Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Machine learning applications in hierarchical time series forecasting: Investigating the impact of promotions

Mahdi Abolghasemi [a,*], Garth Tarr [b], Christoph Bergmeir [c]

[a] School of Mathematics and Physics, The University of Queensland, Australia
[b] School of Mathematics and Statistics, The University of Sydney, Australia
[c] Department of Data Science & AI, Monash University, Australia

ARTICLE INFO

Keywords:
Hierarchical time series
Promotions
Machine learning
Forecasting
Supply chain

ABSTRACT

Hierarchical forecasting is needed in many situations in the supply chain to support decision making. Top-down, bottom-up, and optimal linear combination methods are common in hierarchical forecasting. There is no universally optimal solution for hierarchical forecasting, and each method has some advantages and disadvantages. While top-down and bottom-up methods use only the information at the top and bottom levels, respectively, linear combinations use the individual sales forecasts from all series and levels and combine them linearly, often outperforming the conventional top-down and bottom-up methods. These methods do not directly utilise the explanatory information such as price and promotion status that may be available across different levels in the hierarchy, and their performance may be impacted by these external factors. We propose to use a multi-output regression model that utilises the explanatory variables from across hierarchical levels to simultaneously generate forecasts for all the series at the bottom level. We perform an in-depth analysis of 55 sets of fast-moving consumer goods time series and 3049 products of the M5 forecasting competition data. Our results show that our proposed algorithm effectively utilises explanatory variables from across the hierarchy to generate reliable forecasts for different hierarchical levels, especially in the presence of deep promotional discounts.

1. Introduction

Hierarchical time series represent multiple time series that are hierarchically organised and can be aggregated and disaggregated at different levels with respect to various features (Hyndman, Ahmed, Athanasopoulos, & Shang, 2011). Cross-sectional (Spiliotis, Abolghasemi, Hyndman, Petropoulos, & Assimakopoulos, 2021), temporal (Athanasopoulos, Hyndman, Kourentzes, & Petropoulos, 2017), and cross-temporal (Kourentzes & Athanasopoulos, 2019) hierarchies are three widely developed hierarchical approaches in the literature.

Forecasting in the supply chain is intrinsically hierarchical. Hierarchical forecasting (HF) is often required in the supply chain for making different decisions. For example, from a manufacturer's perspective, forecasting is required at the top level for material resource planning. At a lower level, forecasting may be required for product transhipment to warehouses, order fulfilment, and inventory control at stores. Forecasts can be generated either directly or via different HF methods. If we forecast series directly and independently, they may not add up properly across different levels in the hierarchy because of the hierarchical constraints at different levels. Therefore, various HF methods have been proposed to ensure that forecasts do not violate the constraints and are consistent across different hierarchical levels, i.e., that the forecasts are coherent. Utilising HF methods can potentially improve

[*] Corresponding author.
E-mail addresses: m.abolghasemi@uq.edu.au (M. Abolghasemi), garth.tarr@sydney.edu.au (G. Tarr), Christoph.bergmeir@monash.edu (C. Bergmeir).

https://doi.org/10.1016/j.ijforecast.2022.07.004
0169-2070/© 2022 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

forecast accuracy across different levels in comparison with the direct forecast. However, this depends on the type of HF method chosen and on the particularity of the problem (Nenova & May, 2016).

Top-down (TD), Bottom-up (BU), and linear combination approaches are the common HF methods in the literature. TD and BU methods are the most simplistic and utilise rather simple techniques to generate the final forecasts in the hierarchy. TD approaches only produce forecasts at the top level and disaggregate them to lower levels using some disaggregation factor, whereas BU methods produce forecasts for all bottom-level series individually and aggregate them to compute the forecasts for higher-level series in the hierarchy. Linear combination is a recently proposed method that was subsequently improved (Athanasopoulos, Ahmed, & Hyndman, 2009; Hyndman et al., 2011; Wickramasuriya, Athanasopoulos, & Hyndman, 2019). As the name suggests, the method essentially linearly combines the individual forecasts that are produced for each node in the hierarchy. This method, known as trace minimisation (MinT), minimises the total forecast variance by minimising the variance of the forecasts errors located on the trace of the variance–covariance matrix. MinT has been shown to be a top-performing model in various settings and forecasting problems (Hyndman et al., 2011; Spiliotis et al., 2021).

The efficacy of HF methods depends on the time series features, the level of forecasting in the hierarchy, the forecasting horizon, and the structure of the hierarchy, and one may consider these variables when choosing the most appropriate HF method (Abolghasemi, Hyndman, Spiliotis, & Bergmeir, 2022; Nenova & May, 2016). One important factor that often exists in sales forecasting problems is promotions. Promotions are popular in the fast-moving consumer goods (FMCG) industry to increase sales and drive major changes in the underlying demand behaviour (Nikolopoulos, Litsa, Petropoulos, Bougioukos, & Khammash, 2015). Promotions can enhance sales of different series at different levels of the hierarchy with different scales that are inherently non-linear. That is, different series of the hierarchy may experience different levels of uplifts in sales. For example, the impact of promotional discounts on sales in larger centres may be greater than in smaller stores or regions. Therefore, the level of sales of each series in the hierarchy may significantly vary over time and across different levels of the hierarchy during promotional and non-promotional periods, which consequently may impact the performance of the HF methods. The impact of promotions on sales is an interesting open problem for researchers and practitioners (Abolghasemi, Beh, Tarr & Gerlach, 2020; Kourentzes & Petropoulos, 2016) that has not been investigated thoroughly in the hierarchical time series forecasting literature (Abolghasemi, Hyndman, Tarr, & Bergmeir, 2019). We investigate the impact of promotional discounts on the performance of HF methods to determine the best methods when dealing with hierarchical series that are impacted by promotion.

While each HF method has its pros and cons, the presence or absence of promotional discounts, in particular, can impact their performance. For example, the BU method is shown to be more accurate at the bottom level but it may be prone to overfitting when series are noisy at the bottom levels, such as sales during promotions. TD has proven to be fast and more accurate at the top level of the hierarchy where the forecast is generated directly, but its performance deteriorates as we move to lower levels of the hierarchy, especially when parent–child nodes in a hierarchy depict different patterns. The variations of TD methods do not typically consider the time-varying dynamics of sales and take a static average of historical series as the estimated proportions of future values for lower-level series (Abolghasemi et al., 2019). However, this is not realistic, especially when different series in a hierarchy are impacted by various internal and external variables that can change the dynamics of the series across different levels, and the series may not have a similar pattern. To remedy this issue, Abolghasemi et al. (2019) proposed a dynamic model that forecasts the proportions of child nodes from their parent nodes dynamically. However, their approach does not use all information from across the hierarchy. The MinT method has a number of advantages over BU and TD approaches. Most importantly, it considers the forecasts of all series across all levels of the hierarchy, in contrast to the BU approach, which only uses the bottom level series, and the TD method, which only uses the forecasts at the top level of the hierarchy. MinT improves the forecasts across the entire hierarchy, as opposed to the BU method, which works well at the bottom level, and the TD model, which works better at the top level of the hierarchy. While MinT uses forecasts from all levels, it does not have the flexibility to use explanatory variables directly and benefit from external information that might be available. The MinT method makes some assumptions in estimating the forecasting errors of the variance–covariance matrix that may be unrealistic, especially when series are volatile such as series impacted by promotions. Moreover, it may be difficult and computationally expensive to estimate the variance–covariance matrix in the case of a large hierarchy (Pennings & van Dalen, 2017).

In this study, we propose an approach that benefits from the advantages of the BU, TD, and MinT methods and effectively generates forecasts for series in the hierarchy. We propose to use multi-output machine learning (ML) methods in a TD fashion where we use the forecast only from the top level and explanatory information from across the hierarchy to forecast the series at the bottom level. To ensure that the forecasts are coherent across different levels, we propose to aggregate the obtained forecasts of the bottom-level series to compute the forecasts of the higher-level series. Our approach, in spirit, is similar to the TD method introduced by Abolghasemi et al. (2019): it uses sales forecasts at the top level to dynamically generate forecasts for lower levels, it takes advantage of the BU method to produce coherent forecasts for higher levels, and it mimics the approach of a linear combination method insofar as it uses information from across the hierarchy. However, our approach is unique in three ways. First, we use explanatory variables such as price and promotion type from across the hierarchy that are often available in advance. Second, direct forecasts are only generated at the top level, making it computationally efficient while taking advantage of the information across the hierarchy. Third, we generate forecasts for the bottom level using a multi-input multi-output model that inherently takes into account the correlation between series at the bottom levels and is faster than individually generating forecasts for each series at the bottom level. Our model is dynamic in the sense that we retrain the model over the forecasting horizon and generate forecasts for the series at the bottom level on a rolling-origin basis by using the available information across the hierarchy. This allows for more accurate forecasts, in particular when series are exposed to promotions and show time-varying behaviour.

We use common statistical and ML methods to forecast sales at the bottom level of the hierarchy, whereby extreme gradient boosting (XGB), random forests (RF), and lasso regressions are used in a multi-output fashion. While these multi-output regression models are used in various forecasting tasks (Li, Nan, & Zhu, 2015; Segal & Xiao, 2011; Zhai, Yao, & Zhou, 2020), they have not been sufficiently explored in sales forecasting, and specifically HF. Since in hierarchical time series we have a parent–child structure for series, it is natural to use multi-output regression models where child nodes can be considered as the outputs. We empirically validate our model and compare our approach with current state-of-the-art models.

We conduct a numerical analysis on 55 sets of hierarchical sales time series of a food manufacturing company in Australia with a three-level supply chain, and a hierarchical subset of the M5 forecasting competition with 3049 sets of hierarchical series organised in three levels (levels 10, 11, and 12 of the M5 competition). We generate forecasts for different levels of the hierarchy. The obtained forecasts from this hierarchy are helpful for various decisions across different levels of the supply chain. First, the forecasts are needed at the top level for strategic plans such as material resource planning and demand planning. Second, the forecasts are needed at the middle level (retailers) for production planning and inventory control. Third, the manufacturer needs to forecast at the lowest level of the hierarchy for transportation and logistics purposes (Kahn, 1998). The improvement in forecasting accuracy can potentially generate huge savings on operational activities and improve the supply chain performance considerably (Syntetos, Babai, Boylan, Kolassa, & Nikolopoulos, 2016).

The remainder of the paper is organised as follows. Section 2 reviews the literature. Section 3 reviews the common HF methods in the literature and explains our proposed approach in detail. Section 4 explains the investigated datasets, and the experimental setup is described in Section 5. Empirical results, discussions, and conclusions are presented in Sections 6, 7, and 8, respectively.

2. Literature review

The focus of the studies on hierarchical time series has been on finding an appropriate method to generate accurate forecasts across different levels while perpetrating to the hierarchical constraints. While there is no unique solution for this, studies have shown that different factors—such as the forecasting techniques, demand correlation between different series, the level of the hierarchy at which forecasts are needed, and the forecasting horizon—may contribute to favouring one model over other models in hierarchical time series forecasting (Fliedner, 1999; Gross & Sohl, 1990).

Classical HF techniques generally either disaggregate forecasts from the top of the hierarchy to the lower levels (the TD approach) or aggregate forecasts from the bottom levels to higher levels (the BU approach). One downside of the BU approach is that it can be difficult to model and forecast when the bottom-level series are noisy, which can happen with highly disaggregated series. In the case of a large hierarchy, BU approaches can be labour-intensive and computationally expensive. More importantly, information may get lost when aggregating the lower-level series to forecast the upper-level series (Dangerfield & Morris, 1992). TD models are extremely fast, as they need fewer forecasts, i.e., only one forecast at the top level. Forecasts are almost always wrong. Therefore, it is better to do fewer forecasts (Schwarzkopf, Tersine, & Morris, 1988). TD models often are more appropriate to forecast higher-level series, but their accuracy may deteriorate at lower levels of the hierarchy. This is because estimating the accurate proration factor is a challenging problem, and allocating a disaggregation technique potentially introduces some errors to the forecasts, which is the greatest disadvantage of TD models.

Gross and Sohl (1990) studied different disaggregation methods to determine whether it is better to use disaggregated forecasts or to forecast individually. They forecasted sales for 18 series with different models and compared 21 disaggregation factors. These factors include the simple average, moving average with different orders, lagged proportions, lags correlation, regression, a weighted average based on the variance and covariance of forecasting errors, and a combination of these models. They numerically found that TD models outperformed BU models in two of the three investigated products. Their results indicated that the efficacy of TD models depends on the demand feature, demand forecasting accuracy, and disaggregation factor. Conventional TD methods take an average of proportions when disaggregating series (Athanasopoulos et al., 2009). This is not realistic in the presence of promotions, since the proportions of series and the relationship between parent and child nodes may change dramatically, and because taking an average for disaggregation can introduce large biases to the disaggregation factor.

A limitation of TD and BU models is that they start from opposite ends of the hierarchy but do not use forecasts from other levels of the hierarchy, resulting in the loss of useful information (Pennings & van Dalen, 2017). Middle-out is another approach in practice that tries to take advantage of both the BU and the TD methods (Hyndman et al., 2011). This approach is essentially a combination of BU and TD, where an intermediate level is chosen and forecasts are disaggregated to lower-level series and aggregated to higher-level series.

The MinT approach is a more advanced HF method that covers the drawbacks of single-level TD and BU more effectively. In this method, forecasts are computed at all levels (these are called base forecasts) and then combined using a linear regression model to find the final reconciled forecasts (Hyndman et al., 2011). While the MinT approach is a promising model and has strong theoretical and numerical support, it has some disadvantages. For example, MinT assumes that the out-of-sample forecast error is similar to the in-sample forecast error. However, this may not be the case for sales time series that are highly impacted by promotions. Thus, it may not lead to optimal forecasts across the hierarchy. To remedy this issue, we take advantage of the additional information—i.e., explanatory variables available across the hierarchy, such as price, which is known in advance—and train the forecasting models as we move ahead.

Demand characteristics play an important role in the performance of HF methods (Gross & Sohl, 1990). The performance of HF methods such as TD and BU has been investigated when demand follows AR(1), MA(1), and intermittent patterns (Moon, Simpson, & Hicks, 2013). The impact of other demand characteristics such as seasonality, trend, or entropy on the performance of HF methods has also been investigated in time series hierarchical forecasting. Abolghasemi et al. (2022) investigated 32 time series characteristics from various levels of the hierarchy and used ML classification models for selecting the most suitable reconciliation method according to the time series characteristics.

Promotions are a factor that can significantly change the dynamics of a series, but the impact of promotions on the performance of HF methods in cross-sectional hierarchical time series has not been investigated thoroughly. Essentially, in the presence of promotions, decision makers will have additional information that can be used for forecasting, e.g., the time or type of promotion. Kourentzes and Petropoulos (2016) analysed the impact of promotions in temporal hierarchies and proposed MAPAx to effectively generate temporal forecasts in an empirical study. They empirically showed that including promotional information is useful to improve the forecast accuracy, and that using multiple temporal aggregation levels can make the final forecast robust to model misspecification. In a cross-sectional hierarchical time series study, Mancuso, Piccialli, and Sudoso (2021) proposed to use explanatory variables in a deep neural network model for generating reconciled forecasts for the hierarchy in a TD manner. They used explanatory variables in their proposed method where some series were impacted by promotions, but they did not investigate the impact of promotions in detail. They proposed a custom loss function in the training phase and used a dense layer with n units at the end of their architecture to generate coherent forecasts for all n child nodes. However, the generated forecasts were not fully coherent. Rather, they could regulate their forecasts by a parameter that would compromise between the fitting error and network coherency. Spiliotis et al. (2021) proposed to use ML models in a BU fashion to generate forecasts for hierarchical series. They first generated the base forecasts and then used time series values from across the hierarchy and executed ML models to reconcile the forecasts for each of the bottom-level series. Their proposed model is a single-output ML model that uses actual historical observations of all series in a hierarchy as explanatory variables, but they do not use any external variables, such as price, as explanatory variables.

The other important factor in hierarchical time series is the correlation between series, which has been commonly used in product hierarchical forecasting models. Often, product demands in one family are correlated, for example where products are complementary or substitutes to each other. Therefore, it is reasonable to consider the correlation when forecasting their demand (Chen & Boylan, 2009; Pennings & van Dalen, 2017; Widiarta, Viswanathan, & Piplani, 2007). There are various models that attempt to find the most appropriate approach in product hierarchical forecasting by considering the correlation between their demand. However, existing results remain somewhat inconclusive. For example, Schwarzkopf et al. (1988) analytically compared the forecasting accuracy of TD and BU methods in terms of robustness, precision, and unbiasedness. They found that the TD approach is more robust and accurate than BU when product demands are independent. Dangerfield and Morris (1992) compared TD and BU models by considering the correlation between a family of products of the M competition dataset and found that the BU model is superior to the TD model in most cases. Chen and Boylan (2009) reviewed the correlation between demands on inventory hierarchical forecasting. They analysed different series with simulations and found that a negative correlation makes the TD approach superior to others. Fliedner (1999) designed an experiment and found that either a positive or a negative correlation between item demands leads to accuracy improvements in aggregated time series. However, the lack of this correlation leads to a lower accuracy regardless of the utilised forecasting methodology (direct or derived). He also showed that the direct forecast of the aggregated series is more accurate than the derived forecasts. These models deal with the demand correlation between series and fall out of the scope of this study. In this paper, we develop multi-output models to generate forecasts that inherently account for the correlation of output series.

3. Hierarchical forecasting models

In this section, we discuss the TD, BU, and MinT methods as three well-established HF methods, and then detail our proposed method.

Consider a hierarchical time series with k levels, m series at each level, and n observations for each series in total, where $\boldsymbol{Y}_{i,t}$ depicts the vector of all observations at level i, $\hat{\boldsymbol{Y}}_{i,t}(h)$ is the h-step-ahead forecast at level i, $\boldsymbol{Y}_t$ shows a column vector including all observations, $\hat{\boldsymbol{Y}}_n(h)$ shows the h-step-ahead independent base forecast of all series based on n observations, and $\tilde{\boldsymbol{Y}}_n(h)$ shows the final reconciled forecasts of all series.

The hierarchical time series can be represented with $\boldsymbol{Y}_t = \boldsymbol{S}\boldsymbol{Y}_{k,t}$, where $\boldsymbol{S}$ is a summing matrix of order $m \times m_k$ that aggregates the bottom-level series. For example, for a three-level hierarchy with 12 series at the bottom level (A1, …, A6, B1, …, B6), two series at the middle level (A, B), and one at the top level, we have:

$$\begin{bmatrix} Y_t \\ Y_{A,t} \\ Y_{B,t} \\ Y_{A1,t} \\ \vdots \\ Y_{A6,t} \\ Y_{B1,t} \\ \vdots \\ Y_{B6,t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ & & & & & & & & & & & \\ & & & & & \boldsymbol{I}_{12} & & & & & & \\ & & & & & & & & & & & \end{bmatrix} \times \begin{bmatrix} Y_{A1,t} \\ Y_{A2,t} \\ \vdots \\ Y_{A6,t} \\ Y_{B1,t} \\ Y_{B2,t} \\ \vdots \\ Y_{B6,t} \end{bmatrix}$$

We can represent the final reconciled forecasts with a unified structure $\tilde{\boldsymbol{Y}}_n(h) = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{Y}}_n(h)$, where $\boldsymbol{G}$ is a matrix of order $m_k \times m$ whose elements depend on the type of HF method (Hyndman & Athanasopoulos, 2018).

3.1. Bottom-up

The BU method, as mentioned above, generates forecasts for each series at the bottom level individually and aggregates the forecasts of the bottom series to find the forecast for their parent node and for higher levels. As such, we can show that $\boldsymbol{G} = [\boldsymbol{0}_{m_k \times (m-m_k)} | \boldsymbol{I}_{m_k}]'$, where $\boldsymbol{0}_{i \times j}$ is an $i \times j$ null matrix. This approach often performs well at the bottom level. The downside of the BU model is that it can be difficult to model and forecast when the bottom-level series are subject to noise. In the case of a large hierarchy, BU can be labour-intensive and computationally expensive. More importantly, information such as the seasonality or trend of series may be lost when aggregating the lower-level series to forecast the upper-level series (Dangerfield & Morris, 1992).

3.2. Top-down

The TD method only uses the most-aggregated series at the very top level to directly generate forecasts for the top-level node. In order to generate forecasts for lower-level series, the TD method disaggregates the forecasts to the lower levels with some factors. The TD method is specifically useful when the bottom-level series are noisy and difficult to forecast (Kahn, 1998). We can forecast them more accurately by aggregating series at the upper levels of the hierarchy. Different disaggregation factors, such as historical proportions, can be used to obtain the forecast for the lower-level series (Gross & Sohl, 1990). However, static TD methods can be biased over the horizon, especially for our case study where time series behaviour is impacted by promotions.

Athanasopoulos et al. (2009) proposed a method for disaggregating the top-level forecasts based on the forecasted proportions of lower-level series rather than the historical proportions. In this method, the disaggregation factor, p_j, is calculated as follows:

$$p_j = \prod_{i=0}^{k-1} \frac{\hat{Y}^{(i)}_{j,n}(h)}{\sum(\hat{Y}^{(i+1)}_{j,n}(h))}, \qquad j = 1, \ldots, m_k,$$

where $\hat{Y}^{(i)}_{j,n}(h)$ is the h-step-ahead forecast of the series that corresponds to the node which is i levels above j, and $\sum \hat{Y}_{i,n}(h)$ is the sum of the h-step-ahead forecasts below node i that correspond directly to node i.

These will form the vector $\boldsymbol{P}$, $\boldsymbol{P} = [p_1, p_2, p_3, \ldots, p_{m_k}]$, and the matrix $\boldsymbol{G}$:

$$\boldsymbol{G} = [\ \boldsymbol{P} \mid \boldsymbol{0}_{m_k \times (m-1)}]. \tag{1}$$

This type of TD method aims to consider the dynamics of forecasts but does not take into account the varying nature of proportions that might occur. The other downside of this method, like other conventional TD models, is that it generates biased forecasts even if the base forecasts are unbiased. This method has been applied successfully in tourism forecasting and has been shown to outperform the static TD method (Athanasopoulos et al., 2009).

3.3. Optimal combination

The linear combination method uses a regression model to combine the individually generated forecasts for all series at all levels, $\hat{\boldsymbol{Y}}_n(h) = \boldsymbol{S}\boldsymbol{\beta}_n(h) + \boldsymbol{\varepsilon}_t$, where $\boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{W}_h)$ and $\boldsymbol{\beta}_n(h) = \mathrm{E}[\hat{\boldsymbol{Y}}_{k,n}(h)|\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_n}]$ is the unknown mean of the base forecasts of the bottom level k.

Hyndman et al. (2011), Hyndman, Lee, and Wang (2016) showed that we can use generalised least squares to compute the minimum variance unbiased estimator of $\hat{\boldsymbol{Y}}_{k,t}(h)$ as $\hat{\boldsymbol{\beta}}_n(h) = (\boldsymbol{S}'\boldsymbol{W}_h^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{\dagger}\hat{\boldsymbol{Y}}_n(h)$, where $\boldsymbol{W}_h^{\dagger}$ is the generalised inverse of $\boldsymbol{W}_h$. However, $\boldsymbol{W}_h$ is not known and sometimes is impossible to estimate for large hierarchies. Different approximation methods can be used to estimate it. Those authors showed that weighted least squares can be used to estimate $\hat{\boldsymbol{\beta}}_n(h) = (\boldsymbol{S}'\boldsymbol{\Lambda}_h\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{\Lambda}_h\hat{\boldsymbol{Y}}_n(h)$, where $\boldsymbol{\Lambda}_h$ is a diagonal matrix with elements equal to the inverse of the variances of $\boldsymbol{\varepsilon}_h$. Hence, $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}_h^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{\dagger}$, and the variance of revised forecasts is $\mathrm{Var}[\tilde{\boldsymbol{Y}}_n(h)] = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}_h^{\dagger}\boldsymbol{S})^{-1}\boldsymbol{S}'$.

Wickramasuriya et al. (2019) showed that for any $\boldsymbol{G}$ such that $\boldsymbol{SGS} = \boldsymbol{S}$, the covariance matrix of the h-step-ahead reconciled forecast errors is given by $\boldsymbol{V}_h = \mathrm{Var}[\boldsymbol{y}_{t+h} - \tilde{\boldsymbol{Y}}_t(h)] = \boldsymbol{SGW}_h\boldsymbol{G}'\boldsymbol{S}'$, where $\boldsymbol{W}_h$ is the variance–covariance matrix of the h-step-ahead base forecast errors. They showed that the $\boldsymbol{G}$ matrix that minimises the trace of $\boldsymbol{V}_h$ such that $\boldsymbol{SGS} = \boldsymbol{S}$ is given by $\boldsymbol{G} = (\boldsymbol{S}'\boldsymbol{W}_h\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}$. They called this the minimum trace (MinT). While there are different methods to estimate $\boldsymbol{W}_h$, the shrinkage estimator often returns the most accurate forecasts, and as such we only report the results generated by this estimator and denote it as MinT.

3.4. Dynamic top-down bottom-up algorithm

We propose an algorithm that uses explanatory information from all levels in the hierarchy to generate coherent forecasts for the hierarchy dynamically and fast. More specifically, we propose to use multi-output methods to simultaneously generate forecasts for bottom-level series by using external explanatory information from across the hierarchy and direct sales forecasts at the top level. In contrast to the traditional BU method, which uses only information at the bottom level to forecast the series of the bottom level, we use information from higher levels of the hierarchy to forecast the bottom-level series. We use multiple output models to consider the interaction of the bottom-level series and simultaneously generate forecasts for all of the bottom-level series instead of generating a single forecast for each series. We train our models in a rolling-origin fashion to update the models and information as we roll ahead in time. We build a large multi-input multi-output model for the entire hierarchy. The input features of our model are external explanatory variables across different levels and the actual values of sales time series at the top level at time t. The outputs are the actual values of the time series at the bottom level at time t.

This model is flexible in terms of the input variables that can be used to derive the outputs. One can consider various external features or sales information from various levels that is available and may be useful to improve the performance of the model. Also, our proposed approach is flexible in terms of the level at which the multi-input multi-output model can be built. We build our model by using the direct forecast at the top level of the hierarchy as input and then forecasting the bottom-level series, but the model can be modified to use the direct middle-level forecasts as inputs and generate the output for bottom levels. In terms of the computational time, our approach is significantly faster than MinT and BU, as it requires fewer forecasts, but it is slower than the static TD methods.

We considered including other features, such as the lag values of sales, but they did not improve the performance of the model so they were excluded. The rationale for this is that the time series behaviour for each pair of parent–child nodes changes dynamically, e.g., from promotional times to non-promotional times, and using sales time series from previous periods reduces the accuracy of the model. Therefore, we only use the features for the corresponding forecast period.

Our approach is agnostic to the methods used to generate forecasts. In practice, we utilise common statistical and ML methods whereby XGB, RF, and lasso are used to forecast bottom-level series. Algorithm 1 details the steps of our approach, which is called dynamic top-down bottom-up (DTDBU). We summarise the process in Fig. 1.

Algorithm 1 Proposed dynamic top-down bottom-up approach.

```
for $t = p + t$ to $n$ do
    Split the time series of length $n$ into a training set with length $p$ and a test set with length 1.
    Create a dataset for training the multi-input multi-output models. The training set of the model consists of $p$ observations of each series, where input variables include the actual values of sales at the top level and external explanatory variables, e.g., price and promotion status, for all of the series at time $t$, and the output includes the actual values of the series at the bottom level at time $t$.
    Fit and train the multi-input multi-output model of choice on the training dataset.
    Fit a chosen forecasting model to the training set of the top-level series in the hierarchy, and forecast the next $h$-step periods at the top level.
    Use the models developed in Step 4 to forecast bottom-level series. The inputs of the test set are the direct forecasts of top-level series at time $t + 1$, as generated in Step 5, and explanatory variables of all series as per the training set input features at time $t + 1$. The outputs are the forecasts at the bottom level for all series at time $t + 1$.
    Sum up the forecasts at the bottom level to calculate the forecasts at higher levels at time $t + 1$.
    Calculate the forecast accuracy.
Take the average of accuracy metrics in Step 8, measured equally across all levels.
```

Our methodology is similar to Mancuso et al. (2021) in that we use TD models with explanatory variables to generate multiple outputs for child series in a hierarchical time series setting. However, our approach differs in two key respects. Firstly, our method generates fully coherent forecasts for all levels of the hierarchy. A limitation of Mancuso et al. (2021) is that their forecast coherency is compromised with forecast accuracy. That is, we may generate more accurate forecasts for one series or level that may not add up to another level. In contrast, we use information from across all levels of the hierarchy to generate forecasts for the bottom level only, and then sum them up to generate coherent forecasts for higher levels. So, essentially we use a combination of TD and BU approaches as opposed to a TD model. Secondly, we use a different approach for reconciliation. While Mancuso et al. (2021) use neural network architectures, we propose to use any appropriate parametric or non-parametric models in a multi-output fashion to generate forecasts for all series.

Like us, Spiliotis et al. (2021) similarly use ML models to generate forecasts and benefit from a BU approach to produce coherent forecasts for all levels. However, our approach allows the direct use of external explanatory

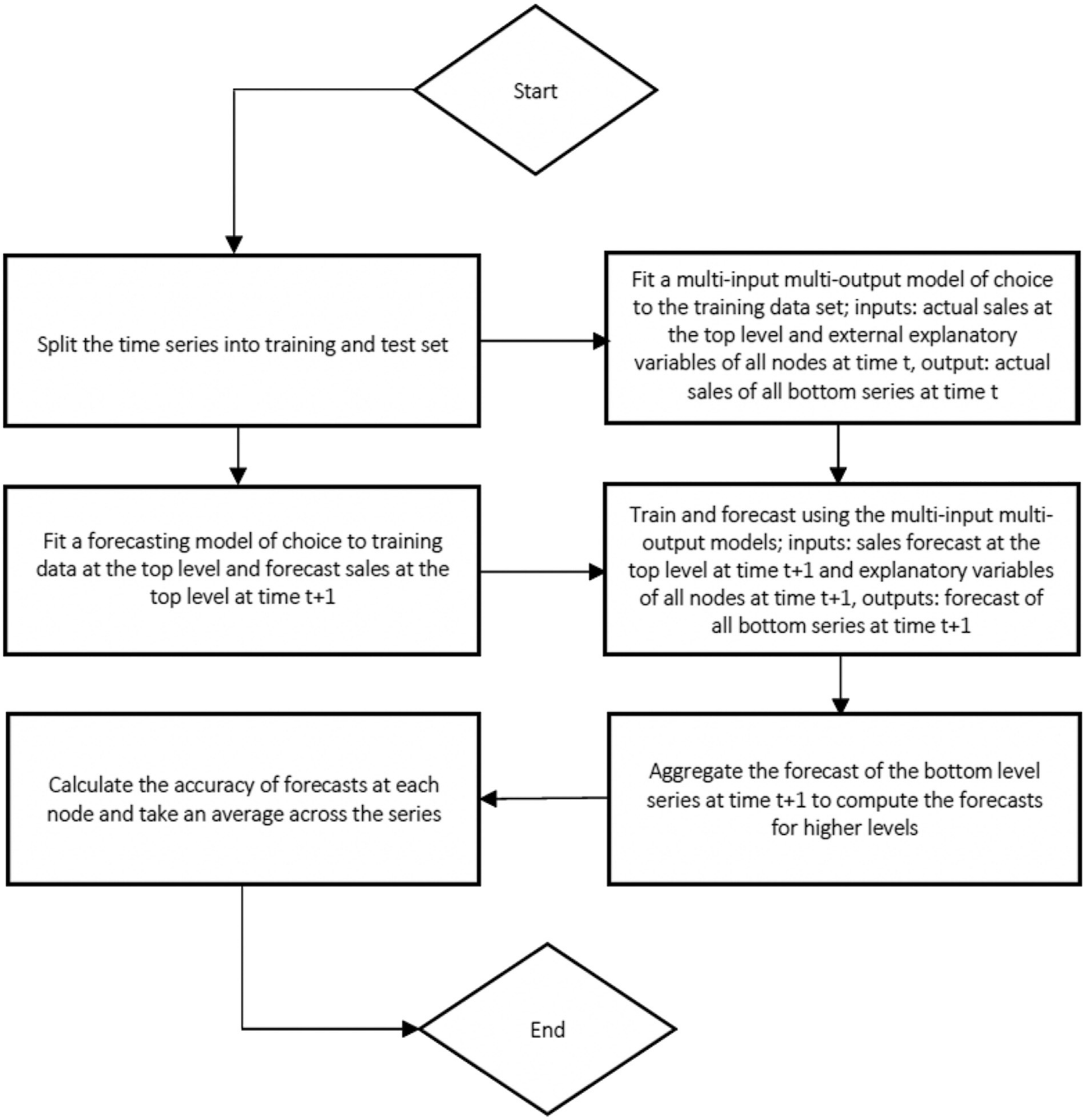
Start
Split the time series into training and test set
Fit a multi-input multi-output model of choice to the training data set; inputs: actual sales at the top level and external explanatory variables of all nodes at time t, output: actual sales of all bottom series at time t
Fit a forecasting model of choice to training data at the top level and forecast sales at the top level at time t+1
Train and forecast using the multi-input multi-output models; inputs: sales forecast at the top level at time t+1 and explanatory variables of all nodes at time t+1, outputs: forecast of all bottom series at time t+1
Calculate the accuracy of forecasts at each node and take an average across the series
Aggregate the forecast of the bottom level series at time t+1 to compute the forecasts for higher levels
End

Fig. 1. Flowchart describing the proposed dynamic top-down bottom-up algorithm.

information across the hierarchy, such as price and promotion status. In contrast, Spiliotis et al. (2021) use ML models to reconcile the bottom-level series and proposed to use the reconciled forecasts in a BU fashion to obtain forecasts for all hierarchical levels. While both studies implement a BU method to obtain fully coherent forecasts across all levels, they take a completely different approach to generate the reconciled forecasts. They use the historical observations of all series in a hierarchy as explanatory variables and then develop one individual ML model to reconcile series for each bottom-level node. Their proposed model requires base forecasts for all of the series across the hierarchy. However, our method only needs to generate the base forecasts for the top-level series and then train only one model for each hierarchy to generate forecasts for all of the bottom-level series simultaneously, thus saving on computational cost.

4. Data

We considered two datasets for this study. The first dataset includes weekly sales data for 55 sets of FMCG products from a food manufacturing company in Australia. Our data include sales, retail price, and promotion time for all products across different locations. The price

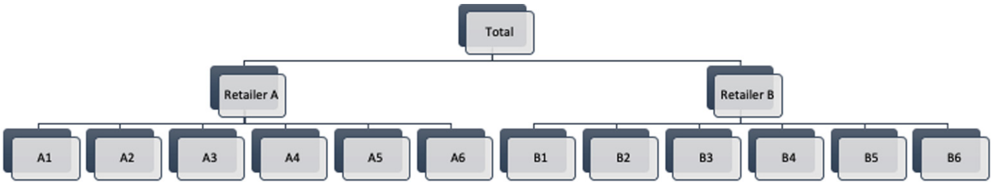

Fig. 2. Structure of the FMCG hierarchical data.

of the products across different locations inherently contains the information for promotional periods, and it is available in advance for the next eight weeks. The data span 120 weeks between September 2016 and December 2018. During a major promotion, the price of a product can be discounted by up to 50%. Sales time series are for cereal and breakfast products that are sold through two major retailers across different states in Australia, but the exact labels of the products are unknown to us.

Fig. 2 shows the hierarchical structure of our time series for one product. The top level (level 0) represents the manufacturer's total sales. There are two main retailers at the middle level (level 1) and 12 different distribution centres (DCs) at the bottom level (level 2). Each DC corresponds to one state in Australia and is responsible for distributing the products among retailer shops in their state. All 55 products have a hierarchical structure identical to that shown in Fig. 2. However, each of these 55 products has different levels of sales, promotions, seasonality, and trend patterns, comprising a comprehensive dataset for empirical analysis.

Fig. 3 shows hierarchical sales for one of the products across different levels of the hierarchy. There are several spikes in time series across different levels which correspond to promotional periods. Our case study supply chain is a coordinated supply chain where retailer A and retailer B have an agreement to put their products on sale over different non-overlapping periods. The spikes in sales occur simultaneously for each main retailer and their corresponding DCs, and they differ from those of the other retailer and DCs. Note that different series across the hierarchy may experience a different level of spike due to promotions. For example, various series across level 2 (across different locations) may react differently to the same promotional offer as per their population, DC size, number of stores, etc. As such, while the bottom-level series and their corresponding retailer have a similar pattern in terms of the promotion time, they experience different levels of promotional spikes and depict different demand patterns.

The second dataset used in this study is a subset of the M5 forecasting dataset (Makridakis, Spiliotis, & Assimakopoulos, 2021). The complete M5 dataset includes daily sales for 3049 products across three states, 10 stores, three categories, and seven departments, comprising 42,840 time series in total that can be hierarchically organised in 10 different cross-sectional levels. The data span from January 29, 2011 to June 19, 2016, where the last 28 days, i.e., from May 23 to June 19, 2016 were used as the test set. We choose a similar subset of M5 data in terms of the hierarchical structure that gives us a large enough dataset to evaluate our proposed approach. We create 3049 cross-sectional hierarchies corresponding to 3049 products that are identical in structure with product sales at the top level, three states at the middle level, and 10 stores at the bottom level, as shown in Fig. 4. Note that one can build a single large hierarchy for all these products. Similar to the FMCG hierarchies, however, we choose to build individual hierarchies for each product to have enough data to evaluate the performance of our methodology. The products correspond to household, food, and hobby categories and are highly intermittent at the lowest level. In addition to historical sales data, promotional information, sales price, and calendar features such as the date, year, month, week number, weekdays, and calendar events are available. Calendar events include special days and holidays that account for 8% of the data. Holidays are categorised into four groups: cultural, sporting, national, and religious. Promotion events are identified by Supplement Nutrition Assistance Program (SNAP) days. SNAP is a United States government initiative that provides purchasing assistance to low-income families. The SNAP dates differ from one state to another, but each state has 10 SNAP days each month and they account for 33% of the data (Makridakis et al., 2021). While we consider these dates as promotions, their impact on sales may vary from one state to another and may be impacted by variables such as SNAP days, holidays, and weekdays.

5. Experimental setup

Data preprocessing was implemented prior to modelling. Since our FMCG products are impacted by promotions and depict volatility across different levels, we used the natural logarithm of sales to stabilise their variance. We also checked for the seasonality of time series. Our investigated series do not have strong seasonality, but they are highly impacted by promotions. The M5 series depict high levels of intermittency and are not strongly impacted by promotions.

For the FMCG data, we considered the first 112 weeks of each hierarchical time series as the training set and the last eight weeks as the test set for validation purposes. Our dataset includes 55 hierarchical series that depict sales for products across different locations. Note that one can build a large hierarchical time series for all products, but that will substantially reduce the number of observations for the training set and reduce the size of the test set. Therefore, we decided to train and test our experiment on 55 sets of hierarchical time series that correspond to 55 products and depict different behaviour across different levels of the hierarchy. Since our time series across different levels are impacted by promotions, we used the regression with ARMA errors (Reg-ARMA)

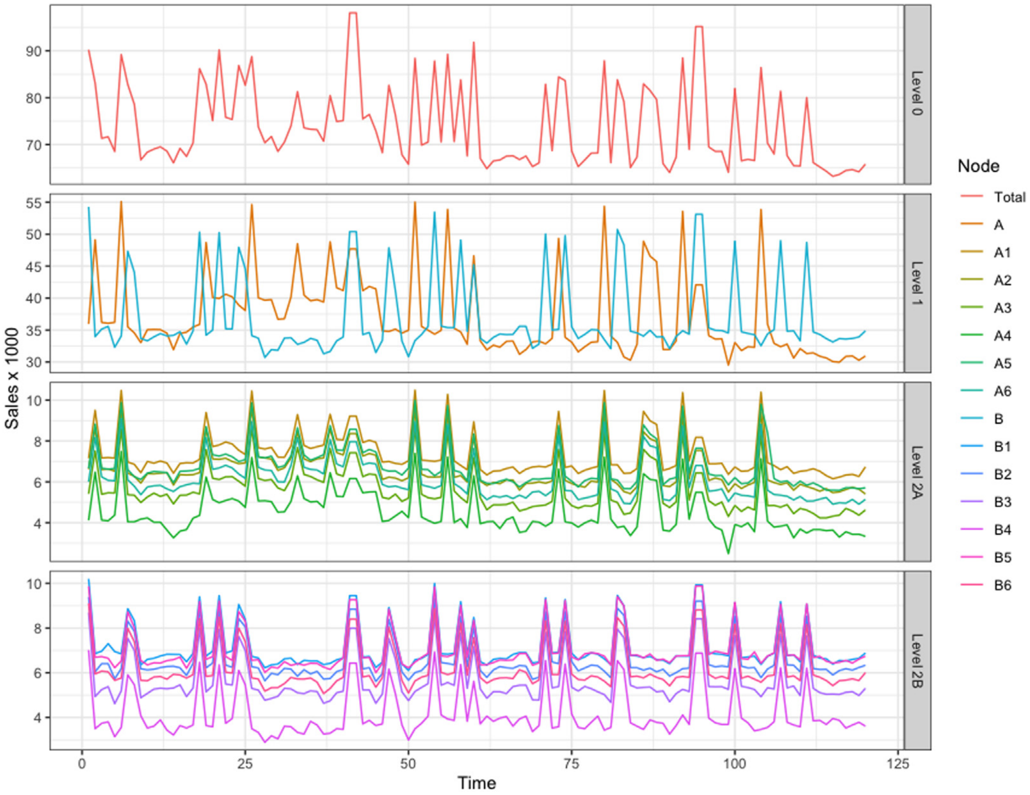

Fig. 3. Sales of a product at different levels of the FMCG hierarchy. Levels 2 A and 2B are both level 2 series that respectively correspond to their parent series A and B at level 1.

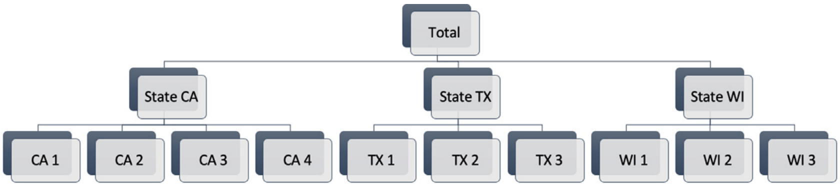

Fig. 4. Hierarchical structure for the subset of M5 series considered in this study.

model from the *forecast* package in R to generate base forecasts for all series where their price is used as an explanatory variable (Hyndman et al., 2020; R. Core Team, 2021). Reg-ARMA benefits from explanatory variables and takes into account the dynamics of errors in series with ARMA terms, making it a suitable candidate for forecasting our series which are impacted by promotions. We generate forecasts on a rolling-origin basis. That is, we train the series until time t and generate a forecast for time $t + 1$. We then move one step ahead and train the series until time $t + 1$ before generating a forecast for the period $t + 2$, and we continue this process until the end of the series.

The price of the products, which inherently captures the impact of promotions, is used as an explanatory variable in the Reg-ARMA model. Price at the bottom level is simply the retail price of sales for the DCs. The price does not differ across DCs and is centrally set by their main retailer. Price at the middle level is the sum of the prices of their corresponding child nodes. Since middle- and bottom-level series have the same policy to set their products on promotion, we simply aggregate the DCs' prices and use those as the explanatory variables for middle-level series. Retailers A and B have similar prices and operate in a coordinated supply chain, so aggregated prices can be used to capture their sales behaviour adequately. At the top level, we choose to use a Reg-ARIMA model with aggregated price as the input for forecasting. In another attempt, we also modelled sales at the middle level using aggregated DCs price as the explanatory variable and then used the MO approach to find out sales at the top and bottom levels. This similarly led to good results, but we opted to continue with the TD approach, as it requires fewer models and is computationally less expensive. We implemented the BU, TD, MinT methods as common benchmark HF methods. All base forecasts for the benchmark methods were generated by Reg-ARMA. We also implemented the proposed DTDBU method, as discussed in Section 3, where we used the XGB, RF, and lasso models in a multi-output fashion to forecast the bottom-level series. Details of these models are discussed in Sections 5.1, 5.2, and 5.3, respectively.

For the M5 series, we follow a similar setting as performed in the M5 forecasting competition. We considered the first 1913 days of data, ranging from January 29, 2011 to April 2, 2016 as the training set, with data from April 25 to May 22, 2016 as a validation set, and the last 28 days from May 23 to June 19, 2016 as the test set. The M5 dataset can be organised in different hierarchical formats with respect to different features. We considered a similar format to the FMCG hierarchical data, where we have sales data for 3049 products aggregated for all states and stores, i.e., level 10 series. Therefore, we have 3049 hierarchies with one total product sale at the top, states in the middle, and stores at the bottom level, as shown in Fig. 4. We develop one model for each hierarchy that corresponds to cross-sectional sales of one product. The inputs of the models are sales at level 12, three binary SNAP variables corresponding to the three states, weekdays, and months. In other attempts, we also used prices as input variables. But that did not improve the accuracy, so we removed them from the explanatory variables. The output variables are sales at the store levels, i.e., sales for 10 stores at the bottom level. In total, we evaluated 1,195,208 forecasts for each model (3049 hierarchical sets × 28 periods of forecasts × 14 series in each hierarchy). We used exponential smoothing top-down (ES-TD), exponential smoothing bottom-up (ES-BU), ARIMA top-down (ARIMA-TD), and ARIMA bottom-up (ARIMA-BU) models as benchmark methods, as per the M5 forecasting competition (Makridakis et al., 2021). We used multi-output ML models to compute the forecasts for lower levels and then aggregated them using the BU methods to ensure the coherence of the final forecasts. We implemented the proposed DTDBU method on M5 data and used XGB, RF, and lasso models in a multi-output fashion to forecast the bottom-level series. We then aggregated these forecasts for the bottom level to obtain the forecasts for higher levels, i.e., states and total.

We assessed the forecasting accuracy using the mean absolute scaled error (MASE), which is a scale-independent accuracy metric and can be used to compare the accuracy of models across different products and categories with different scales (Hyndman et al., 2006). Lower values of MASE correspond to better accuracy, and it is calculated as follows:

$$\text{MASE} = \frac{n-1}{h} \frac{\sum_{t=n+1}^{n+h} |y_t - f_t|}{\sum_{t=2}^{n} |y_t - y_{t-1}|},$$

where n is the number of observations used for training the models, y_t is the actual value of the series at time t, f_t is the forecast for period t, and h is the forecasting horizon. We further validated our results by conducting a statistical test on the obtained results.

5.1. Extreme gradient boosting

XGB is a decision tree-based ensemble method that uses a gradient boosting approach to generate unbiased and robust forecasts (Chen & Guestrin, 2016). XGB has been successfully used in various forecasting problems and has shown promising results, including hierarchical time series forecasting and multi-output forecasting (Abolghasemi et al., 2022; Spiliotis et al., 2021; Zhai et al., 2020). This algorithm uses a number of hyperparameters for generating forecasts. We tuned the parameters using a grid search. We set the values of the learning rate between 0.01 and 0.05 with an increment size of 0.01. The candidate values of the column subsample and subsample size ranged from 0.3 to 1 with an increment size of 0.1. The values for the maximum number of boosting iterations rolled over the range of 100 to 500 with an increment size of 50. The maximum depth of one tree was set between 2 and 20. We used the *squared error regression* as the objective function and choose the best results by minimising the root mean squared error. We used the *MultiOutputRegressor* module from the *sklearn* library in Python 3.5 to implement XGB in a multi-output fashion. The optimal values of hyperparameters vary for different investigated hierarchical time series. In order to save computational time, we optimised the values for a sample from each type of product and shared the optimised values for each hierarchy.

5.2. Random forests

RF is an ensemble method that is constructed from many decision trees. RF has performed well in numerous time series forecasting studies, including hierarchical time series forecasting in single-output form and supply chains in multi-output form (Abbasi, Babaei, Hosseinifard, Smith-Miles, & Dehghani, 2020; Segal & Xiao, 2011; Spiliotis et al., 2021). RF benefits from bootstrap sampling and uses a subset of features to build trees. The final forecast is an average of the target variable for all trees. RF uses a few hyperparameters to generate the final forecasts. We optimised the hyperparameters using a grid search. We set the number of trees between 50 and 100 with an increment size of 10. We ranged the number of variables between 6 and 10 with an increment size of 1, and finally ranged the node size between 10 and 50 with an increment of 5. We used the *RandomForestRegressor* module from the *sklearn* library in Python 3.5 to implement RF in a multi-output fashion. The optimal values of hyperparameters may vary for different investigated hierarchical time series. In order to save computational time, we optimised the values for one sample from each type of product and shared the optimised values for each hierarchy.

5.3. Lasso regression

Lasso regression is a powerful regression method that uses the l_1 regularisation technique in linear regression and imposes a penalty cost λ to control the sparsity of estimated coefficients in the fitted model, thereby regulating the bias and variance of the forecasts (Tibshirani, 1996). The lasso is particularly useful to avoid overfitting the data, especially for series that may depict multicollinearity. The lasso loss function can be defined as

$$L(\beta, \lambda) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

Table 1
Forecasting accuracy of hierarchical forecasting methods measured by the mean of MASE values across all 55 series.

Level	Top level	Middle level	Bottom level	All levels
BU	0.501	0.706	0.677	0.673
TD	0.455	0.724	0.701	0.688
MinT	0.452	0.713	0.684	0.670
Lasso-DTDBU	**0.438**	0.638	0.587	0.585
RF-DTDBU	0.451	0.573	0.591	0.542
XGB-DTDBU	0.453	**0.559**	**0.513**	**0.515**

where y is the actual value of the target variable, $\hat{y}$ is the predicted value of the target variable, λ is the shrinkage estimator, n is the number of observations, β_j denotes parameters used to define the prediction model for $\hat{y}_i$, and p is the number of variables.

The lasso has been successfully implemented in various forecasting applications, variable selections, and multi-output regressions (Li et al., 2015; Ma, Fildes, & Huang, 2016; Zou & Qiu, 2009). We used lasso in a multi-output fashion and applied a five-fold cross-validation technique to obtain the optimal value of λ, following (Bergmeir, Hyndman, & Koo, 2018). We used the *lasso* model from the *sklearn.linear-model* library in Python 3.5 to implement the model in a multi-output fashion. The optimal values of hyperparameters may vary for different investigated hierarchical time series. To save computational time, we optimised the values for a sample of products and shared the optimised values for all series.

6. Empirical results

This section summarises the empirical results obtained with regard to the methods presented in Section 3. We used the *hts* package in R for implementing the BU, TD, and MinT methods used in this study (Hyndman, Lee, Wang & Wickramasuriya, 2020). We developed and implemented the DTDBU models proposed in this study in the Python programming language. In total, for the FMCG series, we evaluated 6600 forecasts for each model (55 hierarchical sets × 8 periods of forecasts × 15 series in each hierarchy). Note that we developed one model for each hierarchy that corresponds to the cross-sectional sales of one product. Table 1 shows the obtained results of the models across 55 sets of hierarchical series at different levels in terms of MASE. We report the results for each level and the average across all levels. To compute the average accuracy across all levels, we take a simple average of three levels. This is plausible, since forecasts at different levels of the hierarchy are used for different purposes but they are equally important for various decisions.

The results indicate that the performances of models differ greatly across the hierarchy and depend on the level of the hierarchy (Shlifer & Wolff, 1979). XGB-DTDTBU outperforms the top-performing benchmark method across the top, middle, and bottom levels by 3%, 20%, and 25%, respectively. The improvement is higher at the bottom level of the hierarchy, followed closely by the middle level. The improvement at the top level is significantly smaller.

At the top level, Lasso-DTDBU obtained the highest accuracy, followed by RF-DTDBU, MinT, and XGB-DTDBU. As noted above, RF-DTDBU, Lasso-DTDBU, and XGB-DTDBU benefit from the advantages of TD and BU methods, but unlike traditional TD and BU models that generate forecast individually for each bottom level series, DTDBU generates forecasts simultaneously for all bottom series using explanatory information from all levels of the hierarchy. This is an interesting finding, since DTDBU outperforms the conventional BU method, indicating that using information from other levels is useful to improve the forecast accuracy of these methods. Moreover, these models outperform the direct forecast at the top level, indicating that we can effectively combine the bottom- and top-level information to generate reliable forecasts at the bottom level that also aggregate accurately and improve the forecast accuracy at the top level of the hierarchy.

The results at the middle and bottom levels show that XGB is superior to other investigated models. Again, our proposed XGB model in the DTDBU algorithm outperforms the well-known TD, BU, and MinT methods. Similar results hold in terms of the average accuracy across all levels. More specifically, XGB as the superior model consistently outperforms MinT, TD, and BU methods by 23%, 25%, and 23%, respectively.

We further looked at the performance of these methods by investigating their accuracy distributions. Fig. 5 depicts boxplots of the accuracy of the models across 55 sets of FMCG hierarchical series at different levels in terms of the MASE. The boxplots show the first quartile, median, third quartile, and outliers. Our results show that XGB-DTDBU consistently generates reliable forecasts at different levels of the hierarchy and has the smallest median MASE across all levels. Whereas RF generates competitive forecasts at the middle and bottom levels, its accuracy deteriorates at the top level. Lasso-DTDBU is not particularly accurate in terms of the median MASE across the bottom and middle levels, but it performs well at the top level.

To validate our findings, we conducted multiple comparisons with the best (MCB) tests to determine whether there is any statistically significant difference between the performance of different HF models at the 5% significance level. Fig. 6 depicts the MCB plot for the investigated methods across all series. XGB-DTDBU overwhelmingly outperformed other methods across the bottom, middle, and average of all levels. This confirms that XGB-DTDBU can use the explanatory information from all hierarchical levels effectively and generate unbiased forecasts for the bottom level without loss of accuracy at higher levels. Our results show that the performance of RF reduces as we move towards the higher levels of the hierarchy. This may be attributed to the current data or settings, and more

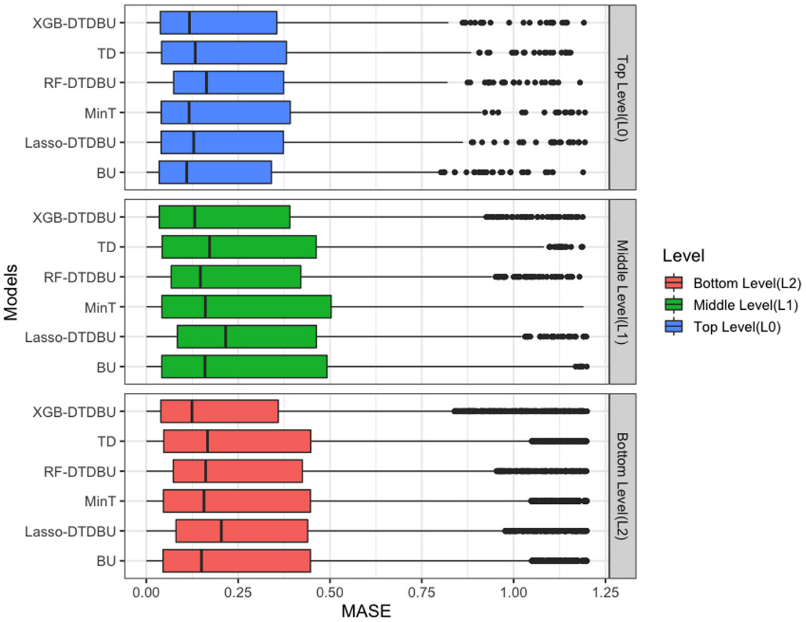

Fig. 5. Boxplots of the accuracy of the models across all FMCG series.

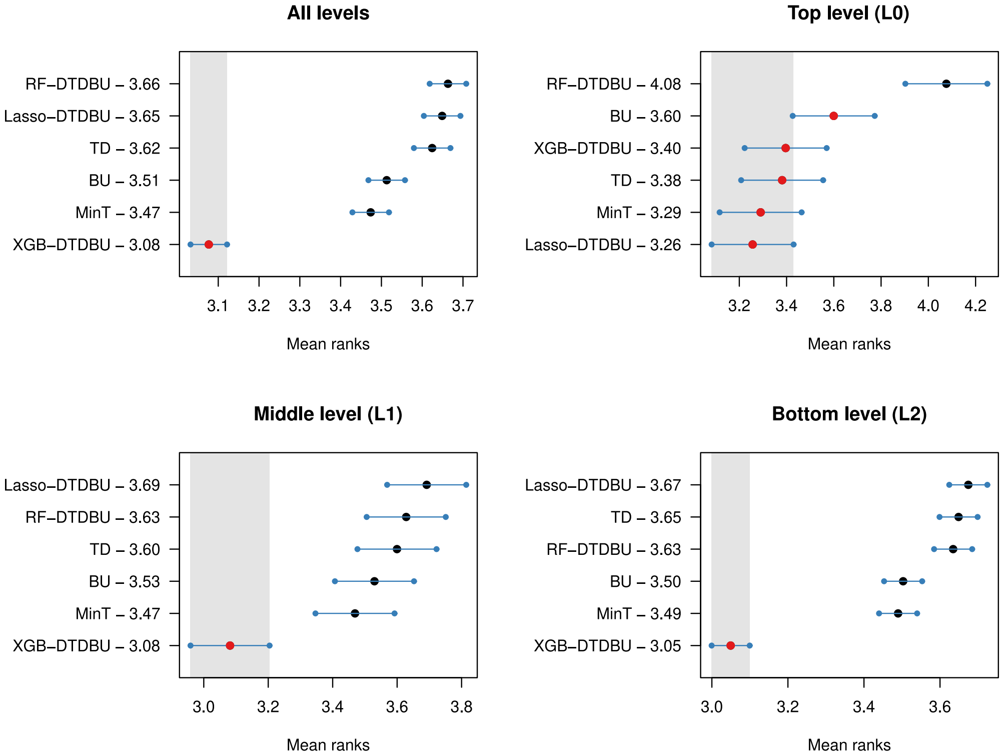

Fig. 6. Statistical significance at the 5% level: MCB tests across all FMCG series.

studies are required to investigate the performance of RF with other settings and datasets.

The above-mentioned results are summarised for all hierarchical series. Our empirical results show the performance of common HF methods and demonstrate that our proposed DTDBU algorithm works well across 55 FMCG hierarchical series comprising series that are impacted by promotions and have different levels of sales, seasonality, and trend. However, the performance of these models might differ depending on promotion status. In Section 7, we discuss the performance of the investigated models for promotional and non-promotional periods.

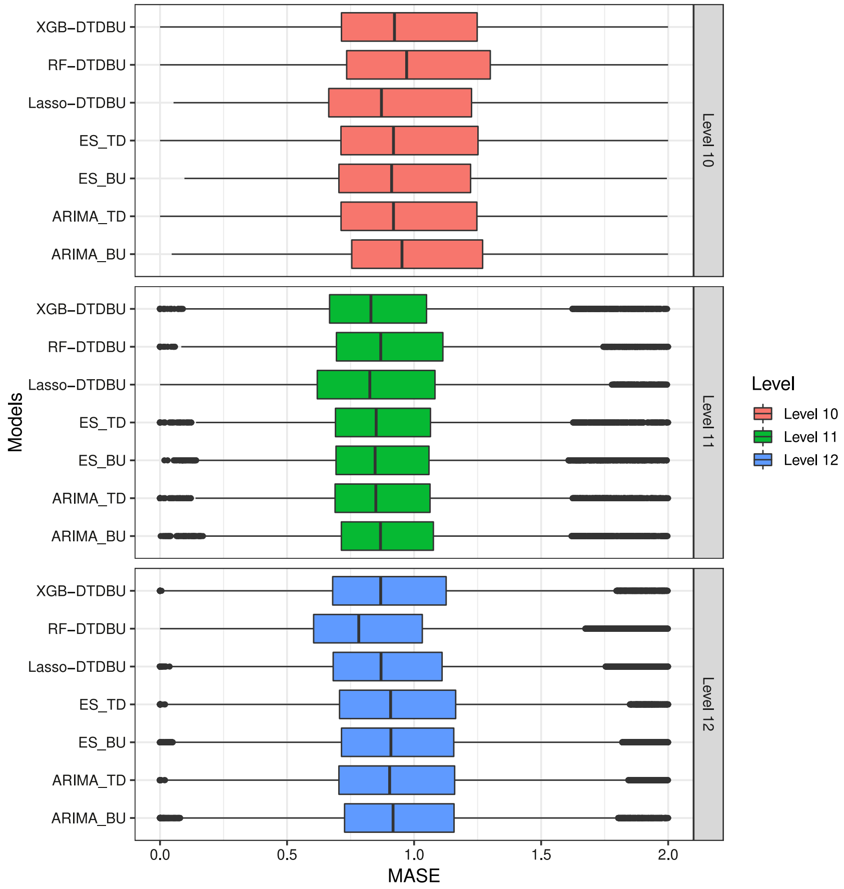

Fig. 7. Boxplot of the accuracy of the models across M5 series.

To further evaluate the performance of our approach, we implemented the proposed DTDBU approach on a subset of the M5 hierarchical series where 3049 products are organised across three-level hierarchical data (levels 10, 11, and 12). Table 2 shows our results in terms of the mean MASE, and Fig. 7 shows the distribution of the forecast accuracy across levels 10, 11, and 12.

The results indicate that ES-BU, XGB-DTDBU, and RF-DTDBU outperform the other methods across levels 10, 11, and 12, respectively. Similar to the FMCG data, the forecasts generated by models using the DTDBU algorithm perform well across different levels. However, unlike the FMCG data, where the DTDBU-generated forecasts outperformed at the top level of the hierarchy, ES-BU is the top-performing model at the top level (level 10) of the M5 series hierarchy. This may be attributed to the heterogeneity and intermittency of the M5 series, where level 12 series are highly intermittent and depict more heterogeneity across household, hobby, and food products. In particular, ARIMA-BU performed poorly at the bottom level even though we generate direct forecasts at the bottom level. This indicates that ARIMA is not a suitable candidate to forecast intermittent series at the bottom level. Lasso-DTDBU and ARIMA-BU show lower accuracy at level 10, but ES-BU is the top-performing model at level 10. This is interesting, as all these models are BU but they use different techniques. They depict different levels of accuracy but confirm that a BU approach may be more accurate than a direct forecast at an aggregated level at higher levels of the hierarchy (see Fig. 8).

Table 2
Forecasting accuracy of hierarchical forecasting methods implemented on a subset of M5 series, measured by the mean of MASE.

Level	Level 10	Level 11	Level 12	All levels
ARIMA-BU	1.359	0.979	1.073	1.073
ARIMA-TD	1.319	0.961	1.054	1.053
ES-BU	**1.308**	0.960	1.060	1.057
ES-TD	1.319	0.963	1.058	1.057
Lasso-DTDBU	1.866	1.000	1.041	1.091
RF-DTDBU	1.423	1.029	**0.955**	**1.004**
XGB-DTDBU	1.326	**0.945**	1.076	1.066

7. Discussion

Now we turn our attention to analysing the impact of promotions on the performance of the investigated HF methods. In total, for FMCG data, we evaluated 6600 forecasts for each model, from which 2310 observations correspond to promotional periods (35%) and 4290 periods (65%) correspond to non-promotional periods. Table 3

Table 3
Forecasting accuracy of models over promotional and non-promotional periods on FMCG data measured by the mean MASE.

Level	Top level		Middle level		Bottom level		All levels	
Series	Pro	No Pro	Pro	No Pro	Pro	No Pro	Pro	No Pro
BU	0.834	0.323	1.296	**0.391**	1.221	**0.386**	1.205	**0.382**
TD	0.710	0.319	1.327	0.401	1.275	0.393	1.244	0.389
MinT	0.696	0.322	1.306	0.396	1.240	**0.386**	1.213	**0.382**
Lasso-DTDBU	**0.588**	**0.317**	0.884	0.439	0.888	0.425	1.145	0.392
RF-DTDBU	0.699	0.357	1.014	0.406	0.822	0.397	0.937	0.402
XGB-DTDBU	0.632	0.357	**0.844**	0.406	**0.739**	0.392	**0.868**	0.404

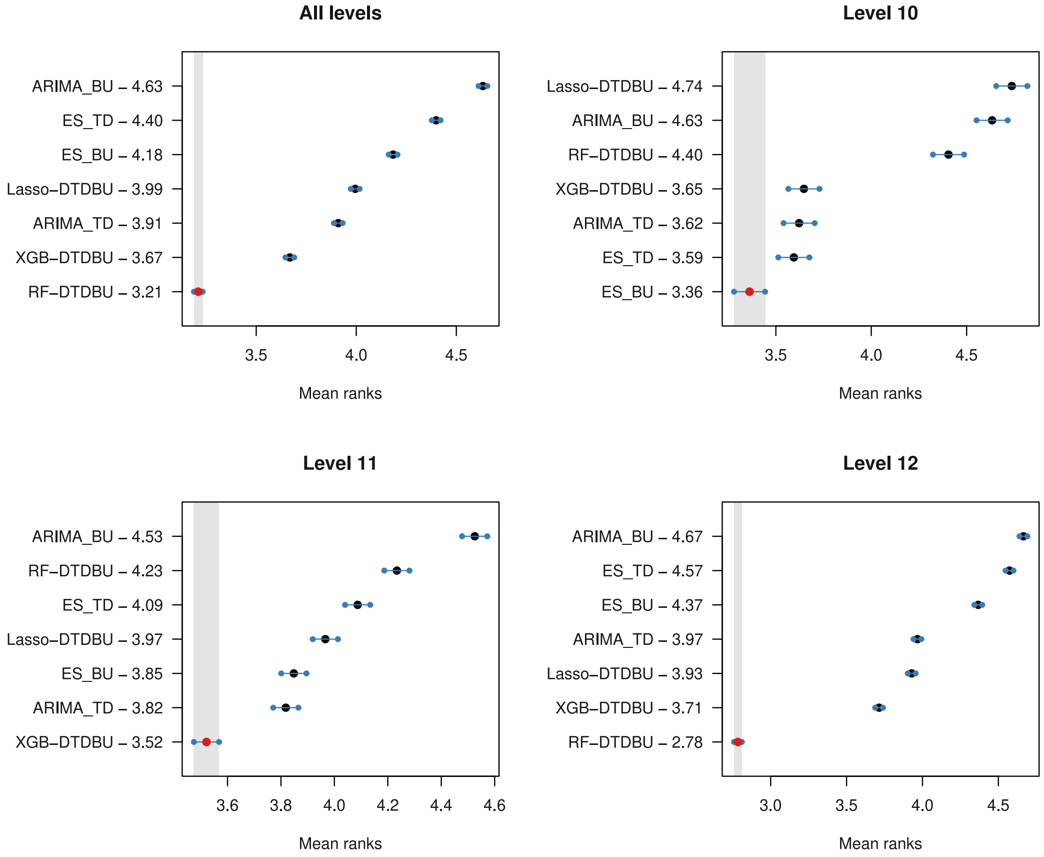

Fig. 8. Statistical significance at the 5% level: MCB tests across a subset of M5 series at levels 10, 11, and 12, and at all levels (referring to the average of these three levels).

summarises the performance of the models in terms of the mean MASE across all levels for promotional and non-promotional periods. Our results indicate that promotions unfavourably impact the performance of the methods across all levels. This is not surprising, as forecasting under promotions is often challenging and can lead to less accurate forecasts (Abolghasemi, Hurley, Eshragh & Fahimnia, 2020). All models perform better in the absence of promotions but their performance differs from one model to another, and from one level to another. On average, XGB-DTDBU is the top-performing model for promotional periods across the middle and bottom levels, but it is minimally outperformed by the Lasso-DTDBU model at the top level. For non-promotional periods, the results are different. At the top level, the Lasso-DTDBU model outperformed other models, followed by TD. The BU and MinT methods outperformed the other methods across the middle, bottom, and all levels, on average.

The performance of each method is relatively similar in the absence of promotions, but differences are observed when we consider promotional periods. For example, during the promotional periods, the difference between the best and worst models across the top, middle, bottom, and average of all levels are 29%, 36%, 42%, and 30%, respectively. However, these differences reduce to 11%, 11%, 9%, and 5% during non-promotional periods. One reason that may contribute to the better performance of the DTDBU approach in comparison to the BU and TD methods is the fact that DTDBU uses information such as sales and prices from across the hierarchy, whereas BU and TD only

Table 4
Forecasting accuracy of models over promotional and non-promotional periods of the M5 series measured by the mean MASE.

Level	Level 10		Level 11		Level 12		All levels	
Series	Pro	No Pro	Pro	No Pro	Pro	No Pro	Pro	No Pro
ARIMA-BU	1.392	1.326	1.000	0.957	1.089	1.056	1.092	1.054
ARIMA-TD	1.360	1.277	0.984	0.938	1.071	1.038	1.073	1.034
ES-BU	**1.354**	**1.261**	0.988	0.933	1.080	1.041	1.080	1.033
ES-TD	1.358	1.280	0.985	0.941	1.075	1.042	1.076	1.037
Lasso-DTDBU	1.896	1.837	1.017	0.983	1.053	1.029	1.105	1.077
RF-DTDBU	1.457	1.388	1.049	1.008	**0.966**	**0.944**	**1.019**	**0.990**
XGB-DTDBU	1.367	1.286	**0.967**	**0.923**	1.089	1.062	1.083	1.048

use the information at one level. Furthermore, most TD methods, like our TD benchmark method, do not dynamically update the disaggregation factor to consider time series variations while disaggregating series in the hierarchy, and thus may not be suitable to capture variations effectively when compelling factors such as promotions impact the dynamics of the series. BU starts from the very bottom of a hierarchy and generates forecasts for each series at the bottom level without considering their possible correlation with each other. Thus, they may miss important information from other series and levels and generally may be prone to overfitting, due to the level of noise at the bottom level. The results of the FMCG series show that the performances of different HF methods are more similar in the absence of promotions, and one can use simpler HF methods to obtain competitive forecasts.

Our empirical results also show that using external information such as price across the hierarchy can significantly improve the forecast accuracy across different levels. The explanatory information may be more useful to improve the accuracy during promotional periods. Our finding confirms the already existing results by the well-established linear combination approach that showed that using information from across the hierarchy is useful to improve the accuracy across different levels. However, we note that while DTDBU uses the price information directly and generates forecasts for the bottom level series simultaneously, the linear combination reconciliation model in our study uses price indirectly. That is, price is used as an explanatory variable to forecast the sales of nodes, and then MinT is used to reconcile the forecasts for all series in the hierarchy.

We conducted MCB tests to analyse whether there is any significant difference between the generated forecasts, and we ranked them according to their performance. Fig. 9 shows MCB plots across all levels for both promotional and non-promotional observations. For promotional periods, the results at the bottom, middle, and across all levels show that XGB-DTDBU generates significantly more accurate forecasts than the other models. However, at the top level, Lasso-DTDBU, RF-DTDBU, MinT, and TD are competitive methods, and there is no statistically significant difference among them. The average forecast accuracy across all levels show that XGB-DTDBU significantly outperforms other methods.

The results are different for non-promotional periods. For non-promotional periods, MinT and XGB-DTDBU performed significantly better than the other methods across all levels, on average. There is no single top-performing model across the top, middle, and bottom levels. Conventional HF methods like BU and TD generate similarly competitive forecasts compared to the more advanced MinT method across the bottom, middle, and top levels. This is an interesting finding which confirms our previous claim and shows that simple HF methods work well when dealing with simple series that are not difficult to forecast, i.e., non-promotional periods with higher forecastability. However, more sophisticated methods may be required when dealing with series that are difficult to forecast, i.e., intermittent series or series impacted by promotions.

Table 4 shows the obtained results, and Fig. 10 shows the MCB tests for the subset of M5 series for both promotional and non-promotional observations. We considered series with at least one SNAP event as promotional periods, which accounted for approximately 33% of observations. Similar to results from FMCG data, we observe that the performance of models is consistently better in the absence of promotions across different levels, but the improvement is not as significant as it is with FMCG data. This might be due to the dynamics of promotions in the M5 data, which are significantly different from FMCG data. In the M5 data, promotion days account for 11% more sales than non-promotion days, whereas this ratio for the FMCG data is 31%, on average, indicting different dynamics of promotion. Our results indicate that ES-BU is the top-performing model at the top level for both promotional and non-promotional periods, followed closely by ES-TD and ARIMA-TD for promotional and non-promotional periods. XGB-DTDBU outperformed the other models at level 11, and RF-DTDBU is the top-performing model at level 11 and across all levels, on average. This indicates that the forecasts generated by the DTDBU approach can significantly improve the forecast accuracy across both promotional and non-promotional periods. The improvement may differ from one dataset to another and across different levels.

In summary, we suggest using the proposed DTDBU approach to generate more accurate forecasts at lower levels of hierarchical data. This approach benefits from the explanatory information across the hierarchy and can improve the forecast accuracy across different levels for promotional and non-promotional periods. Our results based on 55 FMCG and 3049 M5 hierarchical series showed that the improvement is often more significant at lower levels and during promotional periods, but the approach still generates accurate and competitive forecasts at higher levels and during non-promotional periods. The DTDBU

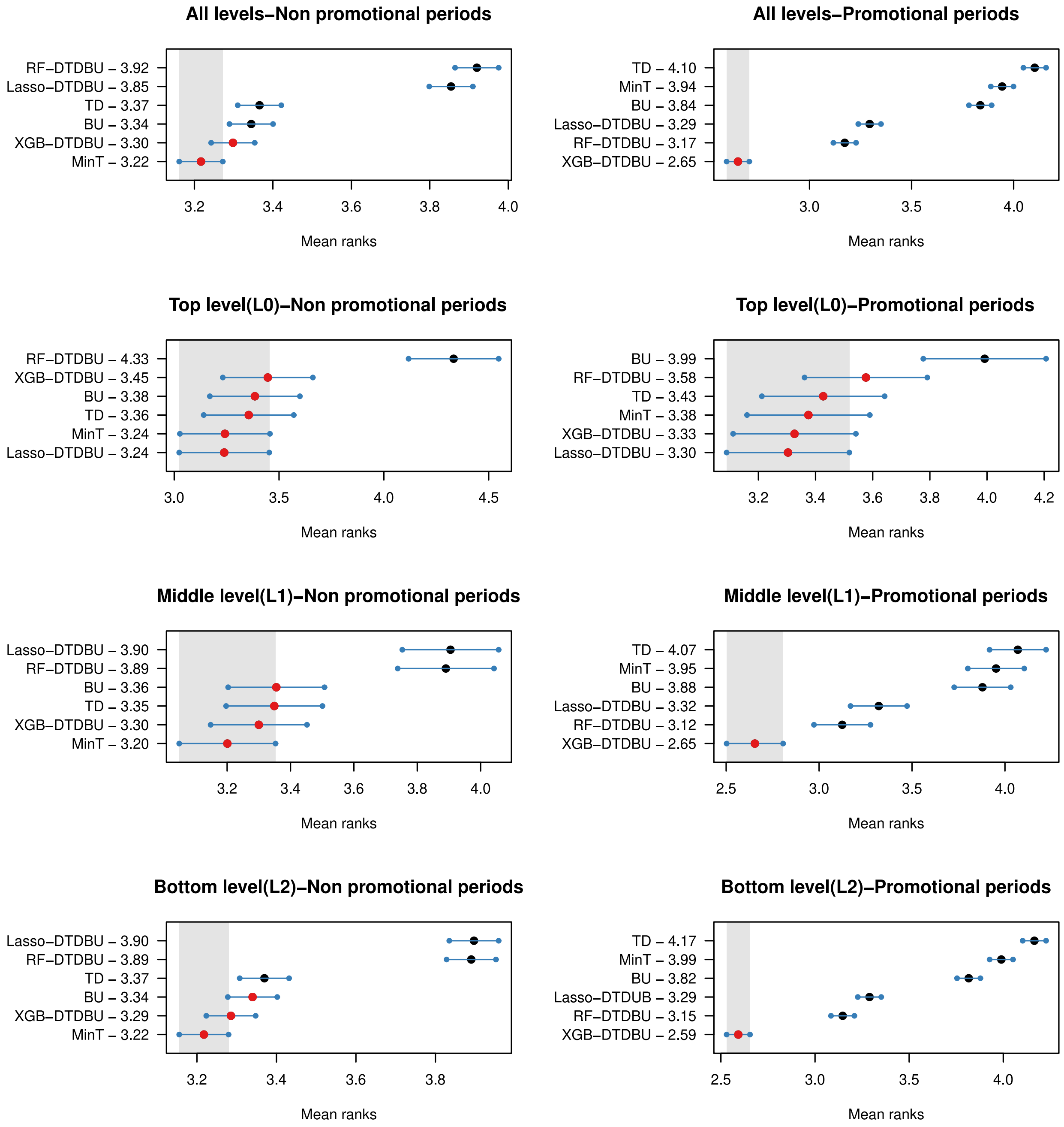

Fig. 9. MCB tests for promotional and non-promotional periods on FMCG data.

approach is particularly more useful when data are not available for the bottom level. For example, the manufacturer at the top level may have delayed access to the bottom-level retailer data, but external explanatory information such as price and promotion status can be used to generate forecasts for lower levels. Promotions may impact the efficacy of the HF models significantly and should be treated with extra care in hierarchical forecasting. Finally, as discussed in Section 3.4, we generated the direct forecasts only at the top level of the hierarchy and used them as an input in the DTDBU algorithm to generate forecasts at the bottom level simultaneously. One can also generate forecasts at an appropriate middle level and use them as input to generate forecasts at the bottom level, or even in reconciliation methods like MinT to reconcile the forecasts in the hierarchy.

8. Conclusion

Hierarchical forecasting methods like TD, BU, and MinT can be used to generate coherent forecasts at different levels of hierarchical time series and to improve the fore-

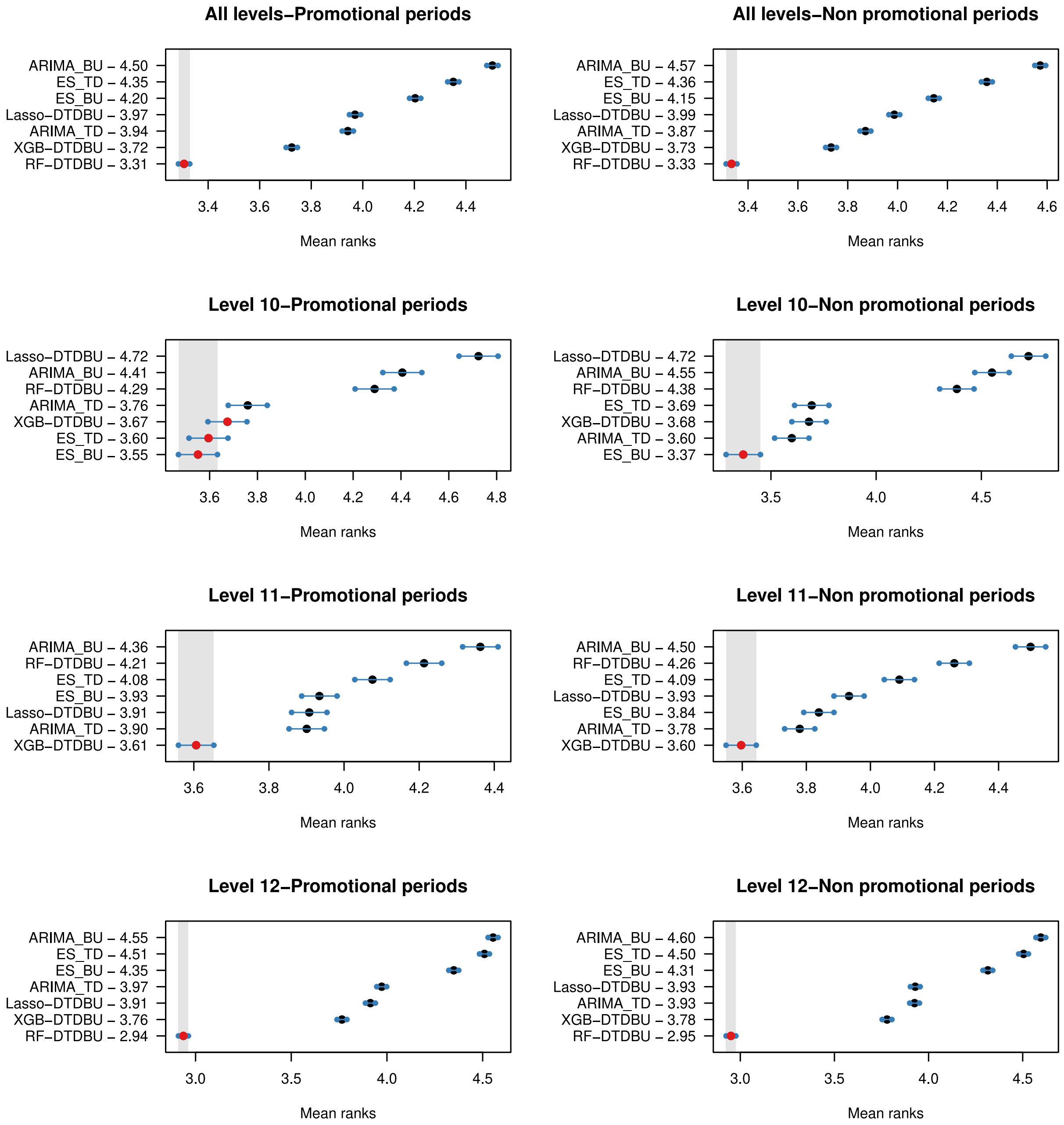

Fig. 10. MCB tests for promotional and non-promotional periods on M5 data.

cast accuracy of time series. The results of numerous empirical studies show that various HF methods have several advantages and disadvantages in various scenarios. While TD methods only use the information from the top level, BU methods only utilise the bottom-level information for generating forecasts. MinT uses the information from all levels by generating forecasts for each series individually and then combining them linearly to obtain coherent forecasts. These methods do not directly use the explanatory variables or information across other levels that may be available in a hierarchy. In this study, we proposed a dynamic top-down bottom-up algorithm that uses multi-output models to forecast series at the top level of the hierarchy and uses explanatory variables from all levels of the hierarchy to simultaneously forecast the series at the bottom level. We used various information, such as price, promotion status, and sales, from higher levels of the hierarchy to forecast the bottom-level series. Then, we aggregated the forecasts of the bottom-level series to find the forecasts at higher levels. We used XGB, RF, and lasso as three common ML methods in the sales forecasting literature. We empirically showed that our proposed multi-output algorithm can outperform the current conventional HF methods.

We used two datasets for empirical analysis: (i) FMCG sales time series, where 55 sets of three-level hierarchical data are highly impacted by promotions, and as a result, show different behaviour across the hierarchy, and (ii) M5 series with 3049 products across levels 10, 11, and 12 that are highly intermittent at level 12 and less significantly impacted by promotions. Our empirical results showed that the DTDBU approach can be effectively used to generate accurate forecasts for both promotional and non-promotional periods. The benefit is often larger at the lower levels of the hierarchy. Promotions may impact the accuracy of the models differently. Our results based on FMCG data, where series are highly impacted by promotions, showed that simple HF methods, such as TD and BU, work well when forecasting hierarchical time series that are not impacted by promotions. However, the results based on the M5 series, where the impact of promotions is not dramatically large, showed that DTDBU improves the forecast accuracy similarly for promotional and non-promotional periods. Our study showed that directly using external explanatory information such as price from other levels of the hierarchy can be useful to improve the forecast accuracy, with their value being maximised in the presence of major promotions that significantly impact sales.

While we proposed a methodology that is capable of generating simultaneous forecasts for the bottom-level series, it still requires BU aggregation to generate fully coherent forecasts. Future studies can focus on finding a customised loss function to generate coherent forecasts in a TD fashion without requiring BU aggregation. There is much debate in the literature about different HF models and their efficacy, and it is not trivial which one works better at different levels (Syntetos et al., 2016). One should choose the most appropriate forecasting model based on the desirable criteria and purpose of forecasting, i.e., the level of interest, time series characteristics, and forecasting horizon. For future research, we also suggest investigating the structure of the hierarchy, i.e., the number of levels, the number of series, and other features such as the correlation of demand series at different levels. These can be embedded as features in ML models to find the most appropriate forecasting and reconciliation method.

Declaration of competing interest

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

Acknowledgment

The authors are grateful to the review team for providing constructive suggestions and comments. We are thankful to Prof. Rob Hyndman and his support throughout this research.

Funding

Garth Tarr was supported by the Australian Research Council (DP210100521).

References

Abbasi, B., Babaei, T., Hosseinifard, Z., Smith-Miles, K., & Dehghani, M. (2020). Predicting solutions of large-scale optimization problems via machine learning: A case study in blood supply chain management. *Computers & Operations Research*, *119*, Article 104941.

Abolghasemi, M., Beh, E., Tarr, G., & Gerlach, R. (2020). Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. *Computers & Industrial Engineering*, *142*, Article 106380.

Abolghasemi, M., Hurley, J., Eshragh, A., & Fahimnia, B. (2020). Demand forecasting in the presence of systematic events: Cases in capturing sales promotions. *International Journal of Production Economics*, *230*, Article 107892.

Abolghasemi, M., Hyndman, R. J., Spiliotis, E., & Bergmeir, C. (2022). Model selection in reconciling hierarchical time series. *Machine Learning*, *111*(2), 739–789,

Abolghasemi, M., Hyndman, R. J., Tarr, G., & Bergmeir, C. (2019). Machine learning applications in time series hierarchical forecasting. arXiv preprint arXiv:1912.00370.

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

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

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70–83.

Chen, H., & Boylan, J. E. (2009). The effect of correlation between demands on hierarchical forecasting. In *Advances in business and management forecasting, vol. 6* (pp. 173–188). Emerald Group Publishing Limited.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGLDD International conference on knowledge discovery and data mining* (pp. 785–794). ACM.

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

Fliedner, G. (1999). An investigation of aggregate variable time series forecast strategies with specific subaggregate time series statistical correlation. *Computers & Operations Research*, *26*(10–11), 1133–1149.

Gross, C. W., & Sohl, J. E. (1990). Disaggregation methods to expedite product line forecasting. *Journal of Forecasting*, *9*(3), 233–254.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579–2589.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice* (2nd ed.). Melbourne, Australia: OTexts, URL http://OTexts.com/fpp2.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., et al. (2020). Forecast: Forecasting functions for time series and linear models. URL http://pkg.robjhyndman.com/forecast, R package version 8.12.

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., Lee, A., Wang, E., & Wickramasuriya, S. (2020). Hts: Hierarchical and grouped time series. URL https://CRAN.R-project.org/package=hts, R package version 6.0.0.

Hyndman, R. J., et al. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, *4*(4), 43–46.

Kahn, K. B. (1998). Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting*, *17*(2), 14–19.

Kourentzes, N., & Athanasopoulos, G. (2019). Cross-temporal coherent forecasts for Australian tourism. *Annals of Tourism Research*, *75*, 393–409.

Kourentzes, N., & Petropoulos, F. (2016). Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*, *181*, 145–153.

Li, Y., Nan, B., & Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, *71*(2), 354–363.

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra-and inter-category promotional information. *European Journal of Operational Research*, *249*(1), 245–257.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2021). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting (in Press)*.

Mancuso, P., Piccialli, V., & Sudoso, A. M. (2021). A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications*, *182*, Article 115102.

Moon, S., Simpson, A., & Hicks, C. (2013). The development of a classification model for predicting the performance of forecasting methods for naval spare parts demand. *International Journal of Production Economics*, *143*(2), 449–454.

Nenova, Z. D., & May, J. H. (2016). Determining an optimal hierarchical forecasting model based on the characteristics of the data set. *Journal of Operations Management*, *44*, 62–68.

Nikolopoulos, K., Litsa, A., Petropoulos, F., Bougioukos, V., & Khammash, M. (2015). Relative performance of methods for forecasting special events. *Journal of Business Research*, *68*(8), 1785–1791.

Pennings, C. L., & van Dalen, J. (2017). Integrated hierarchical forecasting. *European Journal of Operational Research*, *263*(2), 412–418.

R. Core Team (2021). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, URL https://www.R-project.org/.

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.

Segal, M., & Xiao, Y. (2011). Multivariate random forests. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 80–87.

Shlifer, E., & Wolff, R. (1979). Aggregation and proration in forecasting. *Management Science*, *25*(6), 594–603.

Spiliotis, E., Abolghasemi, M., Hyndman, R. J., Petropoulos, F., & Assimakopoulos, V. (2021). Hierarchical forecast reconciliation with machine learning. *Applied Soft Computing*, *112*, Article 107756.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*(526), 804–819.

Widiarta, H., Viswanathan, S., & Piplani, R. (2007). On the effectiveness of top-down strategy for forecasting autoregressive demands. *Naval Research Logistics*, *54*(2), 176–188.

Zhai, N., Yao, P., & Zhou, X. (2020). Multivariate time series forecast in industrial process based on XGBoost and GRU. In *2020 IEEE 9th Joint international information technology and artificial intelligence conference, vol. 9* (pp. 1397–1400). IEEE.

Zou, C., & Qiu, P. (2009). Multivariate statistical process control using LASSO. *Journal of the American Statistical Association*, *104*(488), 1586–1596.