Forecasting system's accuracy: a framework for the comparison of different structures

Carla Freitas Silveira Netto, Vinicius A. Brei, Rob J. Hyndman

Abstract: One of the most challenging aspects for managers when building a forecasting system is choosing how to aggregate the data at different levels. This is frequently done without the manager knowing how these choices can compromise the system's accuracy. This paper illustrates these compromises by comparing different structures and aggregation criteria. Our paper proposes and empirically tests a framework on how to build a coherent and more accurate forecasting system. The framework's first phase compares different time series forecasting methods, including statistical, "standard" machine learning, and deep learning. Results show that one of the statistical methods (autoregressive integrated moving average, or, for short, ARIMA) outperforms machine and deep learning methods. The second phase compares different combinations of aggregation criteria, structures of the forecasting system, and coherent forecast methods (i.e., adjustments to the forecasts at different levels of aggregation). The results show that using different criteria and structures indeed impacts predictions' accuracy. When it is necessary to disaggregate the forecast, our results show that it is best to add more information in a grouped structure, adjusted by a bottom-up method. This combination provides the best performance, i.e., the lowest mean absolute scaled error (MASE) in most nodes, compared to the other structures and coherent forecast methods used. The results also suggest that aggregating the time series further by geographical regions is essential to improve accuracy when forecasting products' and channels' sales.

Keywords: forecasting; sales; aggregation criteria; hierarchical time series; grouped time series

1. Introduction

Handling data structured in different levels of aggregation, either hierarchically or grouped, is ubiquitous in a manager's job. For product evaluation, forecasts by product category are required. One of the most commonly used forecast criteria to evaluate business partner performance is channel type. To define sales territories, it is necessary to have accurate estimates by geographical units [1]. Classical literature already showed that if those forecasts are performed independently, ignoring the aggregation constraints, they will not provide the same insight into the future [2,3]. The problem with aggregation is that forecasts may not naturally add up, following the structure of the forecasting system (from disaggregate to aggregate levels). If managers use these predictions to plan their actions, they can lead to confusion and wrong decisions. All hierarchical levels in an organization must receive the same forecast to ensure aligned decision-making [4]. For that reason, managers need to adjust forecasts estimated at different aggregation levels to become "coherent," being able to add them up, similar to data.

The coherent forecast methods take advantage of the informative signals from all levels of aggregation by combining them, which improves forecasting accuracy [4]. These coherent forecasts, either by level or group, can add precision to the decision to focus efforts on products, partners, departments, regions, and channels that are more likely to be profitable. Therefore, the literature should assist managers by defining which type of data and aggregation strategy produces the best sales forecasting results. Although researchers have studied the aggregation issue in forecasts for a long time [e.g., 2, 3], there is no consensus about the best criteria to determine its levels [5]. It is still unclear whether and how different criteria impact accuracy and which criteria improve it the most. Most of the aggregation criteria studied to date are based on the similarity of the time series. Previous studies used, for example, clustering methods [6, 7], temporal aggregation [8], or product aggregation, based on sales volume or seasonal patterns [see 4, 9, 10]. However, a comparison of criteria based on different market-related variables is still missing. This paper answers a crucial question relevant to management practice by comparing different forecasting system structures using different market-related variables.

Forecasting methods are constantly evolving, and sales forecasting has recently received increased interest in the retail context [see 11 for a comprehensive review of the current forecasting literature in retail, and 12]. However, business-to-business (B2B) sales forecasting has received less attention in the academic literature. Similarly, machine learning applications to the B2B context have seen little improvement [13,14,15,16]. As a rule, modern forecasting methods have not been widely implemented in most businesses' daily routines since, in practice, they still rely on judgmental (qualitative) methods and adjustments [13, 36, 38]. This context should be studied separately from business-to-consumer (B2C) because it is more complex and involves different companies with conflicting interests [13].

In this manuscript, we contribute to the forecasting literature in the B2B context by comparing the performance of 18 statistical, "standard" machine learning, and deep learning methods in five folds of the dataset. Then, we compare the different combinations of aggregation criteria and show that different choices lead to different accuracy results. Using a seven-year manufacturer's monthly time series sales records, we propose a framework for comparing which criterion (product category, channel type, or geographic location) combined with which structure (hierarchical or grouped time series) and which of the seven different coherent forecast methods improves forecast accuracy by the most. In total, 69 different time series performances are compared in our framework. Our results show that, for the dataset used, combining the classical time series method autoregressive integrated moving average (ARIMA), bottom-up method, and a grouped structure that combines all criteria provides the

best results, i.e., lower mean absolute scaled error (MASE) in more nodes. To the best of our knowledge, this comparison has not yet been addressed in the literature. Thus, this paper aims to test and recommend a framework for choosing which aggregation criteria, structure, and coherent forecast method when building a forecast system.

This paper offers contributions to the theory, method, and practice. To theory, we advance the knowledge by providing a framework for building a forecasting system that uses one of the most basic kinds of variables in the field: market-related variables. Since hierarchies have seen limited implementation [11], we expand the study of hierarchies through a large-scale hierarchical and grouped forecast study, including several aggregation criteria. To method, we provide evidence that different criteria and structures impact forecasting accuracy. We also propose and empirically test a framework that uses state-of-the-art time series forecasting methods. To practice, we test different criteria and structures that are easy to implement by organizations of any sector or size, with minimal additional investments. We use automatic forecasting tools performed in open-source software based on data easily accessible to most companies, i.e., their sales records. We also contribute to practice by offering guidance on choosing which aggregation strategy can lead to better decisions regarding budget allocation and more precise actions at the point of sales. We show that managers can use this strategy to improve their sales forecast accuracy by combining readily available information.

This paper is structured as follows. Section 2 presents the relevant literature. First, we review the current discussion on B2B sales forecasting literature. Then, we discuss the different aggregation criteria studied in the literature and describe the coherent forecast methods. Following, we discuss in Section 3 each phase of our analysis, describing the data and method. Section 4 describes the forecast accuracy of the different combinations of structures, aggregation criteria, and coherent forecast methods. Section 4 also presents a discussion of the results. We conclude by presenting implications to the literature and practice, followed by suggestions for future research.

2. Relevant literature

2.1 Sales forecasting in B2B literature

Our review focuses only on the few recent studies that compared statistical techniques with machine learning in the B2B context. Studies that applied machine learning (ML) techniques to the B2B sales forecasting context mainly focused on binary outcomes (i.e., classification problems). Rohaan, Topan, & Groothuis-Oudshoorn [16] estimated whether sales were successful after a quote. Their model that combined natural language processing (NLP) with ML performed better than manual categorization and logistic regression. Bohanec, Robnik-Šikonja, and Borštnar [13] also framed the forecasting problem as a classification problem, estimating the probability of successful closure. These authors compared the performance of five different ML models, with the random forest being the best-performing in both experiments. Rezazadeh [15] used an ensemble of machine learning techniques, again with a binary outcome. The ensemble predictions performed best compared to “user-entered” (judgmental).

However, Bohanec, Robnik-Šikonja, Borštnar [13], and Rezazadeh [15] did not compare their model performances with other statistical models. Lackman [14], on the other hand, did not compare the performance of his proposed model to ML. The author forecasted B2B sales by product line using a simulation model. The focus was not to compare with other methods but to identify variables that could help improve forecasting, balancing parsimony, and comprehensiveness.

Lei, Li, & Yu*'s* [17] study is the most related to our research. The authors also estimate forecasts at different levels of aggregation. However, they did not compare different

structures and criteria but only product aggregations. The authors also chose not to apply machine learning models based on the results of Makridakis, Spiliotis, and Assimakopoulos [18]. The models proposed by Lei, Li, & Yu [17] are based on "associated relationships" (correlation and Granger causality) and achieved better performance at the most aggregate level in the hierarchy. Their model also highly depends on the user's setting of the critical values. To date, the literature has not provided a comparison of machine learning and statistical models to forecast sales using continuous outcomes in the B2B context.

In the forecasting literature, despite the recent hype over machine learning techniques, their accuracy results are mixed [18]. Evidence of the accuracy of machine learning techniques is scarce, and they do not outperform traditional statistical methods. Makridakis, Spiliotis, and Assimakopoulos [18] recommend a few procedures that could allow a fair comparison between statistical and machine learning techniques. For instance, these comparisons should evaluate longer forecast horizons (not only one step ahead), use benchmarks, and be applied to multiple time series, not just one.

2.2 Forecasting systems structure and criteria

Forecasting systems or forecasting support systems (FSS) are defined as a framework or structured process that is repeated in time, allowing for a comparison of the outcomes, not always in the form of software [37, 39]. The key point is that it gives a set of procedures that should be followed in a certain order and creates, in time, a database of previous forecasting performance. In this paper, we study the structure of an FSS fully reliable on quantitative forecasting methods. Judgmental adjustments are out of the scope of our study. We refer to Van den Broeke *et al.* [36] and Fildes *et al.* [37] for studies in the combination of quantitative and judgmental, and Arvan *et al.* [39] for a literature review on the topic.

Estimating time series accurately at different hierarchical levels is not trivial. These forecasts can be challenging since the various levels present different patterns and are not on the same scale [4]. They are, in all senses, different time series, even when they come from the same organizational structure. This difficulty increases with disaggregation, as data become noisy and often intermittent [19]. Another aspect to consider is aggregation structure. Grouped time series are aggregated based on criteria such as product characteristics, geographic regions, or customer characteristics. These criteria can also be represented in a "hierarchical time series" tree structure, as shown in Figure 1. In management contexts, hierarchies commonly appear due to geography (i.e., sales disaggregated by state, region, city, and store) and product classification (i.e., brands disaggregated into groups, subgroups, and finally, products). An example of a grouped time series is when geographic and product hierarchies are used simultaneously, such as when one wants to forecast different products in different regions.

INSERT FIGURE 1 HERE

Following Hyndman and Athanasopoulos' [20] notation, we denote the data at the most aggregate level at time t by y_t ($t = 1,\ldots,T$). Disaggregated data are denoted by $y_{j,t}$, with j corresponding to the "node" (element of the structure) of the observation. In this way, the time series of Figure 1 can be written as follows.

Bottom level: $y_t = y_{AA,t} + y_{AB,t} + y_{AC,t} + y_{BA,t} + y_{BB,t}$ (1)

Middle level: $y_{A,t} = y_{AA,t} + y_{AB,t} + y_{AC,t}$ (2)

$$y_{B,t} = y_{BA,t} + y_{BB,t} \quad (3)$$

Top-level: $$y_t = y_{A,t} + y_{B,t} \, . \quad (4)$$

We let $\hat{y}_h$ denote the vector of forecasts for all nodes (elements of the structure) at horizon h, stacked in the same order as y_t. y_t denotes the time series data in the training set. $\hat{y}_h$ refers to the forecasted time series in a number (h) of steps (e.g., months) ahead. These forecasts can come from any appropriate model. They are created independently for each node without regard to the hierarchical or grouped structure of the data. For example, aggregated forecasts may be obtained from an ARIMA model, while disaggregated series forecasts may come from a Delphi process for each sales division. We call these "base" forecasts.

Several decisions can influence the system's performance when setting up a forecasting system with hierarchical or grouped time series [6]. One of these is what aggregation criteria to use. Fliedner and Mabert [10] studied the influence of different grouping criteria on hierarchical forecast performance. However, their work was based entirely on product type. They compared forecasts by dividing products by unit volume, dollar volume, seasonal index, or historical performance. They concluded that the criteria used to determine the groups for forecasting are determinants of the success of a forecasting system. However, the number or size of the groups did not have a significant impact. Fliedner and Lawrence [6] found no evidence that the added sophistication of clustering methods improved forecast performance. Their study compares different groups based on volume generated by cluster techniques. According to the authors, the grouping of items is responsible for improved forecast performance, not the group formation process (i.e., the clustering technique applied).

Previous studies have also aggregated time series based on their similarities or using clustering methods [e.g., 6, 7, 9, 21] or compared different temporal aggregations [e.g., 8]. However, when the groups do not follow market-related criteria, they are less useful for budget plans and strategy development. While Divakar, Ratchford, and Shankar [22] focused on forecasting by channels, Oliveira and Ramos [4] focused on products. However, these authors did not address the comparison between different criteria based on market-related variables.

Yang and Shang [23] also analyzed other group structures, not in a business context but in a mortality rate forecasting context. The authors found that different group structures affect forecasting performance and that disaggregating by geographical area and gender improved forecasting accuracy. Mircetic *et al. 's* [21] study has explored hierarchical forecast comparison (coherent forecast methods performance at different levels). However, it focuses on the effects of time series characteristics on the performance of hierarchical forecasting approaches. Our study offers a different perspective by analyzing how structures, criteria, and coherent forecast methods impact forecasting accuracy. Table 1 shows how we fill this gap by summarizing the literature on structures and aggregation criteria.

INSERT TABLE 1 HERE

2.3 Coherent forecast methods

An organization's different forecasting needs must be incorporated as part of its forecasting systems, which should be "coherent." That is, disaggregated forecasts should add up to the total forecast in the same way as historical data [19]. Figure 2 shows an example of the problem in a simple hierarchy. The data (light gray in Figure 2) divides correctly, regardless of the aggregation level. However, if forecasts are estimated at different levels independently of the structure (in black), they will not sum up to the same value. To turn these disaggregated forecasts into "coherent" forecasts, one must apply a method to adjust those forecasts. Two of these methods are shown in Figure 2. The Bottom-up method would be to forecast the most disaggregated level and then sum up each level. The Top-down method would be the opposite, forecasting only the aggregate level, and dividing down based

on historical proportions. If the simulated data in Figure 2 were real, the average accuracy of the top-down — Mean Absolute Percentage Error (MAPE)=3.84 — method would be higher than the bottom-up method (MAPE=25). To forecast independently would also lead to lower accuracy (average MAPE = 15.96) and incoherent forecasts between levels.

INSERT FIGURE 2 HERE

That is why forecasts in a forecasting system with different levels of aggregation need to be adjusted to ensure that they add up appropriately. The hierarchical forecasting literature advises on how to distribute the forecasts throughout different levels [20]. Historically, the central theoretical question has been whether to forecast aggregated data and divide them among different levels or produce disaggregated forecasts and add them up. Let $\tilde{y}_h$ denote the adjusted forecasts. Then, they can be expressed as

$$\tilde{y}_h = \boldsymbol{R}\hat{y}_h \tag{5}$$

where $\boldsymbol{R}$ is a matrix that can be decomposed as $\boldsymbol{R} = \boldsymbol{SG}$, and $\boldsymbol{S}$ denotes a summing matrix representing the data's aggregation structure (groups or hierarchies). The matrix $\boldsymbol{G}$ depends on the method to be used. The most common method of obtaining coherent forecasts is bottom-up forecasting. It involves simply summing the most disaggregated forecasts to obtain forecasts for the other series of the structure. This method corresponds to setting $\boldsymbol{G}$ equal to an identity matrix in the right-hand columns and all zeros to the left. That is,

$$\begin{bmatrix} \tilde{y}_h \\ \tilde{y}_{A,h} \\ \tilde{y}_{B,h} \\ \tilde{y}_{AA,h} \\ \tilde{y}_{AB,h} \\ \tilde{y}_{AC,h} \\ \tilde{y}_{BA,h} \\ \tilde{y}_{BB,h} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{y}_h \\ \hat{y}_{A,h} \\ \hat{y}_{B,h} \\ \hat{y}_{AA,h} \\ \hat{y}_{AB,h} \\ \hat{y}_{AC,h} \\ \hat{y}_{BA,h} \\ \hat{y}_{BB,h} \end{bmatrix}.$$

While this method has low computational costs and no information is lost during the aggregation, it usually leads to less accurate forecasts [see 4, 19] since bottom-level series are typically noisy. Sales at the item level are more erratic, with greater variation, and may have insufficient data to construct reliable forecasts [9]. Forecasts perform poorly on the aggregate level and ignore the interrelations between the series [4]. However, a recent study by Yang and Shang [23] provides evidence that bottom-up can perform better than other methods.

Two other traditional methods of obtaining coherent forecasts apply only to hierarchical time series: top-down and middle-out. Top-down begins with the total forecast and then divides it into different levels by (1) average historical proportions, (2) average historical values, or (3) proportions based on forecasts. Each of these three top-down methods has different accuracy results. However, methods based on historical proportions tend to lead to less accurate forecasts since these proportions might change over time [19]. Additionally, top-down methods are usually less accurate at lower levels of the hierarchy, not capturing the series' individual dynamics due to aggregation [4]. These top-down methods also do not result in unbiased coherent forecasts [19]. The middle-out method forecasts some middle-level of the hierarchy. It sums the forecasts to generate predictions for the higher levels and

disaggregates them to obtain predictions for the lower levels. It combines the disadvantages of both bottom-up and top-down methods.

Hyndman *et al.* [19] introduced a fourth method, the "optimal reconciliation" method, later refined by Wickramasuriya, Athanasopoulos, and Hyndman [24]. Unlike the other methods, this method considers the structure of the groups or hierarchies and uses more information than the traditional methods. Therefore, it tends to be more accurate. In this method, the matrix $\boldsymbol{G}$ is estimated by minimizing the forecast error variance of the coherent forecasts. The weights that form the matrix $\boldsymbol{G}$ depend on the hierarchical structure and the covariance matrix of the base forecast errors. Hyndman *et al.* [19] called it "optimal" because the difference between the reconciled forecasts and the incoherent base forecasts is minimized. Wickramasuriya, Athanasopoulos, and Hyndman [24] showed that the trace of the forecast error covariance matrix is minimized using "optimal" reconciliation when the base forecasts are unbiased. This method is called "MinT," or minimum trace. The forecasting literature has established that the "optimal approach" using MinT tends to be more accurate, especially at higher levels of aggregation [4]. However, a known problem is that MinT becomes infeasible for datasets with many repeated values or zeros. For this reason, in this manuscript, we used weighted least squares (WLS) in some of the reconciliations later described. WLS is a variation on MinT that uses only the diagonal of the covariance matrix, setting all other values to zero. The following section describes the data and the methodological procedures.

3. Material and methods

We developed our framework using a dataset provided by a major manufacturer of electrical components in Brazil. The intention was to test the accuracy of our proposed method using complex, real sales data. The dataset consisted of the history of sales records of plugs and light switches. Each record is a stock-keeping unit (SKU) sold from the industry to the channels located in Sao Paulo, Brazil, a megalopolis with more than 12 million inhabitants divided into 96 districts and five zones. The data comprised over seven years of sales records, from July 2010 to September 2017. The total number of observations (channel purchases) was 13,719 B2B orders. We have not used any individual customer information or transactions. The database referred only to channels (points of sale), purchases, and their characteristics (e.g., channel type, size, and revenue). We rescaled all the time series to ensure we met the non-disclosure agreements signed by the authors with the data providers. We linearly transformed the sales records (i.e., multiplying by a constant). The categories of products and channels were created based on the descriptions provided by the datasets. The data cleaning consisted of selecting the variables of interest (location, product type, channel type, and sales amount by time period) and aggregating sales in months. Further aggregation of the products, channels, and geographic regions was also carried out so that the amount of missing data would not make the forecasting infeasible.

The time series were divided into a training set for model estimation and a 12-month test set for post-sample evaluation. We maintained this 12-month test set in all scenarios following Makridakis, Spiliotis, and Assimakopoulos' [18] recommendations for a fair comparison between statistical and machine learning techniques. They suggested evaluating longer forecast horizons (not only one step ahead). The training set for the base forecasts varied from 27 to 75 months. After selecting the best method for the base forecasts, we performed all hierarchical and grouped forecasts considering the complete data available, i.e., 75 months for training and 12 for the test set. Other time horizons can be easily implemented in the code that is available for reproduction (Supporting information S2).

We performed all data analysis, forecasting, and output using R. The single model algorithms were implemented using the following R packages: keras [25], forecast [26],

tsibble [27], nnfor [28], tsfknn [29], NlinTS [30], tsDyn [31], and hts [32] for the different grouped and hierarchical structures' procedures. The link to the packages pages is provided in Supporting information S1. Our framework was divided into two phases, with different steps and outputs explained in the flowchart (Figure 3) and the next sections.

INSERT FIGURE 3 HERE

3.1 Phase 1: Aggregated level forecasts

In phase 1 we performed forecasts using only the most aggregated level. The criteria for choosing which algorithms to compare was their presence on the task view page of the Comprehensive R Archive Network (for short, CRAN). This webpage presents a list of packages for time series forecasting curated by an expert in the field. For this reason, the list was a guide to establishing the best methods available for forecasting time series. In Table 2, we briefly describe each model applied.

INSERT TABLE 2 HERE

The model parameters are fitted automatically by the algorithms using information criteria (such as Akaike's information criterion - AIC). For example, for the ARIMA and Exponential smoothing state-space model (ETS), the forecast package searches for the best model according to AIC and Bayesian Information Criterion (BIC) values. It optimizes the parameters using maximum likelihood estimation, forecasts using the selected model for the nominated horizon, and calculates the associated prediction intervals [33]. The parameters for K-Nearest Neighbors (KNN), Vector Autoregressive Multi-Layer Perceptron (VARMLP), Additive Autoregressive (AAR), linear, Neural Networks for Time Series (nneTs), and Logistic Smooth Transition AutoRegressive model (LSTAR) followed the packages' documentation suggestions.

We separated different folds of the data using a rolling forecasting origin. This is a standard approach in the time series forecasting literature [20] and ensures that the model will have different training and test sets on every fold. We ran the forecasts for the five different datasets (five folds) on each model. Only scaled metrics, such as Mean Absolute Scaled Error (MASE), are recommended for this comparison. Further explanations on MASE estimation are in section 3.2. Figure 4 shows the size of each dataset and how the rolling origin works.

INSERT FIGURE 4 HERE

Next, we compared all of the methods on the full dataset based on different metrics: Mean Error (ME), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Percentage Error (MPE), Mean Absolute Percentage Error (MAPE), Mean Absolute Scaled Error (MASE), Autocorrelation of error at lag 1 (ACF1), and Theil's U. We selected the best-performing method for the aggregation step, detailed in the next section.

3.2 Phase 2: Forecasting system structures

We constructed three structures by aggregating product types, channel types, and geographic regions (Table 3). Figure 5 illustrates these different options of structures, with only some of the nodes we had in the dataset, for concision. A total of 339 products were categorized into three types: plugs, light switches, and others. The database comprised 220 points of sales, categorized into four types: distributor, retail, warehouse, and others. Finally, the geographic hierarchy comprised the Sao Paulo city zones (center, east, west, north, and south) and 53 out of the 96 city districts (the omitted districts had no available sales records

during the database time frame). At the most disaggregated level, there were 339 x 220 = 74,580 product-store combinations. However, such disaggregated data are too noisy to be useful, so we did not consider them. We selected these different criteria based on both theoretical and managerial reasons. For managers, having forecasts on these levels and aggregated by these criteria gives them greater knowledge to plan budgets and strategies with channel partners. Our analysis provides insights into which market-related variables can be used as grouping criteria, improving sales forecast accuracy the most.

INSERT TABLE 3 HERE

INSERT FIGURE 5 HERE

First, we estimated the forecasts based on the three hierarchies (1- geography, 2- products, and 3- channels). Next, we combined them two by two in a grouped structure, resulting in three more structures: 4- product and geography; 5- geography and channels; and 6- product and channels. Finally, the seventh structure integrated all three criteria. The order of the criteria is irrelevant, since the number of sales records on channel x, product y, region z will be the same no matter by which criteria one filters the dataset first. For the hierarchical structures, we compared the bottom-up, the three types of top-down, middle-out, and "optimal" methods using MinT and WLS. For some structures, MinT was not feasible (e.g., for datasets with many repeated values or zeros). Thus, for these cases, we present only the WLS results. Only bottom-up and "optimal" are viable for the grouped structure, as the other methods require a hierarchical structure [32]. Therefore, we focused on comparing the bottom-up and the "optimal" methods in the results section.

To evaluate the different coherent forecast methods and grouping criteria, we used the scaled error proposed by Hyndman and Koehler [34] to remove the effect of the scale of the series at each node. MASE is an evaluation measurement that can compare forecasts with different horizons, time frames, or even different time series [34]. Percentage errors are also unit-free but are not helpful when the number of observations is zero or close to zero. Our study kept the time horizon fixed, but the different grouping criteria generated other time series to be compared.

For seasonal time series, MASE is defined by Hyndman and Athanasopoulos [19] as

$$MASE = mean\left(\left|q_j\right|\right) \tag{6}$$

where $q_j = e_j/Q$, e_j is a forecast error,

$$Q = \frac{1}{T-m}\sum_{t=m+1}^{T} \left|y_t - y_{t-m}\right|$$

is the scaling factor computed on the training data, and m is the number of observations per year. MASE returns a value smaller than one if the out-of-sample forecast error is smaller than the in-sample one-step forecast MAE of the seasonal naïve method. Otherwise, MASE will return a value greater than one. We describe the empirical application results in the following section.

4. Results and discussion

The results and discussion of each part of our proposed framework are presented following the two phases presented in Figure 3.

4.1 Phase 1: Aggregated level forecasts, $\hat{y}_h$

In this stage, we tested the proposed framework using different algorithms and performance measures. Table 4 and Table 5 reinforce the "No Free Lunch" theorem [40] by showing that no algorithm performs best in all scenarios. Because of this theorem, we tested the framework differently in those tables. Table 4 shows the accuracy of each fold using only a scaled measure (MASE) because each fold has a different scale. It means that every time we slice the time series, we change the scale as if they become a new time series. ETS and Theta Method Forecasting (THETAF) had on four out of five folds a MASE lower than one, meaning that these methods were better than the naïve. Seasonal Naive (SNAÏVE), THETAF, ARIMA, and LSTAR had smaller ranges in MASE (.12, .18, .26, .27, respectively), presenting a minor variance in their performance in the different folds. Three (LSTM, NNETAR, and NNETTS) of the five nonlinear methods produced different forecasts each time they were run. They did not provide the same accuracy performance with the same data if one runs the model twice, using default settings. This could be an issue when used by practitioners who are not experts in forecasting methods. We recommend averaging their forecasts over many runs to reduce the problem Another important remark is that users may decide to test a different number of algorithms depending on their computing power and the relevance of the problem at hand. A higher number of tested algorithms may bring more complexity but, at the same time, show more competitive alternatives. For instance, TBATS shows a reasonable performance but a bigger MASE range.

INSERT TABLE 4 HERE

Table 5 presents the results of the models when using the aggregated dataset (i.e., one time series only) in eight different error measures. Table 4 shows that Long Short-Term Memory (LSTM), ARIMA, and THETAF are among the best-performing. ARIMA and THETAF have a higher average rank position in the different folds. Additionally, ARIMA has a lower median MASE. According to Table 5, ARIMA is among the three best-performing methods in six of the eight metrics evaluated. LSTM is tied in the first position with ARIMA in two metrics (MASE, Theil's U). If we consider MAPE, LSTM is the best-performing. However, if we consider MAE, it is a close second place with a difference of only .09. Looking at its performance in the five folds (rank) seems logical to conclude that the advantage of using deep learning models increases with the size of the time series. On the other hand, ARIMA shows the same performance level no matter the time series size.

These results also show that scaled metrics (see MASE and Theil's U in Table 5) are more informative of actual performance. The other metrics show different and conflicting results. Scaled metrics also provide a straightforward and easy-to-understand comparison with the same benchmark method (naïve). After comparing these methods, we performed all analyses in phase two using ARIMA to estimate the base forecasts. The following subsection presents the comparison forecasting systems (i.e., the combination of structures, aggregation criteria, and coherent forecast methods).

INSERT TABLE 5 HERE

4.2 Phase 2: Forecasting systems

In phase two, since we organized the time series in different structures, the scales of each node (each time series forecasted) are different. Therefore, we only compared the performance using a scaled measure (MASE), as recommended by Hyndman and Koehler [34] and explained in section 3.2. Table 6 shows the most aggregated level results in each coherent forecast method and forecasting system structure. All estimations performed better than the naïve method. The product hierarchy and the grouped structure combining products

and geography had the best performances. In the latter structure, contrary to the findings of previous studies [e.g., 4] and in line with Yang and Shang [23], the bottom-up method had the best performance compared to the "optimal" method. For the other five structures, "optimal" performed better. These results provide evidence that the structure of the forecasting system impacts the accuracy of the forecasts and the performance of the coherent forecast methods.

INSERT TABLE 6 HERE

In the middle level of the structures (Tables 7 to 9), we evaluated the performance of each coherent forecast method on each node of the structure (i.e., on each product, channel, or zone). On average, a bottom-up method with a grouped structure that combined geographical information improved the forecasts of each product and channel type. However, three of the five geographical zones performed better in a hierarchical structure, with "optimal" WLS as the coherent forecast method.

For the different product types (Table 7), the bottom-up method in the grouped structure combining products and geographic regions performed the best, on average, and in most nodes. For products one and two, the grouped structure with product and geography combined with an "optimal" method or a structure with all criteria, combined with the bottom-up method, had a good performance. Only for product type 3 did different structures and coherent forecast methods perform better - the grouped structure combining products and channels with the "optimal" and the hierarchical with any of the methods.

INSERT TABLE 7 HERE

For the channel types (Table 8), the bottom-up method in the grouped structure combining all criteria or channel and geographical regions performed better, on average. For channel types 2 and 4, the naïve method has better accuracy than any of the methods compared. None of the structures had a good performance. For channel type 1 the "optimal" WLS performed better in the grouped structure combining products and channels. For this channel, other viable options that obtained similar performance (differences are in the third decimal place) are a hierarchical structure with the "optimal" WLS approach or grouped structure with all criteria and a bottom-up method. For channel type 3, the structure with all criteria combined with the bottom-up method had the best performance.

INSERT TABLE 8 HERE

Finally, in the middle level of the structures that used geography as criteria (the only structures with three levels), the "optimal" WLS in the hierarchical structure performed the best in three of the five zones (Table 9). The bottom-up method in the same structure presented the best average accuracy. In three zones (East, North, and South), the naïve method was better than any method compared. For the center, the bottom-up method performed better in the grouped structure combining all criteria or only product and geography. The center zone is the one that shows better performance. If this is also the zone with the higher number of points of sale and commercial activity in the city, the importance of an accurate forecast in this zone might be higher. For that reason, a manager could favor one of the two grouped structures and a bottom-up method to get the best performance in that particular zone.

INSERT TABLE 9 HERE

As previously explained, the channel and product criteria did not have a third disaggregated level. We compared the performance of the different combinations of structures, with geography as a criterion, in their performance in each district. Table 10 shows that all structures performed similarly when looking at each node (district). The differences between structures were small, and improvements were around .02 (the difference between the best-performing and the second best-performing). Most districts (32) had no difference between the first and second best because more than one combination of structure and coherent forecast method had the same performance. However, the average difference between the best and the worst-performing was higher, .42.

INSERT TABLE 10 HERE

With further disaggregation, most nodes did not perform better than the naïve method. In fact, only 17 of the 53 districts had a better performance than the naïve with one of the combinations (MASE smaller than one). On average, in the bottom level (Table 10) the bottom-up method presented the best average and higher number of best-performing nodes in all structures compared. The hierarchical structures can be further compared using Top-down and Middle-out methods, which are not available for grouped structures. Even adding more coherent forecast methods in those structures, the bottom-up method was the best performing when considering each node. The results can be found in Supporting information S3.

The geographical aggregation added helpful information in a grouped structure, improving the forecasts at the product and channels' levels. Information about different product categories or channel characteristics alone was insufficient to improve the accuracy of the forecasting systems. Nevertheless, when taken with other information regarding the grouping structure, it led to a better sales forecast.

Another result is that no combination of the structure of the forecasting system and coherent forecast method was consistently better in all levels of aggregation. Some compromise had to be made. Figure 6 helps to understand the different compromises necessary at each level when choosing a structure. For Figure 6, we started by choosing the best combination of structures and methods that, on average, performed better for the middle level (product types, channel types, and zones). Then, we evaluated which was also best in the aggregate (or top) level and the most disaggregated (bottom) level, and the impacts of each choice on the accuracy, compared to the best.

INSERT FIGURE 6 HERE

In the product disaggregated level, it is best to combine product and geographical criteria with a bottom-up method. If we move up in the aggregations' levels, this structure also performs best in the most aggregated level. However, on average, it increases the error in the different geographical zones by .06 (middle level), and by .23 in the different districts. At the channel disaggregated level, it is best to combine channels and geographical criteria with a bottom-up method. If we move up in the aggregations' levels, this structure increases the error in the most aggregated level by .16. It also increases on average the error in the different geographical zones by .08 (middle level), and by .04 in the different districts. At the geographically disaggregated level, there are two viable options. The first option is the hierarchical structure combined with a top-down — top-down Gross-Sohl method F (tdgsf) — method (see Supporting information S3 for further results), which presents the best average MASE (3.30). This structure also has the best performance in the middle level (1.41) and only increases the error in the aggregated level by .01. However, the proportion of nodes

in which this combination performs best (23.7%) is lower than the proportion of the structure that combines all criteria with a bottom-up method (67%).

Considering each node's performance individually, it is best to combine all criteria with a bottom-up method. It only increases the average MASE by .01 in the most disaggregated level compared to the best-performing combination. If we move up in the aggregations' levels, this structure increases the error in the most aggregated level by .05. It also increases on average the error in the different geographical zones by .05 (middle level), and by .04 in the different product types. The performance for the different channel types is similar to the best-performing.

Comparing the various structures, the product criteria improve the total level forecast, alone or combined with geography criteria, using a bottom-up method. At the middle level, combining the information of the product and geography either using bottom-up (.72) or "optimal" (.74) is beneficial when the different products are the interest of the forecast. The same phenomenon occurs when considering the channel types since adding geographical information improves the bottom-up method's accuracy. Last, the hierarchical structure with a top-down method performed better for the middle level when the focus moved to the geographical regions. However, Table 10 and Figure 6 show that the grouped structure with all the criteria has the best balance between average MASE values and the number of best-performing nodes. To use it, managers must collect, store, and analyze additional information, so some compromises must be made.

4.3 Discussion

Following the implementation of our framework's phase 1, considering all of the folds and accuracy metrics, ARIMA was the best-performing method. Nonlinear methods and machine learning techniques produced different forecasts each time they are run (see, in italics, the MASE differences from Tables 4 and 5) or outperformed by a standard well-known statistical time series forecasting method—ARIMA. This further supports the recent findings that for the task of forecasting time series without additional features (or covariates) statistical methods display better performance than machine learning techniques [18]. The results also pointed to the advantage of using scaled metrics (MASE and Theil's U). Both were more informative on performance, leading to less conflicting results than other metrics. The results also show the importance of forecasting expertise and comparison of different metrics to allow a manager to choose the best metrics to evaluate accuracy. Another interesting result is the improved performance of the deep learning model with the size of the time series.

Following the implementation of our framework's phase 2, it is clear that the structure and criteria of the forecasting system impact accuracy. Combining product types, channel types, and geographic regions led to more accurate forecasts on each node than choosing only one criterion. In line with the results of Fliedner and Mabert [10] and Fliedner and Lawrence [6], the grouping of items (different aggregation criteria) was responsible for improved forecasting performance. The more sophisticated methods of clustering these time series might not impact accuracy [6]. However, we showed that market-related criteria combined in different structures and applying different coherent forecast methods have an impact.

The criteria and the type of structure used (grouped or hierarchical) make a difference. Our empirical application provides evidence that it is best to aggregate the time series in a grouped rather than a hierarchical structure. It is also interesting to note that the most aggregated data (top-level) have a better performance than the most disaggregated data (bottom-level), in line with previous studies [e.g., 4]. Our results also show that despite previous studies reporting poor performance of the bottom-up method [4, 19] and in line with the recent results reported by Yang and Shang [23], managers should not disregard this

method. It is important to notice that bottom-up works well when the most disaggregated series behave differently from each other and are not too noisy. However, bottom-up ceases to work well with further disaggregation and more noise. In our results, the added sophistication of the "optimal" method did not improve the performance. However, for more complex structures with more nodes and levels, this method is known to work better [4, 19]. The coherent forecast methods should be tested and compared in the same way that different time series forecasting methods are in literature and in practice. Following our proposed framework, divided into two phases, should help practitioners to build a more accurate forecasting system or at least be aware of the compromises made. The results also suggest that aggregating by geographical regions is important in improving sales forecasts' accuracy.

5. Concluding remarks

This manuscript offers several contributions. It contributes to the management literature by focusing on B2B sales forecasting. This context has received less attention from existing academic research [13, 14, 15, 16]. It also provides evidence of the importance of geographical information to improve accuracy. The proposed framework allows firms to implement the forecasting system that best communicates consistent information at all hierarchical levels and departments, leading the efforts in the same direction. The system scales easily and provides consistency for retailers and manufacturers. Many retailers have thousands of stores spread geographically, while manufacturers may have thousands of business partners in different locations selling their products and requiring tailored forecasts. Our strategy allows for a coherent forecast with almost no human effort, and it can be adapted based on goals, horizons, update necessity, and levels of aggregation. Our framework also provides steps to build, compare and decide which market-related variables should be used, alone or combined, to improve forecast systems' accuracy.

Another contribution of our study is that it provides evidence of the accuracy of all methods applied and suggested. Our paper compares the accuracy performance of different forecasting methods (statistical and machine learning); reports accuracy evidence based on a longer forecasting horizon (12 months); and compares different coherent forecast methods, different structures (hierarchies and groups), and criteria based on market-related variables. Our implementation of the framework uses open-source tools and proprietary data that are natural to the process of every organization. The goal was to propose a framework for building and evaluating a forecasting system that favors methods and tools managers can easily apply with minimal investment. Since one of the issues for practice is the limited computational resources, our code is computationally efficient. Our implementation separates the generation of forecasts from their adjustment to become coherent, so it is easily parallelizable. The forecasting for all nodes can be computed in parallel, and the coherent forecast step can then be calculated using sparse matrix algebra. The computational time for generating the forecasts is much more substantial than the time required to adjust them.

Our study has some limitations that future research can still explore. First, replications of this analysis could be done. Different datasets might provide information about other market-related variables, such as promotion or price strategies, that could provide other aggregation criteria. Our goal was not to be exhaustive in the comparisons but to propose a framework that allows managers to test different combinations and to provide evidence that the criteria and structure upon which the forecasting system is built impacts accuracy. Predicting churn or customer lifetime value (CLV) using this framework could also provide further evidence of the impact of the different aggregation criteria and coherent forecast methods on other management problems. Additionally, given the importance of geographic disaggregation, future research could explore the concentration of stores in a specific area. Sales of low-involvement and low-cost durables tend to concentrate in areas of higher retail

activity. Agglomeration theory [35] states that it is more convenient for consumers when stores from the same category are concentrated in a specific geographic area. Thus, consumers can compare alternatives and obtain more product information. The agglomeration of companies in a particular geographic area will, for that reason, positively impact sales, explaining demand more than the agglomeration of consumers [35].

Accurate forecasts can help plan budgets, predict production levels, influence brand image, price perception, customer satisfaction, and other areas of interest. We hope that our study will help address this issue and stimulate future research to focus on the accuracy of predictions and how to build more accurate forecasting systems.

Supporting information

Additional information for this article is available online.

S1 R Packages List
S2 R Code
S3 Further Results on the Hierarchical Structures
Table S1 – MASE of the hierarchical structures at the top and middle levels
Table S2 – MASE of the hierarchical structures at the bottom level

Competing interests: none

References

1. Syam N, Sharma A. Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice. *Ind Mark Manag*. 2018;69:135-146.
2. Dunn DM, Williams WH, DeChaine T. Aggregate versus subaggregate models in local area forecasting. *J Am Stat Assoc*. 1976;71(353):68-71.
3. Shlifer E, Wolff RW. Aggregation and proration in forecasting. *Manag Sci*. 1979;25(6):594-603.
4. Oliveira JM, Ramos P. Assessing the performance of hierarchical forecasting methods on the retail sector. *Entropy*. 2019;21(4):436.
5. Abhishek V, Hosanagar K, Fader PS. Aggregation bias in sponsored search data: The curse and the cure. *Mark Sci*. 2015;34(1):59-77.
6. Fliedner EB, Lawrence B. Forecasting system parent group formation: An empirical application of cluster analysis. *J Oper Manag*. 1995;12(2):119-130.
7. Zotteri G, Kalchschmidt M, Caniato F. The impact of aggregation level on forecasting performance. *Int J Prod Econ*. 2005;93:479-491.
8. Kourentzes N, Rostami-Tabar B, Barrow DK. Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *J Bus Res*. 2017;78:1-9.
9. Dekker M, van Donselaar K, Ouwehand P. How to use aggregation and combined forecasting to improve seasonal demand forecasts. *Int J Prod Econ*. 2004;90(2):151-167.
10. Fliedner EB, Mabert VA. Constrained forecasting: some implementation guidelines. *Decis Sci*. 1992;23(5):1143-1161.
11. Fildes R, Ma S, Kolassa S. Retail forecasting: Research and practice. *Int J Forecast*. Published online 2019.
12. Hoeltgebaum H, Borenstein D, Fernandes C, Veiga Á. A score-driven model of short-term demand forecasting for retail distribution centers. *J Retail*. 2021;97(4):715-725.

13. Bohanec M, Robnik-Šikonja M, Borštnar MK. Organizational learning supported by machine learning models coupled with general explanation methods: A Case of B2B sales forecasting. *Organizacija*. 2017;50(3):217-233.
14. Lackman CL. Forecasting sales for a B2B product category: case of auto component product. *J Bus Ind Mark*. Published online 2007.
15. Rezazadeh A. A generalized flow for B2B sales predictive modeling: An azure machine-learning approach. Forecasting. 2020;2(3):267-283.
16. Rohaan D, Topan E, Groothuis-Oudshoorn CG. Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider. Expert Syst Appl. 2022;188:115925.
17. Lei M, Li S, Yu S. Demand Forecasting Approaches Based on Associated Relationships for Multiple Products. Entropy. 2019;21(10):974.
18. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and Machine Learning forecasting methods: Concerns and ways forward. PloS One. 2018;13(3):e0194889.
19. Hyndman RJ, Ahmed RA, Athanasopoulos G, Shang HL. Optimal combination forecasts for hierarchical time series. Comput Stat Data Anal. 2011;55(9):2579-2589.
20. Hyndman RJ, Athanasopoulos G. Forecasting: Principles and Practice. OTexts; 2018. https://OTexts.org/fpp2/
21. Mircetic D, Rostami-Tabar B, Nikolicic S, Maslaric M. Forecasting hierarchical time series in supply chains: an empirical investigation. Int J Prod Res. 2022;60(8):2514-2533.
22. Divakar S, Ratchford BT, Shankar V. CHAN4CAST: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. Mark Sci. 2005;24(3):334-350. doi:10.1287/mksc.1050.0135
23. Yang Y, Shang HL. Is the group structure important in grouped functional time series? ArXiv Prepr ArXiv211104390. Published online 2021.
24. Wickramasuriya SL, Athanasopoulos G, Hyndman RJ. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J Am Stat Assoc. 2019;114(526):804-819.
25. Chollet F. Allaire, J. keras: R Interface to 'Keras'. R package version 2.2. 4.1. Published online 2019.
26. Hyndman RJ, Bergmeir C, Caceres G, et al. Forecast: Forecasting Functions for Time Series and Linear Models.; 2021. http://pkg.robjhyndman.com/forecast
27. Wang E, Cook D, Hyndman RJ. A new tidy data structure to support exploration and modeling of temporal data. J Comput Graph Stat. 2020;29(3):466-478.
28. Kourentzes N. Nnfor: Time Series Forecasting with Neural Networks.; 2019. https://CRAN.R-project.org/package=nnfor
29. Martinez F. Tsfknn: Time Series Forecasting Using Nearest Neighbors.; 2018. https://CRAN.R-project.org/package=tsfknn
30. Hmamouche Y. NlinTS: Non Linear Time Series Analysis.; 2019. https://CRAN.R-project.org/package=NlinTS
31. Narzo AFD, Aznarte JL, Stigler M. TsDyn: Time Series Analysis Based on Dynamical Systems Theory.; 2009. https://cran.r-project.org/package=tsDyn/vignettes/tsDyn.pdf
32. Hyndman RJ, Lee A, Wang E, Wickramasuriya, S. Hts: Hierarchical and Grouped Time Series.; 2021. https://CRAN.R-project.org/package=hts
33. Hyndman RJ, Khandakar Y. Automatic time series forecasting: the forecast package for R. J Stat Softw. 2008;26(3):1-22.
34. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. Int J Forecast. 2006;22(4):679-688.

35. Liu AX, Steenkamp JBEM, Zhang J. Agglomeration as a Driver of the Volume of Electronic Word of Mouth in the Restaurant Industry. J Mark Res. 2018;55(4):507-523. doi:10.1509/jmr.16.0182
36. Van den Broeke M, De Baets S, Vereecke A, Baecke P, Vanderheyden K. Judgmental forecast adjustments over different time horizons. Omega. 2019;87:34-45.
37. Fildes R, Goodwin P, Lawrence M. The design features of forecasting support systems and their effectiveness. Decis Support Syst. 2006;42(1):351-361.
38. Fildes R, Goodwin P. Against your better judgment? How organizations can improve their use of management judgment in forecasting. Interfaces. 2007;37(6):570-576.
39. Arvan M, Fahimnia B, Reisi M, Siemsen E. Integrating human judgement into quantitative forecasting methods: A review. Omega. 2019;86:237-252.
40. Wolpert D. H., Macready W. G. No free lunch theorems for optimization, in IEEE Transactions on Evolutionary Computation, vol. 1, no. 1, pp. 67-82, April 1997, doi: 10.1109/4235.585893.

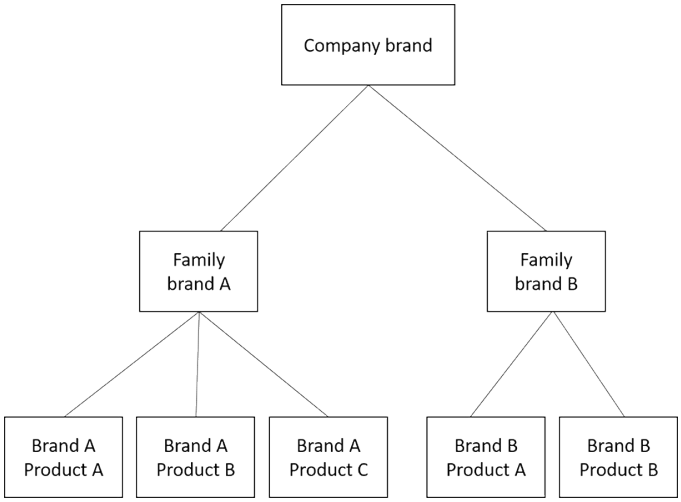

Figure 1 - Hierarchy example

Data	25	25	25
Independent Forecasts	**26**	**21**	**20**
Bottom-up	**20**	**20**	**20**
Top-down	**26**	**26**	**26**

Figure 2 - Coherence problem

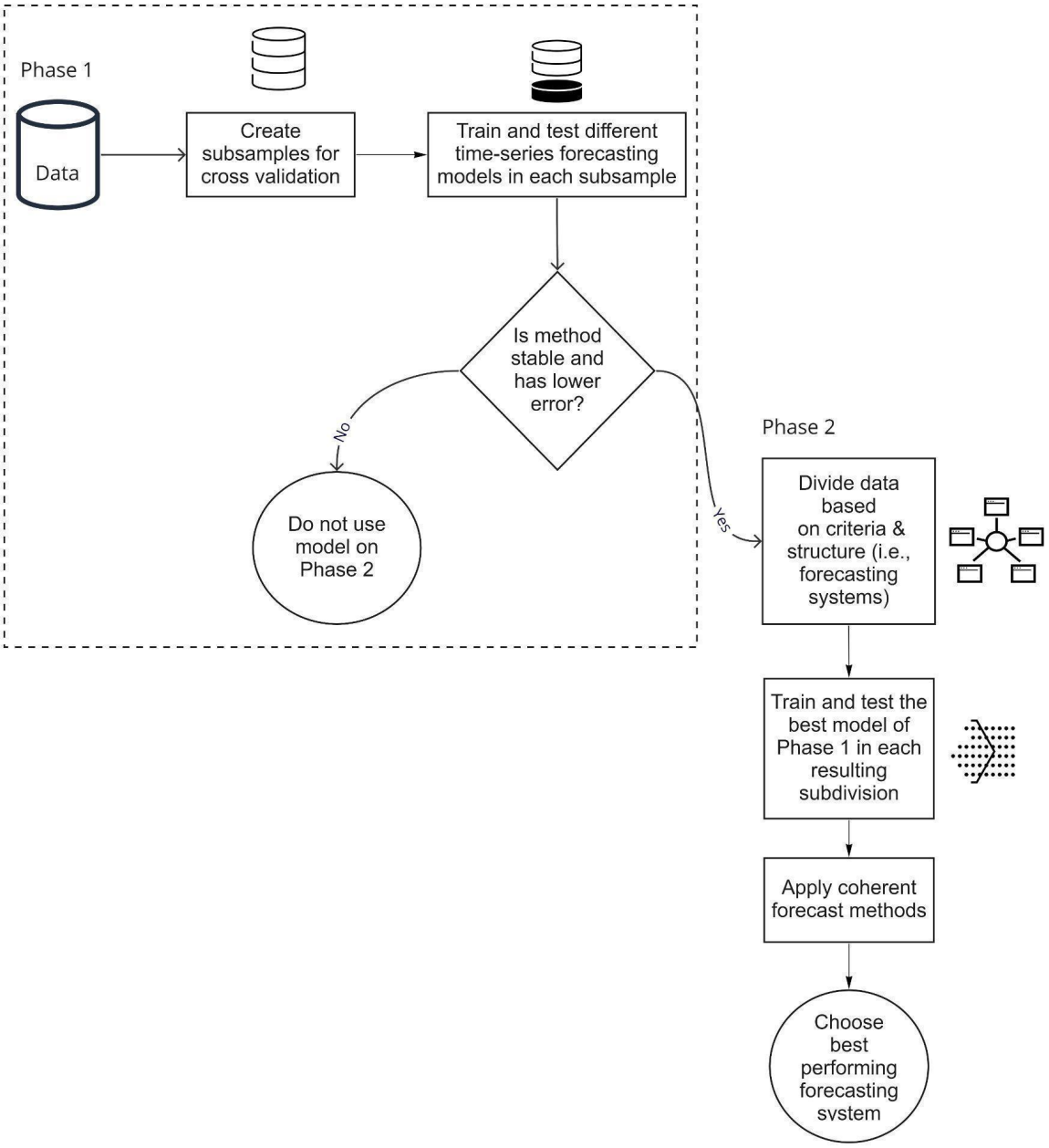

Figure 3 - Proposed framework

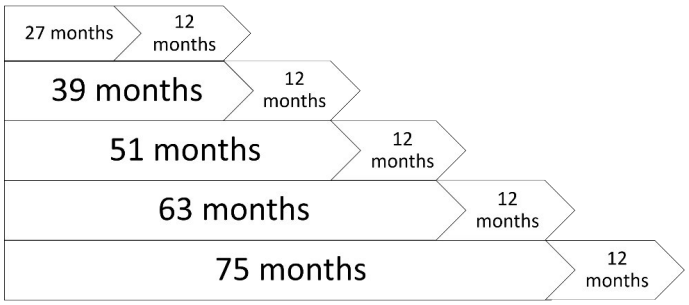

Figure 4 - Evaluation on a rolling forecasting origin

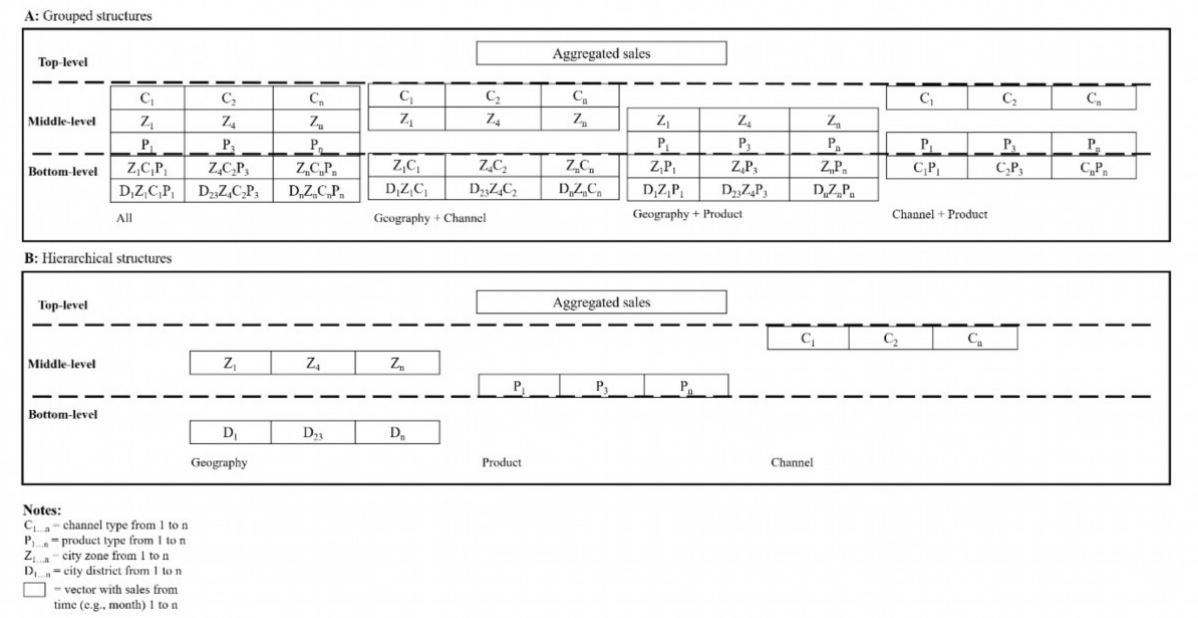

Figure 5 - Illustration of data structures

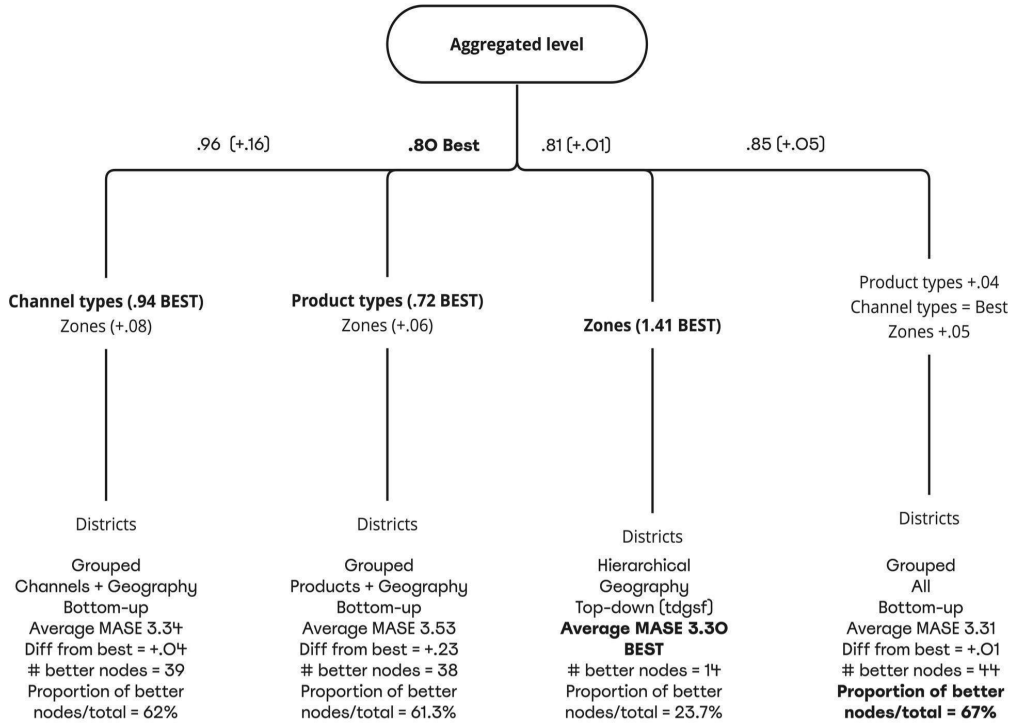

Figure 6 - Compromises by level on average MASE

Paper	Method	By product	By geographical region	By channel type	Comparison of hierarchical and grouped structures
This paper	**Market-related**	✔	✔	✔	✔
Oliveira & Ramos [4]	Market-related	✔	-	-	-
Kourentzes *et al*. [8]	Temporal	-	-	-	✔
Zotteri *et al*. [7]	Clustering	-	-	✔	-
Dekker *et al*. [9]	Clustering	-	-	✔	-
Fliedner & Lawrence [6]	Clustering	✔	-	-	-
Fliedner & Mabert [10]	Clustering	✔	-	-	-

Table 1 – Key literature on structures and aggregation criteria

Base Model	Description
AAR	Additive nonlinear autoregressive model implemented by package tsDyn
ARIMA	ARIMA model implemented by forecast
CROSTON	Croston's method, for intermittent demand, implemented by forecast
CUBIC	Cubic spline, implemented by forecast
ETS	Exponential smoothing state space model implemented by forecast
KNN	K-Nearest Neighbor regression where the number of nearest neighbors and the lags are selected automatically implemented by tsfknn
LINEAR	Linear Autoregressive model implemented by tsDyn
LSTAR	Logistic Smooth Transition Autoregressive model implemented by tsDyn
LSTM	Long short-term memory implemented by Keras
NAÏVE	Equivalent to an ARIMA (0,1,0) implemented by forecast
NNETAR	Neural network time series forecast. Feed-forward neural networks with a single hidden layer and lagged inputs implemented by forecast
NNETTS	Neural Network nonlinear autoregressive model implemented by tsDyn
RWF	Random walk with drift implemented by forecast
SNAÏVE	ARIMA (0,0,0)(0,1,0)m model where m is the seasonal period implemented by forecast
STLM	Applies an Seasonal-Trend decomposition using Loess with memory (STL decomposition), implemented by forecast
TBATS	Exponential smoothing state space model with Box-Cox transformation, ARMA errors, Trend and Seasonal components implemented by forecast
THETAF	Theta Method Forecasting, equivalent to simple exponential smoothing with drift implemented by forecast
VARMLP	Artificial Neural Network VAR (Vector Auto-Regressive) model using a MultiLayer Perceptron implemented by NlinTS

Table 2 - Description of each base model applied

Levels	Series
Geography (3 levels)	
Total	1
Zones	5
Districts	53
Products (2 levels)	
Total	1
Product types	3
Channels (2 levels)	
Total	1
Channel types	4

Table 3 - Aggregation criteria

Forecasting Method	Nonlinear	Diff. results on each run	MASE / Fold 1	2	3	4	5	Median	Mean	Rank/ Fold 1	2	3	4	5	Median	Mean
LSTM	Yes	Yes	.85	**.97**	1.37	1.14	***.78***	.97	1.02	11	3	18	13	**1**	11	9
ARIMA			.79	1.05	1.03	**.88**	**.81**	**.88**	.91	9	11	5	3	2	**5**	**6**
STLM			.75	1.03	1.15	.91	**.86**	**.91**	.94	7	5	15	5	3	5	7
TBATS			.55	1.05	1.03	**.87**	.89	**.89**	**.88**	5	9	4	2	4	**4**	**5**
THETAF			.83	**.91**	**1.01**	.92	.89	.91	.91	10	2	3	6	5	**5**	**5**
NAÏVE			.79	1.04	1.05	1.21	.90	1.04	1.00	8	7	8	15	6	8	9
CUBIC			1.17	1.23	**1.00**	.92	.91	1.00	1.05	14	18	2	7	7	7	10
ETS			.55	1.05	**.99**	.93	.92	.93	**.89**	4	10	**1**	9	8	8	6
NNETAR	Yes	Yes	3.44	1.19	1.13	**.86**	*.93*	1.13	1.51	18	16	14	**1**	9	14	12
CROSTON			**.47**	1.07	1.05	.89	.94	.94	**.88**	2	13	7	4	10	7	7
VARMLP	Yes		1.51	.97	1.07	.92	.94	.97	1.08	16	4	10	8	11	10	10
RWF			1.15	**.91**	1.09	1.37	.96	1.09	1.09	13	**1**	13	18	12	13	11
KNN	Yes		1.65	1.08	1.33	.97	1.01	1.08	1.21	17	14	17	10	13	14	14
AAR			**.47**	1.03	1.08	1.01	1.06	1.03	.93	**1**	6	12	12	14	12	9
SNAÏVE			.96	1.04	1.08	1.00	1.07	1.04	1.03	12	8	11	11	15	11	11
NNETTS	Yes	Yes	**.52**	1.14	1.03	1.15	*1.07*	1.07	.98	3	15	6	14	16	14	11
LSTAR			1.29	1.05	1.07	1.32	1.23	1.23	1.19	15	12	9	17	17	15	14
LINEAR			.56	1.19	1.20	1.26	1.26	1.20	1.09	6	17	16	16	18	16	15

Table 4 – Forecasting performance of the 18 base models in all five folds of rolling forecasting origin[1]

[1] Numbers in bold are the three best-performing methods per fold (each training size) or by considering the median and mean (MASE and ranking position).

Forecasting Method	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U	Times in the top 3
ARIMA	**3.60**	**25.48**	**19.60**	-9.92	**31.95**	**.81**	.07	**.71**	6
LSTM	11.83	**26.83**	**19.69**	5.12	**27.19**	***.81***	.19	**.71**	5
STLM	-4.54	**25.56**	**20.85**	-22.43	38.29	**.86**	.13	.87	3
CROSTON	9.09	28.62	22.77	**-2.47**	**34.38**	.94	-.03	.75	2
TBATS	4.44	27.50	21.59	-10.06	35.03	.89	-.03	**.74**	1
VARMLP	9.17	28.72	22.91	**-2.41**	34.63	.94	-.03	.76	1
NAÏVE	**-3.04**	27.31	21.92	-22.26	39.34	.90	-.03	.78	1
THETAF	**-1.11**	27.35	21.71	-19.39	38.41	.89	-.03	.78	1
ETS	-5.83	28.24	22.30	-27.37	42.06	.92	**-.02**	.87	1
NNETAR	11.17	29.11	23.32	**4.27**	34.79	*.96*	.42	.88	1
RWF	-9.61	29.28	23.27	-33.56	45.34	.96	**-.01**	.92	1
LINEAR	27.92	39.60	30.77	28.61	36.67	1.26	**.02**	1.08	1
SNAÏVE	8.88	29.39	25.98	5.43	39.00	1.07	.21	.77	0
CUBIC	-4.54	27.97	22.09	-25.24	41.07	.91	-.02	.85	0
KNN	-20.09	29.85	24.58	-43.03	46.93	1.01	.23	.86	0
AAR	17.04	32.59	25.81	11.36	34.77	1.06	.18	.87	0
NNETS	28.75	39.53	31.78	29.62	38.91	*1.31*	-.03	1.05	0
LSTAR	17.85	35.87	29.96	10.62	42.17	1.23	.12	1.09	0

Table 5 – Forecasting performance of the 18 base models on the full database

Top-level	Grouped				Hierarchical		
	Product and Geography	All criteria	Product and channels	Geography and channels	Geography	Channels	Products
Bottom-Up	**.80**	.85	.93	.96	.92	.92	**.80**
Optimal MinT	n.a.	n.a.	.86	n.a	n.a.	.88	**.80**
Optimal WLS	.82	.83	.85	.86	.85	.87	**.80**

Table 6 – MASE of the top level with different coherent forecast methods[2]

[2] In the data used, MinT does not work with the geographical information

Middle level		Grouped		Hierarchical
	Product and Geography	All criteria	Product and channels	Products
Product 1				
Bottom-Up	**.90**	.95	1.25	1.06
Optimal MinT	n.a.	n.a.	1.15	1.06
Optimal WLS	.93	.97	1.13	1.06
Product 2				
Bottom-Up	**.43**	**.43**	.49	.49
Optimal MinT	n.a.	n.a.	.47	.49
Optimal WLS	.45	.44	.49	.49
Product 3				
Bottom-Up	.83	.91	.84	**.79**
Optimal MinT	n.a.	n.a.	**.79**	**.79**
Optimal WLS	.84	.86	**.79**	**.79**
Average				
Bottom-Up	**.72**	.76	.86	.78
Optimal MinT			.80	.78
Optimal WLS	.74	.76	.80	.78

Table 7 – MASE of the product's middle level with different coherent forecast methods[3]

[3] In the data used, MinT does not work with the geographical information

Middle level	All criteria	Grouped		Hierarchical
		Product and channels	**Geography and channels**	Channels
Channel type 1				
Bottom-Up	**.25**	.26	.26	.26
Optimal MinT	n.a.	.25	n.a.	.26
Optimal WLS	.26	**.25**	.28	**.25**
Channel type 2				
Bottom-Up	1.35	1.75	**1.29**	1.68
Optimal MinT	n.a.	1.64	n.a.	1.61
Optimal WLS	1.50	1.70	1.40	1.66
Channel type 3				
Bottom-Up	**.68**	.97	.73	.96
Optimal MinT	n.a.	.84	n.a.	.88
Optimal WLS	.69	.81	.71	.86
Channel type 4				
Bottom-Up	1.48	1.48	**1.47**	1.49
Optimal MinT	n.a.	1.49	n.a.	1.49
Optimal WLS	**1.47**	1.49	1.48	1.93
Average				
Bottom-Up	**.94**	1.11	**.94**	1.10
Optimal MinT	n.a.	1.05	n.a.	1.06
Optimal WLS	.98	1.06	.97	1.18

Table 8 – MASE of the channel's middle level with different coherent forecast methods

Middle level	Grouped			**Hierarchical**
	Product and Geography	All criteria	Geography and channels	**Geography**
Center				
Bottom-Up	**.46**	**.46**	.49	.49
Optimal WLS	.48	.48	.48	.47
East				
Bottom-Up	**2.15**	2.27	2.36	2.17
Optimal WLS	**2.15**	2.19	2.20	**2.15**
North				
Bottom-Up	1.01	**1.00**	1.01	1.09
Optimal WLS	**1.00**	1.01	1.04	**1.00**
West				
Bottom-Up	.89	.90	.93	.84
Optimal WLS	.84	.85	.84	**.79**
South				
Bottom-Up	2.82	2.67	2.58	**2.55**
Optimal WLS	2.84	2.78	2.80	2.89
Average				
Bottom-Up	1.47	1.46	1.49	**1.43**
Optimal WLS	1.46	1.46	1.47	1.46

Table 9 – MASE of the geographical middle level with different coherent forecast methods

Bottom level	Grouped						Hierarchical	
	Products and geography		**All criteria**		Channels and geography		Geography	
Districts	Bottom-up	Optimal WLS	**Bottom-up**	Optimal WLS	Bottom-up	Optimal WLS	Bottom-up	Optimal WLS
1	.53	.55	**.53**	.54	.56	.55	.56	.55
2	.50	.49	**.50**	.51	.50	.52	.50	.51
3	1.57	1.61	**1.57**	1.61	2.42	2.41	2.42	2.41
4	.45	.47	**.44**	.47	.44	.48	.45	.49
5	1.22	1.21	**1.21**	1.23	1.22	1.26	1.23	1.24
6	5.21	5.14	**5.21**	5.14	5.03	5.03	5.03	5.03
7	1.66	1.67	**1.66**	1.67	1.69	1.70	1.69	1.69
8	2.26	2.29	**2.26**	2.28	2.33	2.33	2.33	2.33
9	.71	.72	**.70**	.72	.71	.75	.72	.74
10	.26	.25	**.26**	.26	.19	.23	.19	.23
11	1.78	1.86	**1.78**	1.86	2.15	2.18	2.15	2.16
12	37.40	37.31	**37.40**	37.30	37.59	37.38	37.25	37.22
13	1.56	1.57	**1.56**	1.57	1.58	1.59	1.58	1.59
14	.30	.35	**.30**	.36	.45	.47	.45	.47
15	1.73	1.73	**1.73**	1.72	1.73	1.71	1.73	1.72
16	2.33	2.37	**2.33**	2.39	2.52	2.57	2.52	2.55
17	8.44	8.77	**.10**	4.47	.04	4.64	9.10	9.10
18	.36	.40	**.31**	.39	.31	.47	.49	.52
19	1.69	1.68	**1.69**	1.67	1.67	1.67	1.67	1.67
20	1.58	1.58	**1.58**	1.59	1.58	1.60	1.58	1.59
21	1.60	1.60	**1.60**	1.60	1.59	1.61	1.59	1.60
22	3.70	3.66	**3.70**	3.66	3.56	3.56	3.56	3.56
23	1.06	1.07	**1.09**	1.08	1.10	1.09	1.06	1.08

24	.43	.43	**.44**	.45	.45	.47	.43	.45
25	1.89	1.90	**1.89**	1.90	1.89	1.91	1.89	1.92
26	.75	.76	**.75**	.76	.75	.78	.75	.78
27	2.38	2.38	**2.38**	2.38	2.38	2.37	2.38	2.37
28	9.71	9.72	**9.71**	9.72	9.74	9.74	9.74	9.74
29	1.14	1.13	**1.14**	1.15	1.20	1.17	1.20	1.12
30	.46	.47	**.46**	.47	.46	.47	.46	.47
31	2.44	2.68	**1.86**	2.37	1.92	2.68	2.94	3.09
32	1.03	1.04	**1.03**	1.05	1.01	1.04	1.01	1.04
33	2.05	2.06	**2.04**	2.06	2.07	2.08	2.07	2.08
34	.72	.68	**.73**	.69	.73	.69	.73	.69
35	12.97	12.91	**12.89**	12.87	12.91	12.89	12.86	12.86
36	4.74	4.08	**3.11**	2.99	3.91	2.84	2.63	2.61
37	4.63	4.65	**4.68**	4.68	3.69	4.21	4.69	4.69
38	4.25	4.20	**4.25**	4.19	4.24	4.24	4.24	4.24
39	.35	.35	**.32**	.34	.37	.37	.38	.37
40	.15	.17	**.15**	.18	.15	.23	.15	.17
41	3.73	3.68	**2.55**	2.72	2.57	2.90	3.52	3.57
42	5.19	5.12	**4.99**	5.09	4.97	5.15	5.69	5.67
43	.67	.65	**.36**	.45	.38	.50	.47	.58
44	5.10	5.35	**5.10**	5.36	5.72	5.73	5.72	5.73
45	4.99	5.20	**4.99**	5.24	4.85	5.33	4.85	5.56
46	.53	.63	**.53**	.65	.53	.70	.53	.78
47	1.44	1.43	**1.49**	1.47	1.45	1.43	1.43	1.42
48	.31	.29	**.23**	.25	.24	.26	.17	.26
49	1.01	1.15	**1.01**	1.16	1.01	1.26	1.01	1.36

50	.44	.46	**.42**	.46	.37	.43	.49	.50
51	1.81	1.80	**2.31**	2.22	2.31	2.19	2.31	2.10
52	12.70	12.66	**12.72**	12.67	12.60	12.58	12.59	12.57
53	21.32	21.32	**21.32**	21.32	21.32	21.32	21.32	21.33
Average	3.53	3.54	**3.31**	3.42	3.34	3.47	3.56	3.59
Median	1.60	1.61	**1.58**	1.61	1.59	1.67	1.67	1.67
# of best-performing	33	20	39	14	37	16	35	18
% of best-performing	62.3%	37.7%	73.6%	26.4%	69.8%	30.2%	66.0%	34.0%

Table 10 – MASE of the geographical bottom level with different coherent forecast methods