The performance of the global bottom-up approach in the M5 accuracy competition: a robustness check

Preprint · September 2021

CITATIONS
0

2 authors:

Shaohui ma
Nanjing Audit University
22 PUBLICATIONS **236** CITATIONS
SEE PROFILE

Robert Fildes
Lancaster University
179 PUBLICATIONS **7,794** CITATIONS
SEE PROFILE

The performance of the global bottom-up approach in the M5 accuracy competition: a robustness check

Shaohui Ma[a,1]

Robert Fildes[b]

[a] School of Business, Nanjing Audit University, China, 211815

[b] Centre for Marketing Analytics and Forecasting, Lancaster University, UK, LA1 4YX

[1] Corresponding author at: School of business, Nanjing Audit University, Nanjing, 211815, China. E-mail address: shaohui.ma@nau.edu.cn (Shaohui Ma); r.fildes@lancaster.ac.uk (Robert Fildes).

Abstract

The M5 accuracy competition has presented a large-scale hierarchical forecasting problem in a realistic grocery retail setting in order to evaluate an extended range of forecasting methods, particularly those adopting machine learning. The top ranking solutions adopted a global bottom-up approach, by which is meant using global forecasting methods to generate bottom level forecasts in the hierarchy and then using a bottom-up strategy to obtain coherent forecasts for aggregate levels. However, whether the observed superior performance of the global bottom-up approach is robust over various test periods or only an accidental result, is an important question for retail forecasting researchers and practitioners. We conduct experiments to explore the robustness of the global bottom-up approach, and make comments on the efforts made by the top-ranking teams to improve the core approach. We find that the top-ranking global bottom-up approaches lack robustness across time periods in M5 data. This inconsistent performance makes the M5 final rankings somewhat of a lottery. In future forecasting competitions, we suggest the use of multiple rolling test sets to evaluate the forecasting performance in order to reward robustly performing forecasting methods, a much needed characteristic in any application.

Keywords: M-competition; Retail forecasting; Hierarchical forecasting; Global forecasting method; Machine learning; Competition design.

1. The global bottom-up approach in M5

Before the M5 forecasting competition a number of similar Kaggle competitions have focused on the sales of retail grocery products, including 'Rossman Store Sales' in 2015, 'Corporación Favorita Grocery Sales Forecasting' in 2017, and the 'Store Item Demand Forecasting Challenge' in 2018. The M5 accuracy competition adds some new and important dimensions to the task of evaluating the accuracy of alternative forecasting methods. While the other competitions evaluated the forecasts only at the store-product level, M5 also evaluated the accuracy on eleven additional hierarchical aggregate levels using a new measure named Weighted Root Mean Squared Scaled Error (WRMSSE), and this made M5 in nature a large-scale hierarchical forecasting problem.

Hierarchical forecasting has received significant attention in the recent literature. But as far as we are concerned, most have considered only the aggregation problem for general time series; the number of time series to be forecasted at the bottom level and the levels of the hierarchy were in general small compared to the task of M5, and the forecasting methods tested were limited to local methods[2]. So far no research has considered the problem of comparing methods in a large scale retail sales forecasting setting where demand is affected dramatically by many factors, such as price changing, promotions and weather conditions (Fildes, Ma, & Kolassa, 2019). Therefore, for the hierarchical forecasting problem in M5, the existing literature could only provide very limited guidance to M5 players.

In the research literature, when forecasting hierarchical time series, there are four candidate approaches: bottom-up (BU), top-down (TD), middle-out (MO), or optimal reconciliation (ORC). Compared to the BU approach, the TD suffers from information loss due to the use of aggregated series, but can benefit from possible noise canceling out during the aggregation (Fliedner, 1999). Over three decades the research literature has shown mixed results as to a preference between TD or BU forecasting (Syntetos, et al., 2016). In the MO approach, forecasts are generated at a particular level and then aggregated upwards using the BU approach,

[2] Local methods estimate model parameters independently for each time series, in contrast to global methods which estimate model parameters jointly from a group of time series. The global forecasting method usually trains a complex homogenous model to forecast a pool of time series, this enhances the data availability and has the potential to capture cross time series common patterns. For more details on the differences between local and global method see Januschowski, et al. (2018).

and allocated downwards using a TD approach. Optimal reconciliation (ORC) is a post hoc approach, forecasts of all-time series in the hierarchy are generated separately first, then these separate forecasts are combined using a linear transformation to ensure they add up consistently over the hierarchy levels. ORC was originally proposed by Hyndman, et al. (2011), it allows the forecasts in different levels of the hierarchy to be generated with different forecasting methods.

The M5 organizers provided twenty-four benchmark methods for the point forecasting task, five of them used TD approach and the remainder were BU. According to their scores and ranks on the private leaderboard[3], the performance of the TD and BU approaches were depended on what base forecasting method was used, i.e., the forecasting methods that were used for modeling top level or bottom level sales series. For example, when using Exponential Smoothing as the base forecasting method, BU outperformed TD; when using ARIMA, TD performed better than BU. However, for all the benchmarks, their base forecasting methods used the local method of estimation.

An important result of the M5 is that nearly all the top-ranking solutions employed a Global BU approach (GBU), which means using global forecasting methods to generate the bottom level forecasts and then using the BU strategy to obtain coherent forecasts for more aggregate levels. The scores of the top-50 M5 teams show that they all outperformed the best benchmark model significantly, with even the worst WRMSSE 16% less than the best benchmark. The documents and codes provided by the top-ranking solutions showed an obvious difference between the winning solutions compared with the benchmarks is that while most of them also used a BU approach to achieve coherent forecasts, their base forecasting methods all relied on global estimation.

Global methods can learn the common data-generating scheme among different time series

[3] The M5 accuracy competition had two stages: in each stage a test set covering 28 days of sales was used to evaluate submissions. In the first stage, players were allowed to submit their forecasts multiple times, and could obtain their evaluation scores as feedback instantly so that they could then revise and resubmit their forecasts if they so choose. Their scores and ranks were published on the 'public leaderboard' throughout the period. In the second stage, the test data was updated to add the following four weeks to that used in the first stage, and though the participants were still free to (re)submit their forecasts during the given period of time, no feedback was given until the end of the competition. The 'private leaderboard' thus is used at the end of second stage for announcing the final evaluation and the rank of the submissions.

using fewer parameters than the local time series specific alternative. It has been shown to provide superior performance over local methods when forecasting a large number of related time series (Ma & Fildes, 2020). But its performance has so far not been strictly evaluated within a hierarchical forecasting setting as far as we are concerned. According to M5's scoring rule, the score of bottom level forecasting accuracy only accounted for one twelfth of the total. As a consequence, superior performance at the bottom level does not necessarily lead to more accurate forecasts at higher aggregate levels, a phenomenon clearly identified in the M5 summary paper (Makridakis, Spiliotis, & Assimakopoulos, 2020).

The final rankings on the private leaderboard of M5 showed that using a GBU approach achieved significantly more accurate forecasts across all the 12 levels in the hierarchy compared to the organizers' benchmarks. However, whether the superior performance of the GBU is robust[4] over various test periods or only an accidental result arising in the final private leaderboard, effectively on an arbitrary sub-set of the Walmart data, is an important question for retail forecasting researchers and practitioners. The key question is whether the ranking results from this limited test set are robust and generalizable.

2. GBU vs. MO and ORC: a robustness check with M5 data

To check the robustness of the GBU, we conducted a series of forecasting experiments using five rolling test sets with M5 data. The five test sets are in sequence, covering periods from January 1, 2016 to June 19 of the same year, each test set spans twenty-eight days.

We tested two GBU models implemented with a gradient boosting framework named LightGBM (Ke, et al., 2017) which was employed in nearly all the top ranking solutions of M5. Both GBU models were trained using the pooled data which consisted of all the sales time series at the bottom level of the retail hierarchy (30,490 series). The main difference between the two GBU models is that one model employs rolling lag features to generate one day ahead forecasts first and then uses a recursive strategy to generate 2-28 days ahead forecasts iteratively, and the second excludes the rolling lag features so that it can generate 1-28 days ahead forecasts directly (the nonrecursive strategy). Both strategies were widely employed in the M5 top-

[4] Robustness of a forecasting method here refers to how well a method works on alternate segments of the data.

ranking solutions. The features used in both GBU models have been taken from the winning solution.

For a comparison, we also developed three additional hierarchical forecasting models as benchmarks. The first benchmark model used the BU strategy too, but the bottom level forecasts were generated with a local forecasting method based on Lasso regression. The predictors of the regression included the product price, the seasonal dummies (day of week, month of year), and calendar events. The lagged sales were not used as the predictors in the regression so that it could generate 28 days ahead of forecasts directly (a nonrecursive strategy). The second benchmark used the optimal reconciliation (ORC) approach of Hyndman, et al. (2011). Similar to the first benchmark, Lasso regression was employed as the base forecasting model, but the regressions were trained for all 42,840 time series in the 12 levels of the hierarchy separately to generate non-coherent forecasts first, and then the ORC approach was used to obtain coherent forecasts. The predictors in the regression for modeling an aggregate level time series included the average price of the products in the aggregation, the seasonal dummies (day of week, month of year), and calendar event dummies. The third benchmark used the MO approach, which generated forecasts at the ninth level, i.e., the store-department level which consists of 70 aggregated time series. Similar to the above mentioned nonrecursive GBU model, we also used a global method based on LightGBM to train one integrated model for the sales time series of the 70 store-departments, we therefore named this model as Global MO (GMO). The features used in the model were some statistical summaries of the bottom level features (e.g., the price at store department level, which is the average price of all the items of the store department). The top-down decomposition was based on each product's previous four weeks' average sales share in the store-department before the forecasting periods.

Table 1 compares the accuracy of the various models. To save the space, in Table 1 we only report the WRMSSE scores on level 1, level 9, level 12, and the average WRMSSE over the 12 levels. At the bottom level (level 12), the GBU models (Lgb-BU) consistently outperform all the benchmarks on all the five test periods. This shows the accuracy gains from using global forecasting methods at the store-product level are very robust. On comparing the two GBU models, the recursive model (Lgb-Rec-BU) outperformed the non-recursive (Lgb-noRec-BU)

on all the five test periods. This is an interesting result, as the recursive strategy has been criticized as accumulating errors when forecasting horizons are long[5].

Table 1. The performance of the GBUs and the benchmarks evaluated by WRMSSE across five test periods (relative ranks of the five models on WEMSSE are reported in parentheses)

		Lgb-Rec-BU	Lgb-noRec-BU	Lgb-noRec-MO	Las-noRec-ORC	Las-noRec-BU
Level 1 (1)	Private	0.537 (5)	**0.188 (1)**	0.277 (2)	0.446 (4)	0.404 (3)
	Public	**0.213 (1)**	0.565 (5)	0.368 (2)	0.546 (4)	0.484 (3)
	Cv3	0.589 (5)	0.238 (2)	**0.220 (1)**	0.312 (3)	0.354 (4)
	Cv2	0.558 (5)	0.550 (4)	**0.410 (1)**	0.444 (2)	0.471 (3)
	Cv1	0.400 (2)	0.561 (3)	**0.298 (1)**	0.686 (4)	0.780 (5)
Level 9 (70)	Private	0.660 (3)	**0.566 (1)**	0.633 (2)	1.405 (5)	0.719 (4)
	Public	**0.544 (1)**	0.660 (3)	0.636 (2)	0.965 (5)	0.715 (4)
	Cv3	0.670 (3)	0.573 (2)	**0.561 (1)**	0.715 (5)	0.631 (4)
	Cv2	**0.694 (1)**	0.710 (3)	0.710 (2)	0.995 (5)	0.777 (4)
	Cv1	0.641 (2)	0.722 (3)	**0.598 (1)**	1.429 (5)	0.889 (4)
Level 12 (30,490)	Private	**0.867 (1)**	0.872 (2)	0.885 (3)	1.005 (5)	0.953 (4)
	Public	**0.814 (1)**	0.835 (2)	0.842 (3)	0.916 (5)	0.907 (4)
	Cv3	**0.845 (1)**	0.867 (2)	0.875 (3)	0.945 (5)	0.943 (4)
	Cv2	**0.824 (1)**	0.852 (2)	0.853 (3)	0.963 (5)	0.930 (4)
	Cv1	**0.817 (1)**	0.827 (2)	0.819 (3)	0.939 (5)	0.907 (4)
Avg. (42,840)	Private	0.660 (3)	**0.510 (1)**	0.567 (2)	0.881 (5)	0.669 (4)
	Public	**0.477 (1)**	0.652 (3)	0.576 (2)	0.777 (5)	0.661 (4)
	Cv3	0.658 (3)	0.515 (2)	**0.509 (1)**	0.594 (5)	0.592 (4)
	Cv2	0.669 (2)	0.683 (3)	**0.631 (1)**	0.779 (5)	0.706 (4)
	Cv1	0.597 (2)	0.680 (3)	**0.534 (1)**	1.032 (5)	0.852 (4)

Abbreviation of forecasting approaches in the table: Light gradient boosting machine (Lgb), Least absolute shrinkage and selection operator (Las), multiple-step ahead forecasting with recursive approach (Rec), multiple-step ahead forecasting with nonrecursive approach (noRec), hierarchal reconciliation with bottom-up approach (BU), hierarchal reconciliation with middle-out approach (MO), hierarchal reconciliation with optimal reconciliation approach (ORC). 'Private' stands for the test set covering d1942-d1969, which is used for evaluating the private leaderboard of the M5. Similarly, 'Public' covers d1914-d1941, cv3 covers d1886-d1913, cv2 covers d1858-d1885, and cv1, d183-d1857. The number in the brackets of the first column are the number of time series to be forecasted. Bold text in the table shows the best result in the row.

At the top level (level 1) however, the scores of the two GBU models fluctuate dramatically over the five test periods. Recursive GBU (Lgb-Rec-BU) performs extremely well (ranking top) in the test period covering days from no. 1914 to 1941 (evaluated for the public leaderboard), but poorly (ranking bottom) from day 1942 to 1969 (evaluated for the private leaderboard). On the contrary, the nonrecursive BU (Lgb-noRec-BU) performs very well (ranking top) in the

[5] Better strategies are available that overcome the potential problems induced by recursive updating for longer lead times (Ma & Fildes, 2020).

period of the private leaderboard, but poorly (ranking bottom) in the period of the public leaderboard. In the other three test periods, both GBU models show inferior performance compared to the benchmarks. The GMO performance remains stable over the five test periods, with ranking first in CV1, CV2 and CV3, and ranking second in both periods of the private and public leaderboard. For the middle level (level 9), the ranks of those models in terms of WRMSSE are quite similar to their performance on the top level.

The average scores over all twelve levels is also shown in Table 1. As the aggregate levels (level 1-9) take up a large share of the weights in the calculation of the average WRMSSE, the overall ranks are thus quite similar to that at top levels. The nonrecursive GBU performs even better than the M5 winner did in the private leaderboard, though its performance on the other four test sets is not as good as the GMO model. These results explain why the ranks on the Kaggle M5 public leaderboard were so different from those shown on the private leaderboard. Our results of the nonrecursive GBU are consistent with the results reported by the fourth-place team, whose WRMSSEs are 0.630 at CV1, 0.656 at CV2, 0.509 at CV3, 0.613 at the public leaderboard, and 0.535 at the private leaderboard. They used a variant of the nonrecursive GBU and were very surprised by their 4th place result, as their method performed poorly in the test periods so they were not expecting a high rank in the final private leaderboard.

In short, from Table 1 we have learnt that (1) the GBU models work consistently well at the bottom level[6] in various test periods; (2) at the aggregate levels, the performance of the GBU models fluctuated drastically over the different test sets; (3) the performance of the GMO is relatively stable and on average better than the GBU models.

For a BU strategy, the theoretical Mean Squared Forecast Error at the aggregate level can be denoted as

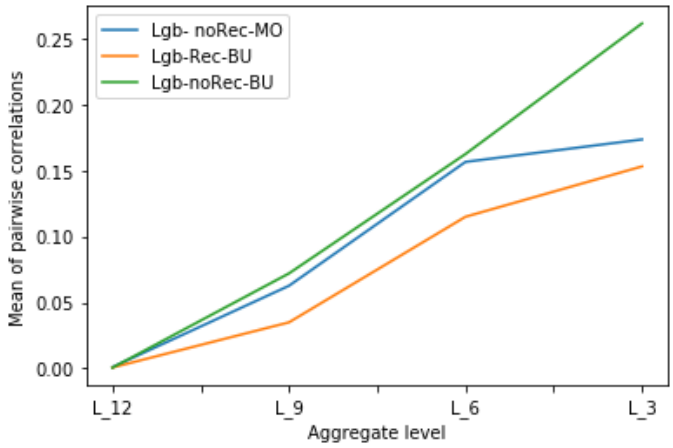
$$MSFE_{BU} = Var\left(D_t - \sum_{i=1}^{N} f_{i,t}\right) = Var\left(\sum_{i=1}^{N} d_{i,t} - \sum_{i=1}^{N} f_{i,t}\right) = \sum_{i=1}^{N} Var\left(d_{i,t} - f_{i,t}\right) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} Cov\left(d_{i,t} - f_{i,t}, d_{j,t} - f_{j,t}\right) \quad (1)$$

where D_t is the aggregated demand in period t; $d_{i,t}$ is the bottom level demand of item i in period t; $f_{i,t}$ is the bottom level forecast of demand item i in period t; N is the number of items in the

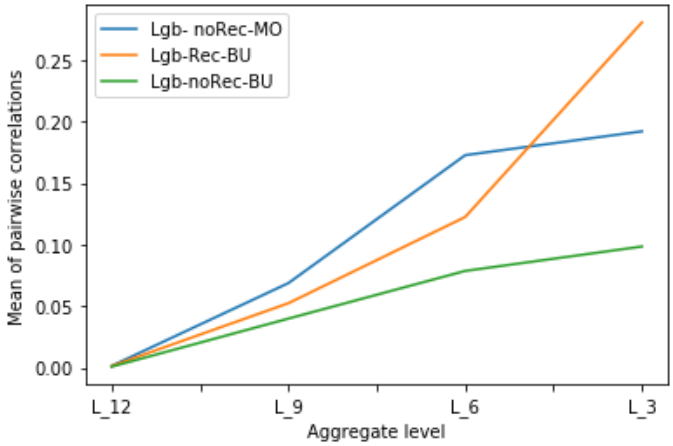
[6] The ranks are nearly the same from level 10 to 12 for the models in Table 1, though so only reported the results on the level 12.

bottom level. Equation (1) shows that the aggregate level theoretical MSFE of a BU model can be decomposed into two parts: a) the sum of bottom level MSFE which is usually used as the loss function when training a GBU model; b) the sum of bottom level cross-series Pairwise Error Covariances (PECs), which is totally neglected during the GBU training. The summation of PECs at the bottom level thus has the potential to explain why the GBU performed consistently well at bottom level (level 12 in Table 1), but erratically at aggregate levels (level 1 & 9 in Table 1). The M5 data sampled multiple products from each store, category, and department, which lead to complex demand correlative patterns at the store-item level. Store or regional level unobservable demand impulses (e.g., extreme weather) can affect the distribution of PECs temporally. Fig. 1 shows how the mean pairwise error correlations[7] change over the different test periods by aggregation level. In the test periods of the public leaderboard, the recursive GBU has lowest pairwise correlations compared to nonrecursive GBU and GMO (Figure 1a); but in the periods of the private leaderboard, it has highest pairwise correlations when the aggregation level is above level 3 (Figure 1b). When calculating the pairwise correlations using the forecast errors covering all five test periods (Figure 1c), the GMO has the lowest average pairwise correlations across the hierarchy. This indicates the temporal changes in the PECs' distribution is the main reason for the inconsistent performance of the GBU models. The MO approach can reduce the chance of bottom level error correlations as the forecasts are decomposed from the aggregate level forecasts according to a predefined proportion, but at the cost of higher bottom level MSFE as the loss of the detailed information during before-training aggregation. Therefore, there is a trade-off to selecting the appropriate middle level for forecasting using the GMO approach.

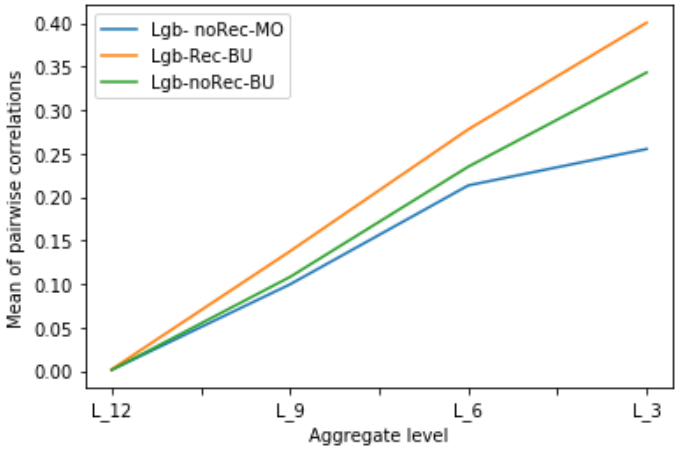
[7] As the scales of the pairwise error covariance are very different for different levels in the hierarchy, in Fig.1 we instead show the pairwise error correlations for a better visualization.



(a) Average forecast pairwise error correlations in the periods of the public leaderboard



(b) Average forecast pairwise error correlations in the periods of the private leaderboard



(c) Average forecast pairwise error correlations over five test periods

Figure 1. Average error pairwise correlations at various levels and test sets

3. Can the performance of the global BU forecasts be improved?

As we have shown in Table 1, the GBU models show inconsistent performance over the

various test periods, and this problem might had been discovered by most of the M5 top players leading them to explore various improvement strategies. In this section, we go on to check whether the further efforts made by the top-ranking players could improve the robustness of the GBU.

Data partition and forecast combination

When using global methods for forecasting many related time series, one question that has often been raised is whether to pool all the time series together to train one integrated model or partition the time series into a number of clusters and then train several global models, one for each of these clusters. Theoretically, global forecasting should benefit from pooling as many related time series as possible. Previous time series Kaggle competitions[8] have shown the superior performance of such fully integrated global models.

Among the M5 top five solutions, however, none of them pooled all the store-product level time series together to train one global model (as we did in Table 1), but instead trained a set of global models per store, category, or department[9]. For example, the first placed player considered an equal weighted combination of recursive and non-recursive LightGBM models that were trained to produce forecasts for the store-item series using the data per store (10 models), store-category (30 models), and store-department (70 models). So the first question addressed here is whether the data partition delivers an improved method for GBU?

Table 2 shows the performance of six GBU models used by the first placed team as well as their equal weights combination on the five test sets[10]. Considering the results from Table 1 and Table 2, we find that at the bottom level (level 12), for both the recursive and nonrecursive approaches, the more products that are pooled together for training, the higher the average forecasting accuracy that can be achieved. For example, the models trained by pooling all the store products together (Table 1) performed on average better than the same models trained per store (Table 2), the models trained per store performed better than the same models trained per store-category, and so on. At the aggregate levels, however, the GBU models' performance are

[8] See Bojer and Meldgaard (2020) for a summary.

[9] A popular reason for data partitioning during the M5 is to cope with the limited amount of RAM available at the Kaggle platform, however, for most of the players, they used their own workstations or employed extra cloud computing resources.

[10] The forecasts are generated using their presented codes.

inconsistently well with the size of data partition used to train the model: sometimes the model trained with smaller clusters of products (e.g., per store-department with recursive model (Lgb-Rec-BU) in the private leaderboard) even provided better aggregate level forecasts than the same model trained with larger clusters (e.g., per store). According to the Equation (1), the explanation as to the benefit of data partition is it may sometimes reduce the error correlations for items in different clusters.

Forecast combination has been proved an effective way to improve forecasting accuracy and robustness in many forecasting competitions. In M5, all the top-ranking teams in their final submissions adopted the idea of combination. For example, as we have mentioned, the first-place solution combined the forecasts of six GBU models.

Table 2. The WRMSSE score of the models at various levels used by the M5 winner

Model		Lgb-Rec-BU	Lgb-Rec-BU	Lgb-Rec-BU	Lgb-noRec-BU	Lgb-noRec-BU	Lgb-noRec-BU	Comb-BU
Data partition		Per store	Per store category	Per store department	Per store	Per store category	Per store department	--
Level 1 (1)	Private	0.540	0.337	0.247	0.208	0.263	0.390	**0.201**
	Public	**0.214**	0.298	0.411	0.784	0.741	0.898	0.540
	CV3	0.561	0.450	0.315	0.329	0.402	0.456	**0.232**
	CV2	0.578	0.512	**0.454**	0.620	0.726	0.729	0.470
	CV1	0.445	**0.354**	0.361	0.682	0.776	0.830	0.455
Level 9 (70)	Private	0.668	0.616	0.601	0.587	0.601	0.626	**0.572**
	Public	**0.536**	0.559	0.604	0.750	0.745	0.836	0.643
	CV3	0.652	0.642	0.621	0.600	0.628	0.658	**0.578**
	CV2	0.712	0.682	0.679	0.764	0.799	0.796	**0.675**
	CV1	0.660	0.636	**0.629**	0.758	0.806	0.825	0.654
Level 12 (30,490)	Private	**0.866**	0.869	0.872	0.876	0.878	0.881	**0.866**
	Public	**0.816**	0.821	0.825	0.841	0.843	0.848	0.825
	CV3	**0.848**	0.855	0.858	0.873	0.877	0.879	0.856
	CV2	**0.831**	0.831	0.835	0.855	0.857	0.860	0.835
	CV1	0.825	0.828	0.828	0.836	0.840	0.842	**0.823**
Avg. (42840)	Private	0.667	0.579	0.548	0.525	0.547	0.599	**0.518**
	Public	**0.474**	0.513	0.569	0.773	0.757	0.852	0.634
	CV3	0.645	0.612	0.564	0.553	0.593	0.623	**0.518**
	CV2	0.686	0.647	**0.626**	0.728	0.779	0.779	0.632
	CV1	0.620	0.580	**0.577**	0.739	0.798	0.824	0.617

Abbreviations in the table: combination of the forecasts of six models on the left (Comb). The others are the same with that in Table 1. Bold text in the table shows the best result in the row.

The last column of Table 2 shows the performance of their combining model on the five test sets. On average, the winner's combination strategy performed well in the private leaderboard and CV3: its WRMSSE being smaller than its best performing component model. On the other three test sets however, its average performance fell between its best and the worst component models. Overall we see that the combining model reduces the WRMSSE fluctuations over test periods and enhances the robustness of the GBU in general, but the combination performs still more poorly than the GMO reported in the Table 1 in four of the test periods.

Judgmental adjustments

The second and fifth place solutions relied heavily on judgmental adjustments to their GBU forecasts. For example, the second-place team adopted a GBU approach to generate bottom level forecasts first, and then used a neural network model named NBeats to forecast time series on the top five levels; then they compared the daily differences between the forecasts of NBeats with that of BU on top five levels by aid of visualization. They adjusted their bottom forecasts day-by-day using judgmental multipliers[11] according to the observed differences. Similarly, the fifth-place team's adjustments were based on the store department level performance of their GBU forecasts in the previous test period[12].

Table 3. The performance of NBeats and benchmarks over top-five levels in terms of WRMSSE

	NBeats		Lgb-Rec-BU		Lgb-noRec-BU		Lgb-noRec-MO	
Level	private	public	private	public	private	public	private	public
Level 1 (1)	0.344	0.585	0.537	0.213	0.188	0.565	0.277	0.368
Level 2 (3)	0.414	0.604	0.549	0.304	0.298	0.567	0.366	0.433
Level 3 (10)	0.464	0.635	0.568	0.398	0.387	0.597	0.457	0.504
Level 4 (3)	0.381	0.595	0.530	0.260	0.247	0.554	0.310	0.392
Level 5 (7)	0.450	0.629	0.548	0.339	0.349	0.566	0.426	0.470

Abbreviations in the table: Neural basis expansion analysis for interpretable time series forecasting (NBeats). The other methods are the same as shown in Table 1. Bold text in the table shows the best result in the column.

The idea of using aggregate level forecasts to adjust the forecasts of the bottom seems promising. The aggregate level forecasts suffer the problem of information loss due to

[11] For more detail, see participants' codes and summary document.

[12] https://github.com/Mcompetitions/M5-methods/tree/master/Code%20of%20Winning%20Methods/A5

aggregation, while some common seasonal or temporal trend signals that are strong at aggregate level may be too weak to be captured by the bottom level forecasts. So the assumption behind such judgmental adjustments is that they can deliver better aggregate level forecasts when using TD than BU. If this assumption is not supported, the adjustments may, on the contrary, worsen the BU forecasts.

For the fifth-place team's approach, from Table 1 and 2, the GBUs' performance on aggregate levels fluctuated drastically on different test panels. Therefore, there is no consistent evidence their adjustments lead to improvements (maybe the winner required no adjustments as we showed in Table 1). In Table 3, we show their NBeats forecasting performance on both the private and public leaderboards on the M5 top five levels. We find that it fails to provide better forecasts than the nonrecursive BU and MO approach in either the public or private leaderboards, though it indeed performs better than the recursive BU in the private leaderboard.

Table 4. Optimal reconciliation on GMO and GBU

Level 1-9		Lgb-noRec-MO	Lgb-noRec-MO	Lgb-noRec-MO
Level 10-12		Lgb-Rec-BU	Lgb-noRec-BU	Lgb-comb-BU
Level 1 (1)	Private	0.277	0.277	0.277
	Public	0.368	0.368	0.368
	Cv3	0.220	0.220	0.220
	Cv2	0.410	0.410	0.410
	Cv1	0.298	0.298	0.298
Level 9 (70)	Private	0.633	0.633	0.633
	Public	0.636	0.636	0.636
	Cv3	0.561	0.561	0.561
	Cv2	0.710	0.710	0.710
	Cv1	0.598	0.598	0.598
Level 12 (30,490)	Private	0.865	0.875	0.865
	Public	0.817	0.838	0.823
	Cv3	0.850	0.866	0.852
	Cv2	0.828	0.853	0.834
	Cv1	0.817	0.829	0.817
Avg (42,840)	Private	0.560	0.564	0.560
	Public	0.568	0.577	0.571
	Cv3	0.500	0.506	0.500
	Cv2	0.621	0.632	0.624
	Cv1	0.532	0.538	0.532

Abbreviations in the table are the same as those in Table 1.

If we could obtain better aggregate levels forecasts with a particular method (in contrast to NBeats), the ORC would be the better technique than the judgmental adjustments, as the ORC approach can reconcile the forecasts on different levels in an automatic way. In Table 4, we use the GMO model reported in the Table 1 to forecast the M5 top nine levels. The bottom three levels are then forecasted using (i) the recursive GBU model, (ii) the nonrecursive GBU model, and (iii) their equal weight combination. Lastly, the ORC was used to reconcile the GMO and GBU forecasts. We can find in the Table 4 that the performance of the three ORC forecasts at the top nine levels are nearly the same as that of the GMO model, while their performance on the bottom three levels are close to that of their respective GBU models. Therefore, the ORC approach is capable of preserving the advantages of both GBU and GMO approaches to provide more robust forecasts.

4. Conclusions

In this commentary we focus on the issue of the robustness of the GBU approach over various test periods. We find that global forecasting models base on the LightGBM in general performed very well for the bottom levels, but lacked robustness when the bottom forecasts were added up to aggregate level forecasts. The reason is due to the complex time-varying error correlative patterns among the thousands of store products, which is not considered in the loss function during the training process of a GBU. The M5 winning teams made efforts to improve the robustness of the GBU approach, including data partition, model combination, and judgmental adjustments. Though these efforts did enhance the robustness of the GBU in some extent, these additional heuristics still failed to deliver consistent performance across time panels compared to the GMO. This lack of robustness made the ranks in the winner board of M5 somewhat arbitrary and undermines any winning claims of the leader.

We found that the GMO approach provided a better solution for aggregate level forecasts when the proper middle level in the hierarchy is selected, while GBU approach can deliver superior performance for the bottom levels in the hierarchy. When using the ORC technique to combine the forecasts generated by the GBU and GMO models, we can obtain better hierarchical forecasts than either GMO or GBU.

The importance of the robustness leads to our final recommendation: in future forecasting

competitions, e.g., M6, we suggest the use of multiple rolling test sets to evaluate forecasting performance instead of only one in order to reward robust forecasting methods.

References

Bojer, C. S., & Meldgaard, J. P. (2020). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*.

Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*.

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

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

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

Januschowski, T., Gasthaus, J., Wang, Y., Rangapuram, S. S., & Callot, L. (2018). Deep learning for forecasting: current trends and challenges. *Foresight: The International Journal of Applied Forecasting, 51*, 42-47.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3149–3157). Long Beach, California, USA: Curran Associates Inc.

Ma, S., & Fildes, R. (2020). Forecasting third-party mobile payments with implications for customer flow prediction. *International Journal of Forecasting, 36*, 739-760.

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*, 245-257.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). *The M5 Accuracy competition: Results, findings and conclusions*.

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