WORKING PAPER JANUARY 2020

Forecasting grouped time series demand in supply chains

Dejan Mircetic[a,a], Bahman Rostami-Tabar[b], Svetlana Nikolicic[a], Marinko Maslaric[a]

[a]*University of Novi Sad, Republic of Serbia;*

[b]*Cardiff Business School, Cardiff University, Cardiff, United Kingdom*

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 these approaches. Results reveal that forecast combinations produce forecasts than are as accurate as individual approaches. Moreover, we develop a model to analyse the impact of time series characteristics on 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, Babai, Ducq, & Syntetos, 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

[a] Correspondence: D. Mircetic, University of Novi Sad, Trg Dositeja Obradovica 6, 21000 Novi Sad, Republic of Serbia, Tel: +381 21 485 2433, Email: Dejan.Mircetic@uns.ac.rs.

among the most important challenges facing modern SCs (Babai, Ali, & Nikolopoulos, 2012; A. Chen & Blue, 2010; Mircetic, Nikolicic, Stojanovic, & Maslaric, 2017; Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016; Teunter, Babai, Bokhorst, & Syntetos, 2018; Trapero, Cardos, & Kourentzes, 2019), and it 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 & Boylan, 2007). Therefore, various levels of forecasts are required for different parts of SC. The level at which forecasting is performed then it will depend on the function the forecasts are fed into (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 & 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 & 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 moment where the forecasting a single SKU is directed toward the hierarchical forecasting (HF) or grouped forecasting (GF)[b].

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 according to the different features (Hyndman, Ahmed, Athanasopoulos, & Shang, 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 & 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

[b] 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.

and planning, reduction of the overall forecast burden and delivery of the high service level (Caplice & Sheffi, 2006; Strijbosch, Heuts, & Moors, 2008; Turbide, 2015). Existing approaches for forecasting hierarchical and grouped times series may involve bottom-up (BU), top-down (TD) and optimal combination (OC) approach. Therefore, in this paper we may use GF or HF approaches to refer to BU, TD and OC approaches. In the TD approach, the univariate forecast is generated at the top level of the forecasting structure and then disaggregated to the bottom level series. Oppositely, the BU generates the multiple univariate forecasts in the bottom level of the forecasting structure and then aggregates these forecasts to the upper levels in the hierarchy. Hyndman et al. (2011) propose the OC approach as a new methodology for HF. OC is using all the information available in the hierarchy by forecasting all of the series independently and then uses a regression model to combine and reconcile created forecasts.

When forecasting demand for a hierarchical/grouped SC network, practitioners need to determine: *i)* the univariate forecasting model to use when generating the base forecasts[c] *ii)* the appropriate forecasting 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 (Rostami-Tabar, 2013). Although this has been studied for decades, however, 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. 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. The study aims to: *i)* evaluate the performance of BU, TD and OC approaches; *ii)* examine the accuracy of forecast combination of different approaches; and *iii)* propose a model to analyse the effect of time series characteristics on the performance of BU, TD and OC; *iv)* demonstrate the application of grouped demand forecasting in the SC. For the simulation study, we generate time series at the bottom level using *sarima.Sim* function in R software. The hierarchical structure in the simulation study is a two-level hierarchy with eight series in a bottom level and 13 series in total. In the empirical study, the grouped structure of the beer

[c] Base forecasts are independent forecasts created at different levels of the forecasting structure by some of the univariate forecasting models.

distribution network contains eleven levels with 56 series in the bottom level and 169 series in total. The empirical study aims to produce forecast required by manufacturing, marketing, finances and logistics. For generating the base forecasts, exponential smoothing state space (ETS) models are used in all levels as a univariate forecasting method.

Our contribution to the literature is threefold: *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 effect of time series characteristics on the forecast performance of different approaches.

In this paper, we use pure historical series in the hierarchical structure and GF performance has been evaluated via statistical accuracy metrics. It is important to highlight that using exogenous variables in the hierarchical structure of the supply chain might improve the forecast accuracy and it needs more investigation. Moreover, the implication of hierarchical forecasting on managerial decisions and consequently monetary saving is crucial for the entire supply chain and it should be prioritized for the future research.

The remainder of the paper is structured as follows: the theoretical background of the HF/GF is introduced in the next section. Section 3 provides forecasting approaches. Section 4 and Section 5 present the simulation and empirical evaluation, subsequently. Section 6 introduces the effect of time series characteristics on the forecasting performance of HF models. We discuss the findings in Section 7 and conclude the paper with future research and final remarks.

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 the forecasters. One of the reasons is because hierarchical or grouped data structures impose 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[d]. That is, an 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 & Athanasopoulos, 2018). Therefore, besides the always present question of accuracy in the forecasting process, forecasters now have to provide forecasts that are accurate and in the same

[d] In literature authors usually refer to this constraint as “aggregate consistency” or “coherent forecasts”.

time consistent through all levels of hierarchy or grouped structure. Accordingly, the HF/GF could be seen as the principle 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 their forecast accuracy. This is essential as an effective planning and operation logistics system require the use of accurate, disaggregated demand forecasts (Caplice & Sheffi, 2006). Errors in forecasting may cause significant misallocation of the resources in inventory, facilities, transportation, sourcing, pricing, and even in information management (Chopra & Meindl, 2007). Forecasting accuracy is directly connected to inventory management, lower errors result in reduced stock-keeping without compromising the service level (Trapero, Kourentzes, & 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 & 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 in the field of an economy (Collins, 1976; Dunn, Williams, & DeChaine, 1976; Dunn, Williams, & Spivey, 1971; Edwards & Orcutt, 1969; Kinney, 1971). Others argue that TD can produce more accurate aggregate forecasts at top levels (Aigner & Goldfeld, 1973; Barnea & Lakonishok, 1980; Grunfeld & Griliches, 1960). 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, & Scott, 2001). On the other hand, researchers suggest that BU should be used the distinction between demand patterns for individual items is important (Dunn et al., 1976; Weatherford et al., 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 in 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 Dangerfield and Morris (1992); Gordon, Morris, and Dangerfield (1997). 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 recent 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 (Athanasopoulos, Ahmed, & Hyndman, 2009; Hyndman et al., 2011; Seongmin, Hicks, & Simpson, 2012). There is still a "dead heat race" between the accuracy of TD and BU in 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; Hyndman & Athanasopoulos, 2018; Hyndman, Lee, & Wang, 2016; Shang & Hyndman, 2017). According to our knowledge, it has not yet been tested on the SC data. Therefore, there is a need to quantify its effectiveness on this data. 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 considering the importance of the forecasting process for the practitioners in SCs, we fill the gap in the literature by comparing the forecast accuracy of the different approaches in a real SC and simulation study. Additionally, we examine the performance of combining the different models by creating a unique forecast and comparing it against individual approaches. Finally, we develop a model to analyse the impact of time series characteristic on the superiority condition of existing 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 the 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 univariate 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. Therefore, 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 $\hat{y}_h$ 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. 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. Additionally, TD methodology can not be used for forecasting 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 $t = 1,\ldots, T$. 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. Accordingly, they proposed minimizing the sum of squared demand errors and determining 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

Bearing in the mind that disaggregation proportions can change over time that could significantly deteriorate the forecast accuracy in the bottom level, it is crucial to capture the dynamic nature of disaggregation proportions. Therefore, there is another direction for obtaining disaggregating proportions, which consists of 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 et al., 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 *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 & Athanasopoulos, 2014).

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 estimator since it provided the most accurate forecasts in the simulation and empirical study[e].

For more details regarding the OC approach and its different estimators, see (Hyndman, Ahmed, & Athanasopoulos, 2007; Hyndman & Athanasopoulos, 2018; Hyndman et al., 2016).

[e] Due to the space restrictions, we here only present the results of the OC with minimum trace estimator. Results of other estimators are presented in the following online Shiny platform and more details about their performance could be obtained by request from the authors via email.

3.4. Combination approaches

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

$$\widehat{\boldsymbol{y}}_{COMB,h} = \frac{1}{\mathrm{m}}\sum_{i=1}^{m}\widehat{\boldsymbol{y}}_{B,h_i}. \quad (8)$$

$$\widehat{\boldsymbol{y}}_{COMBw,h} = \sum_{i=1}^{m} w_i \cdot \widehat{\boldsymbol{y}}_{B,h_i}. \quad (9)$$

Where $\widehat{\boldsymbol{y}}_{COMB,h}$ and $\widehat{\boldsymbol{y}}_{COMBw,h}$ represent the vector of *h*-step-ahead bottom level forecasts, created from the combination of bottom level forecasts of other models, generated for *h*-step-ahead ($\widehat{\boldsymbol{y}}_{B,h_i}$). Additionally, a weighted scheme is determined as following: $w_i = \frac{\frac{1}{E_i}}{\sum_1^m \frac{1}{E_i}}$, E_i - vector of forecasting errors in the bottom level series of the observed model *i* and *m* is the number of models used for combining.

There are three possibilities to combine separate HF/GF models that results in forecasts of combination reconciled: *i)* at the top level, *ii)* 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:

$$\tilde{\boldsymbol{y}}_h = \boldsymbol{S}\cdot\widehat{\boldsymbol{y}}_{B,h}, \text{ where } \widehat{\boldsymbol{y}}_{B,h} = \begin{cases} \widehat{\boldsymbol{y}}_{COMB,h} = \frac{1}{\mathrm{m}}\sum_{i=1}^{m}\widehat{\boldsymbol{y}}_{B,h_i}, no\ weights\ combination; \\ \widehat{\boldsymbol{y}}_{COMBw,h} = \sum_{i=1}^{m} w_i \cdot \widehat{\boldsymbol{y}}_{B,h_i},\ combination\ with\ weights. \end{cases}. \quad (10)$$

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 depending on their individual performances in simulation and empirical study. We also evaluate the performance of the combination approaches via several forecast accuracy measures.

4. NUMERICAL SIMULATION

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

4.1. Experiment design

For evaluating the performance of the different approaches, a simulation study is performed. The simulation hierarchy consists of two 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. Generally, the ARIMA framework of the analysis has been the most useful for research in the SC forecasting (Rostami-Tabar et al., 2015; Syntetos et al., 2016).

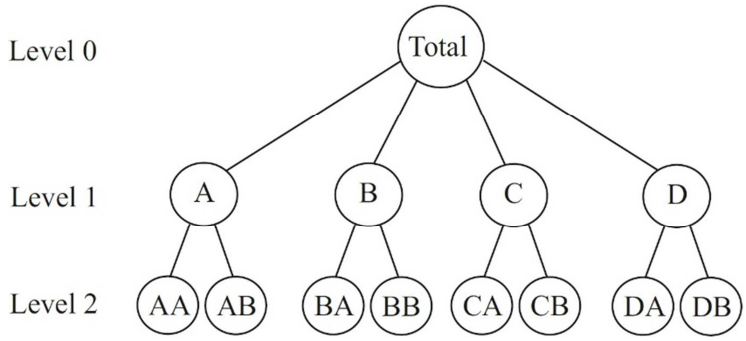

Fig. 1. The hierarchical structure of the simulation study.

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]. 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 the 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 compact form:

$$\boldsymbol{y}_t = S \cdot \boldsymbol{y}_{B,t}. \tag{11}$$

Each generated series has 56 observations, and all 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 in the bottom level. In the literature, there were only two similar simulations studies to look upon (Hyndman et al., 2007; Hyndman et al., 2011). They used 600 and 1000 simulations, respectively. In this paper, we used 500 simulations since, after 300-350 simulations, forecasting errors become stable.

For evaluating 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 12 observations, and the test set with 8 observations. We use ETS forecasting model to produce out of sample base forecasts. ETS models are generated using the automatic identification algorithm from forecast package in R (Hyndman et al., 2018). The algorithm uses the Akaike and the Bayesian information criterion for selecting an appropriate ETS model. Forecasting horizon is set to 8-step-ahead and 1 to 8-steps-ahead forecasts are produced. After that, the out of sample error is determined for every time series from Fig. 1. We use the rolling forecasting procedure for the evaluation. The process is consisted of iteratively adding one observation to the training set and generating 1 to 8-steps-ahead forecasts. The procedure is constantly repeated until training data reached 48 observations. This process yields 32 different error sets for each node in one simulation scenario. After that, the process is repeated with another simulation scenario. We use the Root Mean Square Error (RMSE) to

summarise and report the accuracy by finding the average value of RMSE across all different error sets and the simulation scenarios. Additionally, we test the forecast bias of different models by measuring the Mean Percentage Error (MPE). Moreover, we use the Average Relative Mean Absolute Error (AvgRelMAE) to compare relative forecasting performance between competing models, in terms of overall improvement in Mean Absolute Error (MAE).

4.2. Numerical results

In subsection 4.2, we present the result of simulation investigation on the performance of the HF approaches and HF combinations.

4.2.1. Forecasting performance of the hierarchical forecasting approaches

Table 1 presents the performance of HF models for the hierarchical structure illustrated in Fig. 1.

Table 1. RMSE of different models based on the 500 simulation scenarios[f].

Level	Node	RMSE									
		Base[g]	BU	TD1	TD2	TD3	TD4	OC	TDFP	COMB	COMBw
Level 0	Total	23.59	22.65	23.60	23.60	23.60	23.60	21.60	23.60	21.66	**19.14**
Level 1	A	9.19	9.08	14.50	14.26	17.58	10.12	8.83	9.87	8.83	**8.79**
	B	9.95	9.74	16.22	15.94	19.86	10.36	9.44	10.47	9.38	**9.35**
	C	9.85	9.65	14.99	14.76	18.16	10.79	9.39	10.62	9.34	**9.25**
	D	10.17	10.04	16.17	15.95	19.21	10.84	9.71	10.85	9.66	**9.45**
Level 2	AA	5.39	5.40	8.48	8.32	10.14	5.99	5.32	6.90	5.29	**5.29**
	AB	5.81	5.82	9.61	9.44	12.20	6.38	5.72	7.26	5.68	**5.66**
	BA	5.98	5.98	10.01	9.82	12.60	6.34	5.90	9.10	5.82	**5.80**
	BB	6.26	6.26	10.94	10.75	13.44	6.60	6.15	9.38	6.07	**6.03**
	CA	6.10	6.11	9.87	9.69	12.31	6.86	6.02	11.03	5.95	**5.90**
	CB	5.76	5.76	9.49	9.31	11.56	6.31	5.71	10.81	5.67	**5.61**
	DA	6.00	6.01	9.67	9.53	11.85	6.53	5.93	11.72	5.86	**5.79**
	DB	6.43	6.44	11.42	11.20	13.88	6.92	6.29	12.12	6.22	**6.13**
	Average	8.50	8.38	12.69	12.51	15.11	9.05	8.15	11.06	8.11	**7.86**

The overall results show that the combinations of individual HF approaches (COMBw and COMB) generated the most accurate forecasts in the hierarchy. They performed better than any individual HF approach. We will deal with these findings latter and present the performance of combination

[f] Best results are bolded.

[g] Base forecasts represent independent forecasts obtained by forecasting each time series in hierarchy separately. These forecasts are not aggregate consistent.

approaches more comprehensively in subsection 4.2.2. Here we will continue the discussion with the presentation of the simulation results of the individual HF models.

From all individual HF models (columns 4 to 10 in Table 1), the OC demonstrates the most accurate forecasts in the hierarchy. It performs better than the BU, although the following tests failed to identify any statistically significant difference between these models. OC and BU are closely followed by TD4. TDFP performed well in the upper levels of the hierarchy but failed to produce such accurate forecasts in the bottom level. Results also show the underperformance of TD1, TD2 and TD3 approaches comparing to others in all levels of the hierarchy. The reason could be in the fact that simulated series have varying levels of correlation and participation of the bottom level series in the top aggregate series, which we found to be important in the question of the accuracy of forecasting approaches. Likewise, a great number of the bottom level series has a volatile and dynamic trend, that TD1, TD2 and TD3 are not able to appropriately capture and incorporate in the future forecasts. Therefore, forecasts of these models prove to be unreliable and inaccurate. Moreover, these approaches might be drastically underperformed in some of the simulation scenarios which consequently deteriorate their performance. In Contrast, OC, BU, TD4 and TDFP approaches produce more robust and stable forecasts. This is demonstrated in Fig. 2 were BU, TD4 and OC have a narrow interquartile range, in the box plots which represent their forecasting performance. This suggests that observed models have consistent forecasting errors across all levels and series in the hierarchy. Conversely, TD1, TD2 and TD3 have much larger interquartile range indicating higher dispersion of the forecasting errors, through different levels and series in the hierarchy. Therefore, the outperformance of the TD1, TD2 and TD3 is clear. Differences presented in Fig. 2 and Table 1 are tested on statistical significance. Nemenyi post-hoc test is used to identify the pairs of significantly different forecasts (Pohlert, 2015). The test revealed that TD1, TD2 and TD3 generated statistically indistinguishable forecasts. At the same time, the test identified a statistical difference among forecasts of TD1, TD2 and TD3 and the best performing models OC and BU. Test failed to identify the difference between forecasts of TD1, TD2, TD3 and TD4; and the difference between forecasts of TD1, TD2 and TDFP. Supplementary, the test failed to identify any statistically significant difference among the forecasts of the BU, OC, TD4 and TDFP.

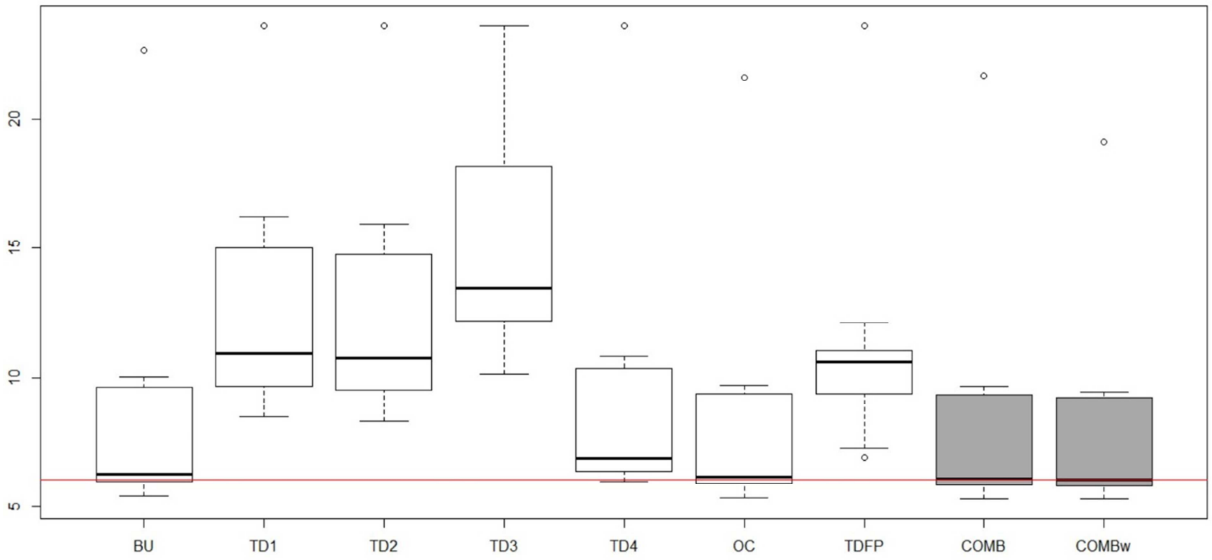

Fig. 2. Box plots for RMSEs of different models tested on simulated data[h].

Fig. 3 displays the performance of different HF models through the hierarchy levels and demonstrates diverging performances by moving from top to the bottom level series. The figure presents that TD1, TD2 and TD3 underperformed compared to other competing models. Moreover, it is noticeable that TD1, TD2 and TD3 perform better only at the highest level of the hierarchy and that their performance deteriorates in all other levels. However, other models show a consistent performance regardless of the hierarchical level, except TDFP which underperformed in the bottom level of the hierarchy. We develop a shiny application[i] that allows readers to perform the comparison between models and their performances.

[h] The red line in the figure represents the median value of the RMSE forecasting error of the COMBw model. White box plots represent the performance of individual HF models, while grey box plots represent the performance of combined HF models (COMB and COMBw).

[i] https://dejanmircetic.shinyapps.io/simulation_study/

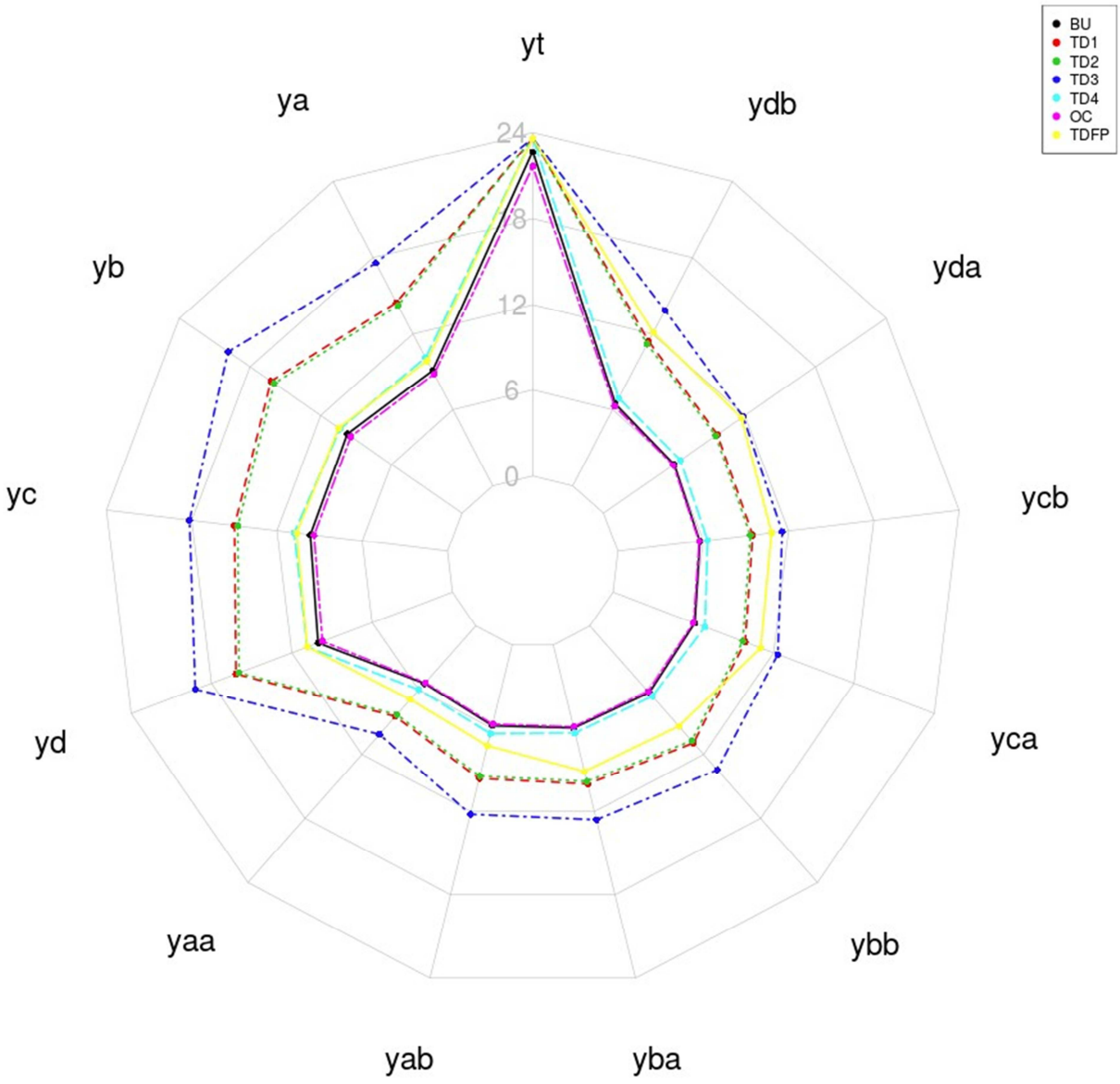

Fig. 3. Performance of different HF models through the hierarchy levels.

The overall performance of BU model was surprisingly good since it managed to be competitive with the OC model. Similar results could be found in Athanasopoulos et al. (2009) and Hyndman et al. (2011). It is possible that BU could demonstrate much poorer results in the hierarchies which have more levels and a higher degree of disaggregation of the data. Forecasting highly disaggregated data (possibly intermittent) could be a challenging task for BU since some of the series features could be undetected and therefore not incorporated in the future forecasts. As a consequence, generating all other forecasts in the hierarchy by aggregating these granular forecasts, could possibly produce misleading and not accurate overall forecasts.

We have also evaluated the forecast bias of different HF models, using the MPE (please refer to Table A.1 in Appendix for details). Results show that OC generates the smallest bias error, and it was

closely followed by the BU, TD4 and TDFP models. Conversely, TD1, TD2, TD3 underperformed and generate highly biased forecasts (Fig. 4).

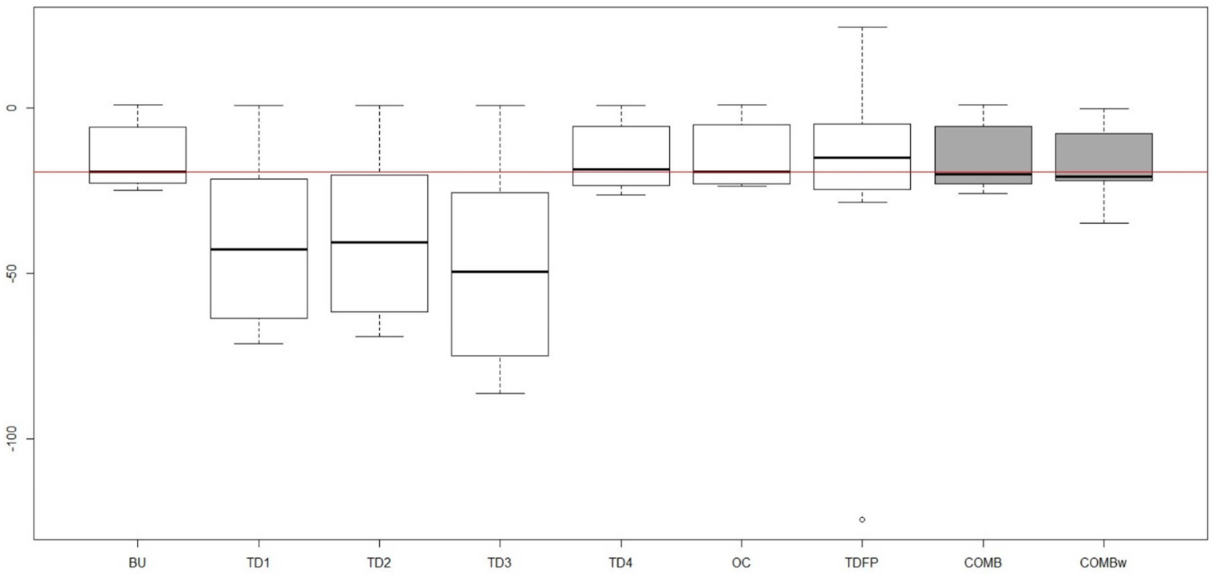

Fig. 4. MPE forecasting error of different models in the simulation study[j].

The Nemenyi post-hoc test failed to identify important differences between forecasts of all models, except for TD3, for which test identified statistically different forecasts from other models. Therefore, the main insights drawn from Table A.1 is aligned with Table 1.

4.2.2. Performance of the combination approaches

Bearing in the mind that several HF models may produce different outcomes (Table 1), combining approaches may be the potential solution to improve the forecast accuracy. This recommendation is usually given for individual (univariate) forecasting. We examine the combination of HF approaches to see whether combining forecasts of HF models produces more accurate forecasts than using a single HF model.

We examined various combinations of HF models in this research. However, we only present the most accurate combination approach. The HF approaches present in the forecast combination approaches are OC, BU and TD4. These models were top-performing individual HF models in Table 1. The combination approaches are described in subsection 3.4. RMSEs and MPEs of the HF combination models (COMB and COMBw) used in the simulation study are presented in the last two columns of Tables 1 and A.1.

[j] The red line in a figure represents the median value of the MPE forecasting error of the OC model. Grey box plots represent the performance of combined HF models (COMBw and COMB).

Fig. 2 displays the summarised performance of COMB and COMBw models (grey boxplots) compared to other HF models. It indicates that combining the forecasts of OC, BU and TD4 reduces forecast errors compared with any individual HF model. Moreover, it indicates that combinations have a thin interquartile range, implying that COMB and COMBw generated reliable forecasts, which resulted in stable forecasting errors through the entire hierarchy. Therefore, COMB and COMBw are robust and consistent for generating forecasts across various hierarchies.

Overall, results show that the COMBw approach provides the most accurate forecasts on average. It is calculated based on the simple average of RMSE across all levels. Moreover, the COMBw approach outperforms all other HF approaches in all nodes of the hierarchy. COMBw was closely followed by COMB, OC, BU and TD4 models. The accuracy of TD1, TD2, TD3 and TDFP was far behind the accuracy of the COMB and COMBw models (Figs. 2 and 3). COMB and COMBw also presented good results in terms of forecast bias but failed to outperform the OC, although the differences were not statistically significant (Fig. 4). We observe that the forecast bias generated by COMB and COMBw approaches is 0.78% and 3.05% higher than the OC model, respectively (Table A.1).

4.2.3. Relative forecasting performance between competing models

From all individual approaches, OC model generates the most accurate forecasts in the simulation study, with least bias (Tables 1 and A.1). Therefore, we use it as a benchmark to compare the forecast accuracy improvement of other HF models and proposed combinations. To that end, we use the average relative Mean Absolute Error (AvgRelMAE) proposed by Davydenko and Fildes (2013). In order 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:

$$AvgRelMAE = \left(\prod_{i=1}^{m} r_i\right)^{1/m} ;\ r_i = \frac{MAE_i^f}{MAE_i^s}. \tag{12}$$

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. For a benchmark model, we choose OC, which implies that $MAE_i^s = MAE_i^{OC}$. The results of the AvgRelMAE comparisons are shown in Table 2.

Table 2. The AvgRelMAE forecasting performance of all HF models compared to the OC model[k].

[k] Best results are bolded.

Level	Node	AvgRelMAE								
		BU	TD1	TD2	TD3	TD4	OC	TDFP	COMB	COMBw
Level 0	Total	1.0470	1.0917	1.0917	1.0917	1.0917	1.0000	1.0917	1.0026	**0.8781**
Level 1	A	1.0163	1.5833	1.5614	1.8753	1.1247	**1.0000**	1.1252	1.0106	1.0063
	B	1.0190	1.6408	1.6142	1.9612	1.1180	**1.0000**	1.1195	1.0051	1.0028
	C	1.0181	1.5340	1.5116	1.8376	1.1245	1.0000	1.1222	1.0052	**0.9972**
	D	1.0180	1.5805	1.5584	1.8670	1.1144	1.0000	1.1129	1.0047	**0.9804**
Level 2	AA	**0.9919**	1.4318	1.4061	1.6798	1.0797	1.0000	1.1969	1.0022	0.9992
	AB	**0.9953**	1.4495	1.4283	1.7549	1.0919	1.0000	1.2035	1.0042	1.0006
	BA	0.9937	1.4683	1.4421	1.7478	1.0778	1.0000	1.2240	0.9991	**0.9935**
	BB	0.9970	1.5305	1.5046	1.8179	1.0799	1.0000	1.2273	0.9983	**0.9903**
	CA	0.9898	1.4151	1.3900	1.7141	1.0731	1.0000	1.2296	0.9979	**0.9861**
	CB	0.9887	1.4444	1.4194	1.6982	1.0822	1.0000	1.2572	0.9997	**0.9855**
	DA	0.9897	1.4284	1.4049	1.7027	1.0813	1.0000	1.1908	0.9994	**0.9816**
	DB	0.9967	1.5154	1.4898	1.7936	1.0853	1.0000	1.1986	1.0009	**0.9815**
	Average	1.0047	1.4703	1.4479	1.7340	1.0942	1.0000	1.1769	1.0023	**0.9833**

Table 2 represents the increase or decrease of MAE forecasting error of different HF models, compared to the forecasts of the OC model in every node. AvgRelMAE is easily interpretable, as it represents the average relative value of MAE adequately, and directly shows how the observed model improves/reduce the MAE compared to the baseline statistical forecast. Obtaining AvgRelMAE < 1 means that on average $\mathrm{MAE}_{\mathrm{i}}^{\mathrm{f}} < \mathrm{MAE}_{\mathrm{i}}^{\mathrm{s}}$, and therefore the observed model improves the accuracy, while AvgRelMAE > 1 indicates the opposite (Davydenko & Fildes, 2013).

Results in Table 2 demonstrate that COMBw approach outperforms the OC and all other HF approaches based on this metrics. COMBw generated consistently good forecasts, outperforming other models in the majority of hierarchy nodes. For easier interpretation of overall forecasting performance in terms of AvgRelMAE, we transform the last row of Table 2 to the percentage scale (Tabel 3). The average percentage improvement in MAE of forecasts is found as (1 − AvgRelMAE) × 100. Positive values in Table 3 indicate the forecast improvement, while negative suggests the opposite. Results demonstrate that by average COMBw reduces MAE forecasting error in the amount of 1.66%, compared to the forecasts of the OC model. COMB and BU also show good results, but on average they increase MAE forecasting error for the amount of 0.23% and 0.47%, respectively. All other models generate forecasts that are significantly less accurate than forecasts of OC in each node of the hierarchy.

Table 3. The average percentage improvement in MAE of all HF models compared to the OC model[l].

	(1 − AvgRelMAE) × 100 (%)								
	BU	TD1	TD2	TD3	TD4	OC	TDFP	COMB	COMBw
Average	-0.47	-47.02	-44.78	-73.39	-9.41	0.00	-17.68	-0.23	**1.66**

Results in Figs. 2 and 3 supplemented with Tables 1, 2 and 3 indicate that combining the forecasts of OC, BU and TD4 provides more accurate forecasts than any of individual HF models. In terms of forecast bias, it appears that there is a still place for improvement of combination approaches since, COMB and COMBw generated forecasts with higher bias, compared to the OC (Fig. 4 and Table A.1). Overall, the results of a simulation study show that HF combinations could offer substantial benefit in terms of HF forecasting.

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. There are only a few examples linking forecasting to various parts of SCs (Mircetic, 2018; Mircetic et al., 2017; Pennings & van Dalen, 2017; Rostami-Tabar et al., 2015; Seongmin et al., 2012; 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. The empirical study is performed to evaluate the effectiveness of different approaches in a real SC network. Additionally, we also examine GF combinations in SC which has never been investigated.

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. 5 illustrates a distribution structure of the brewery company operating in the South-East of Europe. 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 & 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

[l] Best results are bolded.

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 picking to the 14 000 pallets of different brewery products.

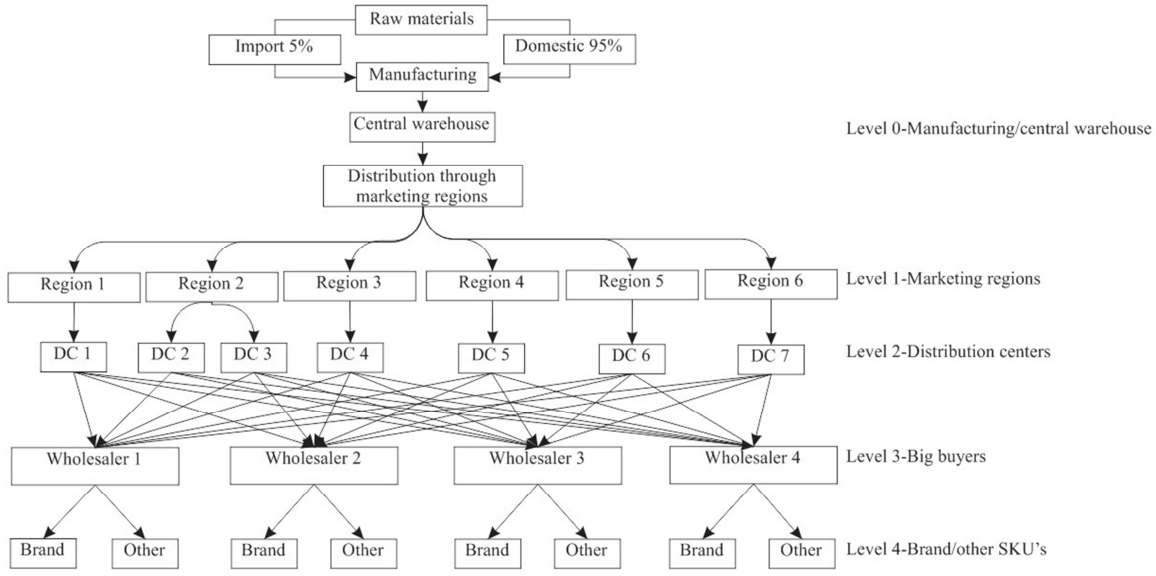

Fig. 5. Multi-echelon brewery distribution network.

The brewery industry is specific in the sense that all of the consumption of the products is accomplished via bars, restaurants and retail shops. There are no direct deliveries and internet 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 is acquiring 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.

Main reasons for using the GF in SCs is to simplify forecasting process, obtain more accurate forecasts, harmonise forecasts from different levels and to provide all information needed for different SC parties. Therefore, in this empirical study, the forecasting structure is designed to

generate forecasts that support the planning and execution of the processes in different parts of SC. Special attention is addressed to the alignment of the time component as well, and not just on 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. The demand in a given multi-echelon brewery distribution chain has the grouped structure. The structure is provided in Table 4, where each row denotes the level of disaggregation.

Table 4. 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_6	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. 6. Results show that series are non-stationary, with a weak trend and pronounced seasonality.

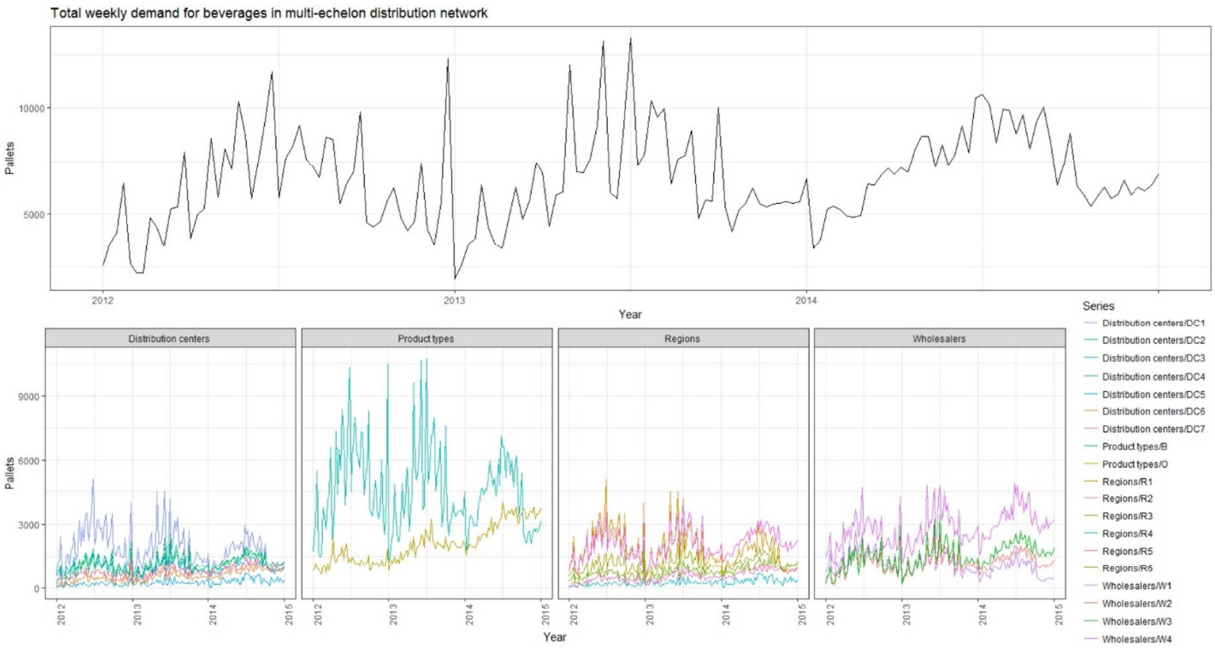

Fig. 6. Total weekly demand, disaggregated by marketing regions, DCs, wholesalers and product types.

5.3. Data and empirical design of experiment

The forecast horizon is equal to the lead time of the decisions driven by the forecast. 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 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 4. To report the forecast performance, we use RMSE, MPE and AvgRelMAE performance metrics[m].

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 in terms of reducing the MAE, compared to the other competing models. For forecasting the grouped brewery demand structure, we only keep BU and OC 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. 7 shows that all models produce similar forecasts, compared by RMSE.

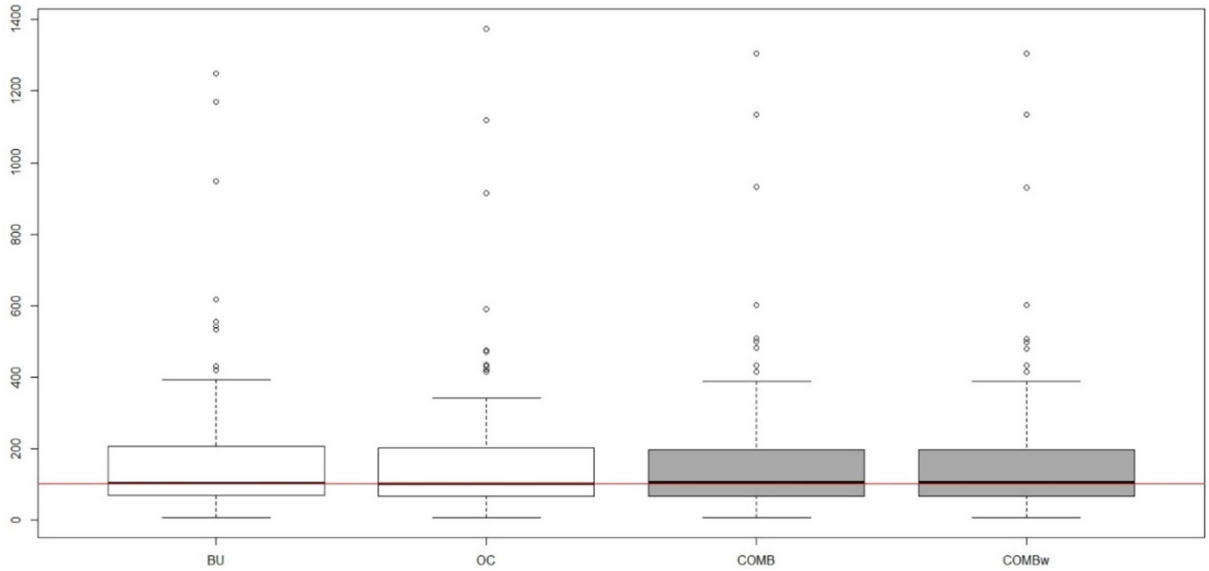

Fig. 7. Box plots for RMSEs of different models tested on the multi-echelon brewery distribution chain[n].

We observe that the OC model performs slightly better than others followed by COMBw, COMB and BU (please refer to Table A.2 in Appendix for details). However, its performance is not significantly

[m] Due to the space restrictions, we here only present RMSE, MPE and AvgRelMAE errors. Additional forecasting error metrics (MAPE, MAE and ME), are also available from the corresponding author on request.

[n] The red line in the figure represents the median value of the RMSE forecasting error of the OC model. Grey box plots represent the performance of combined HF models (COMBw and COMB).

different from other three approaches, as the Nemenyi post-hoc test failed to identify important discrepancy among forecasts of OC, COMBw, COMB and BU.

Fig. 8 presents the performance of BU, OC, COMB and COMBw approaches across all levels in the forecasting structure presented in Table 4. It indicates that all approaches generate similar forecasts in all levels and nodes of the forecasting structure. We develop a shiny platform[o] that allows users to further compare and visualise the performance of these approaches through different forecasting metrics, using real data from the multi-echelon brewery distribution chain.

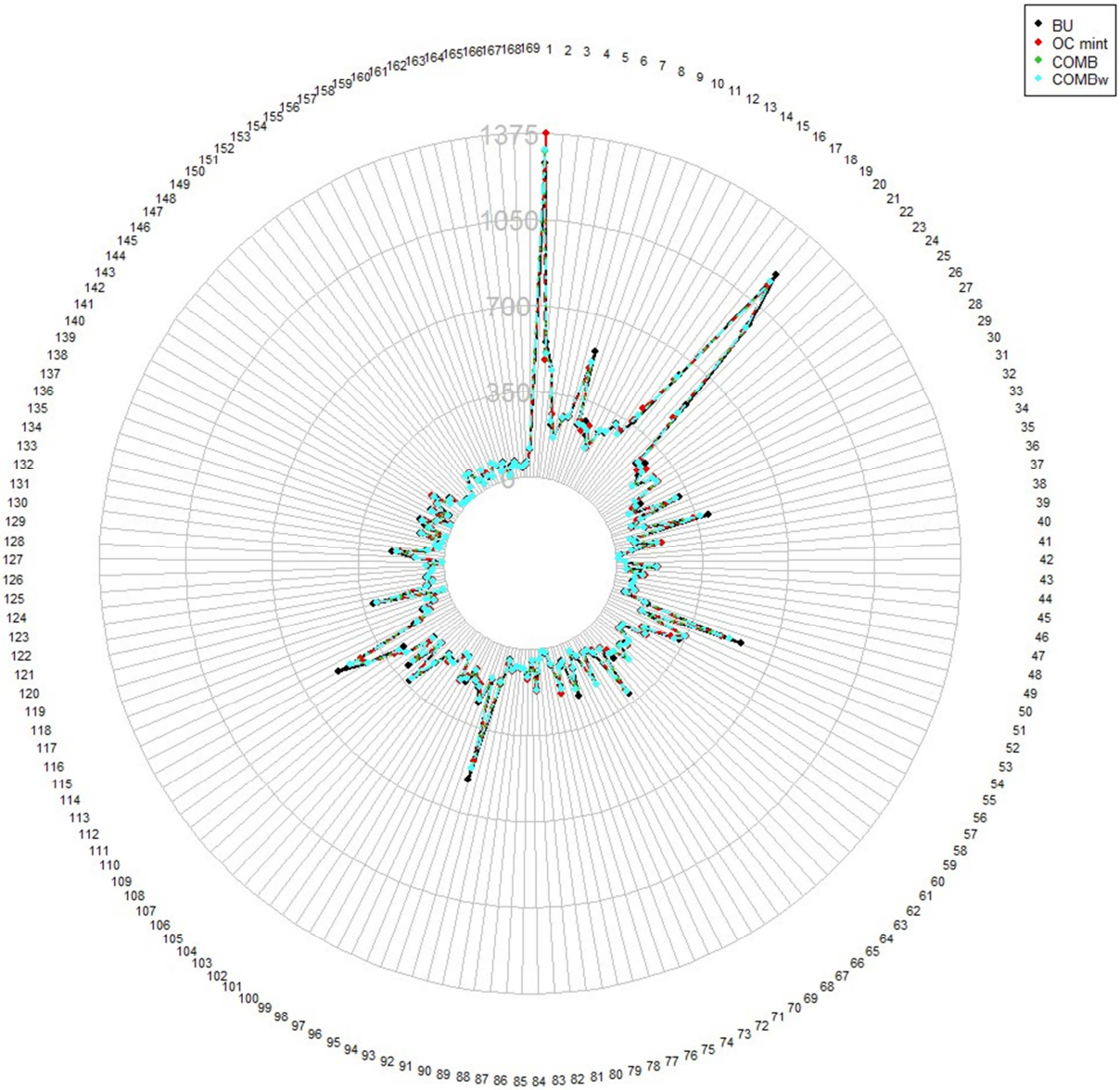

Fig. 8. Performance of different models through the grouped levels while forecasting the demand in the brewery distribution chain[p].

[o] https://dejanmircetic.shinyapps.io/empirical_beverage_study/

[p] Numbers on the diagram perimeter represent the nodes in the grouped forecasting structure. To connect the

The situation is almost identical in a term of forecast bias, where OC generated the smallest bias, measured by the MPE (please refer to Table A.3 in Appendix for details). Fig. 9 presents the performance of the different models measured by MPE.

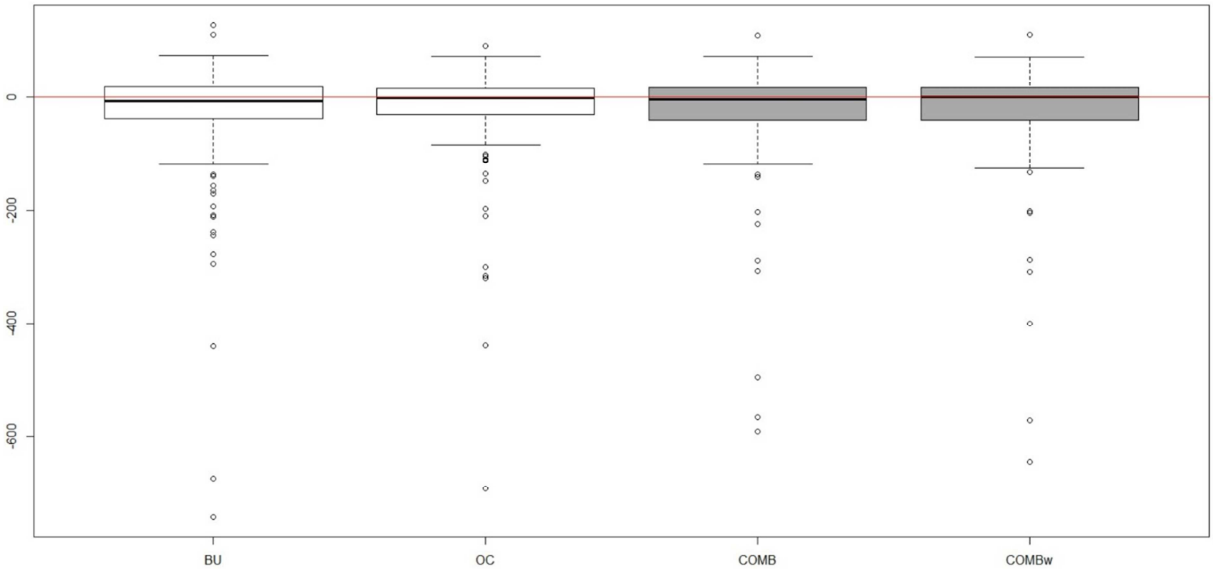

Fig. 9. MPE forecasting error of different models in the empirical study[q].

The figure indicates that COMBw, COMB and BU models closely follow and have similar performance as the OC model. As in the case of RMSE, the Nemenyi post-hoc test failed to identify important differences among MPE forecasts of OC, COMBw, COMB and BU.

Figs. 7, 8 and 9 show that combining the forecasts of OC and BU through two different combination approaches (Eq. 8 and Eq. 9) produce consistently accurate forecasts through all nodes of the considered brewery SC. Moreover, COMB and COMBw forecasts outperform forecasts of BU but failed to beat the OC, although the differences in performance is tiny (please refer to Tables A.2 and A.3 for details).

5.4.2. Relative forecasting performance improvement of the combination approaches

In this subsection, we summarise the accuracy performance of combination approaches in a brewery SC. Due to the specific features of SKU-level demand data, many well-known error measures are not appropriate. In order to overcome the disadvantages of existing measures, Davydenko and Fildes (2013) recommended that an AvgRelMAE should be used.

Since the OC model generated the most accurate forecasts according to RMSE and also produced least biased forecasts based on MPE in the empirical study (please refer to Tables A.2 and A.3 for

label of the node to the number in the diagram, please refer to columns 3 and 4 in Table A.2.

[q] The red line in a figure represents the median value of the MPE forecasting error of the OC model. Grey box plots represent the performance of combined HF models (COMBw and COMB).

details), we use it as a benchmark to compare the forecast accuracy improvement of BU and proposed combination approaches (please refer Table A.4 in the Appendix for details). Table A.4 presents the increase or decrease of MAE forecasting error of different HF models, compared to the forecasts of the OC model in the whole grouped brewery SC.

Results reveal that the OC model has outperformed others, generating the most accurate MAE forecasts, although the Nemenyi post-hoc test failed to identify important differences between models. COMBw and COMB generated consistently accurate forecasts in every node of the group structure and had the closest performance to the OC model. BU also produce good forecasts but it was outperformed by the OC, COMBw and COMB (Table 5). Table 5 presents the summarised average percentage improvement in MAE of competing models, found as (1 − AvgRelMAE) × 100).

Table 5. The average percentage improvement in MAE of all GF models compared to the OC model[r].

	(1 − AvgRelMAE) × 100 (%)			
	BU	OC	COMB	COMBw
Average	-3.9441	0.0000	-0.4961	**-0.4908**

Negative values in Table 5 indicate the reduction in forecast accuracy, compared to the OC model. On average, COMBw and COMB generated approximately 0.5% higher MAE compared to the OC model. BU forecasts were notably more inaccurate than forecasts of OC and on average produce 3.94% higher MAE errors than the OC model. Therefore, forecasts generated by the GF combinations demonstrated accurate results 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.

In this study, we have evaluated the performance of GF models via statistical metrics. The performance of the forecasting methods 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.

[r] Best results are bolded.

6. THE EFFECT OF TIME SERIES CHARACTERISTICS ON THE PERFORMANCE OF FORECASTING APPROACHES

In Section 4, we discuss that the performances 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 showed similar forecasts. In both studies (simulation and empirical), differences between their performances become more apparent in the lower levels of the hierarchy, especially in the bottom level. Therefore, we investigate whether there is any connection between the performance of different HF models and characteristics of the series in the bottom level.

To the best of our knowledge, this is the first study that develops a model to evaluate the effect of time series characteristics on the forecasting performance of different HF models. 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 observed variable Y (Eq. 13).

$$Y = f(X) + \varepsilon. \tag{13}$$

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. 14.

$$Y = \beta_0 + \beta_1 X_1 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \tag{14}$$

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.

This is the first research that investigates the impact of series characteristics and model performance in the context of hierarchical forecasting. In this initial research, we are more interested in developing a model that is easy to understand 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 simple linear model instead of complex nonlinear ones. There are no restrictions in the settings that would limit the inclusion of complex nonlinear models as well, but it will influence on 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 in the bottom levels 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 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 6 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. We use given characteristics and added additional 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:

$$correlation(bts-top) = r_j = \frac{\sum_{t=1}^{T}(y_{j,t}-\overline{y_{j,t}})(y_t-\overline{y_t})}{\sqrt{\sum_{t=1}^{T}(y_{j,t}-\overline{y_{j,t}})^2}\sqrt{\sum_{t=1}^{T}(y_t-\overline{y_{t,}})^2}}; \quad (15)$$

$$aggregate\ share = AS_j = \frac{\sum_{t=1}^{T} y_{j,t}}{\sum_{t=1}^{T} y_t}; \quad (16)$$

$$Coefficient\ of\ variation = \frac{\delta}{\mu} = \frac{\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(y_{j,t}-\frac{1}{T}\sum_{t=1}^{T} y_{j,t})^2}}{\frac{1}{T}\sum_{t=1}^{T} y_{j,t}}. \quad (17)$$

Where $\overline{y_{j,t}}$ represents the mean value of 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 $t = 1,\ldots, T$ and $j = 1, \ldots, n$.

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 bottom level series and the top aggregate series. *Aggregate share* measures the participation of the observed series from the

bottom level in top aggregate series. It provides the information about how "big" or "small" is observed series in the given hierarchy.

Table 6. 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 of curvature)	-0.0053	2.6101	-0.0361	-7.0502	6.9923	14.0424	0.0294	-0.0224
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 6 provides the summary statistics for the time series characteristics that are used as input in the statistical learning algorithms. Fig. A.1 in Appendix provides the distributions for the time series characteristics. Distributions have different shapes, but the majority of the time series characteristics have 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, Witten, Hastie, & Tibshirani, 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 elevated 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. In order to obtain more accurate estimates of the coefficients, 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 7.

Table 7. The effect of time series characteristics on the performance of HF approaches.

	Time series characteristics	BU	TD1	TD2	TD3	TD4	OC	TDFP
1	*Intercept*[s]	5.97	9.93	9.75	12.24	6.49	5.87	9.79
2	*Lumpiness* (Variance of annual variances of remainder)	0.39	-4.36	-4.30		0.49	0.43	
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 7, coefficients are presented only when the effect of each time series characteristics is statistically significant (*p-value* < 0.05)[t]. Each column represents a separate regression model created

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

[t] We restrict extrapolation of the findings regarding the influence of time series characteristics on the

for the different HF model. For different HF models, different characteristics found to be significant. Table 7 shows that *lumpiness Lshift, trend, linearity, curvature, season, correlation, aggregate share* and *coefficient of variation* are among time series characteristics that impact the performance of majority HF models in the bottom level. Positive coefficients indicate that they contribute to the increase of the forecasting error, while negative indicates the effect of decreasing the forecasting error. Therefore HF models have a tendency of producing more inaccurate forecasts while forecasting the time series with a higher values of *Lshift, linearity, curvature, coefficient of variation* and *aggregate share*.

Other characteristics have a mixed influence on different HF models. Accordingly, *lumpines, entropy, Vchange*, *trend* and *correlation* have an increasing effect on the accuracy of standard TD approaches[u] (for those that it was significant), while for the majority of remaining HF models, they have a decreasing effect on the accuracy. In contrast, higher values of *season* variable increase forecast accuracy for the majority of HF models and decrease for standard TD approaches. *Fspots* and *KLscore* have only a significant impact on the standard TD approaches, where *Fspots* decreases and *KLscore* increases the forecasting accuracy. Other characteristics have less effect on the accuracy of HF models and their effect in not holistic since they don't have statistically significant influence for the majority of models.

The last row in Table 7 represents 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.

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. The current practice in SC is to generate separate forecasts using univariate forecasting methodologies to support different decisions. Univariate 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

forecasting performance, only on those time series which have similar or closely related summary statistics to the data provided in Table 6 and Fig. A.1.

[u] By standard TD approcses we mean TD1, TD2 and TD3 models.

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: 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 a combination of these models improves the forecast accuracy (COMB and COMBw). Moreover, we investigate the impact of time series characteristics on the effectiveness of each HF approach.

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

- First, Optimal combination approach (with minimum trace estimator), following by 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 and TDFP outperform other TD methodologies considered in this study. These approaches also seem to be more robust to the hierarchy level and time series characteristics.
- Third, our results show that forecast combination of the existing BU, TD and OC models improve the forecast accuracy. We propose two simple combination approaches based on existing models. They are at least as accurate as forecasts generated by BU and OC approaches. Therefore, when dealing with hierarchical and grouped demand structures in a SC, we recommend practitioners to use a combination of models 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 effect of time series characteristics on 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

impact 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, 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, finances 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 interface between temporal and hierarchical or grouped aggregation has received minimal attention in both academia and industry and this is an issue that we plan to investigate in the next steps of our research. Creating a HF/GF methodology to produce the forecasts in the hierarchies or grouped structures with horizontal connections between nodes within the same level is an interesting avenue for further research. Current models can only produce forecasts in the structures, which have vertical connections between nodes in different levels. These kind of hierarchies or groups with horizontal and vertical connections are useful in the SC.
- 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 impact 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

Aigner, D. J., & Goldfeld, S. M. (1973). Simulation and aggregation: a reconsideration. *The Review of Economics and Statistics*, 114-118.

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting, 25*(1), 146–166.

Babai, M. Z., Ali, M. M., & Nikolopoulos, K. (2012). Impact of temporal aggregation on stock control performance of intermittent demand estimators: Empirical analysis. *Omega, 40*(6), 713-721.

Ballou, R. H. (2004). *Business Logistics/Supply Chain Management-Planning, Organizing, and Controlling the Supply Chain* (Fifth edition ed.): Pearson/Prentice Hall.

Barnea, A., & Lakonishok, J. (1980). An analysis of the usefulness of disaggregated accounting data for forecasts of corporate performance. *Decision Sciences, 11*(1), 17-26.

Boylan, J. (2010). Choosing levels of aggregation for supply chain forecasts. *Foresight: The International Journal of Applied Forecasting*(18), 9-13.

Caplice, C., & Sheffi, Y. (2006). ESD.260J Logistics Systems. *(Massachusetts Institute of Technology: MIT OpenCourseWare), http://ocw.mit.edu. License: Creative Commons BY-NC-SA.*

Chen, A., & Blue, J. (2010). Performance analysis of demand planning approaches for aggregating, forecasting and disaggregating interrelated demands. *International Journal of Production Economics, 128*(2), 586-602.

Chen, A., Yang, K., & Hsia, Z. (2008). Weighted least-square estimation of demand product mix and its applications to semiconductor demand. *International Journal of Production Research, 46*(16), 4445-4462.

Chen, H., & Boylan, J. E. (2007). Use of individual and group seasonal indices in subaggregate demand forecasting. *Journal of the Operational Research Society, 58*(12), 1660-1671.

Chopra, S., & Meindl, P. (2007). *Supply chain management-Strategy, Planning, and Operation* (3rd ed.). New Jersey: Pearson Prentice Hall.

Collins, D. W. (1976). Predicting earnings with sub-entity data: Some further evidence. *Journal of Accounting Research*, 163-177.

Dangerfield, B. J., & Morris, J. S. (1992). Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting, 8*(2), 233-241.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. *International Journal of Forecasting, 29*(3), 510-522.

Dunn, D. M., Williams, W. H., & DeChaine, T. (1976). Aggregate versus subaggregate models in local area forecasting. *Journal of the American Statistical Association, 71*(353), 68-71.

Dunn, D. M., Williams, W. H., & Spivey, W. A. (1971). Analysis and prediction of telephone demand in local geographical areas. *The Bell Journal of Economics and Management Science*, 561-576.

Edwards, J. B., & Orcutt, G. H. (1969). Should aggregation prior to estimation be the rule? *The Review of Economics and Statistics*, 409-420.

Fliedner, G. (1999). An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research, 26*(10), 1133-1149.
Fliedner, G. (2001). Hierarchical forecasting: issues and use guidelines. *Industrial Management & Data Systems, 101*(1), 5-12.
Gordon, T. P., Morris, J. S., & Dangerfield, B. J. (1997). Top-down or bottom-up: Which is the best approach to forecasting? *The Journal of Business Forecasting, 16*(3), 13.
Gross, C. W., & Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting, 9*(3), 233-254.
Grunfeld, Y., & Griliches, Z. (1960). Is aggregation necessarily bad? *The Review of Economics and Statistics*, 1-13.
Hyndman, R. J., Ahmed, R. A., & Athanasopoulos, G. (2007). *Optimal combination forecasts for hierarchical time series*. Department of Econometrics and Business Statistics. MONASH University. Retrieved from http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/
Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis, 55*(9), 2579–2589.
Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*: OTexts.
Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*: OTexts.
Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., . . . Yasmeen, F. (2018). forecast: Forecasting functions for time series and linear models, R package version 8.3.
Hyndman, R. J., Lee, A. J., & Wang, E. (2016). Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis, 97*, 16-32.
Hyndman, R. J., Wang, E., & Laptev, N. (2015). *Large-scale unusual time series detection.* Paper presented at the Data Mining Workshop (ICDMW), 2015 IEEE International Conference on IEEE.
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.
Kahn, K. B. (1998). Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting, 17*(2), 14.
Kinney, W. R. (1971). Predicting earnings: entity versus subentity data. *Journal of Accounting Research*, 127-136.
Mircetic, D. (2018). *Boosting the performance of top down methodology for forecasting in supply chains via a new approach for determining disaggregating proportions.* (Ph.D.), University of Novi Sad, Serbia.
Mircetic, D., Nikolicic, S., Stojanovic, D., & Maslaric, M. (2017). Modified top down approach for hierarchical forecasting in a beverage supply chain. *Transportation research procedia, 22*, 193-202.
Pennings, C. L., & van Dalen, J. (2017). Integrated hierarchical forecasting. *European Journal of Operational Research, 263*(2), 412-418.
Pohlert, T. (2015). The pairwise multiple comparison of mean ranks package (PMCMR).
Rostami-Tabar, B. (2013). *ARIMA demand forecasting by aggregation.* Université Sciences et Technologies-Bordeaux I.
Rostami-Tabar, B., Babai, M. Z., Ducq, Y., & Syntetos, A. (2015). Non-stationary demand forecasting by cross-sectional aggregation. *International Journal of Production Economics, 170*, 297-309.
Schwarzkopf, A. B., Tersine, R. J., & Morris, J. S. (1988). Top-down versus bottom-up forecasting strategies. *The International Journal Of Production Research, 26*(11), 1833-1843.
Seongmin, M., Hicks, C., & Simpson, A. (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, H. L., & Hyndman, R. J. (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics, 26*(2), 330-343.

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

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

Teunter, R. H., Babai, M. Z., Bokhorst, J. A., & Syntetos, A. A. (2018). Revisiting the value of information sharing in two-stage supply chains. *European Journal of Operational Research, 270*(3), 1044-1052.

Trapero, J. R., Cardos, M., & Kourentzes, N. (2019). Empirical safety stock estimation based on kernel and GARCH models. *Omega, 84*, 199-211.

Trapero, J. R., Kourentzes, N., & Fildes, R. (2012). Impact of information exchange on supplier forecasting performance. *Omega, 40*(6), 738-747.

Turbide, D. (2015). How can distribution requirements planning help inventory management? Retrieved 16.04.2016

Villegas, M. A., & Pedregal, D. J. (2018). Supply chain decision support systems based on a novel hierarchical forecasting approach. *Decision Support Systems, 114*, 29-36.

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

Wang, X., Smith-Miles, K., & Hyndman, R. (2009). Rule induction for forecasting method selection: Meta-learning the characteristics of univariate time series. *Neurocomputing, 72*(10-12), 2581-2594.

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