INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
https://doi.org/10.1080/00207543.2025.2607640

Taylor & Francis
Taylor & Francis Group

Enhancing lasso tactical demand forecasts using semantic information

Yves R. Sagaert [a] and Nikolaos Kourentzes [b]

[a]Department of Computer Science, VIVES University of Applied Sciences, Kortrijk, Belgium; [b]Skövde Artificial Intelligence Lab, School of Informatics, University of Skövde, Stockholm, Sweden

ABSTRACT

Leading indicators have been shown to be useful in predicting demand. Nowadays, a large number of potentially interesting variables are available in open databases. This makes it challenging to select the predictively relevant indicators. Lasso regression has become a popular choice when handling many variables, however, its selection is sensitive to changes in the sample and in the presence of correlated variables. This can harm both its predictive performance and the trustworthiness of the forecasts. We consider various approaches to aid variable selection. Specifically, we consider clustering via various statistical based methods, semantic information using a Semantic Bidirectional Encoder Representations from Transformers, or meta-data, such as a popularity index of variables. The resulting groups of variables are evaluated for selection directly, using sequential lasso, or transformed into cluster profiles, or factors using principal component analysis. Using an empirical case, we evaluate the alternative options on their decision and predictive performance, and on the interpretability of the resulting models. Our analysis indicates that there are trade-offs between predictive and decision performance, and interpretability. Semantic information is found to benefit variable selection, although we do not identify a dominant approach. Finally, we emphasise the need for methodological advances for explainability in forecasting.

Highlights

- We compare different groupings of variables to enhance selection with lasso.
- Groupings are based on statistical, semantic, and meta-data information.
- Additional dimensionality reduction by cluster profiles and factors is investigated.
- Semantic information leads to gains in forecast and decision metrics.
- Credibly explainable models should match or outperform non-explainable benchmarks.

ARTICLE HISTORY
Received 31 March 2025
Accepted 10 December 2025

KEYWORDS
Forecasting; variable selection; leading indicators; semantic clustering; explainability; inventory management

1. Introduction

External variables have been successfully used to increase the predictive accuracy of demand forecasts, while uncovering important influences on demand. With numerous external and internal databases of potentially useful predictors, the number of variables can often exceed the number of available observations. Lasso regression has been widely used in these cases, where, apart from its ability to choose predictively useful variables and estimate their contribution, it has provided good accuracy and arguably explainable results. Nonetheless, lasso has well-understood limitations (for a detailed discussion, see Freijeiro-González, Febrero-Bande, and González-Manteiga 2022; Hastie, Tibshirani, and Wainwright 2015). When variables are highly correlated or their number far exceeds the available fitting sample size, the quality of the selection degrades. Both of these limitations are very relevant for practice, where when modelling aggregate demand, both the impact of external variables becomes more pronounced, as well as the number of potential options (Sagaert and Kourentzes 2025).

Our objective is to mitigate these limitations by organising the input variables into semantically cohesive groups. This builds on previous work in the literature, which on the one hand, has evidenced the good performance of lasso demand forecasts for tactical and strategic planning (e.g. Sagaert et al. 2018a, 2019), and from the other hand, has shown that grouping of variables using domain expertise or prior knowledge can improve variable selection and forecast accuracy (e.g. Ma, Fildes, and Huang 2016; Sagaert et al. 2018a; Zou and Hastie 2005).

CONTACT Yves R. Sagaert yves.sagaert@vives.be Department of Computer Science, VIVES University of Applied Sciences, Doorniksesteenweg 145, 8500 Kortrijk, Belgium

In practice, the challenge is that this prior knowledge is often unavailable, or when available, it is difficult to standardise, scale, and replicate, often hinging on particularly skilled individuals (Katsagounos et al. 2021). This motivates us to propose using semantic grouping of variables to improve selection. Moreover, there is evidence that experts have difficulty distinguishing between hype and predictively useful information, due to various judgmental biases (Fildes, Goodwin, and Önkal 2019), a problem that is exacerbated as the number of variables increases, due to cognitive overload (Sroginis, Fildes, and Kourentzes 2023). The proposed methods for systematic semantic grouping can eliminate such judgmental biases.

Using a large number of potential leading indicators we build forecasts for tactical planning for a manufacturer. Using the same variables, we construct semantic and statistical groupings, using information contained in the variables, their meta data, or externally, such as the popularity of variables by analysts. These groups are modelled directly with sequential lasso (Ma, Fildes, and Huang 2016), or after dimensionality reduction by either constructing group profiles or factors (principal components). The latter approaches trade explainability for a simpler variable selection problem, with past evidence of good performance (Ramos et al. 2022; Stock and Watson 2002). The alternative methods are evaluated on forecast and decision metrics, and in terms of explainability. We are interested on explainability from two perspectives, providing insights to stakeholders, and increase the trustworthiness of forecasts.

We find that incorporating semantic information in the variable selection process can lead to decision benefits, and increased trustworthiness and insights. However, our analysis highlights multiple methodological challenges that has limited the focus on explainability and trustworthiness of forecasts so far. The rest of the paper is organised as follows. Section 1 provides a brief overview of the literature in terms of relevant predictive information for demand forecasting, and modifications of lasso. Section 2 presents the investigated methods to enhance variable selection. Section 3 introduces empirical case and the evaluation design, followed by the discussion of the results in Section 4. We conclude with a discussion of methodological challenges that are highlighted in our work.

2. Literature

2.1. Types of external information

The literature has identified multiple sources of information that are predictively valuable for demand forecasting. The most apparent is information related to specific products or customers (Baardman et al. 2023; Fildes, Ma, and Kolassa 2022). For example, price and promotion are relevant in a retail context (e.g. Ma, Fildes, and Huang 2016; Mitra, Saha, and Kumar Tiwari 2024; Ramos et al. 2022). This information is internal and is typically relevant for short-term demand forecasts to support operations.

Customer characteristics (Hooshmand Pakdel, He, and Chen 2025), online search, and social media data can be used to elucidate demand signals, with applications, for instance, in the gaming sector (Schaer, Kourentzes, and Fildes 2022) and fashion forecasting (Fu and Fisher 2023). Schaer, Kourentzes, and Fildes (2019) provides a detailed review of the use of various sources of online customer information and raises two considerations. First, studies that find this information useful can often omit other established control variables, raising questions about whether this information is complementary or replacement. Second, although this information may be available and highly correlated to the demand, it may not be available at the appropriate decision horizons. This problem becomes even more pronounced when focusing on tactical or strategic horizons.

Unstructured data has also been a popular input for demand forecasting. This has often taken the form of contextual information, often incorporated through the use of judgment (Fildes, Goodwin, and Onkal 2021; Sroginis, Fildes, and Kourentzes 2023). A limitation of judgment is that it is not scalable, and can often be biased. Other approaches have included judgmental bootstrapping (Armstrong 2001), where statistical models are used to infer experts' rules, or directly refine expert adjustments (Trapero et al. 2013). More recently, language models have been used to systematise and automate the incorporation of contextual information (Abolghasemi, Ganbold, and Rotaru 2025; He et al. 2024). This source of information is typically focused for short-term operational forecasts.

There is limited work that focuses on tactical and strategic demand forecasts. A line of research has explored the use of experts to formulate demand scenarios (Athanasopoulos et al. 2023; Önkal, Gönül, and Goodwin 2023) to inform decision makers. Such approaches are typically costly and time-consuming. Sagaert et al. (2019) and Sagaert et al. (2018b) have looked at the usefulness of incorporating leading indicators from the macro-economic environment to construct tactical demand forecasts, recognising that the key difficulty is in the appropriate identification of the predictively useful variables, given the very large number of alternatives. They show that lasso regression can outperform various competing approaches.

2.2. Variable selection using lasso regression

Lasso regression has been widely adopted in the literature (e.g. Ma, Fildes, and Huang 2016; Ramos et al. 2022). Nonetheless, it suffers from well known limitations when variables are correlated, or when there is a substantial difference between the sample size and the number of variables to be evaluated. Freijeiro-González, Febrero-Bande, and González-Manteiga (2022) provides a detailed review of the issues, and Su, Bogdan, and Candes (2017) investigate in detail the conditions when lasso regression can pick variables erroneously. This has lead to various modifications of lasso regression. For instance, Zou and Hastie (2005) propose the elastic nets that uses both absolute and quadratic regularisation terms, balancing the pros and cons of ridge and lasso regression. Yuan and Lin (2006) introduce the grouped lasso that can select groups of variables in a single step, which is useful for demand models that include multiple binary indicators. Ma, Fildes, and Huang (2016) propose a multistage lasso (hereafter referred to as sequential lasso) that tackles groups of variables sequentially, to mitigate the spurious inclusion and exclusion of variables.

Of particular interest for demand forecasting is the use of factor models and lasso regression. Factor models resolve the inclusion of a very large number of variables by transforming them with Principal Components Analysis (PCA) and reducing the dimensionality of the problem by eliminating components that explain very limited variance (Stock and Watson 2002). Moreover, the principal components are orthogonal, further simplifying the selection process. They have been used with some success in various demand forecasting tasks (e.g. Trapero, Kourentzes, and Fildes 2015).

Lasso and PCA have been combined in various ways. Gao and Tsay (2025), Tu and Gao (2025), Ma, Fildes, and Huang (2016), and Ramos et al. (2022) are examples where using lasso to select the relevant principal components can improve forecast accuracy. Zhang, Wahab, and Wang (2023) combines the approaches in the opposite way by conducting a PCA on the lasso selection, while Zou and Xue (2018) embeds regularisation in the PCA procedure, resulting in sparse principal components, although not in a forecasting context.

Boivin and Ng (2006) argues that the PCA deteriorates when the number of variables becomes very high. Although Uniejewski and Maciejowska (2023) found that the combination of PCA and lasso outperforms both individual methods, an apparent remedy is the combination of sequential lasso with PCA, which is absent in the literature. The sequential selection limits the number of variables that the PCA is applied on, and thus can overcome this issue. This motivates the proposed methods for enhancing the lasso variable selection in the next section.

3. Variable selection methods

3.1. Least absolute shrinkage and selection operator

Lasso can simultaneously select and estimate the relevant coefficients. Moreover, as it is a shrinkage estimator it mitigates the effects of sampling uncertainty. It achieves these through its shrinkage loss (Hastie, Tibshirani, and Wainwright 2015):

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

where $\boldsymbol{y}$ is the target, $\boldsymbol{X}$ is a matrix of covariates, and $\boldsymbol{\beta}$ is a column vector of coefficients. The hyper-parameter λ calibrates the strength of the shrinkage, i.e. how strongly the coefficients are pushed to zero and minimising $\|\boldsymbol{\beta}\|_1$. If the latter is zero, by setting $\lambda = 0$, the lasso estimator becomes the standard ordinary least squares. For large values of λ the penalty term becomes dominant, pushing most variables out of the model. Typically the value of λ is identified by cross-validation.

The variables in $\boldsymbol{X}$ are augmented by including lagged realisations, autoregressions of the target variable, and binary seasonal dummies to facilitate modelling any seasonality in the time series. We lag the external variables for two reasons. First, we are interested in modelling leading effects, i.e. an early signal from an indicator that can explain the future movement of the demand. Second, to use external variables in a forecasting context, their values in the future periods are needed. For forecasting one period ahead, all lagged variables are available. For forecasting two periods in the future, all 1-period lagged variables will be excluded, as we will not know their future value for 2-periods ahead. As the forecast horizon increases, it restricts the available lags further. This has two consequences: a different model needs to be built for each forecast horizon, and the number of available lagged variables in $\boldsymbol{X}$ decreases for longer forecast horizons. This can eventually exclude leading variables with short leads from the regressions. We use up to 12 lags for the leading indicators (matching the maximum forecast horizon, see Section 4).

3.1.1. Sequential lasso

The correlation of variables in $\boldsymbol{X}$ and with the target $\boldsymbol{y}$ defines which variables remain in the model. This can make spuriously correlated variables to be preferred over more appropriate ones. To introduce more control in the selection procedure, Ma, Fildes, and Huang (2016)

Table 1. Methods for constructing variable groups.

Grouping information	Dimensionality reduction		
	None	Profiles	Factors
None	LASSO	–	pcaLASSO
Pattern similarity	STAT	profSTAT	pcaSTAT
Popularity with analysts	POP	profPOP	pcaPOP
Statistical agency grouping	HIER	profHIER	pcaHIER
Long title of variables	SBERT	profSBERT	pcaSBERT

perform variable selection with lasso sequentially, using groups of variables. Their procedure is as follows. First, a lasso solution is found using the first group of variables. Its residuals are calculated, and a lasso regression is estimated on these residuals using the next group of variables. The process is repeated until there are no more groups left. A drawback is that as the number of groups increases, the computational cost of sequential lasso, will increase, as the λ hyper-parameter is identified at each stage independently.

3.2. Construction of sequential lasso variable groups

An overview of the different proposed methods for variable grouping is provided in Table 1. It summarises the type of information used for the grouping, expanded below, and the type of dimensionality reduction, discussed in Section 3.4. LASSO refers to the direct selection of variables using standard lasso regression, which acts as a benchmark.

3.2.1. Statistical clustering

Without any application context, the simplest way to group variables together is by statistically clustering them. Groups are formed by similarity in shape. Variables are normalised before being used in a lasso regression by removing their mean and dividing them by their standard deviation. Therefore, their scale should not be considered in their clustering. Instead, we use a correlation distance.

The clustering is performed using k-means, where the analyst must select the number of clusters. We trialled two approaches, first, judgmentally set to a value that resulted in medium-sized groups of variables (15 groups), and second, by optimising cluster validation metrics. We considered the Krzanowski-Lai, Calinski-Harabasz, Hartigan, and C indices, as implemented in the *Nbclust* package in R Charrad et al. (2014). The first two indices evaluate the separation between clusters, while the last two focus on cluster cohesion. Optimising the indices resulted in two groups of variables and was subsequently shown to not perform well, so it was excluded from any further analysis. We refer to this method as *STAT*.

3.2.2. Clustering using the popularity index of variables

The leading indicators in our investigation are sourced from the FRED database. The database provides as a meta-data the popularity of a variable. This is stored as an integer, based on which we can group variables according to their popularity, with each group representing a single popularity value. We refer to this method as *POP*.

The use of popularity is premised in that analysts consider a variable to be very informative, however, they have diverse objectives and uses for the variables. Therefore, on the one hand, we obtain a filtering of the variables by human experts, but on the other hand, we accept the risk that this filtering may be without forecasting applications in mind. Therefore, very popular indices are understood in a wisdom-of-a-crowd way to signify that they provide strong signals.

Some popular variables are bound to be relevant to almost any modelling of economic activity, such as inflation or gross domestic product. It is also unlikely that very two similar variables will be equally popular. This provides a reasonable separation of variables into groups, where the sequential lasso can avoid missing important variables due to correlations.

Note that the FRED database by default sorts variables by popularity. Therefore, it is expected that the popularity of the most popular variables is reinforced just by their presentation. This can introduce herding effects, biasing the information in the popularity index.

3.2.3. Clustering on data hierarchies

Another useful meta-data is the organisation of variables in different groups by the statistical agency, for our case in the FRED database, into a hierarchy. Variables are organised into groups logically, according to some economic intuition. We take these groups to organise the variables for sequential lasso, and refer to this method as *HIER*.

This approach has the disadvantage that variables that are logically together can often describe similar attributes and therefore be highly correlated. This can have negative effects for lasso. To mitigate this, we opt to use the lowest grouping level, which comes at the cost of having many groups to model that inflates the computational cost.

3.2.4. Semantic clustering

Sentence-BERT (Bidirectional Encoder Representations from Transformers) or SBERT, is a transformer-based model that can be used for semantic clustering. The model generates a low-dimensional numerical representation of words and sentences, including context and semantic meaning. These numerical representations are called embeddings, on which clustering can take place.

We use SBERT to encode the long title of the indicator into a fixed-size vector of embeddings, which are then clustered using a vector cosine-similarity equal to or larger than 65%. Clusters are only considered if they group at least 20 indicators. Variable selection results were found to be insensitive to small changes in these settings. For our investigation we used the pre-trained language model '*all-MiniLM-L6-v2*' and fine-tuned it on our domain-specific indicator data.

3.3. Availability of meta-data

The grouping information uses meta-data that is publicly available from the FRED database. Alternative data sources, publicly available, such as Eurostat, or commercial or bespoke sources, may not provide the same meta-data. POP that uses the popularity of a variable with analysts may not always be feasible. The meta-data required by HIER is generally available from statistical data providers, and if it is not from a bespoke source, it is likely that the necessary structure information can be sourced from statistics offices. SBERT requires a title or short description for the variables, while STAT does not require any meta-data and is universally applicable.

3.4. Dimensionality reduction

3.4.1. Cluster profiles

Each of the clustering methods above will provide multiple groups of variables, but the total number of variables remains the same. It may be beneficial to reduce the dimensionality of the problem to aid variable selection further. The first approach we consider is to construct for each cluster of variables its profile. This is done by first normalising the data (z-score) and calculating the cluster centroid. The whole cluster of variables is now replaced by the single cluster profile. Once all profiles are constructed, these become the inputs for lasso. There is no need to use the sequential lasso due to the aggressive reduction in the number of variables. This approach is more meaningful for STAT, which groups together similar profiles, than for other methods.

3.4.2. Factors

The other dimensionality reduction approach is based on PCA. We perform a PCA for each group of variables. The resulting components are linear combinations of the original variables, ensuring that the resulting components are orthogonal. This orthogonality facilitates variable selection.

With PCA the dimension of $\boldsymbol{X}$ remains intact. To reduce it, we order the components by the amount of variance of the original variables they explain and retain as many components as needed to explain about 90–95% of the variance within each group of variables. This ends up being the first three principal components, for all groups. Factor models have been popular in the literature (Stock and Watson 2002) and have been shown to perform well in combination with lasso regression (Ramos et al. 2022; Uniejewski and Maciejowska 2023).

For both profiles and factors, we first construct the set of reduced variables from the groups of the original variables and then construct the relevant lags.

3.5. Quantile estimation

As we discuss in Section 4, our focus is on predictions over the lead time. A complication in obtaining the cumulative variance and required quantile is that a different lasso model is built for each forecast horizon. This can substantially complicate the calculation of any covariances between forecast horizons, which may be present since the models are conditioned on the same data and have a similar structure. To overcome this, we follow the recommendation by Saoud, Kourentzes, and Boylan (2022). First, we obtain empirical cumulative error distributions, by calculating the differences between the total observed demand over the lead time and the respective forecasts (from the different models), which was shown to retain all covariance information. The empirical cumulative error distribution is smoothed using kernel density estimation (KDE), from which the desired statistics can be obtained. The use of KDE is found to be beneficial both for smoothing the empirical estimates, reducing the effects of sampling uncertainty, and for retaining any asymmetries in the error distribution (Trapero, Cardós, and Kourentzes 2019).

4. Case study

For the evaluation of the proposed methods, we use an empirical case from a manufacturing company with a global supply chain. The company is headquartered in Europe and produces raw materials needed in the automotive supply chain. Its market portfolio focuses on the US and Europe. The objective is to forecast aggregate demand for strategic and tactical planning, with horizons up to 12 months. We abstract this requirement in short, medium, and long horizons, of 3, 6, and 12 months.

Its production is specialised for materials for cars and trucks separately. Therefore, we are provided with demand data for Europe and the US for the two types, resulting in four series. Additionally, the company looks at country-level demand for operational decision making, ensuring the supply of their goods. In total, this provides 16 country demand series for cars and 12 for trucks.

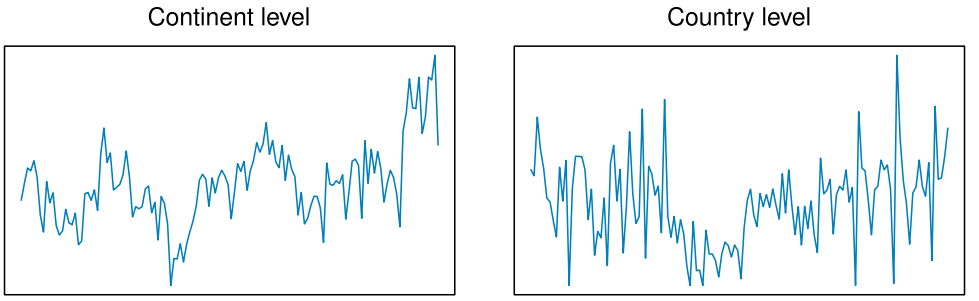


Figure 1. Anonymised examples of a series at the continent and country levels.

Our primary focus is supporting tactical and strategic planning, and therefore modelling the continent-level demand. We use the country-level demand series to provide additional empirical evidence. As the data is confidential, Figure 1 provides two anonymised examples.

Its multinational coverage and upstream position in the supply chain make the macro-economic environment in its markets relevant. Therefore, for leading indicators, we source 1011 different macroeconomic indicators from the FRED database covering the relevant countries. These indicators are lagged up to 12 times for building the lasso models for the different horizons.

The data is monthly and covers the period 2005–2015. The first 84 observations (up to 2011) are used for training, and the last 48 periods are used as a test set. A rolling origin evaluation scheme is used, where after forecasts are produced from an origin, the training set is expanded by one period, and the process is repeated until the test set is exhausted. This facilitates collecting more evidence of the performance of the models, which are updated at each forecast origin.

4.1. Benchmarks and methods

For benchmarks we use an automatically specified exponential smoothing model (ETS) and a lasso regression that directly selects from all possible indicators (LASSO). The type of ETS is identified using the Akaike Information Criterion, corrected for sample size (Hyndman et al. 2008). As ETS is a univariate model, it provides a limit of acceptable performance for any methods that use explanatory variables. The LASSO benchmark helps us gauge whether the inclusion of additional grouping information, or dimensionality reduction, improves performance. We experimented with Long Short-Term Memory (LSTM) neural networks that can model nonlinear and interaction effects, but they did not perform better than the benchmark LASSO. A similar result was found by Sagaert and Kourentzes (2025) when comparing lasso and lightGBM. Therefore, for brevity, together with our interest in explainability, these were excluded (a summary description of the LSTM and their accuracy are provided in Appendix). The investigated methods, STAT, POP, HIER, and SBERT, that pre-group the variables before they are processed by the sequential lasso, are split into three groups, depending on the dimensionality reduction employed, as summarised in Table 1. For the estimation of the quantiles, we rely on in-sample measurements, given the limited sample available.

4.2. Evaluation

We are evaluating the performance of the competing methods across three dimensions: (i) decision, as captured by inventory performance metrics, (ii) forecasting performance, looking at bias and accuracy metrics, and (iii) the explainability of the resulting regressions, which can be useful for validating the models and for providing business insights to stakeholders. As the underlying demand processes are unknown, like all empirical evaluations, any findings from our experiments are conditioned by forecasting methods evaluated, and limited by the sample available by the case company. Therefore, the multifaceted evaluation helps to pool evidence to support any findings. The KDE used in the quantile estimation employs a Gaussian kernel, with its bandwidth determined by the standard Silverman's heuristic.

4.2.1. Decision metrics

As it has been documented multiple times in the literature, forecast accuracy results typically do not translate linearly to gains in the quality of decisions (for a discussion see, Athanasopoulos and Kourentzes 2023; Goltsos et al. 2022). Therefore, when possible, it is preferable to evaluate the performance of competing forecasts in a real or simulated decision environment (e.g. Doganis, Aggelogiannaki, and Sarimveis 2008; Li et al. 2023; Wang and Petropoulos 2016). Demand forecasts are used to manage inventory and set production plans (Ghasemi, Lehoux, and Rönnqvist 2023). We assume flexibility in production and therefore focus on providing forecasts that can result in the desirable fill rates.

We simulate an order-up-to inventory policy with lost sales, as in Silver, Pyke, and Thomas (2016), for target fill rates ranging from 90% to 99%. We track performance across a range of lead times (3, 6, and 12 months), with the policy being reviewed monthly. To avoid biasing the simulation on our initial values, we use a burn-in period of 84 periods. We validated that this was sufficient by ensuring that the simulation had reached a steady behaviour. The test periods start once the burn-in sample is exhausted.

We track the remaining stock-on-hand at the end of the simulation, which is scaled by the mean of the observed demand in the test period. This is done to enable summarising results across time series. We also track the achieved fill rates (FR) in the test set,

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

where $d_i(\alpha)$ is the satisfied demand in period i for a given target α%, and y_i is the observed demand.

4.2.2. Forecasting metrics

Since the supported decision requires accurate demand forecasts over the lead time, we focus on cumulative errors over the lead time L,

$$\epsilon_{Lt} = \sum_{i=1}^{L} y_{t+i-1} - \sum_{i=1}^{L}\hat{y}_{t+i-1}.$$

Index t is used to identify the L periods and is not associated with the forecast origin, with all periods in ϵ_{Lt} being forecasted. As we need scale independent error metrics, we calculate a scaling factor

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

where r is the number of fitting periods, and $p = 2$ for quadratic metrics and $p = 1$ otherwise. For the forecast bias we focus on its magnitude, and measure the Absolute Mean scaled Error,

$$\mathrm{AMsE}_L = \frac{1}{o}\left|\sum_{i=1}^{o}\frac{\epsilon_{Li}}{s^1}\right|.$$

This is done so as not to cancel out biases of opposite direction when we provide the summary metric across series. For the point forecast accuracy we report the Root Mean Squared scaled Error,

$$\mathrm{RMSsE}_L = \sqrt{\frac{1}{o}\sum_{i=1}^{o}\frac{(\epsilon_{Li})^2}{s^2}},$$

and for the quantile forecast accuracy we provide the 95th quantile scaled Pinball,

$$\mathrm{sPIN}_L = \frac{1}{o}\sum_{i=1}^{o}\frac{v_{\alpha i}}{s^1},$$

$$v_{\alpha i} = \max\left(0, \alpha\left(\sum_{j=1}^{L} y_{i+j-1} - \sum_{j=1}^{L}\hat{Q}_{i+j-1}\right)\right) + \max\left(0, (1-\alpha)\left(\sum_{j=1}^{L}\hat{Q}_{i+j-1} - \sum_{j=1}^{L} y_{i+j-1}\right)\right),$$

where $\hat{Q}_i$ is the corresponding α%-quantile forecast. The pinball is calculated on the cumulative errors, matching the other error metrics. Recall that the pinball is the proper score for the news-vendor case with backorders and service level targets (Gneiting and Raftery 2007). Accordingly, for our setup pinball should connect more strongly to the inventory objective than the other metrics, but it is not a proper score here, as we use fill rate and not service level.

4.3. Explainability and validation

Each of the investigated methods will result in a regression with a set of selected variables. Each regression suggests a narrative of what variables drive the observed demand. This can be used to, first, validate the regression where we can contrast the model with our understanding of the demand process, and, second, provide new insights that decision makers may have been unaware of. These two are premised on different, and somewhat strong, assumptions. The first assumes some knowledge of the underlying data generating process, while the second assumes that the estimated model is valid, i.e. that it appropriately abstracts the generating process being modelled. Therefore, empirically assessing a single model presents a circular argument: a model's validity is a prerequisite for its explainability to be meaningful, yet its explainability is often what we use to assess its validity.

This issue is particularly pronounced in business forecasting, where the underlying process is often unknown. Therefore, when validating an estimated model, there is rarely a theoretical model, complete in its important dimensions, to validate against (for discussions of 'white-box' validations see, Fildes and Kourentzes 2011;

Pidd 1997). Any model inference for knowledge discovery assumes that any estimation biases due to omitted variables are negligible (Montgomery, Peck, and Vining 2012, chapter 10), a statement that is equivalent to the fitted model largely matching the underlying data generating process. Moreover, the interpretation of the coefficients of linear models is premised on the corresponding features being independent. Correlations of varying degrees are bound to exist in real data, spurious or not. Moreover, correlations are reinforced by including lags of variables. Implementations of various explainability tools, such as Shapley values, also require a high degree of variable independence to be reliable, when applied to either linear or nonlinear methods (Aas, Jullum, and Løland 2021; Arrieta et al. 2020). Therefore, without a theoretical model to validate against, knowledge discovery on individual estimated models can very easily provide misleading insights. Although it is tempting to rely on domain expertise to validate a model's key features, we need to acknowledge that this knowledge may be false, biased, or constrained by a predominant methodology or epistemology (Fildes and Kourentzes 2011; Pidd 1997), Ultimately, we see multiple limitations in the post-hoc explainability analysis of a single model of unknown validity.

This does not invalidate the need and usefulness of explainability. For example, Ma, Fildes, and Huang (2016), using lasso regression, built promotional models with cross-product effects for retailing. They highlight a small number of captured cross-effects that are dubious, nonetheless, their model is the best-performing amongst benchmarks, with many of the causal connections replicating domain knowledge. Similarly, Fildes and Kourentzes (2011) find that decadal forecasts from a climate model consistently misestimate the long-term trend, yet provides very accurate predictions compared to alternatives. In both cases, the authors identify weaknesses of the models to be improved, while demonstrating that empirically they provide very competitive forecasts, rather than suggesting these models to be unusable.

Since explainability as a post-modelling exercise of a single model can be problematic, knowledge discovery from it can be misleading, and likewise, the validation of any single model can be challenging. Instead, we propose treating explainability as a relative concept, much like we do in model selection. For instance, when using information criteria, the selected model is not assumed to be the true data-generating process; rather, it is the most plausible approximation from the pool of models under consideration (Burnham and Anderson 2002). This is important as even the selected model can still be a poor fit to the data.

By juxtaposing the explainability of competing models, we can assess where their narratives align or diverge. This assessment cannot be done independently of model accuracy. Ideally, we are looking for strong alignment among the most accurate models, as this increases our confidence in the revealed interactions. In contrast, alignment between poorly performing models provides no support for their narratives.

Two additional considerations are important. First, if competing models process information in similar ways, they could produce similar interpretations purely due to the modelling process, much like experts might herd together in a similar narrative. Second, because explainability is a relative statement, the best obtained narrative does not necessarily reveal much about the underlying data-generating process, particularly if an appropriate model was not included in the initial pool. These points are further qualified by the quality and capabilities of the various explainability tools that are employed.

Practically, our recommendation is easy to operationalise. Once the performance of the competing models is ranked, investigate the similarities and differences that are revealed by the employed explainability tools. Models that are clearly dominated in terms of predictive performance will be understood as unable to produce valid predictions or useful insights. When a pool of top-performing models offer aligned narratives, their explanations are credible. Conversely, if those same top-performing models offer highly divergent narratives, based on different included variables, we conclude that their explanations are dubious. As with model selection, if one of these would be the true model, we would have no means of identifying it without some knowledge of the underlying process. Hence, we cannot make claims about the validity of these models in the strict sense. Similarly, if these models cannot substantially outperform extrapolative forecasts that use no external variables, we do not consider the resulting insights credible.

5. Results

We first establish a ranking of the performance of the alternative methods, and subsequently, assess their explainability. The ranking of the methods is based on the inventory outcome of the forecasts, which is discussed first, followed by a discussion of the bias, accuracy, and quantile accuracy of the forecasts. The results are discussed separately at the continent and country levels, as the efficacy of the various methods differs substantially between levels. Pinball results are reported for the $\alpha = 95\%$ quantile, with results for $\alpha = 90\%$ being qualitatively similar.

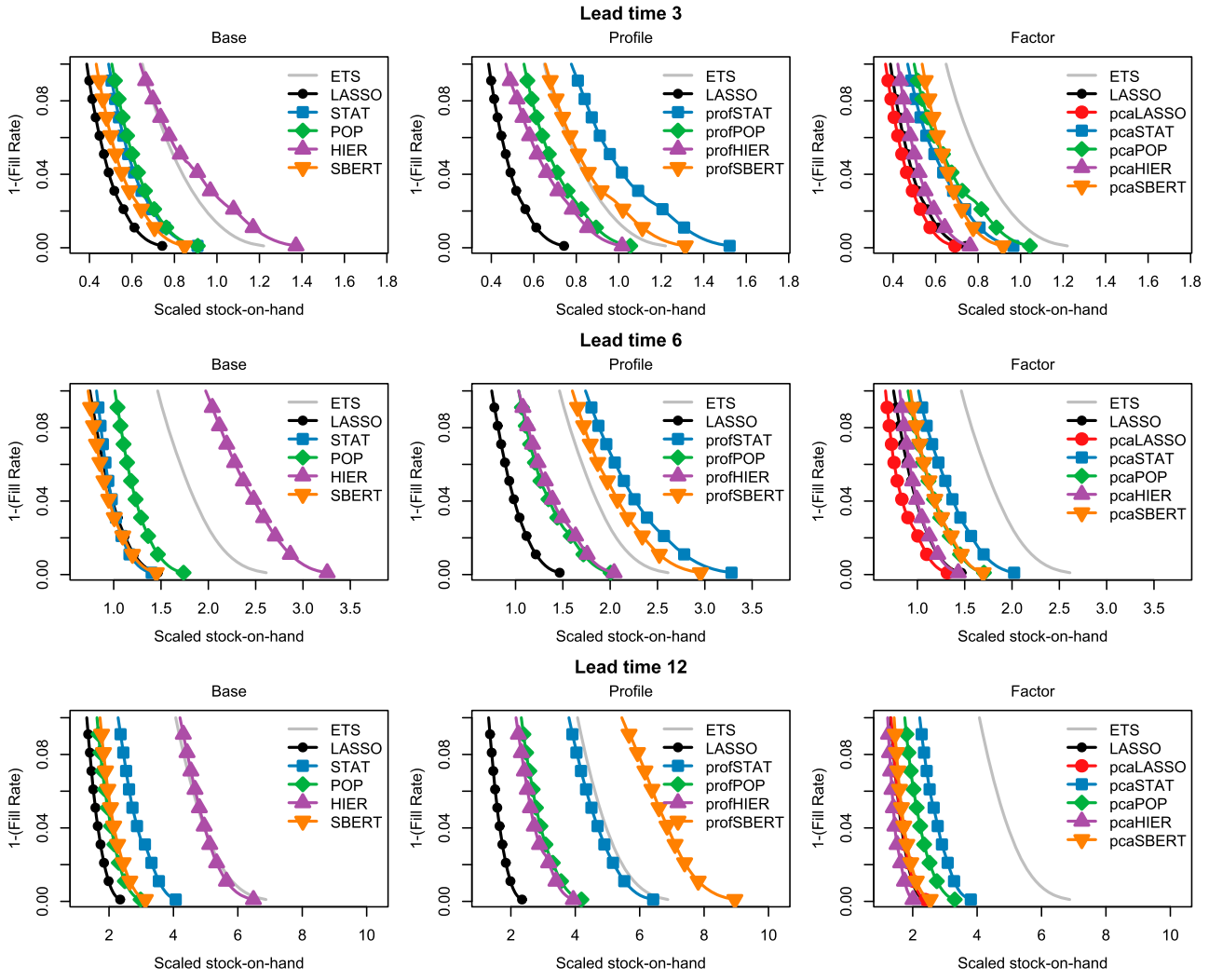


Figure 2. Trade off curves of continent level, between fill rate and scaled stock-on-hand. Each row plots results for a lead time, and each column for each group of methods. ETS and LASSO are provided in all graphs to facilitate comparisons (same location in each row).

5.1. Continent-level results

5.1.1. Inventory performance

Figure 2 provides trade-off curves between fill rates and remaining stock-on-hand at the end of the simulation. Curves closer to the origin dominate other solutions. The results are separated in multiple subplots to aid readability. Each row refers to a specific lead time, with the relevant subplots having the same scale. Each column groups the methods between base variants, profile based variants, and factor based variants. The ETS and LASSO benchmarks are provided in all subplots to facilitate comparisons.

First, we discuss the base methods. Observe for $L = 3$ the location of the ETS and the LASSO solutions. All investigated variable selection alternatives fall between these two, closer to LASSO, apart from HIER which performs worse than the ETS. A similar picture emerges for $L = 6$ and $L = 12$ with the differences that for $L = 6$ the LASSO solution is marginally outperformed by the SBERT and STAT solutions,

The performance of most selection variants based on profile (the middle column in Figure 2) is worse than the base counterparts, with profSTAT and profSBERT being outperformed by ETS: An exception is profHIER, which, however, is always outperformed by LASSO. On the other hand, the factor variants offer on average large gains in performance, with pcaLASSO typically being the best and dominating the LASSO solution. pcaHIER improves substantially over HIER, ranking now amongst the best, and in the case of $L = 12$ first. pcaSBERT becomes very competitive for $L = 12$, but otherwise, its base variant performs better for the rest of the lead times.

Overall, we observe that the semantic clustering and selection via SBERT is useful. Factor variants of LASSO and HIER also perform well. The good performance of POP is somewhat surprising given that the popularity of

Table 2. Forecasting performance statistics across continents.

	Cumulative AMsE						Cumulative RMsSE						Cumulative Pinball					
Method	Lead 3		Lead 6		Lead 12		Lead 3		Lead 6		Lead 12		Lead 3		Lead 6		Lead 12	
ETS	2.54*		4.82	⊘	10.29	✓	2.52		4.86		10.44	✓	0.36	⊘	0.78*	✓	1.95*	✓
LASSO	2.47*		5.01		14.13		2.36*		4.67		12.64		0.43		1.31		4.87	
STAT	2.61		5.16		12.10	✓	2.52		4.85		11.04	✓	**0.28***	✓	0.78*	✓	2.24*	✓
POP	2.42*	⊘	5.46		14.41		2.35*	⊘	5.20		12.96		0.31*	⊘	1.04	⊘	4.54	⊘
HIER	2.63		6.05		14.50		2.56		5.84		13.72		0.35	⊘	1.14	⊘	2.43*	✓
SBERT	2.64		4.80	⊘	12.57	⊘	2.56		4.68		11.41	⊘	0.39	⊘	0.88*	✓	3.22	✓
profSTAT	2.92		5.68		13.09	⊘	2.94		5.71		11.91	⊘	0.41	⊘	0.83	✓	1.80*	✓
profPOP	2.51*		4.66*	⊘	10.13*	✓	2.49		4.65	⊘	9.67	✓	0.33*	⊘	0.75*	✓	**1.76***	✓
profHIER	2.59		5.77		15.30		2.45		5.29		13.60		0.46		1.58		3.72	⊘
profSBERT	2.70		5.68		15.29		2.69		5.50		13.51		0.33	⊘	0.75*	✓	2.36*	✓
pcaLASSO	2.56*		5.49		13.85	⊘	2.43		5.14		12.85		0.46		1.52		3.59	⊘
pcaSTAT	2.47*	⊘	4.73	⊘	11.52	✓	2.40*		4.47	⊘	10.31	✓	0.32*	⊘	**0.66***	✓	1.86*	✓
pcaPOP	**2.19***	⊘	**3.92***	✓	**8.51***	✓	**2.10***	⊘	**3.79***	✓	**7.98***	✓	0.36	⊘	0.94	⊘	2.50*	✓
pcaHIER	2.66		5.66		13.93	⊘	2.57		5.24		12.41	⊘	0.47		1.42		6.36	
pcaSBERT	2.66		5.40		12.41	⊘	2.58		5.09		11.11	✓	0.44		1.12	⊘	2.95	✓

Best per column is in bold. Asterisks indicate no significant statistical differences from best. Better than LASSO is marked with ✓and ⊘ with the latter signifying insignificant differences. Tests at 5% level.

variables is based on users with diverse objectives, albeit not ranking first in any case.

5.1.2. Forecasting performance

Table 2 provides the summary statistics for bias, point, and quantile accuracy, for the three lead times. In each column, the minimum error is highlighted in bold, and other forecasts with no statistically significant differences at 5% level are indicated with an asterisk. All methods that outperform the LASSO are identified with a checkmark (✓), which is circled when the differences are statistically insignificant (⊘]).

Overall, factor methods (pca prefix) perform better, and the gains are more substantive for longer lead times. When exploring the inventory results, we identified SBERT as a good performing method. This is reflected in Table 2, although it appears to be outperformed by ETS in many cases (short lead times and for all pinball cases) even though this is not the case in terms of inventory. pcaLASSO and pcaHIER that were found to perform very competitively in inventory, do not rank well in any bias, point, and quantile accuracy, with the latter ranking worse. We highlighted the surprising performance of POP and its variants. pcaPOP ranks best in terms of bias and point forecasts, and highly in quantile accuracy. Although there is some disconnect in the ranking between inventory and bias/accuracy results, both ETS and LASSO benchmarks are always outperformed by the refined variable selection methods. Finally, observe that for short lead time ($L = 3$), there is limited evidence of statistical differences between the performance of many methods.

5.1.3. Explainability

A tertiary way to compare the alternative methods is in terms of explainability. The profile and factor methods obfuscate any explainability due to the summarisation of the raw variables. Profile methods are also found to perform poorly, so they are ignored here. Since the factor methods are based on principal components, one can reverse the calculation and translate the regression coefficients into coefficients for the original variables. However, since each principal component is a linear combination of the original variables, the independence between variables, which is needed for using regression coefficients for explaining the model, is strongly violated. Moreover, there would be no variable selection once the transformation is reversed, with all original variables contributing to the forecast. Therefore, although factor methods perform well, they cannot provide any additional insights or be validated, much like ETS, even though their decision performance is substantially better.

LASSO, STAT, POP, HIER, and SBERT directly provide a vector of coefficients for the selected variables. It is unreasonable to expect users to balance the influence of a large number of variables (Sroginis, Fildes, and Kourentzes 2023). First, we report the on average number of selected variables (across origins, forecast horizons, and time series): LASSO uses 69.5, STAT 135, POP 116.3, HIER 144.3, and SBERT 153.8. The sequential lasso used in all but LASSO is bound to increase the number of selected variables, as it was designed to overcome the exclusion of variables due to spurious correlations (Ma, Fildes, and Huang 2016). In all cases, the number of selected variables is overwhelmingly large.

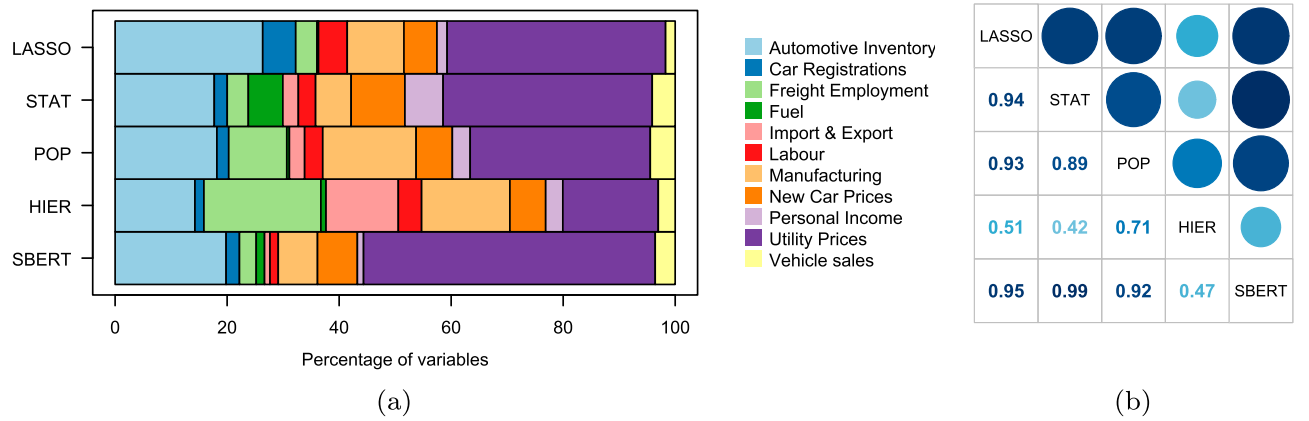


Figure 3. Percentage of variables belonging to each group, as selected by the methods. Percentages are averages across time series, origins, and horizons. The subplot (b) provides the correlations of the percentages between methods. The size of the circles denotes the strength of the correlations, provided numerically in the lower part.

Table 3. Percentage of commonly selected indicators.

		To				
	Method	LASSO	STAT	POP	HIER	SBERT
From	LASSO		69.43%	79.12%	58.47%	94.25%
	STAT	30.10%		57.28%	56.28%	74.60%
	POP	36.84%	62.03%		53.57%	80.36%
	HIER	23.84%	53.02%	46.58%		67.27%
	SBERT	31.43%	57.84%	57.69%	55.49%	

Percentages are based on the number of variables in the 'from' method.

We investigate the similarity of the variable selection between the different methods. Table 3 provides the percentage of variables that are commonly selected between different methods. As the different methods have resulted in a different number of variables, methods with more variables will have a higher probability of increased similarities. Therefore, we measure the percentages from each method to the rest (observe the *from* and *to* in the table). The provided values are averages across time series, and we do not distinguish between lags. All methods exhibit high similarity with SBERT, as it selected the most variables, largely encompassing other methods. The similarity of the various methods with LASSO (with the smallest set of variables) averages around 30%. Between the remaining methods the similarity ranges between 50 and 60%, with some exceptions that have higher similarities.

Arguably, the methods agree on the core narrative, i.e. all identify a core set of external variables that are predictively useful, but still disagree substantially when all selected variables are considered. However, note that on average only around 10% of the possible variables were selected, so even the rather expansive methods, like SBERT that results in large regressions, are in fact skilful in identifying useful predictors and exclude the majority of variables. If we would include in the similarity the agreement in variables that were not selected, then the scores would rise substantially. Reflecting on Figure 2, all methods outperformed the ETS (apart from HIER). This suggests that their included variables are predictively valuable, but due to their number and disagreement between methods, we do not claim that they are fully explainable.

Given the large number of selected variables, to increase the intelligibility of the regressions, and facilitate communication with users and stakeholders, the individual variables can be grouped further (for examples, see Sagaert et al. 2018b; Sagaert and Kourentzes 2025). We do this judgmentally, into 11 groups, as seen in Figure 3(a), which provides the percentage of variables belonging to each group per method, and the correlation of the groupings between methods. On group level, the methods exhibit increased similarity, with correlations over 0.9. An exception is HIER. Its disagreement and poor performance adds to the evidence that the other methods have captured a core of useful variables, while HIER provides an alternative narrative that is not backed by evidence. Investigating Figure 3(a), we can see that HIER has substantial differences when it comes to the importance of employment in freight transportation, import and export activity, and utility prices.

We caution against concluding from the group analysis that the different methods result in similar regression models, and that they are therefore trustworthy. If anything, the contrast between the detailed and group comparisons demonstrates that past comparisons in the literature that looked only at the top contributing variables or grouped representations are potentially giving a more optimistic view than a detailed analysis would.

We summarise the empirical evidence at continent level as follows. When explainability is not a key consideration, enhancing variable selection on massive sets

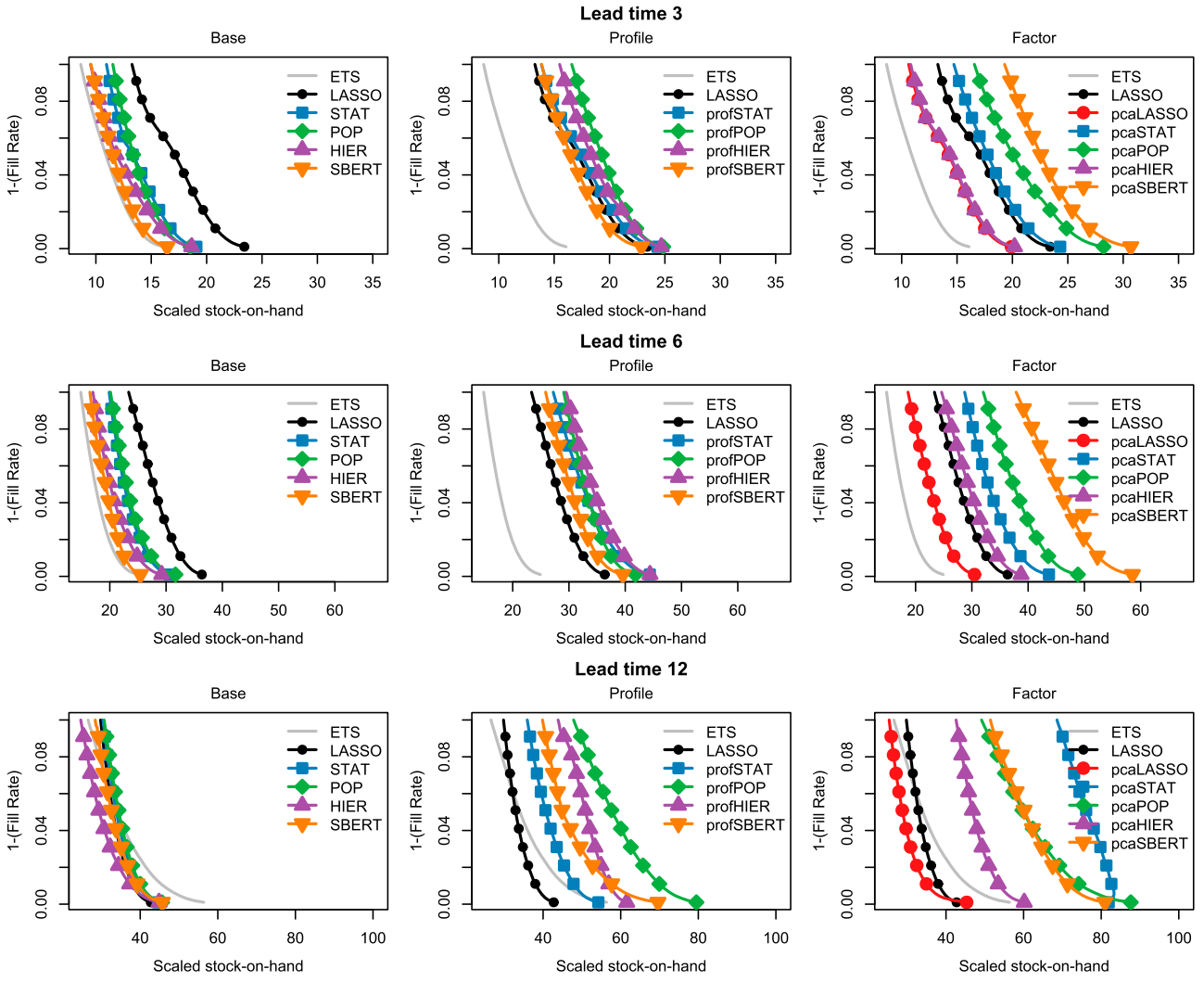


Figure 4. Trade off curves on country level, between fill rate and scaled stock-on-hand. Each row plots results for a lead time, and each column for each group of methods. ETS and LASSO are provided in all graphs to facilitate comparisons (same location in each row).

of variables with factors can be beneficial. This was found to substantially improve in terms of decision performance over ETS and outperform standard LASSO, with pcaLASSO being one of the better-performing methods. When explainability is desirable, embedding semantic information into the variable selection is beneficial. Although this did not result in consistently outperforming LASSO, it resulted in regressions with increased similarity, even at a detailed variable-by-variable analysis.

In the next subsection we contrast the aggregate demand forecasting results with results on more disaggregated data. These forecasts do not directly connect with production planning decisions, but allow us to further understand how methods that embed semantic information behave.

5.2. Comparison with country level demand forecasts

We repeat the same experiment on country level time series. Sagaert and Kourentzes (2025) provide evidence that the usefulness of lasso (and lightGBM) for directly modelling leading indicators on disaggregate data diminishes due to the increased noise in the data. We expect to see a reduction in the performance of methods with explanatory variables compared to ETS. Figure 4 provides the trade-off curves between fill rate and scaled stock-on-hand, in a similar setup to Figure 2.

First, note the location of the ETS and LASSO results for the different lead times. For $L = 3$ and $L = 6$ ETS achieves similar fill rates for much lower stock-on-hand, providing better solutions. For $L = 12$ LASSO outperforms ETS at higher fill rates. Focusing on the base

Table 4. Forecasting performance statistics across countries.

	Cumulative AMsE						Cumulative RMsSE						Cumulative Pinball					
Method	Lead 3		Lead 6		Lead 12		Lead 3		Lead 6		Lead 12		Lead 3		Lead 6		Lead 12	
ETS	2.64*	⊘	5.50*	⊘	12.44*	⊘	2.39*	⊘	5.00*		11.30*	⊘	**0.46***	⊘	**1.05***	✓	**2.93***	✓
LASSO	2.75*		5.69		13.51		2.44*		4.99*		11.63*		0.54*		1.33		4.46	
STAT	2.58*	⊘	5.36*	⊘	12.53*	⊘	2.34*	⊘	4.81*	⊘	11.11*	⊘	0.49*	⊘	1.19*	⊘	3.65*	⊘
POP	2.64*	⊘	5.43*	⊘	12.87*	⊘	2.38*	⊘	4.85*	⊘	11.55*	⊘	0.55*		1.43		4.74	
HIER	2.79*		5.93		14.57		2.48*		5.26		12.59		0.56*		1.41		4.16	⊘
SBERT	**2.55***	⊘	**5.00***	✓	**11.79***	✓	**2.31***	⊘	**4.53***	⊘	**10.52***	⊘	0.48*	⊘	1.14*	⊘	3.71*	⊘
profSTAT	2.97		5.96		12.98*	⊘	2.61		5.24		11.44*	⊘	0.54*	⊘	1.31*	⊘	3.54*	✓
profPOP	3.14		6.45		14.55		2.72		5.56		12.41		0.55*		1.36		3.81	⊘
profHIER	2.94		6.03		13.78		2.55*		5.23		11.98*		0.58		1.51		4.21	⊘
profSBERT	3.06		6.21		14.42		2.69		5.45		12.43		0.57		1.37		3.92	⊘
pcaLASSO	3.10		6.37		14.74		2.67		5.50		12.51		0.61		1.51		3.99	⊘
pcaSTAT	2.85		5.85		13.66		2.50*		5.13*		11.86*		0.56*		1.49		4.42	⊘
pcaPOP	2.94		6.05		14.03		2.57*		5.23*		11.83*		0.59		1.45		4.13	⊘
pcaHIER	2.90		5.99		14.16		2.49*		5.15*		11.98*		0.59		1.61		5.24	
pcaSBERT	2.93		5.88		13.21	⊘	2.55*		5.14*		11.38*	⊘	0.61		1.48		3.92	⊘

Best per column is in bold. Asterisks indicate no significant statistical differences from best. Better than LASSO is marked with ✓and ⊘ with the latter signifying insignificant differences. Tests at 5% level.

methods, it is evident that they all perform better than LASSO, including the HIER that performed poorly on the continent level (see Figure 2). SBERT that relies on semantic information, matches the performance of ETS for short and medium lead times, and performs equally or better for long lead times, depending on the fill rate. STAT and POP perform very similarly to each other, trailing from SBERT and ETS for all cases except $L = 12$. HIER is the best-performing method for lower fill rates for $L = 12$, matched by the rest of the base methods for higher fill rates.

Looking at the results for profile and factor methods, we can see that in most cases they perform poorly, often worse than both ETS and LASSO. pcaLASSO consistently improves upon LASSO, and for $L = 12$ outperforms both ETS and LASSO, providing one of the best inventory outcomes across all methods. This echoes its good performance at the continent level.

Table 4 summarises the bias, point, and quantile accuracies at the country level, and is organised in the same way as Table 2 (continent results). On bias (AMsE) and point accuracy (RMsSE) the SBERT forecasts perform best, being one of the few methods that outperforms ETS. Other base methods, with the exception of HIER, outperform LASSO. This is in agreement with the intuition from Figure 4, where the base methods improve over LASSO. In terms of pinball, ETS is the best performing forecast, with SBERT coming second, with relatively large differences for longer lead times. The results for the profile and factor methods are generally poor, matching the findings from the inventory metrics.

Although the factor methods performed well on the continent level, we argued that it is challenging to validate the resulting regressions. In general, we do not anticipate these methods to accurately capture the nuisances of the unknown demand generating process, and can be seen as a more powerful alternative to ETS for the those series, without claims of explainability or validity. At the country level, these methods do not perform well, underlining that preferring them is predominantly based solely on their performance. This is the crux of the issue when forecasts do not provide explainable outcomes: choosing them becomes a purely empirical question, as is the case for univariate forecasting models. This comes with all the limitations due to data characteristics and sample size. On the other hand, methods like SBERT, where we embed semantic information in the variable selection process, result in more consistent outcomes, where, beyond any evidence of empirical performance, we can seek to validate the resulting regression model on its explainability. The choice of these methods can be based both on empirical evidence and an evaluation of the implied narrative for the demand generating process. This allows both for a more complete evaluation of forecasts and can help highlight structural weaknesses that can be addressed to improve forecasts further (Pidd 1997).

Table 5 provides the average similarity in the selected variables for the base methods, excluding HIER, which exhibited limited explainability in the continent level. The similarity is provided for the car and truck material demand series separately, together with the average similarity across both. The reported values are grouped in two sets, depending on which model we count the number of variables from, similarly to Table 3. Country models are smaller than continent models, explaining the larger similarity values in the Country-to-Continent set. LASSO exhibits the lowest similarity. Grouping the indicators increases the similarity, with the semantic information used in SBERT resulting in the highest similarity. This suggests that the methods can uncover relations

Table 5. Variable selection similarity between continent and country models.

	Continent to Country			Country to Continent		
Method	Cars	Trucks	Average	Cars	Trucks	Average
LASSO	36.30%	37.35%	36.82%	56.69%	60.02%	58.36%
STAT	56.53%	52.48%	54.51%	73.03%	68.03%	70.53%
POP	56.56%	43.71%	50.13%	67.00%	70.27%	68.63%
SBERT	59.49%	51.73%	55.61%	77.72%	72.13%	74.93%

between variables that remain predictively useful across both demand aggregation levels.

6. Conclusions

Our objective has been to investigate whether incorporating semantic and meta-data in variable selection can improve decision and forecasting metrics, and assess their impact on model explainability and trustworthiness. Forecasting metrics are proxies of the usefulness of a forecast. These may not always correlate highly with the objective function of decisions (Athanasopoulos and Kourentzes 2023; Kourentzes, Trapero, and Barrow 2020). A fair critique of the forecasting literature is that it remains overly focused on accuracy comparisons, with two recent reviews of different application areas highlighting that there is limited evidence that the reported gains in accuracy offer equivalent decision improvements (Goltsos et al. 2022; Sonnleitner et al. 2025). The literature lacks clear guidance on how to reconcile divergent accuracy and decision metrics based model rankings. Athanasopoulos and Kourentzes (2023) take the stance that decision performance is more important, as it aligns with stakeholders' objectives. We argue that this is more nuanced, as decision metric evaluations often require substantial simplifications. We see the ranking of the forecasts on different metrics as complementary evidence, and argue that explainability, when relevant, should be considered as well.

In the context of time series forecasting, we argue that explainability should be understood as a relative quality between alternative forecasts and not an absolute one, since the demand generating process is unknown and there is no ground truth to compare against. We should be cautious with models whose explainability is non-credible. We do not argue against methods that do not make claims in terms of explainability, such as factor or univariate methods. These can be seen as performance benchmarks that explainable models should at a minimum match, to make any claims about business insights or knowledge discovery. The initial forecasting competitions (Makridakis et al. 1982; Makridakis and Hibon 2000) have shaped our methodology to be heavily empirically based, and this has led to substantial innovations and improvements in forecasting research (Fildes and Ord 2004). However, abstracting modelling from the application context is often an unnecessary handicap for model selection and validation. Within a problem context, we can query a model on its explainability and validity, providing a supporting dimension for choosing the appropriate forecast.

Overall, we find that there is a trade-off between explainability and decision metrics, with factor methods performing best in supporting tactical planning on aggregate demand. However, the lack of explainability in these methods makes their selection and use purely a matter of empirical evidence, similar to univariate models. We demonstrate the issues with generalising from these results by applying the methods to more disaggregate demand, where they performed poorly. In contrast, incorporating semantic information was found to be beneficial more consistently and supported explainability. The use of semantic information with SBERT for variable selection is found to be promising. This opens new areas of research in embedding semantic modelling in the specification of forecasting models: There are multiple alternatives in using SBERT for variable selection, or leveraging the semantic processing capabilities of newer Large Language Models. Using meta-data, such as variable popularity amongst analysts was helpful, but not as performant. We do not identify a specific approach that is performant in all cases, but this is to be expected given the different nature of the demand data on the continent and country levels. As the evidence in this work is predominantly empirical, additional testing on other case companies is needed.

Another conclusion from our work is that more methodological advances are needed in how to systematically measure and incorporate explainability in forecast evaluation. In this paper, we focus on the selected variables, however, one could focus on the excluded variables, or both, potentially enriching any insights. The analysis focuses on linear models, where, under conditions, reading the model coefficients contributes to their explainability. For more complex models, explainability becomes even more pertinent (Vlachos and Reddy 2025), and there are multiple methodologies and tools for explainability, with their own limitations (Arrieta et al. 2020). Nonetheless, as we illustrate by evaluating our methods at the variable and group levels, different approaches can provide diverse outcomes, which without a ground truth to compare to, can mislead users. Many tools provide explainability from a researcher's perspective, with little evidence that users understand explainability in the same

way (Mohseni, Zarei, and Ragan 2018). This is particularly relevant for machine learning methods, which provide limited mathematical tractability, and we have to rely on various explainability tools, such as SHAP values. As with reading the coefficients of a linear regression, these techniques come with assumptions that often invalidate their use (Verdinelli and Wasserman 2024). Nevertheless, the good performance of these methods, in the presence of multiple input variables, is evidenced in the literature (for example, Liang et al. 2024; Xu et al. 2023), increasing the need for additional research in the area.

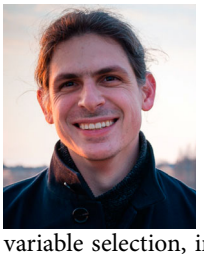
The business value of a forecast is driven by its trustworthiness. Spavound and Kourentzes (2022) propose that forecast trustworthiness has four characteristics: (1) reliability, where the likelihood of large errors is minimised, and their source is understood, (2) stability, where the application of a method is expected to provide similar results to past cases, (3) intelligibility, where the key elements of a forecast are explainable, and (iv) alignment with user objectives. Methods with limited explainability can rank low in the first three criteria. We do not argue against their usefulness, but underline that their choice is empirical and hard to validate. This is where we see variable selection methods enhanced with semantic information as promising. In addition, explainable forecasting models can enable more transparent scenario analyses in business planning. However, this hinges on a rigorous explainability methodology that requires more work. Ultimately, we call for additional research in the area, so as to develop mature methodologies to compliment the empirical evaluation in forecasting research, when possible.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Yves R. Sagaert is head of the research group of Predictive AI and Digital Shift at VIVES University of Applied Sciences in Belgium and director of Belgium's Forecasting Society, BELFORS. He is a researcher at KU Leuven and an Adjunct Professor at the IÉSEG School of Management in Lille (France). He strives to integrate business forecasting with real-world decisions and constraints, bringing forecast models closer to their cost impact. His research focuses on incorporating market intelligence in demand forecasting through leading indicators and the effects on supply chain management, especially with limited historical business data. His expertise lies in supply chain management, leading indicators, business forecasting, inventory, variable selection, and shrinkage methods.



Nikolaos Kourentzes research is in forecasting with statistical and machine learning methods. His current interests are in quantifying model uncertainty and translating forecasts effectively into organisational decisions. He has contributed to hierarchical forecasting, introducing temporal hierarchies, model specification and variable selection, intermittent demand and judgmental forecasting. He has authored a textbook in business forecasting and develops and maintains open-source libraries to facilitate the adoption of forecasting research into practice.

Data availability statement

Due to legal and commercial reasons the supporting data is not available.

ORCID

Yves R. Sagaert https://orcid.org/0000-0002-6322-7160
Nikolaos Kourentzes http://orcid.org/0000-0003-0211-5218

References

Aas, K., M. Jullum, and A. Løland. 2021. "Explaining Individual Predictions When Features are Dependent: More Accurate Approximations to Shapley Values." *Artificial Intelligence* 298:103502. https://doi.org/10.1016/j.artint.2021.103502.

Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin. et al. 2016. "TensorFlow: A System for Large-Scale Machine Learning." In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283. Savannah, GA: USENIX.

Abolghasemi, M., O. Ganbold, and K. Rotaru. 2025. "Humans vs. Large Language Models: Judgmental Forecasting in an Era of Advanced AI." *International Journal of Forecasting* 41 (2): 631–648. https://doi.org/10.1016/j.ijforecast.2024.07.003.

Armstrong, J. S. 2001. "Judgmental Bootstrapping: Inferring Experts' Rules for Forecasting." In *Principles of Forecasting: A Handbook for Researchers and Practitioners*, 171–192. Boston, MA: Springer US.

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI." *Information Fusion* 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Athanasopoulos, G., R. J. Hyndman, N. Kourentzes, and M. O'Hara-Wild. 2023. "Probabilistic Forecasts Using Expert Judgment: The Road to Recovery from COVID-19." *Journal of Travel Research* 62 (1): 233–258. https://doi.org/10.1177/00472875211059240.

Athanasopoulos, G., and N. Kourentzes. 2023. "On the Evaluation of Hierarchical Forecasts." *International Journal of Forecasting* 39 (4): 1502–1511. https://doi.org/10.1016/j.ijforecast.2022.08.003.

Baardman, L., S. B. Boroujeni, T. Cohen-Hillel, K. Panchamgam, and G. Perakis. 2023. "Detecting Customer Trends for Optimal Promotion Targeting." *Manufacturing &*

*Service Operations Management*25 (2): 448–467. https://doi.org/10.1287/msom.2020.0893.

Boivin, J., and S. Ng. 2006. "Are More Data Always Better for Factor Analysis?" *Journal of Econometrics*132 (1): 169–194. https://doi.org/10.1016/j.jeconom.2005.01.027.

Burnham, K. P., and D. R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer.

Charrad, M., N. Ghazzali, V. Boiteau, and A. Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." *Journal of Statistical Software* 61:1–36. https://doi.org/10.18637/jss.v061.i06.

Chollet, F., et al. 2015. Keras. https://keras.io.

Doganis, P., E. Aggelogiannaki, and H. Sarimveis. 2008. "A Combined Model Predictive Control and Time Series Forecasting Framework for Production-Inventory Systems." *International Journal of Production Research* 46 (24): 6841–6853. https://doi.org/10.1080/00207540701523058.

Fildes, R., P. Goodwin, and D. Önkal. 2019. "Use and Misuse of Information in Supply Chain Forecasting of Promotion Effects." *International Journal of Forecasting* 35 (1): 144–156. https://doi.org/10.1016/j.ijforecast.2017.12.006.

Fildes, R., P. Goodwin, and D. Onkal. 2021. "Use and Misuse of Information in Supply Chain Forecasting of Promotion Effects." *International Jouranl of Forecasting* 37 (3): 1329–1329. https://doi.org/10.1016/j.ijforecast.2021.01.024.

Fildes, R., and N. Kourentzes. 2011. "Validation and Forecasting Accuracy in Models of Climate Change." *International Journal of Forecasting* 27 (4): 968–995. https://doi.org/10.1016/j.ijforecast.2011.03.008.

Fildes, R., S. Ma, and S. Kolassa. 2022. "Retail Forecasting: Research and Practice." *International Journal of Forecasting* 38 (4): 1283–1318. https://doi.org/10.1016/j.ijforecast.2019.06.004.

Fildes, R., and K. Ord. 2004. "Forecasting Competitions: Their Role in Improving Forecasting Practice and Research." *A Companion to Economic Forecasting* 2: 322–353. https://doi.org/10.1002/9780470996430.

Freijeiro-González, L., M. Febrero-Bande, and W. González-Manteiga. 2022. "A Critical Review of Lasso and Its Derivatives for Variable Selection under Dependence among Covariates." *International Statistical Review* 90 (1): 118–145. https://doi.org/10.1111/insr.v90.1.

Fu, Y., and M. Fisher. 2023. "The Value of Social Media Data in Fashion Forecasting." *Manufacturing & Service Operations Management* 25 (3): 1136–1154. https://doi.org/10.1287/msom.2023.1193.

Gao, Z., and R. S. Tsay. 2025. "Supervised Dynamic PCA: Linear Dynamic Forecasting with Many Predictors." *Journal of the American Statistical Association* 120 (550): 869–883.

Ghasemi, E., N. Lehoux, and M. Rönnqvist. 2023. "Coordination, Cooperation, and Collaboration in Production-Inventory Systems: A Systematic Literature Review." *International Journal of Production Research* 61:5322–5353. https://doi.org/10.1080/00207543.2022.2093681.

Gneiting, T., and A. E. Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–378. https://doi.org/10.1198/016214506000001437.

Goltsos, T. E., A. A. Syntetos, C. H. Glock, and G. Ioannou. 2022. "Inventory–Forecasting: Mind the Gap." *European Journal of Operational Research* 299 (2): 397–419. https://doi.org/10.1016/j.ejor.2021.07.040.

Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: CRC Press.

He, K., L. Zheng, D. Wu, and Y. Zou. 2024. "Leveraging Large Language Models for Daily Tourist Demand Forecasting." *Current Issues in Tourism* 28: 1–17. https://doi.org/10.1080/13683500.2024.2417712.

Hooshmand Pakdel, G., Y. He, and X. Chen. 2025. "Predicting Customer Demand with Deep Learning: An LSTM-Based Approach Incorporating Customer Information." *International Journal of Production Research* 63: 1–13. https://doi.org/10.1080/00207543.2025.2468885.

Hyndman, R., A. B. Koehler, J. K. Ord, and R. D. Snyder. 2008. *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin, Germany: Springer Science & Business Media.

Katsagounos, I., D. D. Thomakos, K. Litsiou, and K. Nikolopoulos. 2021. "Superforecasting Reality Check: Evidence from a Small Pool of Experts and Expedited Identification." *European Journal of Operational Research* 289:107–117. https://doi.org/10.1016/j.ejor.2020.06.042.

Kourentzes, N., D. K. Barrow, and S. F. Crone. 2014. "Neural Network Ensemble Operators for Time Series Forecasting." *Expert Systems with Applications*41 (9): 4235–4244. https://doi.org/10.1016/j.eswa.2013.12.011.

Kourentzes, N., J. R. Trapero, and D. K. Barrow. 2020. "Optimising Forecasting Models for Inventory Planning." *International Journal of Production Economics* 225:107597. https://doi.org/10.1016/j.ijpe.2019.107597.

Li, L., Y. Kang, F. Petropoulos, and F. Li. 2023. "Feature-Based Intermittent Demand Forecast Combinations: Accuracy and Inventory Implications." *International Journal of Production Research* 61 (22): 7557–7572. https://doi.org/10.1080/00207543.2022.2153941.

Liang, L., B. Liu, Z. Su, and X. Cai. 2024. "Forecasting Corporate Financial Performance with Deep Learning and Interpretable ALE Method: Evidence from China." *Journal of Forecasting* 43 (7): 2540–2571. https://doi.org/10.1002/for.v43.7.

Ma, S., R. Fildes, and T. Huang. 2016. "Demand Forecasting with High Dimensional Data: The Case of Sku Retail Sales Forecasting with Intra-and Inter-Category Promotional Information." *European Journal of Operational Research* 249 (1): 245–257. https://doi.org/10.1016/j.ejor.2015.08.029.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler. 1982. "The Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition." *Journal of Forecasting* 1:111–153. https://doi.org/10.1002/for.v1:2.

Makridakis, S., and M. Hibon. 2000. "The M3-competition: Results, Conclusions and Implications." *International Journal of Forecastings* 16:451–476. https://doi.org/10.1016/S0169-2070(00)00057-1.

Mitra, R., P. Saha, and M. Kumar Tiwari. 2024. "Sales Forecasting of a Food and Beverage Company Using Deep Clustering Frameworks." *International Journal of Production Research* 62 (9): 3320–3332. https://doi.org/10.1080/00207543.2023.2231098.

Mohseni, S., N. Zarei, and E. D. Ragan. 2018. "A Survey of Evaluation Methods and Measures for Interpretable Machine Learning." Preprint, arXiv:1811.11839, 1, 1–16.

Montgomery, D., E. Peck, and G. Vining. 2012. *Introduction to Linear Regression Analysis*. 5th ed. Wiley Series in Probability and Statistics, Vol. 821. Hoboken, NJ: John Wiley & Sons.

Önkal, D., M. S. Gönül, and P. Goodwin. 2023. "Supporting Judgment in Predictive Analytics: Scenarios and Judgmental Forecasts." In *Judgment in Predictive Analytics*, 245–264. Cham, Switzerland: Springer.

Pidd, M. 1997. "Tools for Thinking-Modelling in Management Science." *Journal of the Operational Research Society* 48 (11): 1150–1150. https://doi.org/10.1057/palgrave.jors.2600969.

Ramos, P., J. M. Oliveira, N. Kourentzes, and R. Fildes. 2022. "Forecasting Seasonal Sales with Many Drivers: Shrinkage or Dimensionality Reduction?." *Applied System Innovation* 6: Article 3, 1–17. https://doi.org/10.3390/asi6010003.

Sagaert, Y. R., E. H. Aghezzaf, N. Kourentzes, and B. Desmet. 2018a. "Tactical Sales Forecasting Using a Very Large Set of Macroeconomic Indicators." *European Journal of Operational Research* 264 (2): 558–569. https://doi.org/10.1016/j.ejor.2017.06.054.

Sagaert, Y. R., E. H. Aghezzaf, N. Kourentzes, and B. Desmet. 2018b. "Temporal Big Data for Tactical Sales Forecasting in the Tire Industry." *Interfaces* 48 (2): 121–129. https://doi.org/10.1287/inte.2017.0901.

Sagaert, Y. R., and N. Kourentzes. 2025. "Inventory Management with Leading Indicator Augmented Hierarchical Forecasts." *Omega* 136:103335. https://doi.org/10.1016/j.omega.2025.103335.

Sagaert, Y. R., N. Kourentzes, S. De Vuyst, E. H. Aghezzaf, and B. Desmet. 2019. "Incorporating Macroeconomic Leading Indicators in Tactical Capacity Planning." *International Journal of Production Economics* 209:12–19. https://doi.org/10.1016/j.ijpe.2018.06.016.

Saoud, P., N. Kourentzes, and J. E. Boylan. 2022. "Approximations for the Lead Time Variance: A Forecasting and Inventory Evaluation." *Omega*110:102614. https://doi.org/10.1016/j.omega.2022.102614.

Schaer, O., N. Kourentzes, and R. Fildes. 2019. "Demand Forecasting with User-Generated Online Information." *International Journal of Forecasting* 35 (1): 197–212. https://doi.org/10.1016/j.ijforecast.2018.03.005.

Schaer, O., N. Kourentzes, and R. Fildes. 2022. "Predictive Competitive Intelligence with Prerelease Online Search Traffic." *Production and Operations Management* 31:3823–3839. https://doi.org/10.1111/poms.13790.

Silver, E. A., D. F. Pyke, and D. J. Thomas. 2016. *Inventory and Production Management in Supply Chains*. Boca Raton, FL: CRC Press.

Sonnleitner, B., N. Kourentzes, C. Ehrig, and A. Pflaum. 2025. "Forecasting for Optimization in Road Freight Transport: A Review." *Transportation Research Part E: Logistics and Transportation Review*204:104378. https://doi.org/10.1016/j.tre.2025.104378.

Spavound, S., and N. Kourentzes. 2022. "Making Forecasts More Trustworthy." *Foresight: The International Journal of Applied Forecasting* 66: 21–25.

Sroginis, A., R. Fildes, and N. Kourentzes. 2023. "Use of Contextual and Model-Based Information in Adjusting Promotional Forecasts." *European Journal of Operational Research* 307 (3): 1177–1191. https://doi.org/10.1016/j.ejor.2022.10.005.

Stock, J. H., and M. W. Watson. 2002. "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association* 97:1167–1179. https://doi.org/10.1198/016214502388618960.

Su, W., M. Bogdan, and E. Candes. 2017. "False Discoveries Occur Early on the Lasso Path." *The Annals of Statistics* 45: 2133–2150.

Trapero, J. R., M. Cardós, and N. Kourentzes. 2019. "Empirical Safety Stock Estimation Based on Kernel and Garch Models." *Omega* 84:199–211. https://doi.org/10.1016/j.omega.2018.05.004.

Trapero, J. R., N. Kourentzes, and R. Fildes. 2015. "On the Identification of Sales Forecasting Models in the Presence of Promotions." *Journal of the Operational Research Society* 66:299–307. https://doi.org/10.1057/jors.2013.174.

Trapero, J. R., D. J. Pedregal, R. Fildes, and N. Kourentzes. 2013. "Analysis of Judgmental Adjustments in the Presence of Promotions." *International Journal of Forecasting* 29 (2): 234–243. https://doi.org/10.1016/j.ijforecast.2012.10.002.

Tu, S., and Z. Gao. 2025. "A Supervised Screening and Regularized Factor-Based Method for Time Series Forecasting." Preprint, arXiv:2502.15275.

Uniejewski, B., and K. Maciejowska. 2023. "Lasso Principal Component Averaging: A Fully Automated Approach for Point Forecast Pooling." *International Journal of Forecasting* 39 (4): 1839–1852. https://doi.org/10.1016/j.ijforecast.2022.09.004.

Verdinelli, I., and L. Wasserman. 2024. "Decorrelated Variable Importance." *Journal of Machine Learning Research* 25: 1–27.

Vlachos, I., and P. G. Reddy. 2025. "Machine Learning in Supply Chain Management: Systematic Literature Review and Future Research Agenda." *International Journal of Production Research* 63: 1–30.

Wang, X., and F. Petropoulos. 2016. "To Select or to Combine? The Inventory Performance of Model and Expert Forecasts." *International Journal of Production Research* 54 (17): 5271–5282. https://doi.org/10.1080/00207543.2016.1167983.

Xu, Q., Z. Wang, C. Jiang, and Y. Liu. 2023. "Deep Learning on Mixed Frequency Data." *Journal of Forecasting* 42 (8): 2099–2120. https://doi.org/10.1002/for.v42.8.

Yuan, M., and Y. Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68 (1): 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x.

Zhang, Y., M. Wahab, and Y. Wang. 2023. "Forecasting Crude Oil Market Volatility Using Variable Selection and Common Factor." *International Journal of Forecasting* 39 (1): 486–502. https://doi.org/10.1016/j.ijforecast.2021.12.013.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2): 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Zou, H., and L. Xue. 2018. "A Selective Overview of Sparse Principal Component Analysis." *Proceedings of the IEEE* 106:1311–1320. https://doi.org/10.1109/JPROC.2018.2846588.

Appendix. Description of the LSTM

The LSTM are implemented using the Keras library (Chollet et al. 2015) and Tensorflow (Abadi et al. 2016). They have a first layer of 6 nodes, fully connected to a 12-node output layer, corresponding to a 12-step ahead forecast horizon. All neurons use the Rectified Linear Unit activation function. The topology was calibrated using 12-month rolling origin forecasts on a validation set of 18 months, testing up to 128 nodes per layer and deeper architectures, including dropout layers. Training is done using the Adaptive Moment Estimation optimiser, with early stopping (20 epochs) and MSE loss. All other settings use the default values in Keras. To reduce forecast variance and the sensitivity to the network initialisation, the final prediction is the median of 10 reinitialised LSTMs (Kourentzes, Barrow, and Crone 2014).

Table A1 provides comparable RMsSE summary statistics to Tables 2 and 4. The LSTM produces viable forecasts, but they are not dominating the LASSO equivalent. In most cases, LASSO remains more accurate, except for Lead 12 at the continent level.

Table A1. Cumulative RMsSE of LSTM.

Level	Lead 3	Lead 6	Lead 12
Continents	2.78	5.37	11.76
Countries	3.57	6.76	13.64