive. If a series contains negative values, we add a constant positive number to make the entire series positive. The constant is chosen in a way that all observations become positive. The simulation is repeated for 500 times, producing 500 different scenarios of the series at the bottom level. In the literature, there were only two similar simulations studies to look upon (Hyndman et al. 2011; Hyndman, Ahmed, and Athanasopoulos 2007). They used 600 and 1000 simulations, respectively. In this paper, we used 500 simulations since, after 300-350 simulations, forecasting errors become stable.

In order to evaluate the forecasting performance of each model, we divide each simulated series into in-sample/training and out-of-sample/test sets. Training data are initially set with 88 observations, and the test set with 12 observations. The exponential smoothing state space (ETS) models are used to produce out of sample forecasts. ETS models are applied using the automatic identification algorithm implemented using the *forecast* package for R (Hyndman et al. 2018). The algorithm uses the Akaike information criterion for selecting an appropriate ETS model. Forecasting horizon is set to 12-step-ahead and 1 to 12-steps-ahead forecasts are produced. After that, the out of sample error is determined for every time series based on the structure depicted in Fig. 1. We use the Root Mean Square Scaled Error (RMSSE) to summarise and report the accuracy by finding the average value of RMSSE across all different error sets and the simulation scenarios (Hyndman and Athanasopoulos, 2018). The measure is calculated for each series as follows:

$$RMSSE = \sqrt{mean(q_j^2)},$$

where

$$q_j = \frac{e_j^2}{\frac{1}{T-m}\sum_{m+1}^{T}(y_t - y_{t-m})^2},$$

y is the actual value at period t and e_j is the forecast error, T is number of observations, and $m = 1$ for non-seasonal series. Additionally, we estimate the forecast bias of different models by measuring the Mean Percentage Error (MPE).

4.2. Numerical results

Due to the space restrictions, in this subsection we only present the most important results of the simulation study. Please refer to the Appendix A for comprehensive analysis and more details about the performance of the HF approaches and HF combinations in the simulation study.

The overall results of the simulation study show that the forecast combinations of individual HF approaches , i.e. COMBw and COMB, generat accurate forecasts in the hierarchy. Moreover, COMBw is more accurate than any individual HF approach in all nodes of the hierarchy measured by by RMSSE (Table A1, Figs A1 and A2). COMBw is closely followed by BU, OC and COMB models. The accuracy of TD approaches (TD1, TD2, TD3, TD4 and TDFP) is far behind the accuracy of the COMBw model. Furthermore, TD1, TD2, TD3 and TDFP demonstrate diverging performances by moving from the top to the bottom level series. Moreover, it is noticeable that TD1, TD2, TD3 and TDFP perform better only at the highest level of the hierarchy and their performance deteriorates in all other levels. This is especially pronounced by TDFP approach which failed to produce reliable forecasts at the bottom level series. The results are almost the same in terms of the forecast bias evaluated via MPE. Again, the most accurate approach is COMBw followed by OC, BU and COMB models (Fig. A2). All TD models demonstrate poor results and underperformed in terms of forecast bias.

5. EMPIRICAL EVALUATION

In Section 5, we assess the empirical validity of the main findings of this research using real time series of a SC distribution network from a European brewery company. There is a lack of studies evaluating the performance of the BU, TD and OC in the SCs, specially with large datse sets. There are only a few recent studies in he literature that link forecasting to various parts of SCs (Abolghasemi et al., 2019; Mircetic, 2018; Mirčetić et al., 2017; Pennings & van Dalen, 2017; Punia et al., 2020; Rostami-Tabar et al., 2015; Seongmin, Hicks, & Simpson, 2012; Spiliotis et al., 2020; Villegas & Pedregal, 2018). To the best of our knowledge, this is the first study that examines a comprehensive grouped demand forecasting in a SC network with a large data set. The empirical study is performed to evaluate the

effectiveness of different approaches in a real SC network. Additionally, we also examine the forecast combinations of GF approaches in SCs, which has never been investigated before..

In subsection 5.1, we first provide details of the real SC distribution network and the empirical data available for the purposes of our investigation along with the experimental structure employed in our work. We then present the actual empirical results in subsection 5.2.

5.1. Supply chain distribution network

Fig. 2 illustrates the distribution structure of the brewery company operating in the European market. The scale economies in the transport of freight, combined with the market requirement to provide fast and reliable delivery times, drive most large firms to operate multi-echelon distribution inventories (Caplice and Sheffi 2006). For the same reasons, the observed brewery company has a multi-echelon distribution structure, and its distribution network spreads over several distribution centers (DC) located across various geographical regions. Different DCs are designated to serve only particular market regions. The distribution starts from the central warehouse, which is directly connected to the manufacturing plant. The plant produces more than 200 different beer product families. The annual output from the central warehouse varies, and it is usually between 250,000-300,000 pallets. Highest demand peaks occur in the spring/summer months (May, June and July) with the demand peak of 14,000 pallets/week of different brewery products.

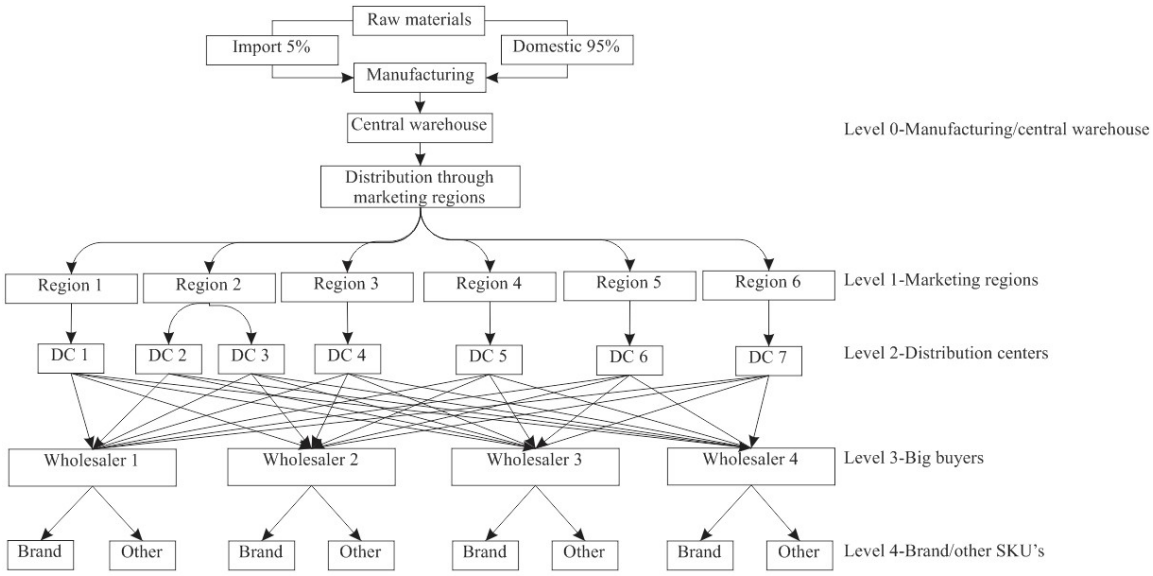

Fig. 2. Multi-echelon brewery distribution network.

The brewery industry has a particular distribution network in a sense that all product consumption is accomplished via bars, restaurants and retail shops. There are no direct deliveries and online sales of products. In order to provide products to a wide consumer network, observed distribution chain is divided into six marketing regions. These marketing regions are supplied through seven DCs. Each

region has one designated DC, except region 2, which is served through two DCs. Further distribution of goods is carried out through wholesalers. There are four big wholesalers which are dominating in the observed market. Some of those wholesalers are big retail chains, while others act as agents between manufacturers and small retailers, bars and restaurants. Nevertheless, each wholesaler recieves brewery products from the DCs and makes the further placement of goods on the market. Goods that are provided to the wholesalers are classified as a brand and other products. Brand products represent the most important products for the company since they are providing the majority of revenue in the market. It is a top-selling beer which comes in different packaging types. In observed brewery distribution chain, there is no further feedback from the wholesalers regarding the point of sale data, therefore the visibility of the customer data is limited.

There are multiple reasons to use the GF in SCs including: i) simplifying forecasting process, ii) obtaining more accurate forecasts, iii) harmonising forecasts from different levels and iv) providing all information needed for different SC parties. Therefore, in this empirical study, the forecasting structure is designed to generate forecasts to support the planning and execution of the processes in different parts of SC. Special attention is given to the alignment of the time component in addition to the cross-sectional alignment of the grouped structure.

5.2. Grouped structure for forecasting the brewery demand

The demand dataset available for the purpose of our research includes 56 weekly time series (SKUs) for the period from 2012 to 2015; from a brewery company. The unit of observation is a pallet. Managers in the Brewery company strive to create an information platform to support the planning and the execution of mutually harmonized and coordinated processes in the manufacturing, marketing, finance and logistics departements. The main input to such a platform is the forecast that informs the desired decision-making level. Therefore, the demand in a given multi-echelon brewery distribution chain has the grouped structure and it is designed to be aligned with the decision-making process in the company. The structure is provided in Table 1, where each row denotes the level of disaggregation.

Table 1. Grouped structure of brewery demand.

Disaggregation level	Level	Labels	Number of series
0	Total	Total	1
1	Regions (marketing regions)	R_1, R_2…R_6	6
2	Distribution centers	DC_1, DC_2…DC_7	7
3	Wholesalers	W_1, W_2…W_4	4
4	Product types	B and O	2
5	Regions x Distribution centers	R_1DC_1, R_2DC_2,…R_6DC_7	7
6	Regions x Wholesalers	R_1W_1, R_1W_2…R_6W_4	24

7	Regions x Product types	R_1B, R_1O…R_6O	12
8	Distribution centers x Wholesalers	DC_1W_1, DC_1W_2…DC_7W_4	28
9	Distribution centers x Product types	DC_1B, DC_1O…DC_7O	14
10	Wholesalers x Product types	W_1B, W_1O…W_4O	8
11	Distribution centers x Wholesalers x Product types	$R_1DC_1W_1B$, $R_1DC_1W_1O$…$R_6DC_7W_4O$	56
	Total number of series		169

At the top level, the total aggregate demand for brewery products is presented. Demand is further divided by marketing regions, DCs, wholesalers, product types and their accompanying interactions. This division provides information related to the manufacturing with total demand, marketing by region demand and product types of demanded products (brand or other products), a financial sector with the large buyers demand and logistics with a spatial fragmentation of demand. The total node represents the central warehouse, while nodes from R_1 to R_6 represent the marketing regions. Nodes at level 2 (from DC_1, DC_2…DC_7) represent the DCs. In level 3, four wholesalers are represented with nodes from W_1 to W_4. In level 4, B and O nodes represent the product types. Further levels represent the interactions between observed disaggregating features. Levels 5, 6, 7 represent the demand disaggregation of different marketing regions by DCs, wholesalers and product types (nodes from R_1DC_1 to R_6O). In levels 8 and 9 demand of DC is further subdivided by the wholesalers and product types (nodes from DC_1W_1 to DC_7O). Nodes in level 10 (from W_1B to W_4O), represent the demand of each wholesaler subdivided by product types. The most disaggregated data arise when we consider the two product types that are supplied through seven different DCs to four different wholesalers, giving a total of 2 x 7 x 4 = 56 bottom level series in the observed grouped structure. These series represented by the nodes from $R_1DC_1W_1B$ to $R_6DC_7W_4O$. Time plots of time series for the first four levels are presented in Fig. 3. Results show that series are non-stationary, with a weak trend and pronounced seasonality.

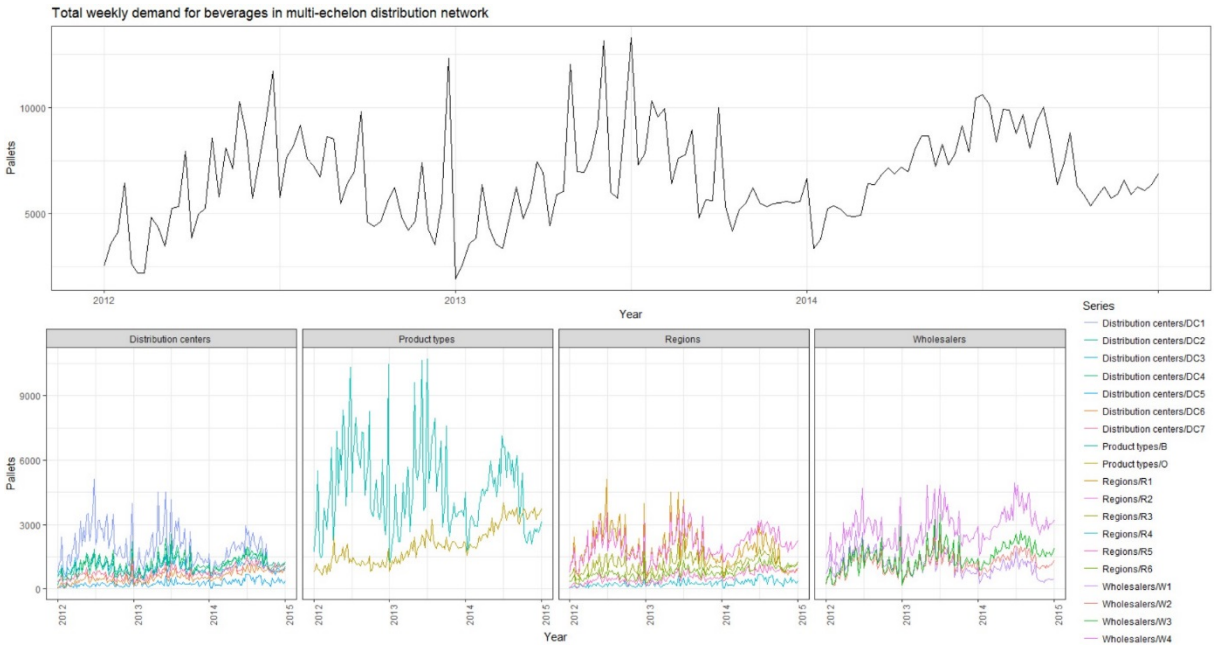

Fig. 3. Total weekly demand, disaggregated by marketing regions, DCs, wholesalers and product types.

5.3. Data and empirical design of experiment

The forecast horizon is equal to the lead time of the decisions driven by the forecast (Nikolopoulos, Syntetos, Boylan, Petropoulos, & Assimakopoulos, 2011). Since the replenishment orders are required every week and manufacturing needs annual forecasts for creating the production and procurement plan, the demand is forecasted on the weekly level for one year ahead. All other sectors require forecasts between these two periods, so they can be easily determined by looking at the forecasts for the period of their interest (monthly, quarterly, semi-annually and annually).

For evaluating the forecasting performance of HF/GF approaches using real time series, we divide each series at each level into training/in-sample and test/out-of-sample sets. Training data is set to 104 weekly observations and includes the period from 2012 to 2014. The test data is set to 52 weekly observations and it represents the period from 2014 to 2015. As in the simulation study, we also use ETS forecasting models from *forecast* package in R, to produce the out of sample base forecasts for the brewery SC data. Forecasting horizon is set to 52-steps-ahead (one year ahead), and 1 to 52-steps-ahead forecasts are generated. After that, the out of sample error is determined for every series in the grouped structure from Table 1. To report the forecast performance, we use RMSSE, MPE and AvgRelMAE performance metrics[b].

[b] Additional forecasting error metrics (RMSE, MAPE, MAE, MASE and ME), are also available from the corresponding author on request and through Shiny platform: https://supplychainanalytics.shinyapps.io/empirical_beverage_study/

5.4. Empirical results

In this section, we present the results of the empirical investigation. In the subsection 5.4.1, we jointly evaluate the effectiveness of GF approaches and GF forecast combinations using real data from a brewery SC, while in the subsection 5.4.2 we examine whether GF forecast combinations improve the forecast accuracy or not. Subsection 5.4.3. summarises the key implications for practitioners. For forecasting the grouped brewery demand structure, we only keep BU, OC, COMB and COMBw approaches. We exclude TD approaches because of their unfeasibility with the empirical data structure. Shang and Hyndman (2017) also suggest that BU and OC are the only approaches that are currently suitable for forecasting the grouped demand structures.

5.4.1. The performance of the grouped forecasting approaches and its combinations

Results of the empirical study confirm the simulation results. Fig. 4 shows that all models produce similar forecasts, compared by RMSSE. The red line in Fig. 4 represents the median value of the RMSSE forecasting error of the OC model, while grey box plots represent the performance of combined GF models (COMBw and COMB).

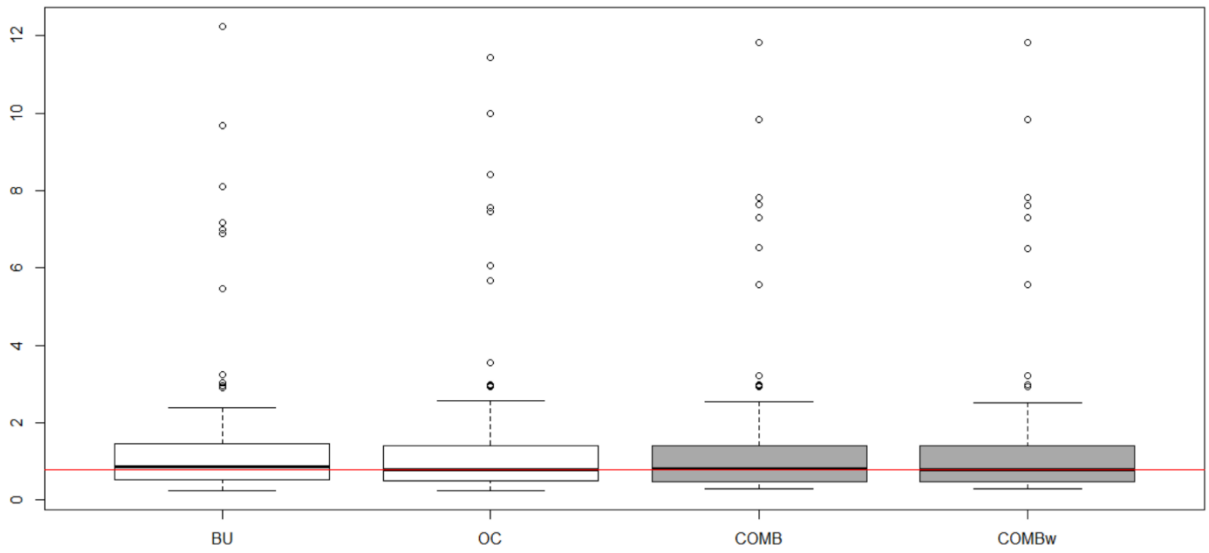

Fig. 4. Forecasting performance of different models tested on the multi-echelon brewery distribution chain via RMSSE.

We observe that the COMB and COMBw models generated almost identical forecasts which were slightly better than forecasts of OC and BU (please refer to Table B1 in Appendix for details). The difference in the performance of models is tested using MCB (Multiple Comparison with the Best method) (Koning et al., 2005). The test also confirms that COMB and COMBw outperform others (Fig. 5). However, their performance is not significantly different from the other two approaches, as the MCB test failed to identify important discrepancy among forecasts of COMB, COMBW, OC and BU.

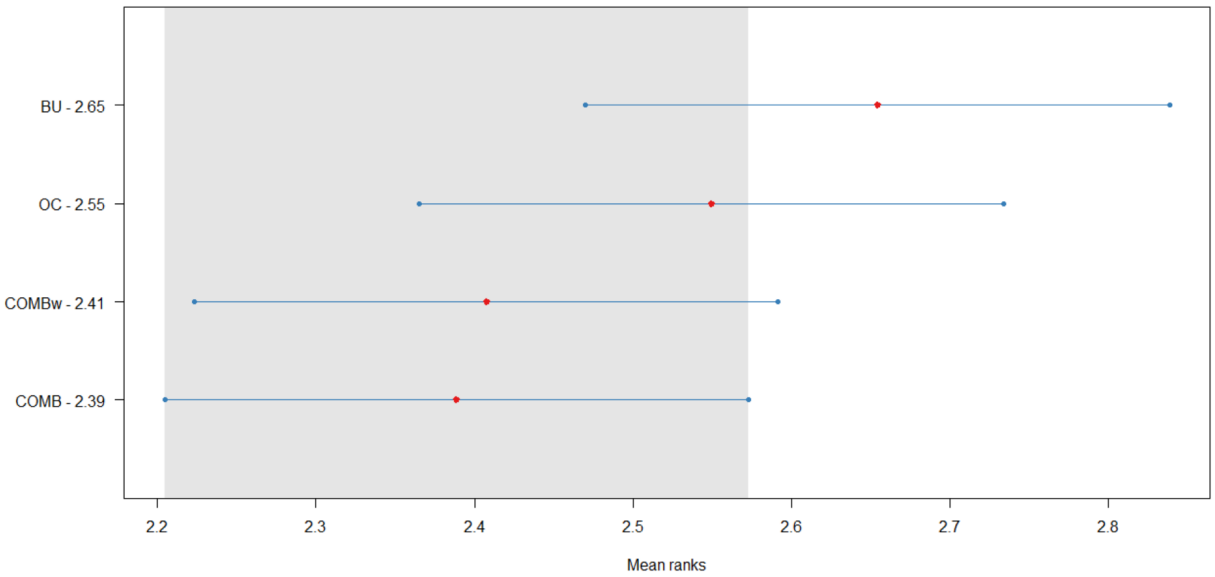

Fig. 5. Performance of different models evaluated through MCB test. RMSSE errors are used for computing the ranks and a 95% confidence level.

The situation is almost identical in a term of forecast bias, where all models demonstrated similar performance, measured by the MPE (please refer to Table B2 in Appendix for details). Fig. 6 presents the performance of the different models measured by MPE. The COMBw generated the lowest median value of the MPE forecasting error and it is represented by the red line in Fig. 6.

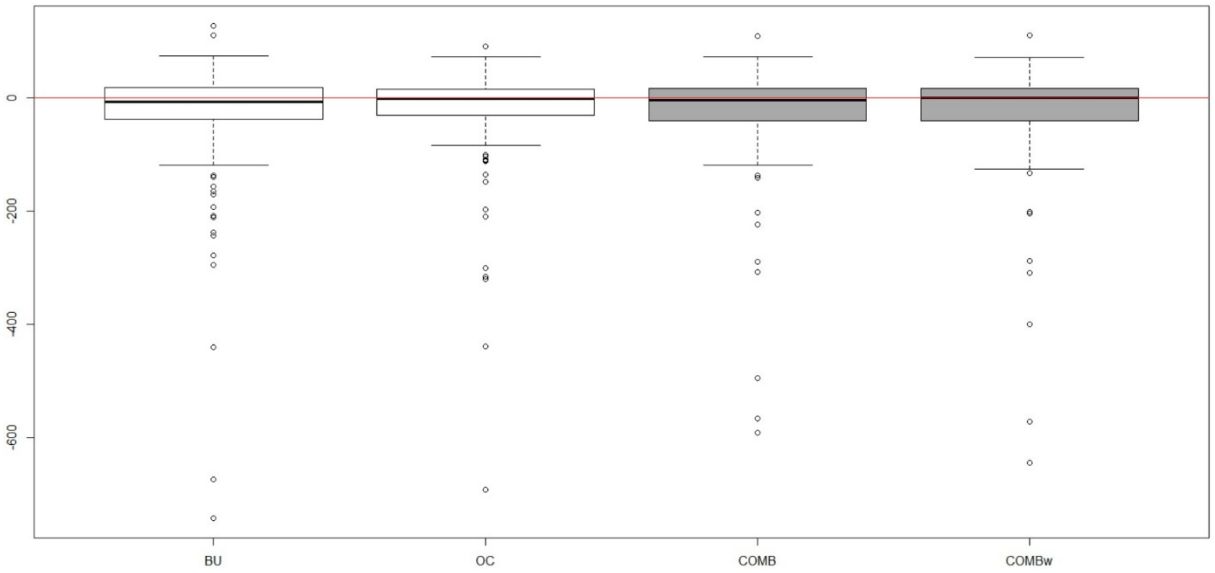

Fig. 6. MPE forecasting error of different models in the empirical study.

Althoug the results indicate the similar performance, but COMB and COMBw model has the smallest outliers. The post hoc MCB test also confirms the accurate results of COMB and COMBw which are closely followed by the OC model as illustrated in Fig. 7. The test also revealed that COMB model

generated statistically different forecasts from the BU, which underperformed in terms of forecast bias compared to other models.

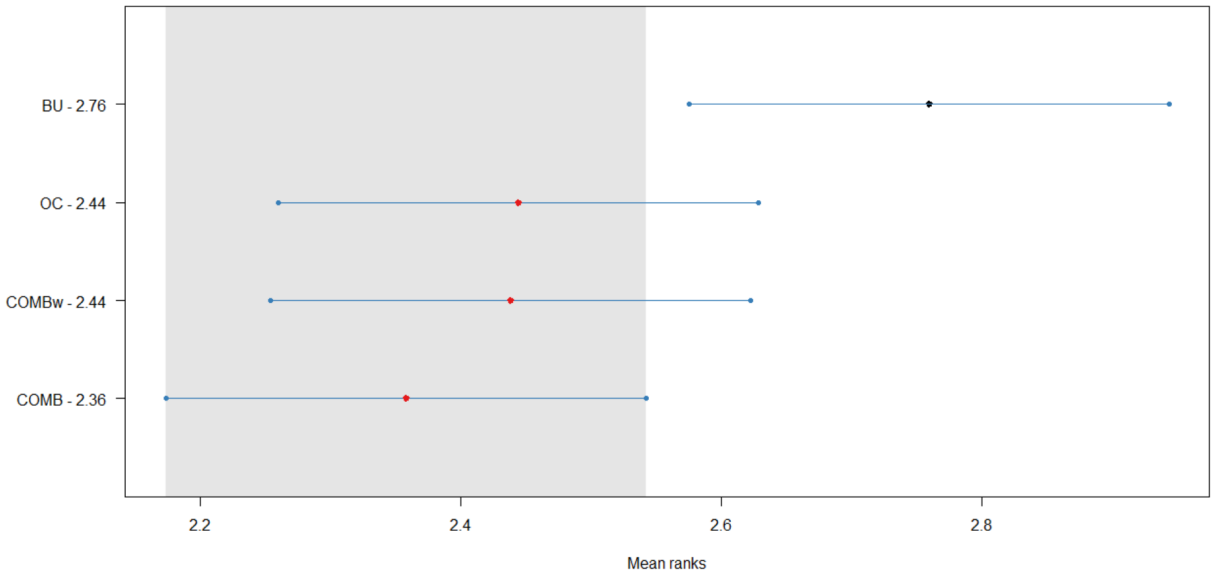

Fig. 7. Bias performance of different models evaluated through MCB test. MPE errors are used for computing the ranks and a 95% confidence level.

Figs. 4, 5, 6 and 7 show that combining the forecasts of OC and BU through two different combination approaches (Eq. 8 and Eq. 9) produce consistently more accurate forecast through all nodes of the brewery SC. Moreover, COMB and COMBw forecasts outperform forecasts generated by OC and BU approaches.

5.4.2. Relative forecasting performance improvement of the combination approaches

In this subsection, we summarise the accuracy performance of GF and forecast combination approaches in a brewery SC. Due to the specific features of SKU-level demand data and different scale of time series in the levels of the forecasting structure, many well-known error measures are not appropriate. In order to overcome the disadvantages of existing measures, Davydenko and Fildes (2013) recommended that the average relative Mean Absolute Error (AvgRelMAE) should be used. To determine improvement/reduction in forecasting performance between competing models, AvgRelMAE uses a geometric mean of MAE ratios between models. The procedure of calculating AvgRelMAE used in the paper is the following:

$$AvgRelMAE = \left(\prod_{i=1}^{m} r_i\right)^{1/m} ; \; r_i = \frac{MAE_i^f}{MAE_i^s}. \tag{13}$$

Where MAE_i^s is the MAE of the baseline statistical forecast for the series *i*, MAE_i^f is the MAE of the competing model for the series *i* and *m* is the total number of time series. Since the COMB model generated the most accurate forecasts according to RMSSE and also produced least biased forecasts based on MPE in the empirical study (please refer to Tables B1 and B2 in the Appendix for details), we use it as a benchmark model i.e $MAE_i^s = MAE_i^{COMB}$. The results of the AvgRelMAE comparisons are shown in Table B3 in the Appendix. Table B3 presents the increase or decrease of MAE forecasting error of different models, compared to the forecasts of the COMB in the whole grouped brewery SC. AvgRelMAE is interpretable, as it represents the average relative value of MAE adequately, and directly shows how the observed model improves the MAE compared to the baseline forecast. If AvgRelMAE < 1, it means that on average $MAE_i^f < MAE_i^s$, and therefore the observed model improves the accuracy, while AvgRelMAE > 1 indicates the opposite (Davydenko and Fildes 2013).

Our results reveal that the COMB model outperforms others, generating the most accurate MAE forecasts. COMBw generates consistently accurate forecasts at every node of the group structure and has the closest performance to the COMB model. For easier interpretation of overall forecasting performance in terms of AvgRelMAE, we transform the last row of the Table B3 to the percentage scale. The average percentage improvement in MAE is determined by (1 − AvgRelMAE) × 100, which is summarised in Table 2. Positive values indicate the forecast improvement compared to COMB model, while negative suggests the opposite.

Table 2. The average percentage improvement in MAE of all GF models compared to the OC model.

	(1 − AvgRelMAE) × 100 (%)			
	BU	OC	COMB	COMBw
Average	-2.7013	-0.3134	0.0000	**-0.0049**

Note. The best result is bolded.

On average, COMBw generates almost identical forecasts, while OC produces approximately 0.3% higher MAE compared to the COMB model. BU forecasts are notably less accurate, BU produces, on average, 2.7% higher MAE errors than the COMB model. Therefore, forecasts generated by the GF combining forecast of individual HF approaches demonstrat more accurate fresults and we recommend further development and usage of GF combinations in contrast to using individual GF approaches to generate grouped time series forecasts in SCs.

5.4.3. The implications for the practitioners

In this study, we evaluate the performance of GF models via statistical metrics. The performance of the forecasting models should ideally be evaluated through utilities such as cost reduction or service

improvement. However, this is a challenging task as forecasting is used at various levels of the hierarchy to support different type of decisions in finance, logistics, marketing or transportation planning. This will require the knowledge on how these functions are implemented as well as their relevant monetary parameters. Moreover, evaluating the performance of GF models on the entire hierarchical or grouped time series needs more research which we will be considered in our future works. Nevertheless, insights drawn from the forecast performance evaluation by statistical metrics is important for practitioners. Our result suggests that when dealing with hierarchical and grouped time series forecasting, combining forecasts generated from individual models could improve the forecast accuracy and reduce the bias. The comprehensive evaluation of different HF approaches, initially performed on 500 simulation scenarios, indicate that BU, OC, COMB and COMBw approaches outperform all TD approaches. Moreover, forecasts from TD approaches proved to be unreliable at the bottom level of the hierarchy. Additionally, COMBw generates more accurate (evaluated through RMSSE) and less biased forecasts (evaluated through MPE) than all other models according to the MCB test. The findings from the simulation study were cross-validated through the empirical SC data set. Results of the empirical investigation are aligned with the insights driven from the simulation study. Forecast combination approaches (i.e. COMB and COMBw) proved to be a useful tool in forecasters' toolbox since they outperform all individual HF models examined in this research. These findings could be extremely useful for decision makers when choosing the right approach for hierarchical or grouped time seies forecasting.

6. THE ASSOCIATION OF TIME SERIES CHARACTERISTICS AND THE PERFORMANCE OF FORECASTING APPROACHES

In Section 4 and Appendix A, we discuss that the performance of different models are increasingly diverging by moving from the top aggregate level to the bottom level in the hierarchy. At the top level, all HF models show similar forecasts. In both simulation and empirical studies, the differences between their performance becomes more apparent at lower levels of the hierarchy, more specifically at the bottom level. Abolghasemi et al. (2020) suggest that the forecast accuracy of different HF methods is closely related to the characteristics of the individual series. Therefore, we investigate whether there is any association between the performance of different HF models and characteristics of the series at the bottom level. To that end, we develop an additive multiple linear regression model. Multiple linear regression represents a model for forecasting cross-sectional data. It assumes that there is a linear relationship between input features $X=(X_1, X_2,\ldots, X_p)$, and the response variable Y (Eq. 14).

$$Y = f(X) + \varepsilon. \tag{14}$$

where ε is a random error term, independent from *X*, with mean zero. It is the irreducible part of forecasting error of the model, therefore we are only interested in estimating the relationship $f(X)$ shown in the Eq. 15.

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \quad (15)$$

where X_j represents the *j* predictor, and β_j represents the average effect of X_j predictor while holding all other predictors fixed.

The algorithm provides insights into the interaction among characteristics of time series and the accuracy of different HF models. The idea is to measure and extract different characteristics of time series (that comprise the bottom level of the hierarchy), which will then be used as independent features (*X*) in the multiple linear regressions. RMSEs of different HF models are set as dependent variables (*Y*). The main goal is to identify the influential (i.e. statistically significant) time series features, rather than creating the most accurate statistical learning algorithm on a given set of data. For every HF model, a separate regression model was created. Therefore, we create seven regression models.

In this study, we are mainly interested in developing a simple and interpretable model for managers than a complex one, which is hard to interpret. Therefore, we put more emphasis on the model's interpretability than on its predictive power, i.e. we chose a multiple linear regression model instead of complex nonlinear ones. There are no restrictions in the settings that would limit the inclusion of complex nonlinear models, however it will influence the model's interpretability which is an important feature for the end-users. Generally, if the aim is to develop an algorithm in which interpretability is not a concern, this research can be easily extended to more flexible models, such as Generalized Additive Models, Ridge regression, Lasso regression, Classification and regression trees, Random Forests and Boosting trees.

For evaluating the statistical learning algorithm, we form a database which contains the results of the simulation study. For each node at the bottom level of the simulation study, 19 different time series characteristics and RMSE forecast errors of HF models, are extracted, scaled and recorded in the database. Therefore, the database contains 4000 (8 series at the bottom level * 500 scenarios) different entries for time series measures and the HF forecasting errors. We use 70% of the data for training and 30% for the test. Therefore, 19 different time series characteristics are used as independent variables in regression models which are presented in the first column of Table 3 from number 1 to 19.

First, 16-time series characteristics are described in detail by Hyndman, Wang, and Laptev (2015) and Wang, Smith-Miles, and Hyndman (2009). These characteristics might provide insights into why some

HF models perform better than others on the same data. Moreover, we include the following three characteristics: *i*) correlation between observed bottom level series and the top aggregate series (*correlation* (*bts-top*)); *ii*) the participation of observed series from the bottom level to the top aggregate series (*aggregate share*); and *iii*) *coefficient of variation*. Additional characteristics are calculated by the following equations:

$$correlation(bts-top) = r_j = \frac{\sum_{t=1}^{T}(y_{j,t}-\overline{y_{j,t}})(y_t-\overline{y_t})}{\sqrt{\sum_{t=1}^{T}(y_{j,t}-\overline{y_{j,t}})^2}\sqrt{\sum_{t=1}^{T}(y_t-\overline{y_{t,}})^2}}; \quad (16)$$

$$aggregate\ share = AS_j = \frac{\sum_{t=1}^{T} y_{j,t}}{\sum_{t=1}^{T} y_t}; \quad (17)$$

$$coefficient\ of\ variation = \frac{\delta}{\mu} = \frac{\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(y_{j,t}-\frac{1}{T}\sum_{t=1}^{T} y_{j,t})^2}}{\frac{1}{T}\sum_{t=1}^{T} y_{j,t}}. \quad (18)$$

Where $\overline{y_{j,t}}$ represents the mean value of the observed *j* bottom level series ($y_{j,t}$) and $\overline{y_t}$ is the mean value for the top level series (y_t), observed in the historical period *t* = 1,..., *T* and *j* = 1, …, *n*.

The *coefficient of variation* measures the volatility of the time series, while *correlation* (*bts-top*) is measuring the strength and direction of the linear relationship amongst the bottom level series and the top aggregate series. *Aggregate share* measures the participation of the observed series from the bottom level at the top aggregate series. It provides information about how "big" or "small" is the observed series in the given hierarchy.

Table 3. Summary statistics for the observed time series characteristics.

	Time series characteristics	Mean	Standard deviation	Median	Min	Max	Range	Skew	Kurtosis
1	*Lumpiness* (Variance of annual variances of remainder)	0.1072	0.1464	0.0525	0.0001	2.0815	2.0814	3.3441	20.3256
2	*Entropy* (Spectral entropy*)*	0.8230	0.1531	0.8772	0.5378	0.9990	0.4612	-0.436	-1.3808
3	*ACF1* (First order of autocorrelation)	0.4536	0.4205	0.5491	-0.8612	0.9809	1.8421	-0.484	-1.0181
4	*Lshift* (Level shift)	0.9115	0.3358	0.8768	0.1635	1.9714	1.8079	0.3221	-0.7585
5	*Vchange* (Variance change)	0.4028	0.1619	0.3937	0.0507	1.3470	1.2963	0.5684	0.8635
6	*Cpoints* (The number of crossing points)	14.8788	10.5203	14.0000	1.0000	49.0000	48.0000	0.2970	-1.0800
7	*Fspots* (Flat spots)	5.3063	4.2595	4.0000	1.0000	39.0000	38.0000	2.4953	8.2992
8	*Trend* (Strength of trend)	0.5771	0.3786	0.7216	0.0000	0.9985	0.9985	-0.341	-1.5784
9	*Linearity* (Strength of linearity)	-0.0084	3.9174	0.0012	-7.7978	7.6329	15.4306	0.0057	-0.7903
10	*Curvature* (Strength	-0.0053	2.6101	-0.0361	-7.0502	6.9923	14.0424	0.0294	-0.0224

	of curvature)								
11	*Spikiness* (Strength of spikiness)	0.0003	0.0007	0.0001	0.0000	0.0151	0.0151	5.3342	74.6329
12	*Season* (Strength of seasonality)	0.4114	0.2293	0.4474	0.0000	0.9819	0.9819	-0.239	-0.9794
13	*Peak* (Strength of peaks)	4.8182	4.1722	3.6155	0.0663	37.8018	37.7355	1.4785	3.1022
14	*Trough* (Strength of trough)	-4.7130	3.9976	-3.5012	-28.453	-0.0648	28.3889	-1.335	2.1290
15	*KLscore* (Kullback-Leibler score)	1.2269	1.5811	0.7708	0.0702	41.7759	41.7057	8.7920	162.069
16	*Change.idx* (Index of the maximum KL score)	25.0863	11.6418	24.0000	12.0000	44.0000	32.0000	0.2237	-1.4699
17	*Correlation* (*bts-top*)	0.2367	0.4281	0.1781	-0.9174	0.9931	1.9105	-0.036	-0.5774
18	*Aggregate share*	0.1250	0.1215	0.0732	0.0080	0.7405	0.7325	1.6579	2.6632
19	*Coefficient of variation*	0.8040	1.5286	0.2676	0.0065	14.1916	14.1851	3.8992	19.1592

Table 3 provides the summary statistics for the time series characteristics that are used as input in the statistical learning algorithms. Fig. B1 in the Appendix B provides the distributions of the time series characteristics. Distributions have different shapes, but the majority of the time series characteristics has a right-skewed distributions.

In order to determine the best subset of predictors, we use the best subset selection combined with a validation test set (James et al. 2013). The best subset selection procedure is applied to the train data in order to determine the best model (i.e. combination of predictors) for each subset size (1 to 19 variables in the model). Following that, each model is evaluated through the validation test set. The model with the smallest test error (i.e. one that minimizes the mean square error) is chosen as the most appropriate in the given situation. To obtain more accurate estimates of the coefficients, the best subset selection procedure is repeated on a whole data set and previously determined optimal subset size. The final model is determined through evaluation of the best model determined from the described selection procedure and the variance inflation factor (VIF). VIF factor reveals the presence of possible multicollinearity in the predictors. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity (James et al. 2013). Therefore, each model is tested on the presence of multicollinearity, and predictors with VIF factor higher than 10 are excluded from the regression. The resulting models are presented in Table 4.

Table 4. The effect of time series characteristics on the performance of HF approaches.

	Time series characteristics	BU	TD1	TD2	TD3	TD4	OC	TDFP
1	*Intercept*[c]	5.97	9.93	9.75	12.24	6.49	5.87	9.79
2	*Lumpiness* (Variance of annual	0.39	-4.36	-4.30		0.49	0.43	

[c] This is not a time series characteristic. It is the intercept of the regression model.

	variances of remainder)							
3	*Entropy* (Spectral entropy*)*	0.68			-6.39	0.67	0.75	
4	*ACF1* (First order of autocorrelation)							
5	*Lshift* (Level shift)	1.05	0.61	0.63		1.22	0.95	
6	*Vchange* (Variance change)	0.57			-0.27	0.68	0.55	
7	*Cpoints* (The number of crossing points)							
8	*Fspots* (Flat spots)		1.76	1.70	1.61			
9	*Trend* (Strength of trend)	1.25	-1.06	-1.06	-0.83	1.09	1.23	
10	*Linearity* (Strength of linearity)	0.53	0.85	0.79	1.03	0.49	0.48	
11	*Curvature* (Strength of curvature)	0.54	0.53	0.52	0.29	0.54	0.55	
12	*Spikiness* (Strength of spikiness)				0.33			
13	*Season* (Strength of seasonality)	-0.44	0.50	0.49	1.12	-0.34	-0.41	
14	*Peak* (Strength of peaks)				-0.38		0.11	
15	*Trough* (Strength of trough)				0.37			
16	*KLscore* (Kullback-Leibler score)		-0.50	-0.48	-0.34			
17	*Change.idx* (Index of the maximum KL score)				-0.62			
18	*Correlation* (*bts-top*)	0.29	-1.93	-1.85	-2.46	0.35	0.31	
19	*Aggregate share*	4.21	5.01	4.90	5.12	4.63	4.07	6.82
20	*Coefficient of variation*	0.47	0.84	0.84	0.53	0.50	0.46	
	Adjusted R^2	0.6626	0.7546	0.7584	0.6855	0.5425	0.6565	0.015

In Table 4, coefficients are presented only when the effect of each time series characteristics is statistically significant (*p-value* < 0.05)[d]. Each column represents a separate regression model created for different HF models. For different HF models, different characteristics found to be significant. Table 4 shows that *lumpiness Lshift, trend, linearity, curvature, season, correlation, aggregate share* and *coefficient of variation* are among time series characteristics that can be associated with the performance of majority HF models at the bottom level. Positive coefficients indicate that they possibly contribute to the increase of the forecasting error, while negative indicates the effect of decreasing the forecasting error. Therefore, HF models tend to produce more inaccurate forecasts while forecasting the time series with higher values of *Lshift, linearity, curvature, coefficient of variation* and *aggregate share*.

Other characteristics can be associated with a mixed influence on different HF models. Accordingly, *lumpiness, entropy, Vchange*, *trend* and *correlation* have an increasing effect on the accuracy of standard TD approaches (TD1, TD2 and TD3) (for those that it was significant), while for the majority of remaining HF models, they has a decreasing effect on the accuracy. In contrast, higher values of *season* variable increases forecast accuracy for the majority of HF models and decreases for standard

[d] We restrict extrapolation of the findings regarding the association of time series characteristics on the forecasting performance, only on those time series which have similar or closely related summary statistics to the data provided in Table 3 and Fig. B1.

TD approaches. *Fspots* and *KLscore* have only a significant impact on the standard TD approaches, where *Fspots* decreases and *KLscore* increases the forecast accuracy. Other characteristics have less effect on the accuracy of HF models and their effect is not holistic since they don't have statistically significant association for the majority of models.

The last row in Table 4 presents the adjusted R^2, i.e. the percentage of the RMSEs variability of HF error models explained by given additive linear regression models. The adjusted R^2 ranges from 1.5% of the explained error variability for the TDFP model, to the high 75.84% of the explained error variability for TD2.

In addition to investigating the association of the time series characteristics with the accuracy of HF models, managers can use Table 4 to estimate the possible forecasting errors at the bottom level series and therefore select between competing HF approaches. To do so, managers first need to estimate and scale the time series characteristics at the bottom level and then multiply characteristics with coefficients from columns of Table 4. Since each column in Table 4 represents a separate regression model, the result of this process will be the estimation of the forecast error for each HF approach.

7. CONCLUSION AND FURTHER RESEARCH

Various levels of forecasts are required in SC to support decisions in different departments such as logistics, marketing, manufacturing and finance. One of the main current practices in SC is to generate separate forecasts at each level using a forecasting model to support different decisions. Seperate forecasting methodologies can only provide accurate forecasts for the unit for which forecasting is performed such as particular level, sector or echelon in the SC. They are not able to provide coherent forecasts across all levels or echelons of the SC. Consequently, these forecasts might be more damaging for the efficiency of the SC, than having perhaps less accurate HF/GF forecasts that are coherent across different parties of the SC. We argue that SC and HF/GF are naturally matched. In this paper, we demonstrate the application of GF methodology in a multiple-echelon distribution network of a major European brewery industry. Special emphasis is given to the design of the forecasting structure to ensure that generated forecasts are aligned with the need of all parties involved in delivering final products in the brewery distribution network. The forecasting structure consists of eleven levels and 169 nodes in total. It provides forecasts for the planning and the execution of processes in key parts of SC including manufacturing, marketing, finance, and logistics. In this paper, we also considered the fact that there is no agreement on which HF approach provides the most accurate forecast. Therefore, we evaluate the effectiveness of BU, TD and the OC approaches in the simulation study, and examine whether the combination of forecasts from these models reduce bias and improve forecasting

accuracy. Moreover, we investigate the association of time series characteristics with the effectiveness of each HF approach.

The main findings of this paper can be summarised as follows:

- First, Optimal Combination approach (minimum trace with shrinkage estimation of the variance-covariance matrix) and BU outperform all TD approaches on average and across all levels. Therefore, we recommend practitioners to use these approaches when generating demand forecasts across various levels of hierarchical or grouped data structures. Moreover, we notice that OC and BU approaches demonstrate robustness and consistency in producing stable and accurate forecasts, regardless of the hierarchy level and time series characteristics. This is a very important result for practitioners as BU shows to be very competitive given its simplicity.
- Second, in comparing various variations of TD approaches, we observe that TD4 outperform other TD methodologies considered in this study. This approach seems to be more robust to the hierarchy level and time series characteristics.
- Third, our results show that combining forecasts of the existing BU, TD and OC models improve the forecast accuracy and reduce bias. We propose two simple combination approaches based on individual forecasts of existing models (COMB and COMBw). Using both simulation and empirical studies, these combinations outperform BU, TD and OC models in terms of forecasting accuracy and bias. Therefore, when dealing with hierarchical and grouped demand structures in a SC, we recommend practitioners to use a combination of models forecasts instead of using individual approaches. This is an important result for practitioners as there is no concern about selecting the best method.
- Fourth, we examine the association of time series characteristics with the forecasting performance of different approaches at the bottom level of the hierarchy. The most influencing time series characteristics on the accuracy of HF models are *lumpiness Lshift, trend, linearity, curvature, season, correlation, aggregate share* and *coefficient of variation.* We also show that higher values of the *Lshift, linearity, curvature, coefficient of variation* and *aggregate share* may have a negative association and deteriorate the forecast accuracy performance of HF models. Additionally, *lumpiness, entropy, Vchange*, *trend* and *correlation* could have an increasing effect on the accuracy of standard TD approaches (for those that it was significant), while for the majority of remaining HF models they could contribute to decreasing of the forecasting accuracy. In contrast, a higher presence of the seasonality in data (*season*) may increase forecast accuracy for the majority of HF models and decrease accuracy for standard TD approaches. Finally, *Fspots* and *KLscore* have only a

significant impact on the standard TD approaches, where *Fspots* could contribute to decreasing and *KLscore* to increasing the forecasting accuracy.

- Finally, we demonstrate the application of grouped forecasting in SC using a multi-echelon distribution network of a major European brewery company. We empirically present the holistic approach for designing the forecasting structure while forecasting the demand in the SC. The structure serves as the information platform to support the planning and execution of processes in manufacturing, marketing, finance and logistics.

In order to guarantee the reproducibility principle in our research, we provide the R codes for the simulation and empirical experiments used in this paper. Our codes will also be available in an open-source R package.

As far as the next steps of research are concerned, further work into the following areas would appear to be merited:

- The value of using exogenous variables in a hierarchical/grouped structure is an important avenue for future research. The external variables that might be used in these structures might take three forms: 1) variables that are independent of the aggregation level such as weather variable 2) variables that are hierarchical in the same manner as the data such as population and 3) variables that are independent or perhaps unique to each level such as GDP. Determining the conditions under which causal models provide more accurate results and also which type of exogenous variables should be used in a hierarchical structure has important implications in practice.
- The association of time series characteristics from all levels in the forecasting structure and creating an algorithm to link the time series characteristics of the entire forecasting structure to the accuracy of each HF model is another interesting avenue for the future research.
- The extension of the work described here to cover utility metrics such as monetary savings would allow linkage between forecasting and managerial decisions. Moreover, it may lead to new methodologies to evaluate the forecast accuracy across a hierarchical structure.

REFERENCES

Abolghasemi, Mahdi, Rob J Hyndman, Garth Tarr, and Christoph Bergmeir. 2019. "Machine learning applications in time series hierarchical forecasting." *arXiv preprint arXiv:1912.00370*.

Abolghasemi, Mahdi, Rob Hyndman, Evangelos Spiliotis, and Christoph Bergmeir. 2020. "Model selection in reconciling hierarchical time series." *arXiv preprint arXiv:2010.10742*.

Aigner, Dennis J, and Stephen M Goldfeld. 1973. "Simulation and aggregation: a reconsideration." *The Review of Economics and Statistics*:114-8.

Athanasopoulos, George, Roman A. Ahmed, and Rob J. Hyndman. 2009. "Hierarchical forecasts for Australian domestic tourism." *International Journal of Forecasting* 25 (1):146–66.
Babai, M Zied, Mohammad M Ali, and Konstantinos Nikolopoulos. 2012. "Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis." *Omega* 40 (6):713-21.
Ballou, Ronald H. 2004. *Business Logistics/Supply Chain Management-Planning, Organizing, and Controlling the Supply Chain*. Fifth edition ed: Pearson/Prentice Hall.
Barnea, Amir, and Joseph Lakonishok. 1980. "An analysis of the usefulness of disaggregated accounting data for forecasts of corporate performance." *Decision Sciences* 11 (1):17-26.
Boylan, John. 2010. "Choosing levels of aggregation for supply chain forecasts." *Foresight: The International Journal of Applied Forecasting* (18):9-13.
Caplice, Chris, and Yossi Sheffi. 2006. "ESD.260J Logistics Systems." *(Massachusetts Institute of Technology: MIT OpenCourseWare), http://ocw.mit.edu. License: Creative Commons BY-NC-SA*.
Chen, Argon, and Jakey Blue. 2010. "Performance analysis of demand planning approaches for aggregating, forecasting and disaggregating interrelated demands." *International Journal of Production Economics* 128 (2):586-602.
Chen, Argon, Kyle Yang, and Ziv Hsia. 2008. "Weighted least-square estimation of demand product mix and its applications to semiconductor demand." *International Journal of Production Research* 46 (16):4445-62.
Chen, Huijing, and John E Boylan. 2007. "Use of individual and group seasonal indices in subaggregate demand forecasting." *Journal of the Operational Research Society* 58 (12):1660-71.
Chopra, Sunil, and Peter Meindl. 2007. *Supply chain management-Strategy, Planning, and Operation*. 3rd ed. New Jersey: Pearson Prentice Hall.
Collins, Daniel W. 1976. "Predicting earnings with sub-entity data: Some further evidence." *Journal of Accounting Research*:163-77.
Dangerfield, Byron J, and John S Morris. 1992. "Top-down or bottom-up: Aggregate versus disaggregate extrapolations." *International Journal of Forecasting* 8 (2):233-41.
Davydenko, Andrey, and Robert Fildes. 2013. "Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts." *International Journal of Forecasting* 29 (3):510-22.
Dunn, Douglas M, William H Williams, and TL DeChaine. 1976. "Aggregate versus subaggregate models in local area forecasting." *Journal of the American Statistical Association* 71 (353):68-71.
Dunn, Douglas M, William H Williams, and W Allen Spivey. 1971. "Analysis and prediction of telephone demand in local geographical areas." *The Bell Journal of Economics and Management Science*:561-76.
Edwards, John B, and Guy H Orcutt. 1969. "Should aggregation prior to estimation be the rule?" *The Review of Economics and Statistics*:409-20.
Fliedner, Gene. 1999. "An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation." *Computers & Operations Research* 26 (10):1133-49.
Fliedner, Gene. 2001. "Hierarchical forecasting: issues and use guidelines." *Industrial Management & Data Systems* 101 (1):5-12.
Gordon, Teresa P, John S Morris, and Byron J Dangerfield. 1997. "Top-down or bottom-up: Which is the best approach to forecasting?" *The Journal of Business Forecasting* 16 (3):13.
Gross, Charles W, and Jeffrey E Sohl. 1990. "Disaggregation methods to expedite product line forecasting." *Journal of Forecasting* 9 (3):233-54.
Grunfeld, Yehuda, and Zvi Griliches. 1960. "Is aggregation necessarily bad?" *The Review of Economics and Statistics*:1-13.
Huber, J., Gossmann, A., & Stuckenschmidt, H. (2017). Cluster-based hierarchical demand forecasting for perishable goods. *Expert systems with applications, 76*, 140-151.
Hyndman, Rob J, Roman A Ahmed, and George Athanasopoulos. 2007. "Optimal combination forecasts for hierarchical time series." In, 23. MONASH University.

Hyndman, Rob J, Roman A Ahmed, George Athanasopoulos, and Han Lin Shang. 2011. "Optimal combination forecasts for hierarchical time series." *Computational Statistics and Data Analysis* 55 (9):2579–89.

Hyndman, Rob J, and George Athanasopoulos. 2018. *Forecasting: principles and practice*: OTexts.

Hyndman, Rob J, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. 2018. "forecast: Forecasting functions for time series and linear models, R package version 8.3."

Hyndman, Rob J, Alan J Lee, and Earo Wang. 2016. "Fast computation of reconciled forecasts for hierarchical and grouped time series." *Computational Statistics & Data Analysis* 97:16-32.

Hyndman, Rob J, Earo Wang, and Nikolay Laptev. 2015. Large-scale unusual time series detection. Paper presented at the Data Mining Workshop (ICDMW), 2015 IEEE International Conference on IEEE.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An introduction to statistical learning*. Vol. 112: Springer.

Kahn, Kenneth B. 1998. "Revisiting top-down versus bottom-up forecasting." *The Journal of Business Forecasting* 17 (2):14.

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*(3), 397-409.

Kinney, William R. 1971. "Predicting earnings: entity versus subentity data." *Journal of Accounting Research*:127-36.

Mircetic, Dejan. 2018. "Boosting the performance of top down methodology for forecasting in supply chains via a new approach for determining disaggregating proportions." University of Novi Sad.

Mircetic, Dejan, Svetlana Nikolicic, Djurdjica Stojanovic, and Marinko Maslaric. 2017. "Modified top down approach for hierarchical forecasting in a beverage supply chain." *Transportation research procedia* 22:193-202.

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., & Vasilakis, C. (2020). Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. European journal of operational research, 290(1), 99-115.

Nikolopoulos, Konstantinos I, M Zied Babai, and Konstantinos Bozos. 2016. "Forecasting supply chain sporadic demand with nearest neighbor approaches." *International Journal of Production Economics* 177:139-48.

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. *Journal of the Operational Research Society, 62*(3), 544-554.

Pennings, Clint LP, and Jan van Dalen. 2017. "Integrated hierarchical forecasting." *European Journal of Operational Research* 263 (2):412-8.

Punia, Sushil, Surya P Singh, and Jitendra K Madaan. 2020. "A cross-temporal hierarchical framework and deep learning for supply chain forecasting." *Computers & Industrial Engineering* 149:106796.

Rostami-Tabar, Bahman. 2013. "ARIMA demand forecasting by aggregation." Université Sciences et Technologies-Bordeaux I.

Rostami-Tabar, Bahman, Mohamed Zied Babai, Yves Ducq, and Aris Syntetos. 2015. "Non-stationary demand forecasting by cross-sectional aggregation." *International Journal of Production Economics* 170:297-309.

Schwarzkopf, Albert B, Richard J Tersine, and John S Morris. 1988. "Top-down versus bottom-up forecasting strategies." *The International Journal Of Production Research* 26 (11):1833-43.

Seongmin, Moon, Christian Hicks, and Andrew Simpson. 2012. "The development of a hierarchical forecasting method for predicting spare parts demand in the South Korean Navy—A case study." *International Journal of Production Economics* 140 (2):794-802.

Shang, Han Lin, and Rob J Hyndman. 2017. "Grouped functional time series forecasting: An application to age-specific mortality rates." *Journal of Computational and Graphical Statistics* 26 (2):330-43.

Singh, S., Kumar, R., Panchal, R., & Tiwari, M. K. (2020). Impact of COVID-19 on logistics systems and disruptions in food supply chain. International Journal of Production Research, 1-16.

Smith, Andrew K. 2019. "Combined Model Selection Criteria (CombMSC), R package."

Spiliotis, Evangelos, Mahdi Abolghasemi, Rob J Hyndman, Fotios Petropoulos, and Vassilios Assimakopoulos. 2020. "Hierarchical forecast reconciliation with machine learning." *arXiv preprint arXiv:2006.02043*.

Strijbosch, L. W. G., R. M. J. Heuts, and J. J. A. Moors. 2008. "Hierarchical estimation as a basis for hierarchical forecasting." *IMA Journal of Management Mathematics* 19 (2):193-205. doi: 10.1093/imaman/dpm032.

Syntetos, Aris, Zied Babai, John Boylan, Stephan Kolassa, and Konstantinos Nikolopoulos. 2016. "Supply chain forecasting: Theory, practice, their gap and the future." *European Journal of Operational Research* 252 (1):1-26.

Teunter, Ruud H, M Zied Babai, Jos AC Bokhorst, and Aris A Syntetos. 2018. "Revisiting the value of information sharing in two-stage supply chains." *European Journal of Operational Research* 270 (3):1044-52.

Trapero, Juan R, Manuel Cardos, and Nikolaos Kourentzes. 2019. "Empirical safety stock estimation based on kernel and GARCH models." *Omega* 84:199-211.

Trapero, Juan R, Nikolaos Kourentzes, and Robert Fildes. 2012. "Impact of information exchange on supplier forecasting performance." *Omega* 40 (6):738-47.

Turbide, David. "How can distribution requirements planning help inventory management?", Accessed 16.04.2016.

Villegas, Marco A, and Diego J Pedregal. 2018. "Supply chain decision support systems based on a novel hierarchical forecasting approach." *Decision Support Systems* 114:29-36.

Vogel, Sebastian. 2013. *Demand fulfillment in multi-stage customer hierarchies*: Springer Science & Business Media.

Wang, Xiaozhe, Kate Smith-Miles, and Rob Hyndman. 2009. "Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series." *Neurocomputing* 72 (10-12):2581-94.

Weatherford, Lawrence R, Sheryl E Kimes, and Darren A Scott. 2001. "Forecasting for hotel revenue management: Testing aggregation against disaggregation." *Cornell hotel and restaurant administration quarterly* 42 (4):53-64.

Widiarta, H., Viswanathan, S., & Piplani, R. (2009). Forecasting aggregate demand: An analytical evaluation of top-down versus bottom-up forecasting in a production planning framework. *International Journal of Production Economics, 118*(1), 87-94.