Omega 136 (2025) 103335

Contents lists available at ScienceDirect

Omega

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

Inventory management with leading indicator augmented hierarchical forecasts☆

Yves R. Sagaert [a,*], Nikolaos Kourentzes [b]

[a] VIVES University of Applied Sciences, Doorniksesteenweg 145, 8500 Kortrijk, Belgium
[b] Skövde Artificial Intelligence Lab, School of Informatics, University of Skövde, Sweden

ARTICLE INFO

Keywords:
Forecasting
Inventory management
Leading indicators
Hierarchical reconciliation
Variable selection

ABSTRACT

Inventory management relies on accurate demand forecasts. Typically, these are univariate forecasts extrapolating patterns from past demand. The disaggregate nature of demand at the Stock Keeping Unit (SKU) level makes the incorporation of external information challenging. Nonetheless, such leading information can be critical to identifying disruptions and changes in the demand dynamics. To address the inventory planning needs of a global manufacturer we propose a methodology that identifies predictively useful leading indicators at an aggregate demand level, and translates that information to SKU-demand by leveraging on the hierarchical structure of the problem. Therefore, the proposed methodology provides probabilistic forecasts enriched by leading indicator information at SKU-level, as inputs for inventory management. The methodology automatically adjusts the choice of indicators for different required lead times, with some being more informative about the short-term demand dynamics and others for the long-term. We demonstrate the benefits both in the case of backorders and lost-sales, for a variety of lead times. We further benchmark the solution against solely using leading indicators or hierarchical forecasts, demonstrating that the benefits appear primarily by the proposed blending of the modelling approaches. The outcome is demonstratively better forecasts and inventory management for the case company. Additionally, management gains insights into the main drivers of their short and long-term demand, and the ability to adjust inventory replenishment accordingly. The ability to account for diverse macro and market information in operations is paramount for firms with a global reach that face different market conditions across countries. Additionally, the transparency of which leading indicators are influencing forecasts of different lead times is conducive to increased forecast trustworthiness.

1. Introduction

Inventory management is facilitated by accurate forecasts. Typically, these forecasts are based on univariate models, extrapolating from the historical demand. Univariate models, such as exponential smoothing, have demonstrated reliable performance, relatively limited data requirements, and ease of use [1], explaining their widespread adoption in the practice and software [2,3, chapter 13]. A main advantage of these models is their scalability and low computational cost, enabling their use in organisations that require a large number of forecasts to support inventory management decisions, for instance, in retailing [4,5]. More recently machine learning methods have demonstrated their effectiveness in similar conditions [6], with lightGBM, an advanced boosting method building on decision trees, being one of the top performers [7]. Although these models are adequate in normal conditions, they may fail in situations where there are disruptions, or rapid changes in the market conditions, due to their extrapolative nature. In the literature there are examples of this in the context of the supply chain from the 2008 financial crisis [8] and tourism forecasting from the more recent Covid-19 pandemic [9]. In practice, this is often addressed by incorporating contextual information through the use of judgemental adjustments of model forecasts [10,11]. However, the performance of these adjustments is found to be inconsistent, due to various judgemental biases [11,12], and do not scale. Furthermore, there is ample evidence that experts can find it difficult to identify predictively valuable cues in the available contextual information, resulting in ineffective or even damaging adjustments [13,14].

An alternative is to build models that incorporate leading indicators. A leading indicator is an external variable that contains predictive information for a target variable with a sufficient lead. The lead has to be at least as long as the forecast horizon to be useful [15]. Leading

☆ Area: Supply Chain Management The manuscript is processed by Associate Editor Lars Magnus Hvattum.

* Corresponding author.
E-mail address: yves.sagaert@vives.be (Y.R. Sagaert).

https://doi.org/10.1016/j.omega.2025.103335
Received 25 March 2024; Accepted 29 March 2025
Available online 12 April 2025

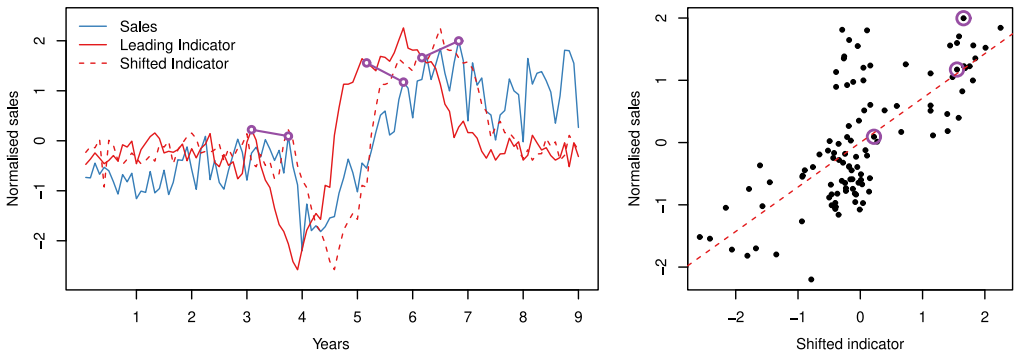

Fig. 1. An example of an 8-month leading indicator from our case company. The scatter plot provides a linear regression fit, together with the highlighted points of the time series plot.

indicators can be macroeconomic and other market variables that are leading changes in the target demand series [8]. The difficulty is identifying the appropriate indicators and estimating their effect on demand. The disaggregate nature of demand at Stock Keeping Unit (SKU), and when this is further separated to stock-keeping locations that inventory decisions are taken, e.g., SKU/store, results in increased noise in the demand signals that make identifying leading indicators problematic. Sagaert et al. [16] provide an implementation blueprint of how to incorporate leading indicators in tactical demand forecasting, using lasso regression. In their case, they focus on aggregate sales across multiple SKUs and locations, to aid lasso regression to identify the relevant leading indicators among the several thousands of alternatives. They also argue that for each forecast horizon, a different set of leading indicators may be relevant, as different signals may have predictive value for different periods in the future. This adds to the computational complexity and cost that are not congruent with the scalability requirements of the typical inventory management context.

In this work we propose a methodology to incorporate leading indicators at the aggregate company demand data and transfer these to disaggregate SKU/location forecasts. The objective is to obtain forecasts that remain accurate in the face of supply chain disruptions and changes, while retaining scalability, and therefore being practically relevant. To do this we leverage the recent advances in hierarchical forecasting [17,18]. We go beyond typical implementations of hierarchical forecasting, where all aggregate and disaggregate time series are modelled by the same forecasting model family, and typically using the same information. Instead, we rely on univariate models for the disaggregate predictions, and on methods with leading indicators for the aggregate levels. We consider statistical and machine learning approaches, namely lasso regression and lightGBM. We address the various complications this introduces and demonstrate the efficacy of the proposed approach on inventory performance, showing that it provides superior results than both conventional modelling using only univariate per SKU forecasts and hierarchical forecasts that rely on a single model family.

We evidence the benefits of our approach using a manufacturing firm that operates on multiple continents and therefore faces multiple macroeconomic environments. Fig. 1 provides an example of the connection of a leading indicator with their sales. The firm produces materials necessary for tyre manufacturing, for private and industrial use vehicles. The leading indicator is tracking the purchases of two and four-wheel vehicles in Sweden and exhibits a strong 8-month lead to sales. Vehicle sales lead tyre sales, and, in turn, lead sales for the case company. The relatively short lead of the indicator to the expected lifespan of tyres is due to the firm supplying tyre manufacturers and therefore the implicit lead to the end consumer is substantially longer. This narrative was validated by the management in our case company.

We demonstrate the synergies of the various components of the proposed methodology. First, we evidence that identifying leading information at SKU-level data is ineffective, requiring a hierarchical treatment. Second, although solely using hierarchical forecasting provides performance gains over standard SKU-level forecasting, these are smaller than gains obtained when leading information is used.

The paper is structured as follows. In Section 2 we provide an overview of the relevant literature. Section 3 details the proposed methodology. Section 4 introduces the company case, the available data, and presents the experimental design that is used to evaluate the proposed methodology. Section 5 presents our findings, followed by a discussion of why modelling indicators at aggregate data is advantageous in Section 6, managerial implications in Section 7, and concluding remarks in Section 8.

2. Literature review

The idea of leveraging external information to demand forecasting and inventory management is long established in the operations management literature. Kouvelis et al. [19] underscore the significance of including market intelligence in demand forecasting to increase effective response capability in inventory management. Aviv [20] advocate that the decision-maker should obtain information about the current market conditions at the beginning of each period, to improve its demand forecasting and inventory decisions. Fildes et al. [13] argues that effectively integrating cross-functional information can improve demand forecasts, but that a judgemental incorporation of such information is often prone to bias, for example, due to too much attention on information relating to a single, isolated past event. Sroginis et al. [14] postulate that information overload is an issue for demand planners and forecasters when selecting relevant information. This motivates the use of models to evaluate a large number of variables, subsequently enhanced by selective judgemental intervention. Nonetheless, when the number of potential variables becomes too large, this can reduce the efficacy of statistical selection methodologies. Chuang et al. [21] stresses the importance of starting with a good set of predictor variables prior to any selection and advices including variables grounded on theory. In agreement, Sagaert et al. [8] show that experts can successfully reduce a larger set of variables to improve the result of a subsequent statistical variable selection, however how to best do this remains an open question.

2.1. Exogenous information

Many different types of external information have been reviewed to improve demand forecasting. One obvious type of information is product-specific information. A common example is price and promotional information, which is known to improve the demand forecasts [22,23] and is particularly relevant to retailing [5,24]. The scope of this information is often short-term, and although it is increasingly incorporated into models, it has been one of the main drivers for expert adjustments of model forecasts [10,11]. To incorporate additional market intelligence in demand forecasts, the literature has looked into

using other inputs, such as social media data. For example, [25] use these inputs in fashion forecasting. Similarly, geographical location information of the point-of-sales [26] and clickstream data [27] have been used to enhance inventory management.

Schaer et al. [15] demonstrate that often consumer behaviour related online data can have a limited time-scope, being less helpful in modelling longer-term dynamics. Therefore, although this information can be leading, in many cases the lead times that a firm faces can be longer than the predictive lead that such variables carry [28]. This introduces two challenges that further complicate the selection of variables. On the one hand, many predictors are excluded when the decision lead time is considered, forcing analysts to look for alternative variables that may contain that signal. This can increase spurious selection of variables, which has been partially evidenced both with models [24] and judgement [14]. On the other hand, as different decision lead times result in a different pools of variables, with potentially different lags of the same information, multiple models may need to be build, and the dimensionality of variable selection can increase substantially, as we discuss in the methodology section below.

Longer-term leading indicators can be sourced from the macroeconomic environment of the firm. Sagaert et al. [8] shows that lasso regression can effectively identify such among a very large pool of potential indicators, improving forecasts up to a year ahead, and being able to capture exogenous disruptions in the sales of a firm. Further work has evidenced the usefulness of this for tactical forecasting [16, 29], however with the limitation that working with aggregate demand data is a precondition for lasso to identify informative macro-economic indicators.

2.2. Variable selection

From the above it is apparent that the analyst is often faced with the situation where the potential number of predictors can exceed the number of historical observations, complicating variable selection. In the literature, most methodologies fall under three approaches, (i) factor models, (ii) shrinkage estimators, and (iii) use of models that can operate without variable selection. Factor models use some dimensionality reduction method, such as principal component analysis, to reduce the initial set of variables to a smaller set that can then be modelled with usual variable selection approaches. It became popular in the forecasting literature following its application in macroeconomic modelling by Stock and Watson [30], building on a large body of previous literature [31], with applications in demand forecasting [32, 33]. A limitation of factor models is that it becomes challenging to identify the influence of individual variables, which can be important both for providing business insights and validating the model [34]. Subsequent work has proposed various approaches to mitigate this, for example, Bernanke et al. [35] develop models that some variables are not transformed into factors and evaluated in their original form.

Shrinkage estimators retain the initial set of variables and focus on addressing the estimation challenge. They operate by regularising sample estimators towards zero, therefore reducing the estimation variance, with ridge and lasso formulations being the most widely used [36]. Both are able to provide estimates when the number of regressors exceeds the available sample, with the latter being able to select variables as well. One of the main advantages of lasso regression is that it achieves a sparse solution [37], while trying to uncover the true signal. Various other estimation methodologies that aim to reduce the estimation or forecast variance have been shown to have equivalences with shrinkage [38]. In the forecasting literature, lasso has numerous successful applications, for example, in macroeconomics [39], retailing [24], and tactical forecasting [8].

Both factor models and shrinkage estimators have shown good performance. Roth Tran [40] uses lasso to select external predictors for sales forecasting and inventory optimisation, with lasso found to be preferable over other techniques such as principal component analysis. Their focus is on shorter-term forecasting, as the variable of interest is weather information. [33] contrast the two in a retailing setting and find that although both are effective in handling a large number of variables, factor models perform better during promotional periods, in agreement with [32], and lasso is preferable otherwise. During promotional periods, the demand uplift is important, which shrinkage estimators will likely underestimate. Otherwise, shrinkage is more reliable as it mitigates sampling uncertainty [37]. These findings motivate our focus on shrinkage methods for variable selection.

Other methodologies, such as lightGBM [7], can avoid variable selection altogether. These are typically ensemble learning methods, such as random forests, or overparametrised methods, such as neural networks. While these have shown superior accuracy in demand planning tasks [for example, lightGBM in the recent M5 forecasting competition, [6]], they lack transparency and interpretability. These can be important for the trustworthiness of the forecasts [34], business insights [8], and model validity. To reduce their opaqueness, it is common to use SHAP values [41], or similar [42], to elucidate which variables have higher importance for the trained model. However, these are often misinterpreted, particularly when it comes to inference for correlated variables [43] and can have substantial computational cost. Nonetheless, when inference is not the primary objective and the focus is solely on forecast accuracy, approaches that side-step variable selection can be very useful. We take this stance in our work, investigating both lightGBM and lasso regression that offer contrasting benefits.

2.3. Hierarchical information

In a demand planning and inventory management context, a firm needs to deal with multiple time series, which themselves may be interconnected. Often these are organised in hierarchies of items that match the organisation of the operations of a firm. For example, different variants of a product are grouped, as this has implications for various operations, such as production, warehousing, etc. Likewise, there is the implicit constraint that the subaggregate series add up to the aggregate one, i.e., the total of the variants of a product match its total sales. Although this is a given for historical data, it is not so for demand forecasts and the decisions they support. Forecasts may violate this constraint. To achieve this coherency there is a large body of literature on hierarchical forecasting [for recent reviews, see18,44]. Traditionally, this problem was seen as a top-down or bottom-up, that is, forecasting at an aggregate level and disaggregating, or at a disaggregate level and summing up, to achieve coherency [45]. The aggregation and disaggregation follow the organisational hierarchy, therefore offering coherent forecasts at decision-making relevant levels. However, these approaches have been criticised for introducing biases [46] or increasing modelling risk [47]. Instead, nowadays hierarchical forecasting is seen as an operation of reconciliation of the forecasts generated across the different levels of the hierarchy, which, apart from coherency, can bring accuracy benefits [46,48,49]. Furthermore, [4] remarks that clustering similar time series can improve the use of data and help reduce the computational needs. [50] note that grouping of the time series improves forecastability, and highlight that using interpretable relationships, such as the use of hierarchical information, is preferred over data-driven clustering.

2.4. Contributions

We contribute to the literature by proposing a methodology to augment the SKU-level forecasts and improve inventory decisions with intelligence from leading indicators, which is identified at an aggregate level of sales. Our work is at the intersection of forecasting with leading indicators and hierarchical forecasting, and specifically, we provide

1. the theoretical motivation and the empirical evidence for using exogenous variables at aggregate levels of hierarchical forecasting. This also contrasts with the standard hierarchical forecasting implementations in the literature where a single model family is used across cross-sectional hierarchies.
2. a methodology to account for the impact of different model parameter estimation methods and their implied effect on model variance on hierarchical reconciliation. We focus on shrinkage and maximum likelihood estimators.

These address gaps identified by Athanasopoulos et al. [18] in their review of hierarchical forecasting, and show the limitations of previous research in using leading indicators for demand planning and inventory management [for example,8] and how to overcome these.

3. We contrast the business insights provided by lasso and lightGBM, judgemental and model based approaches to summarise these, and the interaction between the selection of leading indicators for different forecast horizons and hierarchical reconciliation.

These are important aspects to increase the trustworthiness of forecasts, which is an emerging consideration in the literature [34], their managerial usefulness, and facilitate any additional adjustments by experts by communicating what information is already captured by the models [13].

3. Methodology

The proposed methodology has two elements, first, the generation of predictions for the disaggregate and aggregate demand, and then the combination of the two views using hierarchical forecasting. The reconciliation of the forecasts in the hierarchical forecasting step is a post-forecasting step that can increase the performance of forecasts further by combining the information across the independently specified forecasting models. We rely on this to transfer the leading indicator information across the hierarchy. These two steps together ensure that the resulting forecasts that inform inventory decisions incorporate any available leading information about potential changes in the market. Our methodology handles probabilistic forecasts that are necessary for inventory management. This way, we provide both mean forecasts, as well as safety stocks, as needed, with both reflecting the increased information due to the leading indicators. We detail how these are done in the following subsections.

3.1. Generation of forecasts

3.1.1. SKU-level forecasts

Two challenges at SKU-level forecasts influence the choice of the forecasting approach. First, at the disaggregate level, typically, we are faced with a large number of time series, and therefore automatic specification of methods and scalability are desirable. Second, we anticipate that noise in the time series can make the identification of exogenous variables difficult.

Therefore, at the disaggregate level, we generate forecasts using univariate forecasting models. Although these forecasts can be obtained by any forecasting method, we focus on exponential smoothing and lightGBM,

We rely on the state space exponential smoothing family of models [51]. Exponential smoothing models can tackle a variety of time series patterns that may include trend and seasonality, interacting in an additive or multiplicative way. The state space formulation goes beyond the classic exponential smoothing methods in that it uses maximum likelihood estimation for the smoothing parameters and initial values of the models, as well as permitting automatic model selection using information criteria [1]. This offers substantial benefits in terms of automation and reliability [51,52].

The lightGBM is an optimised Gradient Boosting Machine [GBM,53] proposed by Ke et al. [7]. Given an initial set of predictions, with boosting one can achieve better performance by re-training a model on the residuals from these predictions. This step can be repeated iteratively on the residuals of the previous step. The final boosted forecast is the ensemble of all these forecasts. GBMs rely on weak predictors, such as decision trees, making very few assumptions about the data. However, GBMs come with a multitude of calibration options, which have led to different methods being proposed in the literature. The lightGBM method is one of the more widely used ones due to various optimisations in terms of computation and the inclusion of regularisation options to increase the reliability of the outputs. In the recent M5 forecasting competition, most of the well-performing machine learning methods made use of lightGBM, outperforming standard statistical models [6]. However, the boosting step obfuscates how different inputs are used by the GBM, making the method less transparent to users.

Our methodology does not require the use of these two alternative forecasting methods, and other methods could be used. We select these two given the evidence in favour of their forecasting performance. Furthermore, exponential smoothing is widely available in various forecasting software, while lightGBM is incorporated in multiple machine learning environments. The model-agnostic nature of the SKU-level forecasts is valuable, as it facilitates the incorporation of forecasts that may have been judgmentally adjusted by the users to incorporate contextual information [11]. Furthermore, it is simple to enrich these forecasts with promotional and pricing information that may be available at the disaggregate SKU level [22].

3.1.2. Forecasts with leading indicators

Aggregating the demand of different SKUs and locations, changes the focus from individual demand dynamics to overall ones. Individual demand patterns are averaged out, and market dynamics become more prominent. We rely on macroeconomic leading indicators to model these. As with the SKU-level forecasts, we build a statistical model that incorporates leading indicators and a machine-learning counterpart. For the statistical model, we rely on lasso regression, building on the work by Sagaert et al. [8]. For the machine learning method, we formulate a lightGBM that can use the leading indicators with the considerations prescribed below for the lasso regression. First, we introduce the modelling approach, and then we discuss the motivation for using both methods.

The lasso regression model is a popular shrinkage estimator for regression models [54] that can simultaneously select variables and provide appropriately shrunk regression coefficients. This shrinkage mitigates the effect of sampling uncertainty on the estimates and also allows the selection of variables when the number of candidate variables exceeds the available fitting sample size. This is a very useful property for selecting leading indicators, as the typical case is that they will be in the thousands, while the available sample size will be much smaller.

To achieve these, lasso regression uses a modified loss function from the usual ordinary least squares

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \tag{1}$$

where $\boldsymbol{y}$ is the target time series of sample size n, $\boldsymbol{X}$ is a matrix of p covariates, and $\boldsymbol{\beta}$ is a p-dimensional column vector with the regression coefficients. The shrinkage hyper-parameter λ controls how strongly the coefficients in $\boldsymbol{\beta}$ are shrunk to zero, with $\lambda = 0$ resolving in the usual ordinary least squares solution. For the fitting errors, an ℓ_2 ($\|\cdot\|_2$) norm is used, while for the shrinkage penalty an ℓ_1. In Eq. (1), the ℓ_1 penalty enables coefficients to become zero, effectively removing the impact of the corresponding covariates from the regression. The selection of λ is typically done by minimising the cross-validated in-sample error.

This facilitates automatic model building. However, Sagaert et al. [8] also finds that lasso reaches a better selection of variables once

human experts pre-filtered the starting large groups of indicators. This results in better forecasting performance. This way, one can leverage the expertise of the company, without being too onerous for the experts, since they only select from large groups of indicators. Groupings provided by statistics bureaus were found to be sufficient. This can also lessen the computational load, as it can greatly reduce the pool of indicators for lasso to select from.

Having access to our case company, we rely on their expertise in their market to pre-filter indicators. This was done by interviewing management. The interview was based on the latest quarterly reports from external parties that identified factors that could influence sales on a tactical–strategic level. A list of keywords was curated to select a set of macroeconomic indicators.

In our case, the covariates are classed into three categories: (i) autoregressive inputs, (ii) seasonal dummies, and (iii) exogenous leading indicators. The autoregressive inputs are tasked with capturing time series dynamics present in the target series. The seasonal dummies allow the model to opt between a deterministic and a stochastic representation of seasonality, the latter modelled by sufficiently lagged autoregressions. Although in the long-term we anticipate stochastic seasonality to be more apt for business time series, for shorter samples the deterministic representation often performs better and it is statistically indistinguishable [55]. These two sets of inputs aim to explain as much of the target series variance as possible before any external variables are used.

With leading indicators there are two complexities to overcome. First, connections between the target and explanatory variables can be spurious, often due to trends or seasonalities. These co-movements will manifest as increased correlations, even when no causal connections exist. We address this by (i) first modelling these components as univariate information (autoregressive terms and seasonal dummy variables), and (ii) by using differenced versions of the leading indicators, where trend and seasonality are removed. In most cases, differencing will reduce correlations. This can also help lasso regression, which in the presence of highly correlated explanatory variables will only choose the maximally correlated one and ignore the rest [36].

Second, there is a question of what is the lead-order of the indicators, i.e., how many periods in advance they contain predictive information for the target variable. In building a forecasting model we must ensure that the conditionality on time is retained, i.e., that no information from the future is used. Consider an input with a one-period lead. When forecasting one-period ahead, the value of this input is available. When forecasting two periods ahead, the value of the input is not available, as it is not sufficiently leading. Instead, an input with a two-period lead could be used. Note that the two-period lead is available when modelling the one-period ahead, along with all longer leads. The implication for modelling is that for forecasting h-steps ahead only the leading indicators of h or more periods are available. Therefore, the forecasting task defines the minimum lead-orders that should be investigated. In our case, as the forecast horizon for the case company is 12 months, we create variables from 1 to 12 periods of lead. This results in building h different regression models, each corresponding to a specific forecast horizon, with an increasingly smaller pool of inputs. For instance, for the 12-step ahead forecasting model only the 12-period lead indicator variables are available. No such restrictions are needed for the autoregressive inputs and the seasonal dummies. One can extend to additional lead orders, beyond the 12 of our case. However, an investigation of the correlations between longer leads and the target series showed that there were minimal remaining connections and were excluded from our analysis. Some indicators with lead lower than 12 periods already exhibited low correlations but were kept in the analysis to not violate the conditionality on time.

An alternative to imposing these restrictions would be to forecast the indicators in $\boldsymbol{X}$. However, this is undesirable, as $\boldsymbol{X}$ will typically contain a large number of variables, require building several forecasting models, and, crucially, the objective of the leading indicators is to capture the effect of disruptions and other structural shifts that are hard to forecast. Therefore, producing forecasts of those would introduce substantial forecasting errors in the covariates, and eventually in the forecasts of the target variable.

For the lightGBM the same inputs are considered, with the aforementioned transformations. Here, it is helpful to contrast the two forecasting approaches. Lasso, being a linear regression, provides a transparent output, where the user can identify the most important leading indicators. This can both help with validating the model and with providing additional insights to the user. Nonetheless, captured interactions are linear, and in the presence of multicollinearity, lasso will only pick the most correlated variable of each group of covariates. This may not necessarily be the most causally connected variable [for an example of this effect, see24, where the authors warn about some of the selected variables]. Furthermore, lasso is very efficient when it comes to variable selection and estimation. LightGBM exhibits opposite behaviour in almost all of these dimensions. Due to the boosting algorithm, it obfuscates how the different inputs are used, as the final output is the result of an ensemble of different trees and not a single model. Being a universal approximator, it can model nonlinear interactions, overcoming the linear restriction of lasso. However, this exact property weakens the use of sensitivity analysis and similar approaches to disentangle how its various inputs are used. Multicollinearity is not an estimation concern for decision trees, and likewise for GBMs. However, it will influence the measured importance of variables, reducing the transparency of the method further. Finally, lightGBM is a computationally intensive method.

3.2. Hierarchical reconciliation

The SKU-based and aggregate forecasts can be arranged in a hierarchical structure, with SKUs aggregating by product category, geographical location, or other demarcations that may be appropriate. These separate aggregations can be applicable at the same time, for instance, SKUs aggregating by both product categories and geographical location. An example, from our case study company, is provided in Fig. 3. Forecasts generated at different nodes of the hierarchy will generally not be aggregate consistent, that is, summing up forecasts to their aggregate node will not numerically agree with the forecast generated at that node. The forecasts are therefore incoherent, which is undesirable in practice, as it does not support aligned decision-making [56]. This motivates forecast reconciliation, i.e., to modify the forecasts so that they become coherent.

The current thinking in hierarchical forecasting is one of forecast reconciliation, where forecasts are generated across all nodes of the hierarchy, and then reconciled so that they become coherent [18]. Given a collection of bottom-level time series y_{it}, with $i = 1, \dots, m$ indexing the time series, and t the time period, we collect them in column-vector $\boldsymbol{b}_t$ of all bottom-level observations at period t. From $\boldsymbol{b}_t$ we can construct the observations of the complete hierarchy as

$$\boldsymbol{y}_t = \boldsymbol{S}\boldsymbol{b}_t,$$

where $\boldsymbol{S}$ is the summing matrix of dimensions $n \times n$ that maps the bottom level to a column vector of the observations of all time series in the hierarchy at period t, $\boldsymbol{y}_t$, of size n. For example,

$$\boldsymbol{S} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ & & \boldsymbol{I}_m & & \end{bmatrix}$$

where $\boldsymbol{I}_m$ is the identity matrix of dimension $m = 5$, maps to the hierarchy in Fig. 2. $\boldsymbol{S}$ can map arbitrary complex hierarchies. Note that this hierarchy stems from the decision structure of the forecast user [18] and is not data-driven.

With hierarchical forecasting we generate forecasts for each node of the hierarchy and arrange them in a similar fashion, where $\hat{\boldsymbol{y}}_{t+h}$

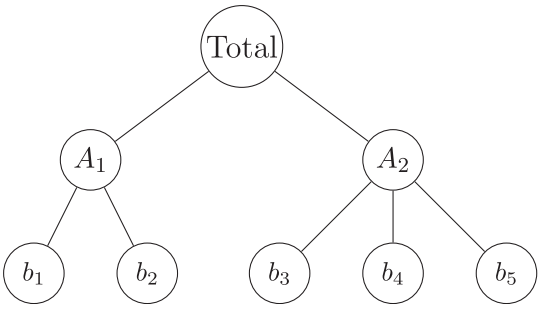

Fig. 2. A hierarchy with 5 series in the bottom level, two aggregate nodes A_1 and A_2, and the top-level *Total* node.

contains the h-steps ahead forecasts for all nodes, which can be incoherent. [46] show that we can obtain a coherent set of forecasts $\tilde{\boldsymbol{y}}_{t+h}$ by

$$\tilde{\boldsymbol{y}}_{t+h} = \boldsymbol{S}\boldsymbol{G}\hat{\boldsymbol{y}}_{t+h}, \tag{2}$$

where

$$\boldsymbol{G}_h = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}, \tag{3}$$

and $\boldsymbol{W}_h$ is the variance–covariance matrix of the forecast errors for horizon h, $\boldsymbol{W}_h^{-1}$ is its inverse, and $\boldsymbol{S}'$ the transpose of $\boldsymbol{S}$.

3.3. Estimation of $\boldsymbol{W}$

Matrix $\boldsymbol{W}_h$ is $n \times n$ and can often be challenging to estimate, due to its size [46,57]. Developing reliable approximations is what made the forecast reconciliation approach feasible [46], and in the literature, there are multiple successful approximations that observe various simplifying assumptions [46,57,58]. Furthermore, it is common practice to use a single approximation for all h, where all $\boldsymbol{W}_h = \boldsymbol{W}_1$. In our case, this is undesirable, as we have different lasso models for each forecast horizon, therefore it is critical to obtain all $\boldsymbol{W}_h$, while balancing the number of parameters that need to be estimated.

To this end, we rely on the variance approximation $\boldsymbol{W}_h$, which prescribes a diagonal matrix with its elements being estimates of the forecast variance, typically obtained from the in-sample Mean Squared Error (MSE) of the forecasts. This approximation imposes that off-diagonal elements are zero, and therefore assumes that there are no cross-effects between time series. Although this assumption appears strong, [57] shows that the imposed restriction has a positive effect on the resulting quality of the reconciled forecasts due to the lessening of estimation uncertainty in the elements of $\boldsymbol{W}_h$ that is propagated via Eq. (2) and Eq. (3) to the final forecasts. This is to the extent that can make the misspecification of $\boldsymbol{W}_h$ of limited consequence when the alternative is a weakly estimated $\boldsymbol{W}_h$.

In populating the diagonal of $\boldsymbol{W}_h$, we follow a different procedure depending on the method that is used to generate the forecasts, with the intent to minimise the number of parameters that need to be estimated. For exponential smoothing models there are analytical expressions of h-step ahead forecast variances [51] which rely solely on the estimated smoothing parameters. For lasso we rely on empirical estimates of the h-step ahead MSE_h. Since the objective of lasso models is to avoid a minimum-MSE fit to the data, we sample the empirical forecast errors in a quasi-out-of-sample fashion. We conduct a rolling origin evaluation in the in-sample data. Using $v \leq l-h$ observations, where l is the length of the in-sample data, we fit the lasso and produce forecasts for periods $v+h$ and collect the respective forecast errors. This is repeated until all samples have been exhausted. Naturally, longer horizons result in a fewer number of forecast errors, eluding to another motivation for the majority of hierarchical applications to use $\boldsymbol{W}_h = \boldsymbol{W}_1$. For lightGBM we repeat the process outlined for the lasso. This is necessary for the case where it uses leading indicators (at the top level of the hierarchy) and simplifies the methodology otherwise by keeping the approach consistent.

A different way to understand the connection between $\tilde{\boldsymbol{y}}_{t+h}$ and $\hat{\boldsymbol{y}}_{t+h}$ is to see the forecast reconciliation as a forecast combination, where all $\hat{\boldsymbol{y}}_{t+h}$ are combined into $\tilde{\boldsymbol{b}}_{t+h}$. Subsequently, by multiplying by $\boldsymbol{S}$ the $\tilde{\boldsymbol{y}}_{t+h}$ are obtained (see Eq. (2)). The forecast combination weights are restricted to conserve forecast coherency, according to Eq. (3). In this framing, we can see that hierarchical forecasting can be beneficial for forecast accuracy, as it is a forecast combination, and in our case, will transfer the information from the leading indicators to the bottom-level univariate forecasts. Likewise, as the estimation uncertainty of the forecast combination weights can harm the performance of the combined forecasts [59], so does the estimation uncertainty of $\boldsymbol{W}_h$, which motivates the choice of a restricted approximation.

3.4. Probabilistic reconciliation

For inventory management we need to calculate the safety stock, for which we require the appropriate quantile of the predicted demand distribution. Therefore, the point predictions in $\tilde{\boldsymbol{y}}_{t+h}$ are insufficient. To obtain probabilistic expressions we follow the methodology by Panagiotelis et al. [17]. Given a forecasting model, one can generate estimates of the predictive distribution by simulating different forecast traces in the future by perturbing the innovation term in models. This can be done by bootstrapping the forecasts, which results in a distribution of forecast traces in the future and corresponds to the desired forecast distribution. Panagiotelis et al. [17] rely on this to generate hierarchical probabilistic forecasts. Forecast traces are simulated for all nodes of the hierarchy, and each set of traces is reconciled to give a coherent simulated set of traces. By constructing a collection of these, we can obtain the hierarchical predictive distribution. The reconciliation of each set of forecast traces is done using the aforementioned methodology, and therefore is fairly simple to implement.

Finally, note that since the lasso models are using only lagged values of the leading indicators, there is no additional uncertainty originating from these, and therefore the simulation of future forecast traces of the lasso is based on simply bootstrapping the observed forecast errors.

4. Case study

The proposed methodology is used to support the inventory management of a manufacturing company with a global supply chain. The company has its headquarters in Europe, but maintains production sites globally, and serves several major tyre manufacturers in multiple markets and countries. Given this global reach and diverse market portfolio, there is a large number of potential leading indicators that can be relevant in improving demand forecasts.

In the empirical evaluation that follows we evaluate both the impact on forecasting and inventory performance. Given the highly upstream position in the supply chain of the case company, it maintains a relatively small number of SKUs, which makes it ideal for extensive testing of our methodology. From a computational standpoint, it is feasible to build forecasting models with leading indicators for all SKUs, and not only for the aggregate level. Therefore, we can evaluate several different setups, forecasting the SKU demand using (i) exclusively univariate forecasts; (ii) exclusively forecasts with leading indicators; (iii) mixed approaches that leverage hierarchical forecasting, where all levels can be based on univariate or leading indicator forecasts, or the proposed blended approach. Forecasts are generated to support various planning requirements, for up to one year ahead.

For the period of the provided data, the case company had been using the Holt-Winters method. This corresponds to an exponential smoothing model with additive errors and slope and multiplicative seasonality. Following the classification by [51] this model belongs to Class 5, which is characterised by numerical issues and infinite forecast

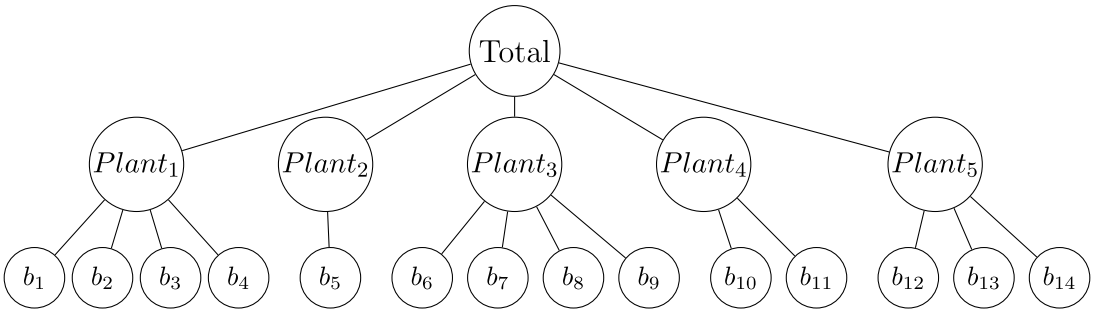

Fig. 3. The hierarchy of the case company time series, with b_i denoting the different SKUs at the bottom level of the hierarchy.

variance, due to the mixing of additive errors and multiplicative seasonality. To illustrate this, consider non-positive values that are possible from the combination of the additive errors and slope, being multiplied by the seasonal component. Another issue is that not all time series are well approximated by this method. Therefore, we replace this with the complete exponential smoothing family of methods, that enables model selection and will also provide forecast variance expressions that are needed for the inventory management calculations.

4.1. Data

Data originate from five global plants that in total produce 14 SKUs, and are organised as illustrated in Fig. 3. The time series are sampled at a monthly frequency and span 11 years. The first 9 years are used for the training of the models (108 observations), and the remaining data are used as a test set (26 observations). The company requires a forecast of 12 months in the future, and this is done for the test period in a rolling origin setting (13 iterations). The time series typically exhibit some trend and seasonality, although it is not always strong.

We source potential leading indicators from the Federal Reserve Economic Data (FRED). The indicators consist of different types of data, covering different aspects of the macroeconomic dynamics. The complete set of possible indicators is 67 851 variables. However, as discussed in Section 3.1.2 we rely on pre-filtering by one of the plant managers, reducing the complete set to 1 011 indicators. Lagged versions of these are constructed, covering up to 12 lags, aligning with the maximum forecast horizon, resulting in 12 132 total variables. Note that this set is reduced for models of higher h. For instance, for $h = 2$ only 11 121 variables are available, while for $h = 12$ only 1 011 remain.

4.2. Methods

All exponential smoothing models use a maximum likelihood estimation, and the appropriate model form is selected using the Akaike Information Criterion, corrected for sample size. We rely on the smooth package [60] for R [61] for the implementation. For lightGBM, we use the similarly named R package implementation of [62] with inputs of auto-regressive components, 100 training rounds for each model and minimise the Mean Squared Error (MSE). All forecasts with leading indicators, lasso and lightGBM, use the aforementioned leading indicators, autoregressive lags, 11 binary seasonal indicator variables, and appropriately differenced versions of the indicators. For lasso, the λ is obtained via cross-validation. The glmnet package for R [37] is used for the implementation of the lasso models. For the hierarchical reconciliation in estimating the $\boldsymbol{W}_h$ we start with 108 observations from the in-sample data and expand with the rolling origin.

We use the following naming convention. Forecasts from exponential smoothing are denoted by *E*, lightGBM by *G*, and from lasso by *L*. When a hierarchical setting is used, then the forecasts for each level are given from the top to bottom level. For example, *LEE* uses lasso on the top level, and exponential smoothing in the middle and bottom levels. LightGBM uses leading indicators only when applied to the top level of the hierarchy. at lower levels of the hierarchy, univariate lightGBM forecasts perform best, due to the increased variability of the time series.

4.3. Inventory simulation

To inform the inventory performance evaluation, we implement an order-up-to inventory policy with lost sales [63]. The review period is set to 1, while for the lead time 3, 6, and 12 months are considered. The inventory simulation is run for different target fill rates (from 90% to 99.9%), which are used to construct trade-off curves. This is done to provide a more informative view of the performance of the proposed methodology, rather than focus on a specific fill rate.

To mitigate the impact of the initialisation of the inventory levels and the orders in the system, we use 108 periods as a burn-in sample, preceding the test set. The inventory policy is left to run on that sample, but the performance is not used in the evaluation. Visual inspection demonstrated that this was sufficient for the ordering cycles to exhibit canonicity with minimal influence from the initial values.

4.4. Evaluation metrics

We rely on three categories of evaluation metrics, (i) forecast performance metrics; (ii) cumulative forecast performance metrics that align closer to the inventory performance; and (iii) inventory performance metrics. These are detailed below.

Beyond evidencing the performance of the proposed methodology, we report the performance across these alternative metrics to contrast the conclusions from each. Furthermore, the evaluation is focused on the bottom level of the hierarchy, where the inventory decisions take place, following the arguments by [64] that recommend restricting the evaluation in a hierarchical setting to the decision-relevant levels.

4.4.1. Forecast error metrics

We are interested in tracking the accuracy and bias of the point forecasts, and the pinball loss that is the proper score for quantiles [65]. The pinball also corresponds to the appropriate loss for the order-up-to policy when there are backorders.

Let $e_t = y_t - \hat{y}_t$ be the error for period t, and a scaling factor

$$s^p = \frac{1}{r-1}\sum_{t=1}^{r-1}\left(|y_{t+1} - y_t|\right)^p, \tag{4}$$

where $p \in \{1, 2\}$ is the order of the scaling factor and r the sample of the training set. This changes across rolling origins, as methods have additional fitting samples available to them. Note that s^1 corresponds to the mean absolute error of an 1-step ahead in-sample naive forecast, and s^2 is the equivalent mean squared error. We use the Root Mean Squared scaled Error (RMSsE), the Absolute Mean scaled Error (AMsE), and the scaled Pinball (sPIN) that are defined as

$$\text{RMSsE}_h = \frac{1}{o}\sqrt{\sum_{i=1}^{o}\frac{\left(y_{i+h} - \hat{y}_{i+h}\right)^2}{s^2}},$$

$$\text{AMsE}_h = \frac{1}{o}\left|\sum_{i=1}^{o}\frac{(y_{i+h}-\hat{y}_{i+h})}{s^1}\right|,$$

$$\text{sPIN}_h = \frac{1}{o}\sum_{i=1}^{o}\frac{w_i}{s^1},$$

$$w_i = \begin{cases}(y_i-\hat{Q}_i)\alpha, & \text{if } y_i \geq \hat{Q}_i\\ (\hat{Q}_i-y_i)(1-\alpha), & \text{if } y_i < \hat{Q}_i,\end{cases}$$

where o is the number of forecast origins in the out-of-sample for the given horizon h, and $\hat{Q}_i$ is the prediction for the α% quantile for period i. We report results for $\alpha = 0.95$. All metrics are scaled to facilitate the calculation of summary metrics, as the different time series have different scales. The scaling factor can be seen as an estimate of the time series variance, assuming non-stationarity, and matching the loss order of the error metrics to avoid introducing any biases [64]. Therefore the metrics can be interpreted as normalised versions of the usual Root Mean Squared Error, Absolute Mean Error, measuring the size of the bias, and Pinball. We report the average errors for 1 to 3, 1 to 6, and 1 to 12 periods ahead.

4.4.2. Cumulative error metrics

As we are interested in covering the demand over the lead time, the conventional way of measuring forecasting accuracy, as above, does not align with our objective. Instead we can replace the errors with the cumulative errors over the lead time as

$$ce_{Lt} = \sum_{i=1}^{L} y_{t+i-1} - \sum_{i=1}^{L}\hat{y}_{t+i-1}, \tag{5}$$

where $L \in \{3, 6, 12\}$ is the lead time. This gives us the RMSsCE and AMsCE. For the pinball, its cumulative counterpart, sCPIN, is based on summed actuals, and quantiles of the cumulative predictive demand distribution. The latter is calculated by collecting a distribution of the cumulative forecast errors, as in (5), up to that point, estimating empirically from these the desired quantile, and adding the summed forecasted demand. Given that forecasts are generated for each period ahead, up to the required lead time, a challenge is to correctly account for the covariances between the predictive distributions of the different forecast horizons, which are necessary for estimating the predictive distribution over the lead time. [66] show that omitting these covariances can have a significant negative effect, and in the absence of analytical formulas for these covariances for the majority of models and methods, the above procedure provides good results. Note that the same is used for the calculation of the predictive demand quantiles that are needed for the safety stock of the order-up-to policy.

4.4.3. Inventory metrics

We use the average on-hand inventory, I_α^+, and the fill rate, FR_α, achieved over the lead time, in the test set, for a target rate α. The I_α^+ is scaled with the mean of the observed demand over the test period, to make it scale independent. This scaling factor is chosen as the demand series are non-stationary, and therefore the in-sample mean demand can be substantially different. Similarly, we avoid scaling with the per-period demand, akin to a percentage metric, due to potential numerical issues. The fill rate is calculated as

$$\text{FR}_\alpha = \frac{1}{o}\sum_{i=1}^{o}\frac{d_i(\alpha)}{y_i}, \tag{6}$$

where $d_i(\alpha)$ is the satisfied demand in period i for a given target α%. Note that the review is at one period, and therefore the fill rate is calculated in every period before summarising.

5. Results

In our analysis of the results, we are interested in evidencing the various elements that make our proposed approach necessary. First, we show that the direct use of leading indicators at SKU-level series is problematic and does not yield accuracy or inventory benefits. Second, we show that although hierarchical forecasting is advantageous in mitigating the impact of weakly performing forecasts, it is the diversity of information available to forecasts across the hierarchical levels that maximises the benefits. This results in the proposed mix of leading indicators at the upper levels of the hierarchy, with univariate methods at the lower ones, closer to where inventory decisions are made. Although the specific mix of forecasting methods, statistical or machine learning, seems to have a small impact on the performance statistics, the lasso regression provides more consistent results and sufficient transparency to the users to identify principal leading effects in their business environment. Below, we first evidence these from a forecasting accuracy perspective, and then from an inventory performance point of view.

Table 1
RMSsE summary for base methods.

Method	Non-cumulative			Cumulative		
	t+3	t+6	t+12	t+3	t+6	t+12
	SKU-Level					
E	**0.824**	0.874	**0.930**	2.048	3.931	8.005
L	0.837	**0.873**	0.985	**1.962**	**3.473**	**7.632**
G	2.772	2.860	2.922	8.274	17.063	34.863
	Total					
E	**0.935**	1.045	0.912	**2.339**	5.189	7.622
L	1.092	**0.985**	**0.882**	2.872	**3.730**	**3.507**
G	3.997	4.039	4.091	11.991	24.235	49.086

5.1. Forecasting performance

5.1.1. Base methods

Table 1 reports the RMSsE for the base methods, the univariate exponential smoothing (E), the lasso regression with leading indicators (L), and the lightGBM (G), for both the SKU-level and aggregate total (see Fig. 3). Recall, that for G at the SKU-level there are no leading indicators. These are used only at the aggregate level. Errors are provided in their non-cumulative and cumulative form. We are interested in validating that the leading indicators add value, with the expectation of observing a stronger effect on the aggregate level. We are expecting a similar effect for longer horizons, as the influence of leading indicators would become more pronounced as we move further away from the forecast origin. In each column, the method with the lowest error is highlighted in bold. We only provide the RMSsE for brevity, with the other metrics providing similar insights. Detailed reporting follows in Table 2 that focuses on the SKU-level.

For the non-cumulative errors, for the SKU-level, E remains very competitive, on average outperforming L. For the aggregate total the opposite is true, with L gaining over E for all but the shortest horizons. This validates the understanding that at SKU level univariate methods remain competitive, as the noise in the data makes it difficult to build models with predicatively valuable leading indicators. This further supports our motivation to identify effective ways to transfer the information from the leading indicators at the aggregate level to the SKU-level series. For the cumulative errors, L exhibits better performance, with the difference between E and L increasing for the aggregate level. To understand why the ranking is inversed for the SKU-level, we need to consider how the cumulative errors are calculated. From (5) we can see that it implies a summation of observations, that acts as a filter, reducing the focus on individual observations, and therefore the impact of noise. The cumulative errors connect more closely to the inventory decisions, as we are interested in the best forecast over the lead time.

The performance of G is weak throughout the Table 1. For both cases, this is due to limited data available for training the lightGBM. Furthermore, at the top-level, lightGBM has to contend with a single series and a substantially larger number of inputs, neither of which is conducive to its performance.

Table 2
Forecasting performance at the SKU-level.

Method		Non-cumulative			Cumulative		
		t+3	t+6	t+12	t+3	t+6	t+12
		RMSsE					
B	E	0.824	0.874	0.930	2.048	3.931	8.005
	L	0.837	0.873	0.985	1.962	**3.473**	7.632
	G	2.772	2.860	2.922	8.274	17.063	34.863
H	EEE	**0.815**	0.868	0.937	2.007	3.999	7.928
	LLL	0.912	0.969	1.076	2.188	4.161	8.422
	GGG	1.218	1.202	1.163	3.211	6.040	10.802
LU	LLE	0.879	0.931	1.019	2.024	3.920	7.979
	LEE	0.817	**0.849**	**0.891**	**1.936**	3.620	**6.817**
	LLG	1.022	1.035	1.063	2.511	4.675	8.239
	LGG	1.154	1.148	1.134	2.996	5.655	10.106
		AMsE					
B	E	1.316	1.397	1.486	3.286	6.315	12.905
	L	1.318	1.372	1.549	**3.077**	**5.461**	11.967
	G	4.529	4.684	4.799	13.53	27.977	57.332
H	EEE	**1.298**	1.381	1.487	3.222	6.426	12.780
	LLL	1.451	1.560	1.733	3.476	6.817	13.649
	GGG	1.921	1.894	1.836	5.065	9.516	17.046
LU	LLE	1.406	1.503	1.651	3.247	6.383	13.109
	LEE	1.307	**1.354**	**1.419**	3.121	5.794	**10.890**
	LLG	1.596	1.643	1.705	3.884	7.433	13.260
	LGG	1.809	1.799	1.785	4.681	8.833	15.716
		sPIN (95%)					
B	E	0.180	0.201	0.237	0.457	0.962	2.011
	L	0.198	0.207	0.234	0.599	1.364	2.683
	G	0.275	0.266	0.270	0.846	1.789	3.880
H	EEE	0.175	0.196	0.235	0.451	0.944	2.053
	LLL	0.205	0.219	0.246	0.544	1.227	2.224
	GGG	0.272	0.268	0.275	0.840	1.797	3.909
LU	LLE	0.174	0.187	0.232	0.438	0.967	**1.603**
	LEE	**0.167**	**0.181**	**0.217**	**0.434**	**0.864**	1.771
	LLG	0.203	0.210	0.248	0.550	1.124	2.450
	LGG	0.259	0.251	0.262	0.771	1.558	3.560

B, H, and LU, stand for base, hierarchical, and leading indicators at upper hierarchical levels.

5.1.2. Hierarchical methods

Table 2 reports all forecasting performance metrics for the SKU-level, which is of interest for the inventory decisions. The summary across series is provided for all metrics, in their non-cumulative and cumulative version, with the latter showing the over-the-lead-time performance. The best-performing method in each column is highlighted in bold. The table organises the methods according to 'B' for base, 'H' for hierarchical, and 'LU' for forecasts with 'leading indicators at upper hierarchical levels. This helps understand where accuracy gains stem from. The RMSsE results for the base methods are the same as those reported in Table 1 for the SKU-level.

First, we compare the hierarchical methods (EEE, LLL, GGG) with their base counterparts. Starting from the RMSsE, we observe that EEE is on average more accurate than E, although in many cases the differences are small. The same improvements are not observed when comparing LLL and L, with LLL being in all cases less accurate. In the case of GGG and G, there are large improvements throughout. The results exhibit some consistency when we consider how hierarchical forecasting operates. At its core, it is a forecast combination. The lightGBM forecasts in the different levels of the hierarchy perform poorly, suggesting some model misspecification. This is ameliorated by the combination, where forecasts that fail in a complementary way can be improved upon. This is observed to a lesser extent in the case of exponential smoothing. In contrast, there is little that the lasso regression on the disaggregate time series can add to the lasso on the total sales, where the effect of leading indicators can be captured best. In this case, there is little for the forecast combination to improve on, with LLL performing worse than L. A similar understanding emerges from the remaining metrics, AMsE and sPIN.

Next, we consider the hierarchical forecasts that use leading indicators at the top level, and univariate forecasts at the lower levels of the hierarchy. Namely, these are the LLE, LEE, LLG, and LGG, listed within the group LU. Between the methods that combine lasso with exponential smoothing and lasso with lightGBM, the former are more accurate, leveraging the better performance of exponential smoothing over lightGBM. Nonetheless, the LLG, and LGG uniformly outperform GGG and G across all metrics and horizons. Focusing on LLE and LEE, the latter exhibits lower RMSsE, AMsE, and sPIN, across all horizons. We already saw that applying lasso to the SKU-level (L) performs poorly. Applying lasso to the intermediate level of the hierarchy results in relatively better performing forecasts, which nonetheless do not improve upon simply using univariate forecasts that at that level, as LEE does. LEE is one of the best-performing methods throughout and when it does not rank best, it is a close second. This is validated across all metrics.

The conclusions we can draw so far are the following. Hierarchical forecasting can reduce large forecast errors, for instance, those exhibited by the lightGBM forecasts, but does not always lead to better results. When lasso forecasts on SKU-level are included, which incorporate leading indicators, then the performance of the hierarchical forecasts suffers. However, when we use leading indicators only on the upper levels of the hierarchy, then we obtain the best-performing results throughout metrics and horizons. This is due to the univariate and the leading indicator forecasts having complementary errors that can be used by the forecast combination implied in hierarchical forecasts to obtain better results (the quality of the forecast combination depends on the covariance between the errors of the forecasts combined). The leading indicators help capture changes in the observed sales that are not explained by their univariate structure. Instead, univariate models throughout the hierarchy err similarly, leaving less space for the forecast combination to improve the outcome.

Considering the differences between cumulative and non-cumulative results, the use of leading indicators results in larger differences in the cumulative case, which also connects more closely to the inventory decisions. The non-cumulative errors ignore any error covariances across horizons, as the performance for each horizon is calculated separately and then averaged to provide the figures in the table. The intuition between these covariances is that large errors will negatively affect future performance. Cumulative errors account for this [66]. The value of the leading indicators is seen not only in reducing long-term errors but also in reducing these covariances, suggesting that forecasts that use them can capture more information making errors that are less correlated and more random.

We observe improvements across all metrics. In terms of bias (AMsE), L and LEE rank highly (except for non-cumulative t+3, where EEE ranks first, followed by LEE). We anticipate this to translate to favourable inventory performance, as there is empirical evidence that bias connects to that [e.g.,67]. The differences in bias become starker for longer lead times, and therefore we expect to see larger gains in inventory there. The sPIN in a newsvendor setting with no lost sales corresponds to the inventory performance. Our inventory setting includes lost sales, so the connection in our simulations will be weaker but sPIN is still indicative. LEE dominates for most cases, except for the cumulative t+12 where LLE, another method with leading indicators, exhibits lower sPIN.

5.1.3. Top level leading indicators methods

Table 3 compares the LEE that overall resulted in the top performing forecasts with GEE. GEE uses a lightGBM with leading indicators at the top level and univariate exponential smoothing forecasts for the middle and SKU-levels. Although LEE remains the overall best, the differences between the two methods are small. Even though lightGBM suffers from a limited training sample, its forecasts complement the univariate (and well-performing) disaggregate forecasts. This is in agreement with the previous intuition from the forecast combination, of why using leading indicators at the aggregate levels of the hierarchy is beneficial.

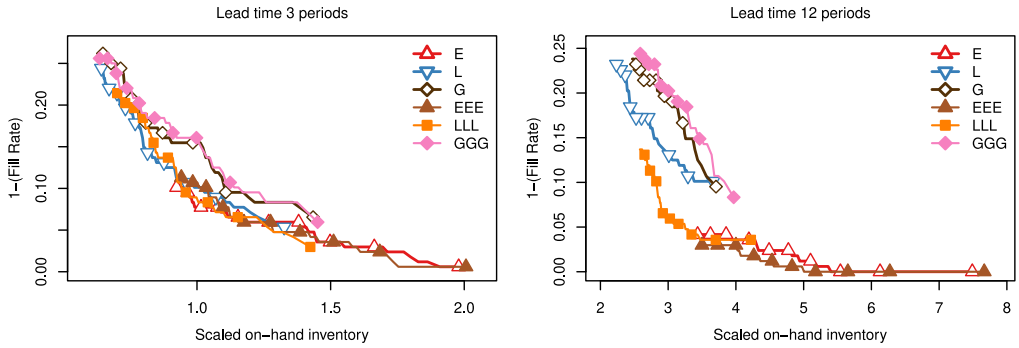

Fig. 4. The effect of hierarchical reconciliation on inventory performance curves.

Table 3
Forecasting performance at the SKU-level.

Method	Non-cumulative			Cumulative		
	t+3	t+6	t+12	t+3	t+6	t+12
	RMSsE					
LEE	0.817	**0.849**	**0.891**	**1.936**	**3.620**	**6.817**
GEE	**0.810**	0.856	0.914	1.981	3.877	7.589
	AMsE					
LEE	1.307	**1.354**	**1.419**	**3.121**	**5.794**	**10.890**
GEE	**1.291**	1.361	1.451	3.180	6.230	12.213
	sPIN (95%)					
LEE	**0.167**	**0.181**	**0.217**	**0.434**	**0.864**	**1.771**
GEE	0.173	0.193	0.231	0.439	0.922	1.988

5.2. Inventory performance with lost sales

Fig. 4 plots the inventory performance in terms of realised fill rate, as calculated in (6), against the remaining scaled stock on-hand, for the shortest and longest lead times. The results of the omitted lead time are proportional and excluded for brevity. We plot the 1-(fill rate), focusing on the unmet demand, which connects to the lost sales. Naturally, as the inventory increases, there will be more leftover on-hand and the 1-(fill rate) will tend to zero. Therefore, the best-performing method is the one that can achieve the lowest amount of lost sales for the minimum amount of stock on-hand, which means that curves closest to the lower-left corner dominate other solutions.

In Fig. 4 we first compare base forecasts with their hierarchical counterparts. For the $t+3$ lead time, the differences in performance are relatively small, with only G and GGG being substantially different and worse, reflecting the accuracy results reported in Table 2. The more challenging $t+12$ lead time separates the performance of the methods further. In the longer term, leading indicators will have more opportunity to capture changes in sales, as in the very short term there is limited time for disruptions to develop to their full effect. Nonetheless, observe that L performs rather weakly and is grouped with G and GGG. L is attempting to model leading effects directly at the SKU-level and is found to be ineffective. In case of E and EEE both are strong performers, with EEE having an edge over E. From these comparisons we can see that hierarchical forecasting does not lead to consistent benefits for inventory, echoing our findings from the various accuracy metrics.

In Fig. 5 we compare the results of E, L, LEE, and LLE. Even though in the short term we do not anticipate large differences, we can see that for high fill-rates both LLE and LEE outperform E. L is not able to achieve these high fill rates. For the $t+12$ lead time the differences become more pronounced, with LEE leading with small differences from LLE and with substantial gains from E. Including leading indicators at the upper level of the hierarchy results in substantial inventory gains at the SKU-level over standard univariate forecasts.

We note that both LEE and LLE perform well. We explore this further in Fig. 6 where we compare LLL, LLE, LEE, and EEE, i.e., all the possible different options for a cutoff of the inclusion of leading indicators. In the shorter lead time, the various curves cross and overlap. LEE (marginally) leads for higher fill rates, while LLE does so for the upper part of the figure. In agreement with previous evidence, for the short lead times, there is limited space for leading indicators to provide a clear benefit. For the longer lead time, LEE leads consistently, although different alternatives are close to it for different fill rates. Therefore, we see that when there is a sufficient horizon for the leading indicators to be impactful, including them at the top level of the hierarchy provides consistent gains, while including them directly at the SKU-level (see LLL, and L in Fig. 4) harms inventory performance. Finally, in Fig. 7 we compare the performance of LEE and GEE which use lasso or lightGBM base forecasts at the top level that include leading indicators. Both perform very similarly, suggesting that it is the inclusion of the information that provides the gains.

The results of the inventory simulation, where lost sales are considered, largely agree with the outcome from the sPIN evaluation reported in Tables 2 and 3. In the short term leading indicators offer gains, but these are typically small. On the other hand, they offer more substantial benefits for longer lead times. When it comes to how to include these indicators, doing so directly at the SKU-level can be damaging, with standard exponential smoothing (E) performing best in almost all cases. However, when these are included at the upper levels of the hierarchy, they result in more substantive gains. There is some robustness in terms of how these are included. Although the LEE method is the one that leads in terms of overall performance, LLE and GEE are fairly close. Finally, although hierarchical forecasting can leverage the complementary information included even in poorly performing forecasts, without using leading information solely at the upper levels of the hierarchy and univariate forecasts at the lower ones, its maximal gains cannot be attained.

5.3. Influence of aggregate forecast across horizons

Matrix $\boldsymbol{G}_h$ contains the combination weights of all base forecasts to the reconciled bottom-level forecasts. Each row in the table contains the combination weights of the based forecasts that correspond to a single reconciled SKU-forecast. Fig. 8 plots the average contribution of each hierarchical level across horizons for the LEE case. The average is calculated across forecast origins, where $\boldsymbol{G}_h$ is re-estimated, and across SKUs. Only the weights of nodes that are directly connected to an SKU are used for this plot. For example, in Fig. 3 for b_1 only the weights for b_1, Plant_1, and Total are considered. Other nodes in the hierarchy have a residual contribution, which is excluded here for clarity.

As the forecast horizon increases, the influence of the top-level, which is based on the model that includes the leading indicators, increases as well. The mid-level's contribution remains fairly constant, while the bottom-level, i.e., the base SKU forecast's, contribution decreases. Considering that in the calculation of $\boldsymbol{G}_h$ the variance of the

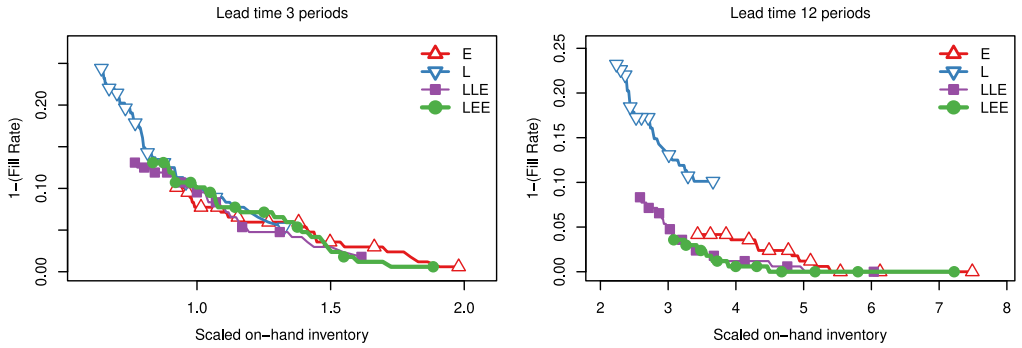

Fig. 5. Inventory performance when leading indicators are used at the upper levels of the hierarchy.

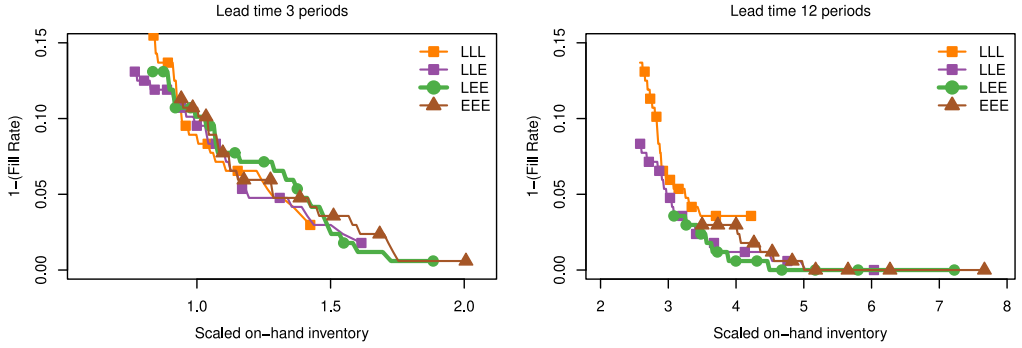

Fig. 6. Inventory performance curves for different levels including leading indicators.

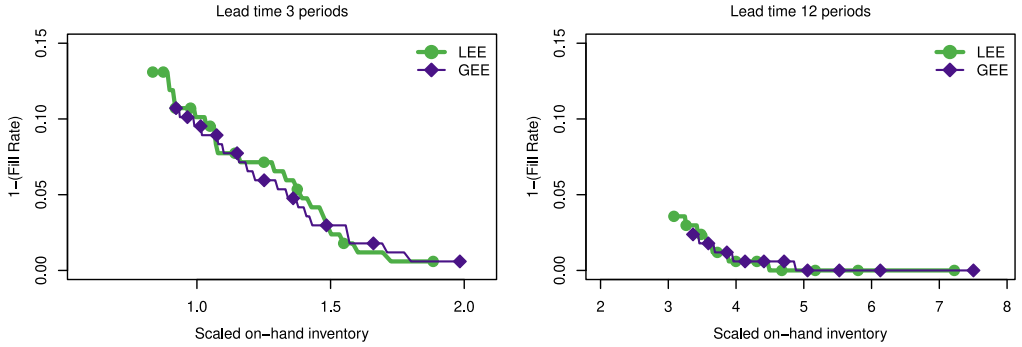

Fig. 7. Inventory performance curves (target fill-rates range from 90.0% to 99.9%).

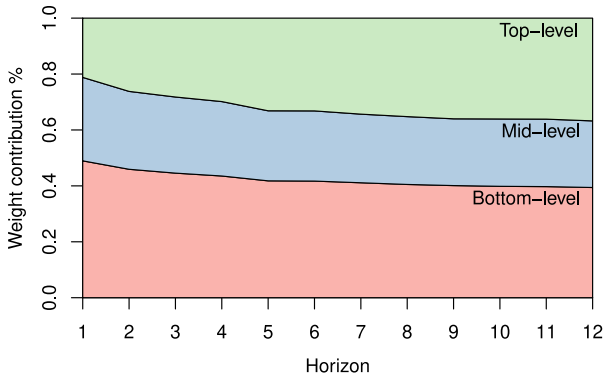

Fig. 8. Average percentage contribution of each hierarchical level across forecast horizons.

forecast errors is used for $\boldsymbol{W}_h$, the usefulness of lasso becomes clear. Lasso, using leading indicators can obtain relatively smaller errors than the univariate forecasts for longer horizons, making use of the information contained in the leading indicators. For shorter horizons, there is limited space for the leading indicators to have an important role. Conversely, the mid-level, which takes advantage of some noise filtering due to the aggregation, but not of leading indicators, does not exhibit an increasing contribution as the horizon increases.

These insights help us understand further the reported results in Table 2 and Fig. 5. For longer horizons, the usefulness of E drops, corresponding to poorer forecasting and inventory results. This also explains the increasing difference in the performance of LEE from the methods that rely exclusively on univariate forecasts as the horizon increases.

5.4. Choice of leading indicators

Having demonstrated the positive impact of including leading indicators in the demand forecasts and their increasing importance over longer lead times, we investigate the nature of the selected leading

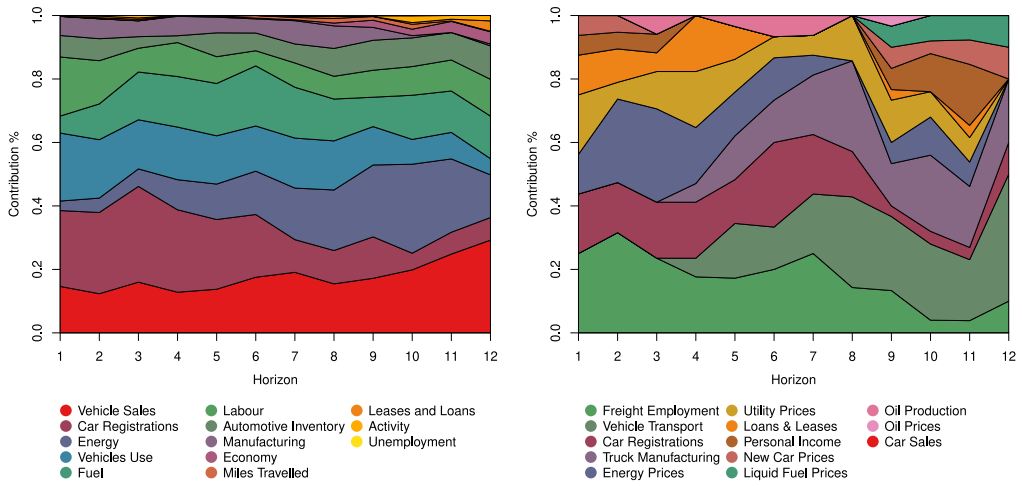

Fig. 9. Average percentage contribution of each group of indicators across forecast horizons via judgemental grouping (left) and via clustering with machine learning (right).

indicators. We rely on LEE for this analysis, as lasso regression gives a direct view of how each input variable contributes. We use two approaches to group the selected indicators. Our first approach is to judgementally group the indicators. Although this approach is manual, it has the advantage that market expertise can be used in the semantic grouping of the leading indicators. The second approach is model-based and therefore scalable. We rely on sentence-BERT by Reimers and Gurevych [68] (Sentence-Transformers package) to semantically cluster the names of the indicators. In brief, sentences are tokenised in numerical vectors. These are then clustered using a k-means algorithm and the Euclidean distance. For naming each cluster we trialled using Large Language Models to automatically obtain short descriptions, however, this was ineffective, and we refined those judgementally.

Fig. 9 summarises the percentage that each group contributes to the pool of selected leading indicators per forecast horizon, ordered from biggest to smallest. Judgemental grouping is on the left and the sBERT clustering on the right. Matching categories are plotted with matching colours. These groups cover multiple related indicators, across different countries, and all the relevant lags.

For the judgemental grouping there are some dominant groups, such as *Vehicle Use*, *Vehicle Registration*, *Energy*, and *Fuel*, which on average contribute from 70 to 80% of the selected indicators. All the remaining groups account for the rest of the selected variables. The contribution of the groups does not remain constant over time, with some clear dynamics being evident. The indicators in the group *Car Registrations* are dominant for short horizons, but become replaced by *Energy* and *Vehicle Sales* for longer horizons. We similarly observed a gradual reduction of the importance of the *Vehicle Use* group over increasing horizons. On the other hand, groups like *Fuel* and *Labour* have a more consistent behaviour throughout the different forecast horizons.

For the SBERT clustering, we can see that the group *Car Registrations* has a similar shape compared to the group formulated based on judgements. For other groups, the comparison is less clear, as both the naming and the coverage of the groups differs, There are similar tendencies, such as the increased contribution over the forecast horizon of *Vehicle Sales* (for both cars and trucks) on the left, and the *Truck Manufacturing* on the right plot, which contributes in the medium and long term. The *Freight Employment* cluster has a much narrower scope: employment specifically in freight transportation. Its contribution is clear as more employment means more road transport and more need for tyres. The group overlaps with the *Vehicle Use* and *Labour* from the left plot. In both cases, we can see this has a large contribution in the short term and a smaller contribution for longer horizons. Although there is an overlap, the similarities are less clear for *Energy & Fuel* (left) and *Energy Prices*, *Utility Prices & Liquid Fuel Prices* (right). For the right plot in Fig. 9, we can see that *Freight Employment* and *Vehicle Transport* represent 20%–40% of the contributions, together with *Car Registrations*, *Truck Manufacturing*, *Energy Prices* and *Utility Prices* contribute on average from 70 to 80%.

Table 4
Number of leading indicators in lasso and lightGBM.

Horizon	Lasso	LightGBM	Common
1	23.2	226.5	2.0
2	31.4	219.7	3.6
3	32.0	224.4	3.4
4	30.7	220.1	3.4
5	36.5	214.5	4.2
6	33.5	214.8	5.8
7	34.5	204.2	5.5
8	35.4	194.2	7.0
9	36.8	185.8	7.6
10	37.3	168.5	9.6
11	33.2	155.4	9.1
12	30.3	138.4	12.8

In conclusion, both approaches have some complementarity. We argue that the judgemental approach provides more intuitive groups, that also exhibit a more natural evolution over time. However, this is a time consuming process, that sBERT can automate. The latter provides narrower groups with more volatile contributions over time. Ultimately, it is a question of resources, with a preference towards the judgemental approach that can incorporate contextual expertise.

Finally, this exploration can support the qualitative validation of the models, where the inclusion of specific leading indicators can be challenged.

We use SHAP values [41] to elucidate how the different inputs are used by the lightGBM. These track the contribution of each input to the output of the method. We find that lightGBM uses a far larger number of inputs compared to the lasso. Table 4 presents the average number of indicators used, across forecast origins, by the two methods, and the number of indicators present in both (we use the judgemental clustering of lasso indicators). LightGBM uses at minimum more than four times and at maximum almost ten times more the number of indicators that lasso does. Furthermore, there is little overlap in the number of commonly selected indicators.

This illustrates a fundamental difference between the two modelling approaches. Lasso provides a sparse selection that makes it easier for managers to both validate and gain new insights into market drivers. The number of chosen indicators for the lightGBM can be overwhelming, substantially reducing the ability of users to correctly account for how all the different leading indicators contribute to the predictions.

This can lead to simplified mental models that can be erroneous and lead to misinterpreting the lightGBM modelling of the business-relevant indicators. Additionally, the lasso prescribes linear contributions, while the exact nature of how each indicator affects the output of the lightGBM is challenging to disentangle due to the boosting and ensembling used. The performance metrics indicated a small advantage of lasso over lightGBM, which together with the interpretability of the former, make a compelling case in its favour for our case.

6. Discussion

We found it is more beneficial to introduce leading indicators at the top level of the hierarchy, rather than at the disaggregate level where the inventory decisions are taken. To better understand why this is the case, and the applicability of this finding more widely to other cases, we focus on the aggregation implications for the time series due to the hierarchical treatment. We consider unknown demand-generating processes, where each time series y_i has a conditional mean μ_{it} and a conditional variance σ_{it}^2, conditioned on all information up to period t. Naturally, we do not expect these to be constant over time, reflecting the existence of various time series patterns. For each time series we can devise the signal-to-noise ratio as μ_{it}/σ_{it}. Since regressions model the conditional expectation of a target variable, the higher the signal-to-noise ratio of a variable, the more information a regression can capture. Conversely, higher σ_{it} corresponds to higher standard errors for the coefficients of the regression that makes it harder to elucidate the effect of explanatory variables. This case would make modelling leading indicators challenging.

Aggregating across all m series at the bottom level of the hierarchy, the conditional mean of the aggregate top level series y_{Tt} is

$$\mu_{Tt} = \sum_{i=1}^{m} \mu_{it},$$

and its variance is

$$\sigma_{Tt}^2 = \sum_{i=1}^{m} \sigma_{it}^2 + 2 \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \mathrm{Cov}\,(y_{it}, y_{jt}) = \sum_{i=1}^{m} \sigma_{it}^2 + 2 \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \rho_{ij}\sigma_{it}\sigma_{jt}, \tag{7}$$

with covariances expressed as the product of the correlation between two series and their standard deviations. As long as the relative increase of μ_{Tt} is larger than that of σ_{Tt} the signal-to-noise ratio of the aggregate series will increase, which is beneficial for regression modelling. From (7) we can see that this depends on the correlation of the bottom-level series. When $\rho_{ij} = 0$ the increase in variance is only due to $\sum_{i=1}^{m} \sigma_{it}^2$. The intuition is that for independent demand series μ_{Tt} will increase faster than σ_{Tt}. To see why this is the case, if there are only two series at the bottom-level then $\sigma_{Tt}^2 = \sigma_{1t}^2 + \sigma_{2t}^2$ which prescribes a triangle with sides σ_{1t}, σ_{2t}, and hypotenuse σ_{Tt} whereby the triangular inequality $\sigma_{1t} + \sigma_{2t} \geq \sigma_{Tt}$. This is known to hold for the m-dimensional case. Therefore, when we contrast this to μ_{Tt} we can see that when the bottom-level series are independent it is expected that the signal-to-noise ratio of y_{Tt} will be favourable since the increase in μ_{Tt} will be larger. Negative correlations strengthen this further as they reduce σ_{Tt}^2, while strong positive correlations between time series are detrimental, with the sizes of σ_{it}^2 and σ_{jt}^2 becoming relevant in finding how large ρ_{ij} can be for $\sum_{i=1}^{m} \sigma_{it}^2 \geq \sigma_{Tt}^2$ to still hold.

In general, at the bottom level of the hierarchy we anticipate observing diverse correlations between products, with all correlations being strongly positive being unlikely. This is understood intuitively as the noise of the bottom-level time series being partially cancelled out at the more aggregate levels, and suggests that it is highly likely that our proposed approach of modelling the effect of leading indicators at the top level of the hierarchy and transferring their effect to the disaggregated ones through the hierarchical mechanism will be beneficial. Fig. 10 plots the correlations between the bottom-level time series for our case company, where it is evident that many series have low or

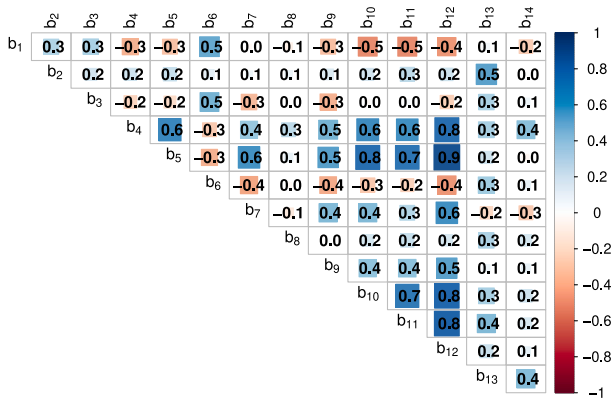

Fig. 10. Correlations between bottom level series (see Fig. 3). The squares (size and colour) and the values within represent correlations.

negative correlations, in agreement with the observed ability of lasso regression to model the effect of the leading indicators at the top level of the hierarchy. Exploring the correlation structure of the bottom-level demand series can help guide analysts to decide whether to adopt the proposed approach.

7. Managerial implications

Our collaborating firm operates globally, facing multiple macroeconomic environments as well as being susceptible to global disruptions. Leading indicators have the potential to give early signals for management to act upon. Although we are not the first to make this observation [e.g.,8], previous work provided this modelling on a strategic and tactical level, leaving it up to management to identify how to best act upon this information. Instead, our focus has been on providing a tool to support the operations of the manufacturer. We demonstrate that the proposed approach can provide automatic and objective modifications of the otherwise univariate forecasts that support inventory decisions. This lifts the burden of management to judgmentally intervene to account for novel tactical and strategic information due to leading indicators. Such adjustments are both time-consuming and potentially biased [11,13], our approach frees up management from this task, allowing them to focus on alternative value-adding activities that humans perform better than statistical models, such as the incorporation of contextual information not available to models, for instance, due to its unstructured nature. There is evidence that humans having to deal with multiple aspects of adjusting forecasts can lead to information overload with negative effects on performance [14].

In terms of implementation, we rely on publicly available leading indicators, sourced from statistics bureaus. This limits data costs, although it does not preclude sourcing data from proprietary databases. Furthermore, our approach does not require a complete overhaul of the forecasting process. The SKU-level forecasts can continue operating as is, with the leading indicator model connecting to those through hierarchical forecasting. Hierarchical forecasting can be seen as a follow-up step after the generation of the base forecasts. However, in the case of our collaborating firm, the SKU-level forecasts were found to be unnecessarily constrained to a specific type of exponential smoothing, and the use of other types of exponential smoothing is advised. We investigate the use of both statistical and machine learning approaches, namely lasso regression and lightGBM. Although performance-wise LEE and GEE are comparable both in terms of forecast accuracy and inventory performance, lasso is much simpler and more transparent. This transparency can aid in having trustworthy forecasts, which themselves are becoming of increasing importance for firms [21,34]. On the one hand, being able to see which leading indicators are selected by the model allows managers to validate the knowledge captured by the

model, and on the other hand, helps them to better understand the intricacies of the markets they operate in. From an implementation point of view, there are further forecasting process design and software considerations, beyond our case firm, that we expand on in the conclusions below.

Ultimately, this work shows that leading indicators are important for operations, beyond tactical and strategic predictions and decision making. This paper provides the toolset to achieve this, reusing existing operational forecasting infrastructure. By showing how leading indicators enhance the otherwise univariate forecasts, managers at various functions of the organisation can have a more aligned view of key drivers in their market.

8. Conclusions

Inventory management is typically supported by univariate forecasts, due to the inherent noise of SKU-level demand data, as well as the scale of the forecasting challenge that can impose computational restrictions in a practical setting [4]. On the other hand, such models cannot anticipate substantial changes in demand patterns. One can use leading indicators to model these changes. However, these models are difficult to deploy on SKU-demand data. This research proposes a methodology that can (i) automatically identify predictively useful leading indicators at a macro-level and (ii) relay this information to SKU-level demand forecasts. In doing so, we address the various technical challenges that arise. We find that the inclusion of the leading indicator information is more critical than the forecasting method used to achieve that.

The proposed methodology translates macro and global market information to the level of operational decision-making at SKU demand. The different elements in our methodology are investigated separately to assess the source of improvement in forecasting and inventory. The gains cannot be attributed solely to any single of the following elements: leading indicators, lasso methodology, lightGBM, or hierarchical modelling. Even though each element has a positive effect, we provide evidence that the maximal improvements come from the proposed methodology of joining these elements. We provide a theoretical explanation of why this is the case and guidelines of when to anticipate the reported benefits.

We deploy this to aid the inventory management of a global manufacturing company. We demonstrate the benefits for lead times relevant to their operations, both with and without lost sales, reflecting different supply arrangements with their customers. We show that hierarchical models with leading indicators can incorporate market externalities and they gain increasing importance for longer horizons. Our methodology can provide management with insights on the most important leading effects, and how they differ for shorter or longer lead times, facilitating a better understanding of the forces that affect the market they operate in.

Furthermore, the proposed approach is relatively easy to implement, as the more expensive model with the leading indicators is necessary only at the most aggregate level for our case company. The rest of the forecasts can remain exponential smoothing based.

The selection of the leading indicators was largely automated, which simplifies the implementation of the proposed methodology to other firms. The process involved an initial step where experts from the company selected from large categories of indicators, from which the model made the final selection. The incorporation of human expertise was done in an ad-hoc fashion, but indicates that there may be additional benefits in identifying how to best support experts in refining their choices. This is operationally relevant as the model remains in use over long periods, where periodic calibration will be necessary. Future research on refining the incorporation of human expertise, and optimising the recalibration of the model, from a process perspective, can further enhance the usefulness of the proposed methodology for practice.

Our work has a number of limitations that invite future investigation. Although we are favourable to the use of lasso due to its transparency, we highlight that increasing the number of potential leading indicators can become challenging for the model and computationally demanding. We evidence that alternative methods, such as lightGBM, are viable, and therefore future research should explore additional modelling alternatives. This also relates to the inclusion of additional classes of leading indicators that may be relevant in other application contexts. Furthermore, here we rely on the organisational hierarchy to guide the reconciliation process. Building on the argument that aggregation of series can be beneficial for the identification of relevant indicators, future research can investigate data-driven groupings of series to maximise this effect. We evidence the benefits of the methodology on fill rates. However, our simulation does not fully reflect the various ordering and supply chain costs that organisations face. Overcoming this, together with additional empirical evidence are welcome extensions of this work. Finally, in this study we do not consider the effect of revisions of the macroeconomic variables. Arguably, if the models are recalibrated at every forecasting cycle this is not an issue, however, it may make the selection of the leading indicators volatile, reducing the trustworthiness of the solution.

Finally, in this research, we demonstrate the benefits of relying on cross-sectional hierarchies, i.e., aggregating across products or other natural demarcations for a company. We do not consider temporal hierarchies, where data are aggregated over longer time intervals. Beyond any modelling implications, this can provide access to additional leading indicators that are available at lower sampling frequencies than the inventory review frequency. Future directions of research could explore the usefulness of cross-temporal hierarchies [49] that incorporate aspects of both cross-sectional and temporal hierarchies to expand the pool of leading indicators and also potentially improve the aggregate model accuracy and by extension the SKU-level inventory decisions.

CRediT authorship contribution statement

Yves R. Sagaert: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Nikolaos Kourentzes:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Conceptualization.

Declaration of competing interest

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

Acknowledgment

Nikolaos Kourentzes was funded for this research by Riksbankens Jubileumsfond, Sweden, Ref no. SAB22-0073.

Data availability

The data that has been used is confidential.

References

[1] Gardner Jr ES. Exponential smoothing: The state of the art—Part II. Int J Forecast 2006;22(4):637–66.
[2] Fildes R, Schaer O, Svetunkov I, Yusupova A. Survey: What's new in forecasting software? Oper Res Manag Sci Today 2020;47(4).
[3] Ord JK, Fildes R, Kourentzes N. Principles of business forecasting. 2nd ed. Wessex Press Publishing Co.; 2017.
[4] Seaman B. Considerations of a retail forecasting practitioner. Int J Forecast 2018;34(4):822–9.

[5] Fildes R, Ma S, Kolassa S. Retail forecasting: Research and practice. Int J Forecast 2022;38(4):1283–318.
[6] Makridakis S, Spiliotis E, Assimakopoulos V. M5 accuracy competition: Results, findings, and conclusions. Int J Forecast 2022;38(4):1346–64.
[7] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.
[8] Sagaert YR, Aghezzaf E-H, Kourentzes N, Desmet B. Tactical sales forecasting using a very large set of macroeconomic indicators. European J Oper Res 2018;264(2):558–69.
[9] Kourentzes N, Saayman A, Jean-Pierre P, Provenzano D, Sahli M, Seetaram N, et al. Visitor arrivals forecasts amid COVID-19: A perspective from the Africa team. Ann Tour Res 2021;88:103197.
[10] Fildes R, Goodwin P, Lawrence M, Nikolopoulos K. Effective forecasting and judgmental adjustments: an empirical evaluation and strategies for improvement in supply-chain planning. Int J Forecast 2009;25(1):3–23.
[11] Perera HN, Hurley J, Fahimnia B, Reisi M. The human factor in supply chain forecasting: A systematic review. European J Oper Res 2019;274(2):574–600.
[12] Trapero JR, Pedregal DJ, Fildes R, Kourentzes N. Analysis of judgmental adjustments in the presence of promotions. Int J Forecast 2013;29(2):234–43.
[13] Fildes R, Goodwin P, Önkal D. Use and misuse of information in supply chain forecasting of promotion effects. Int J Forecast 2019;35(1):144–56.
[14] Sroginis A, Fildes R, Kourentzes N. Use of contextual and model-based information in adjusting promotional forecasts. European J Oper Res 2023;307(3):1177–91.
[15] Schaer O, Kourentzes N, Fildes R. Demand forecasting with user-generated online information. Int J Forecast 2019;35(1):197–212.
[16] Sagaert YR, Aghezzaf E-H, Kourentzes N, Desmet B. Temporal big data for tactical sales forecasting in the tire industry. Interfaces 2018;48(2):121–9.
[17] Panagiotelis A, Gamakumara P, Athanasopoulos G, Hyndman RJ. Probabilistic forecast reconciliation: Properties, evaluation and score optimisation. European J Oper Res 2023;306(2):693–706.
[18] Athanasopoulos G, Hyndman RJ, Kourentzes N, Panagiotelis A. Forecast reconciliation: A review. Int J Forecast 2024;40(2):430–56.
[19] Kouvelis P, Chambers C, Wang H. Supply chain management research and production and operations management: Review, trends, and opportunities. Prod Oper Manage 2006;15(3):449–69.
[20] Aviv Y. A time-series framework for supply-chain inventory management. Oper Res 2003;51(2):210–27.
[21] Chuang HH-C, Chou Y-C, Oliva R. Cross-item learning for volatile demand forecasting: An intervention with predictive analytics. J Oper Manage 2021;67(7):828–52.
[22] Kourentzes N, Petropoulos F. Forecasting with multivariate temporal aggregation: The case of promotional modelling. Int J Prod Econ 2016;181:145–53.
[23] Baardman L, Boroujeni SB, Cohen-Hillel T, Panchamgam K, Perakis G. Detecting customer trends for optimal promotion targeting. Manuf Serv Oper Manag 2023;25(2):448–67.
[24] Ma S, Fildes R, Huang T. Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra-and inter-category promotional information. European J Oper Res 2016;249(1):245–57.
[25] Fu Y, Fisher M. The value of social media data in fashion forecasting. Manuf Serv Oper Manag 2023;25(3):1136–54.
[26] Bertsimas D, Kallus N, Hussain A. Inventory management in the era of big data. Prod Oper Manage 2016;25(12):2006–9.
[27] Huang T, Van Mieghem JA. Clickstream data and inventory management: Model and empirical analysis. Prod Oper Manage 2014;23(3):333–47.
[28] Kourentzes N, Sagaert YR. Incorporating leading indicators into sales forecasts. Foresight: Int J Appl Forecast 2018;(48).
[29] Sagaert YR, Kourentzes N, De Vuyst S, Aghezzaf E-H, Desmet B. Incorporating macroeconomic leading indicators in tactical capacity planning. Int J Prod Econ 2019;209:12–9.
[30] Stock JH, Watson MW. Forecasting using principal components from a large number of predictors. J Amer Statist Assoc 2002;97(460):1167–79.
[31] Stock JH, Watson MW. Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In: Handbook of macroeconomics, vol. 2, Elsevier; 2016, p. 415–525.
[32] Trapero JR, Kourentzes N, Fildes R. On the identification of sales forecasting models in the presence of promotions. J Oper Res Soc 2015;66(2):299–307.
[33] Ramos P, Oliveira JM, Kourentzes N, Fildes R. Forecasting seasonal sales with many drivers: Shrinkage or dimensionality reduction? Appl Syst Innov 2022;6(1):3.
[34] Spavound S, Kourentzes N. Making forecasts more trustworthy. Foresight: Int J Appl Forecast 2022;(66):21–5.
[35] Bernanke BS, Boivin J, Eliasz P. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. Q J Econ 2005;120(1):387–422.
[36] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. CRC Press; 2015.
[37] Friedman J, Tibshirani R, Hastie T. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1–22. http://dx.doi.org/10.18637/jss.v033.i01.
[38] Stock JH, Watson MW. Generalized shrinkage methods for forecasting using many predictors. J Bus Econom Statist 2012;30(4):481–93.
[39] Li J, Chen W. Forecasting macroeconomic time series: LASSO-based approaches and their forecast combinations with dynamic factor models. Int J Forecast 2014;30(4):996–1015.
[40] Roth Tran B. Sellin'in the rain: Weather, climate, and retail sales. Manag Sci 2023;69(12):7423–47.
[41] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
[42] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115.
[43] Verdinelli I, Wasserman L. Decorrelated variable importance. J Mach Learn Res 2024;25(7):1–27.
[44] Babai MZ, Boylan JE, Rostami-Tabar B. Demand forecasting in supply chains: a review of aggregation and hierarchical approaches. Int J Prod Res 2022;60(1):324–48.
[45] Fliedner G. Hierarchical forecasting: issues and use guidelines. Ind Manag Data Syst 2001;101(1):5–12.
[46] Wickramasuriya SL, Athanasopoulos G, Hyndman RJ. Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. J Amer Statist Assoc 2019;114(526):804–19.
[47] Kourentzes N, Athanasopoulos G. Elucidate structure in intermittent demand series. European J Oper Res 2021;288(1):141–52.
[48] Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice. 3rd ed. Melbourne, Australia: OTexts; 2021, URL OTexts.com/fpp3.
[49] Kourentzes N, Athanasopoulos G. Cross-temporal coherent forecasts for Australian tourism. Ann Tour Res 2019;75:393–409.
[50] Zhu X, Ninh A, Zhao H, Liu Z. Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry. Prod Oper Manage 2021;30(9):3231–52.
[51] Hyndman R, Koehler AB, Ord JK, Snyder RD. Forecasting with exponential smoothing: the state space approach. Springer Science & Business Media; 2008.
[52] Hyndman RJ, Koehler AB, Snyder RD, Grose S. A state space framework for automatic forecasting using exponential smoothing methods. Int J Forecast 2002;18(3):439–54.
[53] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29(5):1189–232.
[54] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol 1996;58(1):267–88.
[55] Ghysels E, Osborn DR, Sargent TJ. The econometric analysis of seasonal time series. Cambridge University Press; 2001.
[56] Kourentzes N. Toward a one-number forecast: cross-temporal hierarchies. Foresight: Int J Appl Forecast 2022;67:32–8.
[57] Pritularga KF, Svetunkov I, Kourentzes N. Stochastic coherency in forecast reconciliation. Int J Prod Econ 2021;240:108221.
[58] Athanasopoulos G, Hyndman RJ, Kourentzes N, Petropoulos F. Forecasting with temporal hierarchies. European J Oper Res 2017;262(1):60–74.
[59] Claeskens G, Magnus JR, Vasnev AL, Wang W. The forecast combination puzzle: A simple theoretical explanation. Int J Forecast 2016;32(3):754–62.
[60] Svetunkov I. smooth: Forecasting using state space models. 2023, R package version 3.2.0. URL https://CRAN.R-project.org/package=smooth.
[61] R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023, URL https://www.R-project.org/.
[62] Shi Y, Ke G, Soukhavong D, Lamb J, Meng Q, Finley T, et al. Lightgbm: Light gradient boosting machine. 2024, R package version 4.3.0. URL https://CRAN.R-project.org/package=lightgbm.
[63] Silver EA, Pyke DF, Thomas DJ. Inventory and production management in supply chains. CRC Press; 2016.
[64] Athanasopoulos G, Kourentzes N. On the evaluation of hierarchical forecasts. Int J Forecast 2023;39(4):1502–11.
[65] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Amer Statist Assoc 2007;102(477):359–78.
[66] Saoud P, Kourentzes N, Boylan JE. Approximations for the lead time variance: A forecasting and inventory evaluation. Omega 2022;110:102614.
[67] Kourentzes N, Trapero JR, Barrow DK. Optimising forecasting models for inventory planning. Int J Prod Econ 2020;225:107597.
[68] Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 conference on empirical methods in natural language processing. Association for Computational Linguistics; 2019, URL https://arxiv.org/abs/1908.10084.