A new SVM-based ensemble approach for time series forecasting.

Marco A. Villegas[a,*], Diego J. Pedregal[a], Juan R. Trapero[b]

[a]Department of Business Administration. ETSI Industriales. Universidad de Castilla-La Mancha.
[b]Department of Business Administration. Faculty of Chemical Science and Technology. Universidad de Casilla-La Mancha.

Abstract

Time series analysis has remained as an extremely active research area for decades, receiving a great deal of attention from very different domains like econometrics, statistics, engineering, mathematics, medicine and social sciences. To say nothing about its importance in real-world applications in a wide variety of industrial and business scenarios. However, as hardware becomes ubiquitous, the amounts of data collected is more and more overwhelming, bringing us all the so-called big data era. It is in this context where automatic time series analysis deserves especial attention as a mean of making sense of such enormous databases.

Nevertheless, the automatic identification of the appropriate data modelling techniques stands in the middle as a compulsory stage of any big data implementation. Research on model selection and combination points out the benefits of such techniques in terms of forecast accuracy and reliability. This study proposes a novel ensemble approach for automatic time series forecasting as a part of a big data implementation. Given a set of alternative models, a Support Vector Machine (SVM) is trained at each forecasting origin to select the best model, according to the computed features and the past performance. The feature space embeds information of the time series itself as well as responses and parameters of the alternative models. This approach will help to reduce the risk of misusing modelling techniques when dealing with big datasets, and at the same time will provide a mechanism to assert the appropriateness of the underlying models used to analyse such data. The effects of the proposed approach are explored empirically using a set of representative forecasting methods and a dataset of 229 weekly demand series from a leading household and personal care UK manufacturer. Findings suggest that the proposed approach results in more robust predictions with lower mean forecasting error and biases than base forecasts.

Keywords: Decision support system, ensemble learning, support vector machines, time series analysis, model selection.

1. Introduction

Companies have traditionally adopted business intelligence solutions to support decision making on a consistent daily basis, bringing data from different sources into a common data infrastructure. However, the primal focus was mainly the creation of reporting tools (Davenport and Harris, 2007). In recent years, the so called business analytics (BA) introduced a new approach in this domain by leveraging on the latest progress on both computer science (e.g. data mining algorithms) and hardware technology (e.g. cloud computing, in-memory technology), making possible the integration of data sources and business operation on a higher stage of abstraction (Sheikh, 2013).

Although the complexity involved in a real implementation is even higher in a company-wide scope, organizations will maximize value from BA by integrating those models and tools into a broader architecture that ultimately unify data, operations and business. This approach will enable the “consistent source of truth” (Davenport, 2006) throughout the organization, as well as a more seamlessly and productive use of data by the staff. Nevertheless, research on BA has primarily focused on either solving the technical issues implied (Plattner, 2009) (Lenzerini et al., 2003)(e.g. data warehousing and IT infrastructure) or optimizing the final user-level productivity from the business perspective (Barone et al., 2010b) (Jeston and Nelis, 2008) (Taylor, 2011) (Barone et al., 2010a). In spite of that, the automatic identifica-

*Corresponding author.
Email address: marco.villegas@uclm.es (Marco A. Villegas)

tion of the appropriate data modelling techniques stands in the middle as a compulsory stage of any BA implementation.

In this sense, forecasting models are of strategic nature given that they gear business decisions ranging from inventory scheduling to strategic management (Petropoulos et al., 2014). Focusing on a supply chain context, automatic model selection is a necessity due to the high number of products whose demand should be forecast (Fildes and Petropoulos, 2015). Forecasting and operational research literature has faced the problem with different approaches. A first approach could be aggregate selection, where a single source of forecasts is chosen for all the time series (Fildes, 1989), instead of individual selection, where the particular method appropriate for each series is selected. However, aggregate selection cannot distinguish the individual characteristics of each time series (such as trend and/or seasonality) and, in general terms, individual selection outperforms aggregate selection, although with an associated higher complexity level and computational burden (Fildes and Petropoulos, 2015).

Regarding individual selection, different criteria to choose the most adequate model can be found in the literature. For instance, information criteria like Akaike Information Criteria (AIC) or Schwarz's Bayesian Criteria (SBC) are typically used. These information criteria produce a value that represents the compromise between goodness of fit and the number of parameters. Billah et al. (2006) compare different information criteria to select the most appropriate exponential smoothing model on simulated data and a subset of the time series from the M3 competition database, where the AIC slightly outperformed the rest of information criteria considered.

The identification of the best forecasting model has also been addressed depending on the time series features. Initially, Pegels (1969) presented nine possible exponential smoothing methods in graphical form taking into account all combinations of trend and cyclical effects in additive and multiplicative form. Collopy and Armstrong (1992) developed a rule-based selection procedure model (RBF) based on a set of 99 rules for selecting and combining between methods based on 18 time series features. In order to automatize this procedure, Adya et al. (2001) developed and automated heuristics to detect six features that had previously been judgmentally identified in RBF by means of simple statistics achieving a similar forecasting accuracy performance. Petropoulos et al. (2014) analysed via regression analysis the main determinants of forecasting accuracy involving 14 popular forecasting methods (and combinations of them), seven time series features and the forecasting horizon as a strategic decision. Wang et al. (2015) propose a rather different approach for long-term forecasting based on dynamic time warping of information granules. Instead, Yu et al. (2016) focus on finding an empiric decomposition (intrinsic mode functions) to aggregate the individually forecast components later into an ensemble result as the final prediction. An alternative for selecting among forecasts is the performance evaluation of the methods in a hold-out sample (Fildes and Petropoulos, 2015; Poler and Mula, 2011), where forecasts are computed for single or multiple origins (cross-validation) usually via a rolling-origin process (Tashman, 2000)

Finally, another option is to explore combination procedures (Clemen, 1989). In fact, Fildes and Petropoulos (2015) concluded that combination could outperform individual or aggregate selection for non-trended data. Different combination operators (mode, median and mean) to compute neural network ensembles are analyzed by Kourentzes et al. (2014), where the mode is found to provide the most accurate forecasts.

Apart from forecasting models considering time series, it should be noted the automatic identification algorithms developed for causal models. For instance, marketing analytics models to forecast sales under the presence of promotions have been analyzed by Trapero et al. (2015). Additionally, models capable of incorporating data from other companies in a supply chain collaboration context with information sharing have been explored by Trapero et al. (2012).

In addition to traditional time series modelling techniques, Artificial Intelligence (AI) algorithms have proved to be quite effective as a mean to build higher level methodologies to face big data challenges in an effective way, gearing upon both traditional and AI low-level techniques. An initial attempt has been carried out by Garcia et al. (2012), where multiple time series have been classified according to its ACF and PACF values to reduce the number of forecasting ARIMA models to be designed. However, the forecasting implications of that procedure in terms of out-of-sample accuracy was not described. Li and Hu (2012) propose a combination of ARIMA models using fuzzy logic rules and particle swarm optimization. Another efforts focus on building hybrid models incorporating AI models as an intrinsic component. In a different context, Wang et al. (2013) describe a successful application of SVM in a multiple classifier ensemble, where a bunch of one-class SVM classifiers are trained on different sub-features vectors. An additional approach consists on finding homogeneous groups of time series, and model each group separately. For example, Lu and Kao (2016) focus on clustering the time series to model each cluster via extreme learning machine. Other approaches tend to give more importance to the hints coming from the application domain, using models that allow embedding such meta information. An example of this is given by Homaie-Shandizi et al. (2016) who use decision trees to predict monthly pilot reserve hours.

In this work we aim at building an Automatic Forecasting Decision Support System for the 229 Stock Keeping Units (SKU) of a leading household and personal care UK manufacturer. The data is highly volatile and with small serial correlation. The system consists of fitting a set of models coming from the Exponential Smoothing and ARIMA family of models with different levels of complexity. The key issue is that the model selection is not based only on information criteria, Schwarz's (SBC), but it is rather more sophisticated by adding a number of additional features with the aid of a multi-class SVM. Initially, potential features found in the literature like SBC statistics, ACF and PACF values, unit root tests, etc., were evaluated and only 19 were kept. Then, the SVM is trained to select the most adequate alternative from a set of models, including Exponential Smoothing and ARIMA models with different levels of complexity. The results show that the proposed approach improves the out-of-sample forecasting accuracy with respect to single or combined models.

The key contributions of this paper are as follows: i) propose a novel ensemble approach for time series forecasting based on SVM classification, ii) compare base and ensemble forecast error characteristics out-of-sample, iii) investigate the effects of the ensemble on forecasting errors, as measured in terms of median, mean, bias and variance.

The rest of the paper is organized as follows: Section 2 introduces the forecasting models and the use of the SVM for automatic model selection. Section 3 presents an empirical evaluation of the approach in a demand planning case study with real data. Section 4 analyses the results followed by some final considerations and afterthoughts.

2. Methods

2.1. Forecasting models

Let z_t be the mean-corrected output demand data sampled at a weekly rate, a_t a white noise sequence (i.e. serially uncorrelated with zero mean and constant variance), θ_i a set of parameters to estimate and B the backshift operator in the sense that $B^l z_t = z_{t-l}$. Then, taking into account the fact that no seasonality is present in the data, the forecasting models considered in this paper are the following:

$$\textbf{M1}: \quad z_t = a_t \tag{1}$$

$$\textbf{M2}: \quad z_t = (1 + \theta_1 B + \theta_2 B^2) a_t \tag{2}$$

$$\textbf{M3}\text{ (ETS)}: \quad (1 - B) z_t = (1 + \theta_1 B) a_t \tag{3}$$

$$\textbf{M4}: \quad (1 - B) z_t = (1 + \theta_1 B + \theta_2 B^2) a_t \tag{4}$$

$$\textbf{Mean}: \quad \text{Mean of forecasts } \textbf{M1} \text{ to } \textbf{M4} \tag{5}$$

$$\textbf{Median}: \quad \text{Median of forecasts } \textbf{M1} \text{ to } \textbf{M4} \tag{6}$$

Model **M1** is white noise, model **M2** is a MA(2), model **M3** is an IMA(1,1) that is actually treated as a Simple Exponential Smoothing model or a ETS(A,N,N) in (Hyndman et al., 2008) nomenclature (where E, T, S, A and N stand for Error, Trend, Seasonal, Additive and None, respectively), **M4** is an IMA(1,2), and **Mean** and **Median** are combination methods. In essence, two stationary, three non-stationary models and two combinations of models are considered.

It is important to note that some models are nested versions of others. For example, model **M1** is a parametrically efficient version of the rest of models if $\theta_1 = 0$ and $\theta_2 = 0$ in model **M2**, $\theta_1 = -1$ in **M3**, or $\theta_1 = -1$ and $\theta_2 = 0$ in **M4**. Similarly, ETS model **M3** is a particular version of **M4** with $\theta_2 = 0$. Finally, models **M2** and **M4** are not nested with any other models because of the difference operator.

These sort of constraints have been taken into account in the estimation procedure, since when approximate constraints are found the models preferred are the most parsimonious ones. This is particularly important when dealing with estimated roots close to unity. Specifically, when $\theta_1 < -0.992$ in model **M3**, the model is switched to **M1** for forecasting purposes. Similarly, if any root in the MA polynomial of model **M4** is smaller than -0.992 the model is switched to a MA(1), that is not any of the models considered above. No unit roots where detected when estimating model **M2**.

For the 229 SKU time series considered, at least one of the models **M1** to **M4** above is correct in statistical terms in the sense that one of them filters out all the serial correlation present in the data. Figure 1 shows the minimum for the four models **M1-M4** of the Ljung-Box Q statistic to test the absence of serial correlation for eight lags, approximately two months of data (Ljung and Box, 1978). Bearing in mind that the maximum number of parameters is two, a conservative value for degrees of freedom to perform the test is 6.

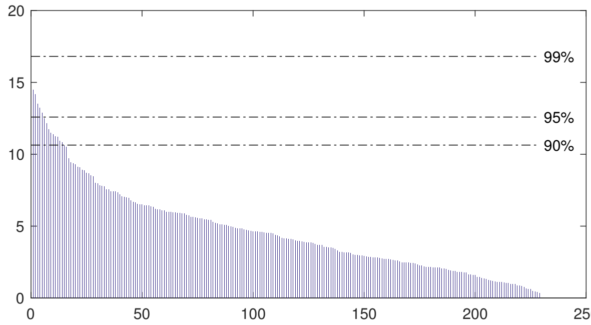

Figure 1: Minimum Ljung-Box Q statistic for each product

The critical values for the Q test at confidences levels of 90%, 95% and 99% on a Chi-Squared distribution with 6 degrees of freedom are 10.64, 12.59 and 16.81, respectively, marked by dotted, dashed and solid horizontal lines in Figure 1. Therefore, most of the values are well below the 90% confidence limit. This means that, depending on the level of confidence, for 93.89%, 97.82% and 100% of the SKU series at least one of the models is correctly specified, in the sense of not leaving significant serial correlation below the confidence limits mentioned before. This means that the models fulfill the rule of being a sufficient representation of the data at the same time of preserving parsimony.

2.2. *Support vector machines*

The support vector machine classifier is basically a binary classifier algorithm that looks for an optimal hyperplane as a decision function in a high-dimensional feature space (Shawe-Taylor and Cristianini, 2004). Consider the training data set $\{\mathbf{x}_k, y_k\}$, where $\mathbf{x}_k \in \mathbb{R}^n$ are the training examples and $y_k \in \{-1, 1\}$ the class labels. The training examples are firstly mapped into another space, referred to as the feature space, eventually of a much higher dimension than $\mathbb{R}^n$, via the mapping function Φ. Then a decision function of the form $f(\mathbf{x}) = \langle \mathbf{w}, \Phi(\mathbf{x})\rangle + b$ in the feature space in computed by maximizing the distance between the set of points $\Phi(\mathbf{x}_k)$ to the hyperplane parameterized by ($\mathbf{w}$,b) while being consistent on the training set. The class label of $\mathbf{x}$ is obtained by considering the sign of $f(\mathbf{x})$. In the non-separable case, the misclassified examples are quadratically penalized scaling by a constant C, the cost parameter, and the optimization problem takes the form $\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{m}\xi_k^2$ under the constraint $y_k f(\mathbf{x}_k) \geq 1 - \xi_k^2$, $\forall k$. Using Lagrangian theory, the optimal vector $\mathbf{w}$ is known to have the form $\mathbf{w} = \sum_{k=1}^{m}\alpha_k^* y_k \Phi(\mathbf{x}_k)$ where α_k^* is the solution of the following quadratic optimization problem:

$$\max_{\alpha} W(\alpha) = \sum_{k=1}^{m}\alpha_k - \frac{1}{2}\sum_{k,l}^{m}\alpha_k\alpha_l y_k y_l\left(K(\mathbf{x}_k, \mathbf{x}_l) + \frac{1}{C}\delta_{k,l}\right) \quad (7)$$

subject to $\sum_{K=1}^{m} y_k\alpha_k = 0$ and $\alpha_k \geq 0, \forall k$, where $\delta_{k,l}$ is the Kronecker symbol and $K(\mathbf{x}_k, \mathbf{x}_l) = \langle\Phi(\mathbf{x}_k), \Phi(\mathbf{x}_l)\rangle$ is the Kernel matrix of the training examples. The extension for the case of multiclass-classification with j levels, $j > 2$, could be done by the "one-against-one" approach in which $j(j-1)/2$ binary classifiers are trained; then the appropriate class is found by a voting scheme Meyer et al. (2015).

The function K is also known as the kernel function, which computes inner products in the feature space directly from the inputs x. It is supposed to capture the appropriate similarity measure between the arguments, while being computationaly much less expensive than explicitely computing the mapping Φ and inner product. Although the design of kernel functions is a very active research area, there are some popular kernels that have been tested in a variety of domains and applications with good results. The *polynomial kernel* is defined as $K(x, z) = p(\langle x, z\rangle)$ where $p(\cdot)$ is any polynomial with positive coefficients. In many cases it also refers to the special case $K_d(x, z) = (\langle x, z\rangle + R)^d$ where R and d are parameters. *Gaussian kernels* (also known as Radial Basis Functions kernels) are the most widely used kernels and have been studied in many different applications. It is defined by

$$K(x, z) = exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \quad (8)$$

where σ is a parameter that controls the flexibility of the kernel. In this study gaussian kernels are extensively used, and the parameter σ is estimated via *cross-validation* as explained in Section 2.3.

Though it was extended to regression problems since its early days Müller et al. (1997), SVM were originally designed as a classification algorithm Cortes and Vapnik (1995) and have been extensively exploited in a huge variety of classification contexts, e.g. hand-written digit recognition, genomic DNA Furey et al. (2000), text classification Joachims (2002), sentiment analysis Pang et al. (2002). But surprisingly SVM have not been applied to the problem of model selection in the context of multiple forecasting models. This paper constitutes an contribution in the area of possible applications of SVM in this scenario.

2.3. Feature selection and extraction

Reliable results in SVM and other *data driven* modelling techniques are considerable conditioned to the quality of the data available for training. Apart from the correctness of the data itself, there are some other aspects regarding the dimensionality of the dataset. In fact, it is well known that as the number of variables increases, the amount of data required to provide a reliable analysis grows exponentially Hira and Gillies (2015). Many feature selection (removing variables that are irrelevant) and feature extraction (applying some transformations to the existing variables to get a new one) techniques have been discussed to reduce the dimensionality of the data Kira and Rendell (1992), to say nothing about some other approaches based on linear transformation and covariance analysis like PCA and LDA Cao et al. (2003) Duin and Loog (2004). For the experiments carried out in this work, the initial dataset contained 14885 records (65 origins and 229 products), and 39 features including SBC and Q statistics, ACF and PACF, fitted parameters for each model (if any), gaussianity and heterokedasticity tests over residuals, unit tests, relative differences and ranking of the alternative models. After a process of feature selection and extraction via cross-validation, the number of variables was reduced to 19 resulting in a matrix $\mathcal{W}$ of dimension 14885×19. The final features are:

- Four (4) last SKU values available at time t.
- Relative (6) differences among the predictions provided by the alternative forecasting methods (**M1** to **M4**).
- Four (4) predictions provided by the alternative forecasting methods (**M1** to **M4**).
- Parameters (5) used by the forecasting methods.

The vector of labels $\mathcal{L}_i$ is formed as a categorical variable indicating the model with lower forecasting error at horizons $t+i$ for $1 \leq i \leq 4$. For horizons $t+2$, $t+3$ and $t+4$ the model with lower forecasting error is selected as the one that minimizes the total sum of squared errors in all the spanned weeks $t+1, \cdots, t+i$.

For each week k, with $4 < k < 65$, a SVM with a radial basis function kernel (RBF) is trained using the training set $\mathcal{W}_{train}$ and the corresponding vector of labels $\mathcal{L}_i$ for each forecasting horizon. $\mathcal{W}_{train}$ is form as a partition of matrix $\mathcal{W}$ including up to four weeks of history ($h = 4$), i.e., gathering records for weeks $k-4$ to $k-1$. Similar considerations were done in shaping vector $\mathcal{L}_i$. Different values for h were also empirically tested, resulting in $h = 4$ as the optimal for the current dataset. For optimizing the σ and *cost* parameters a 5-fold cross-validation was performed.

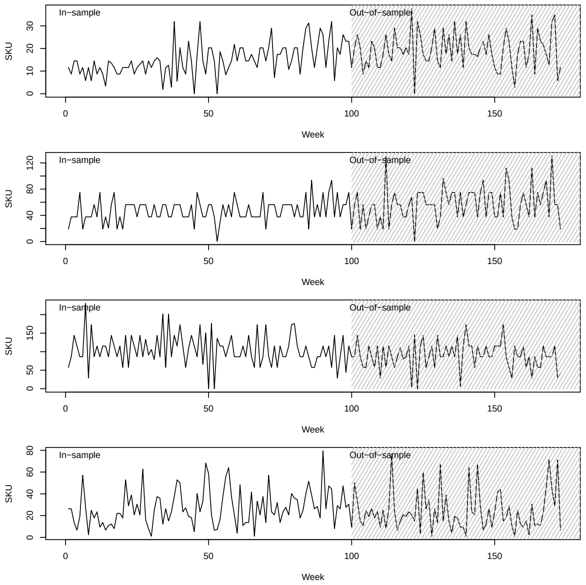

Figure 2: Example of some SKU from the dataset

3. Case study

The evaluation of the models proposed is carried out on a set of 229 demand series from a leading household and personal care UK manufacturer. For each product, there are 173 weekly sales observations, from which 101 observations are used as in-sample and the rest are reserved for out-of-sample evaluation. Therefore, a set of 69 forecast rounds of 4 weeks ahead were carried out for each product. Figure 2 shows some examples of the time series in the dataset, where the shaded area shows the out-of-sample period. Neither seasonality is visible in the sample SKUs by eye inspection, nor any strong correlation pattern.

A rolling forecasting experiment is carried out by expanding the in-sample span one week at a time. All forecasting models are fitted in the in-sample partition using the available data up to time T with $101 \leq T \leq 169$, and tested in the out-of-sample partition using observations $T+1, \cdots, T+4$. The out-of-sample forecasting errors are therefore calculated for the four forecasting horizons on each of the 229 products. We measure the forecast error using scaled mean squared error (sMSE) and scaled median squared error (sMdSE) that come from computing the scaled error (*sE*) and scaled squared error (*sSE*) of the lead time forecast according to the following formulae:

$$sE_{T+l} = \frac{\sum_{j=1}^{l} z_{T+j} - \sum_{j=1}^{l} \hat{z}_{T+j}}{\frac{1}{T}\sum_{i=1}^{T} z_i}, \tag{9}$$

$$sSE_{T+l} = \frac{(\sum_{j=1}^{l} z_{T+j} - \sum_{j=1}^{l} \hat{z}_{T+j})^2}{\frac{1}{T}\sum_{i=1}^{T} z_i}, \tag{10}$$

where the denominator is the mean of the time series, $\hat{z}_{T+j}$ stands for the forecast at time $T+j$ and $l = 1, 2, 3, 4$. Using these metrics has the advantage of allowing zero values at some periods of the series, and makes the results scale independent and therefore we can summarize them across products and forecasting horizons. Scaled absolute errors were also calculated in addition to squared errors in (10), but results were very similar and are not reported. Such results are available from the authors.

Models are estimated by Exact Maximum Likelihood using the ECOTOOL toolbox written in MATLAB (Pedregal and Trapero, 2012), except **M3** that was handled in SSpace (Pedregal and Taylor, 2012). SVM were treated by using the R package e1071 Meyer et al. (2015).

4. Results and discussion

One key issue in this study is the agnostic point of view, by which we assume that there is not necessarily a stochastic process that underlies the actual data, especially for this case study, where there is little correlation structure in the data. Being this true, one might expect that the best model in forecating terms changes with the forecast origin and/or horizon.

Some evidence emerges in the in-sample properties of the models. For example, computing the SBC for all the SKUs with 101 observations and the full sample we see a different model selection in 37% of the time series (Table 1). Detailed information about model selection is shown in the first two columns of Table 1, where SBC tends to select models with higher number of parameters and unit roots as the sample size increases. For the small sample size, in 55% of the cases the simplest model **M1** is chosen, i.e. for more than half of the SKUs the best model is that there is no model! Such proportion is reduced with the full sample, but still **M1** is the best model according to SBC in 39.30% of the cases, followed by the Exponential Smoothing with 29.26% of the cases.

The third column of Table 1 also shows the best models that would have been selected with the full sample based on a pure forecasting criterion used with the forecasting results obtained between samples 101 and 173. The disagreement with the SBCs selection are of 69% and 44% of the SKUs for the small and full sample sizes, respectively. Taken the information in Table 1 altogether, it shows evidence of the little correlation structure seen in the data, that tends to become more important with longer time series.

Table 1: Percentage of SKU for which each model is best according to SBC on different data-partitions and the out-of-sample forecast performance

	SBC(101)	SBC(173)	Out-of-sample
M1	55.46%	39.30%	17.03%
M2	14.41%	9.61%	16.16%
M3	13.97%	29.26%	34.93%
M4	16.16%	21.83%	31.88%

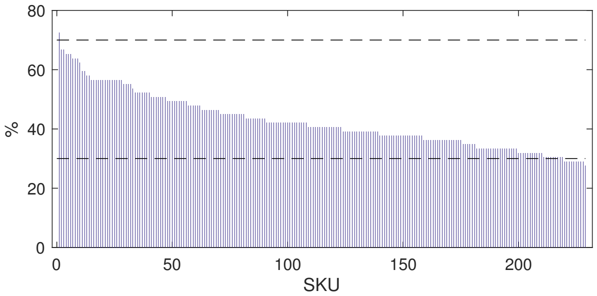

Figure 3: Proportion of time origins at which the best model is best for all SKU

Figure 3 shows, for each SKU, the proportion of times out of 69 forecasting origins in the rolling experiment that the best model is actually best according to the forecasting errors. For example, a 50% for a single SKU in that figure means that the best model was best in 35 of the forecasting origins. Only in 4 SKUs the winner model was best in more that 60% of the forecasting origins and only in 28 SKUs the proportion where superior to 50%. This means that, even when a model is best minimizing the forecasting error for a single SKU, rarely it is the best at more than 50% of the forecasting origins.

In order to get a deeper insight into the complexity of the problem, Figure 4 shows a single SKU, where it may be seen that, taken up to observation 101 it may be considered stationary and therefore either **M1** or **M2** may be appropriate candidates. However, the fact that a trend appears afterwards implies that such model might not be optimal any more.

Such intuitions are supported by Table 2, which shows the SBC for all models with samples up to 101 and full sample, in addition to the sMSEs. The pre-

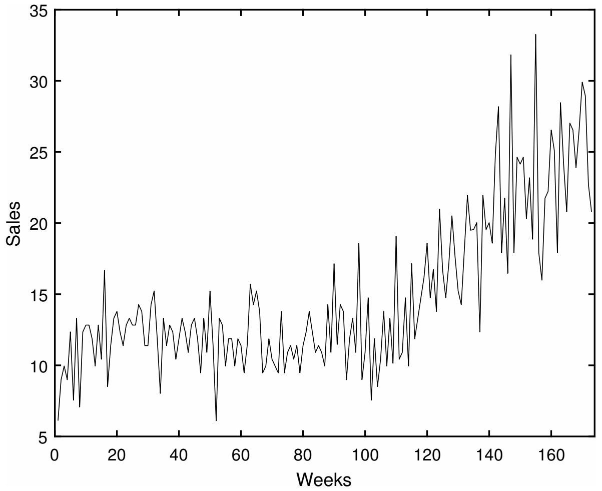

Figure 4: Example of SKU

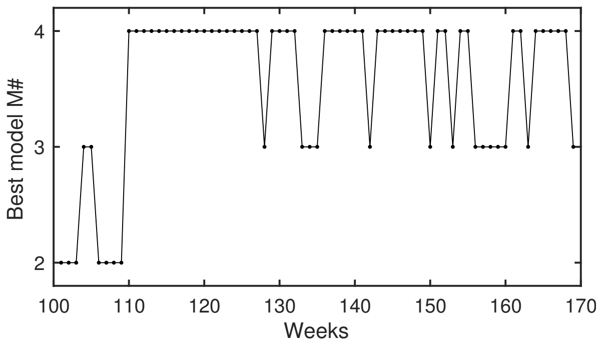

Figure 5: Best model for each out-of-sample forecast origin for SKU in Figure 4

ferred model according to SBC in the small sample is **M1**, with a Q(8) statistic of 3.72 indicating that there is no correlation left on the residuals. This model is the worst for the full sample. Additionally, the best model for the full sample switches to **M4** (Q(8) is 9.54), while the forecasting criterion suggests that **M4** is the best, with little margin over **M3**. Interestingly, the model that is the best considering all forecasting origins is model **M4** with just 53% of the time.

This evidence is complemented with Figure 5, which shows the best model according to its forecasting performance for each forecast origin in the out-of-sample span for the same SKU. At the very beginning the best models tended to be **M2** or **M3**. But afterwards, as the trend becomes more prominent, the best model switches to **M4** most of the time, though not always.

The previous evidence shows that there is not best model outperforming the rest for all SKUs, all forecasting origins and all forecasting horizons. Moreover, even for a single SKU there is not consistent best model along time. At this point the SVM-based ensemble approach is introduced in order to test the hypothesis that there is some pattern that would allow to improve the forecast accuracy over all models and possible combinations of them. In this sense, the proposed approach might be considered a sophisticated combination method in itself.

Table 2: SBC for all models in two different data-partitions and sMSE for out-of-sample for SKU in Figure 4

	SBC(101)	SBC(173)	sMSE
M1	1.52	3.31	0.053
M2	1.61	2.89	0.038
M3	1.59	2.26	0.013
M4	1.64	2.19	0.012

Table 3 shows the scaled mean (median) squared errors for all forecasting models and methods, including a Naïve model that serves as a benchmark. The last row corresponds to the errors generated by selecting the best possible model for every forecasting-step out of the models set considered.

Several facts emerge from Table 3. Firstly, taken as a whole, all models outperform the Naïve by a wide margin, implying that all models capture, at least at some part of the experiment, the correlation structure of the data. Secondly, for individual models **M1** to **M4** the best is consistently the Exponential Smoothing **M3** model, with an advantage that grows with the forecasting horizon. Thirdly, combinations of methods (mean and median) do not manage to outperform the Exponential Smoothing and both provide virtually the same results. Finally, and most importantly, the SVM-based ensemble approach is the overall best for all forecasting horizons with errors that fall between the Exponential Smoothing and the baseline minimum forecasting errors. Once more, the advantages of the proposed method are appreciated more clearly for higher forecast horizons.

The power of the proposed approach is considerably enhanced when the bias is considered, based on the sE_t measurements in equation (9), see Table 4. All biases are small bearing in mind that the highest bias in the table is 0.124 and the normalization imposed on the data implies a mean of 1. Conclusions about bias is quite different depending on whether we rely on SME or sMdE, but due to the robustness of the median it is safer to use the sMdE values in parenthesis. In essence, the bias replicate what was seen in squared errors, the models with the smallest squared errors are at the same time the models with the smallest bias. The best is the SVM-

Table 3: Forecast accuracy for out-of-sample sets in sMSE (sMdSE).

	Out t+1	Out t+2	Out t+3	Out t+4
Naive	0.184 (0.041)	0.558 (0.128)	1.114 (0.266)	1.856 (0.437)
M1	0.115 (0.032)	0.278 (0.075)	0.486 (0.134)	0.743 (0.195)
M2	0.109 (0.030)	0.255 (0.072)	0.447 (0.130)	0.689 (0.192)
M3	0.100 (0.026)	0.221 (0.054)	0.363 (0.087)	0.533 (0.123)
M4	0.102 (0.027)	0.230 (0.059)	0.380 (0.096)	0.555 (0.139)
mean	0.101 (0.027)	0.226 (0.060)	0.374 (0.101)	0.549 (0.150)
median	0.101 (0.027)	0.225 (0.059)	0.373 (0.101)	0.549 (0.150)
SVM-based ensemble	**0.099 (0.026)**	**0.212 (0.052)**	**0.334 (0.081)**	**0.471 (0.110)**
Baseline	0.071 (0.011)	0.149 (0.022)	0.234 (0.034)	0.327 (0.049)

Table 4: Forecast bias multiplied by 10^2 for out-of-sample sets in sME (sMdE).

	Out t+1	Out t+2	Out t+3	Out t+4
Naïve	0.352 **(0.000)**	0.857 **(0.000)**	1.010 (7.335)	1.167 (12.447)
M1	0.317 (-4.690)	0.785 (-5.252)	0.903 (-6.199)	1.024 (-7.295)
M2	**-0.242** (-4.598)	-0.345 (-5.185)	**-0.208** (-5.336)	**-0.068** (-7.914)
M3	-1.259 (-4.006)	-2.367 (-3.909)	-3.826 (-3.101)	-5.282 (-3.627)
M4	-2.041 (-4.508)	-4.230 (-4.693)	-6.769 (-5.004)	-9.307 (-5.672)
mean	-0.807 (-4.474)	-1.539 (-4.853)	-2.475 (-5.546)	-3.408 (-6.134)
median	-0.832 (-4.589)	-1.715 (-5.068)	-2.669 (-5.094)	-3.614 (-6.241)
SVM-based ensemble	0.309 (-3.013)	**0.155** (-2.403)	-0.610 **(-1.310)**	-2.365 **(-2.652)**
Baseline	0.002 (-1.246)	-0.713 (-1.123)	-1.961 (-0.939)	-3.193 (-1.635)

based ensemble, followed by the Exponential Smoothing, then model **M4** and combinations of models.

The SVM-based ensemble approach is allowed to select among the different forecasting models at each forecast origin, and therefore, it is rather more flexible to adapt to stochastic or structural changes in the SKUs. This fact explains why SVM-based ensemble outperform all the considered alternatives in forecast accuracy.

5. Conclusions

This study proposes a novel SVM-based ensemble approach for automatic time series forecasting. Since forecasting models shape business decisions at different levels within companies, this paper aims at enhancing the power of forecasting techniques by proposing a new approach blending standard criteria, like the Schwartz Bayesian Criterion (SBC), with AI techniques, SVM in particular, in a context of supply chain forecasting.

The procedure consists of selecting the best forecast available from a menu of choices at each point in time by means of a SVM trained in a feature space that embeds the most recent information, forecasts, the relative performance and parameters of the models involved. As far as the authors are concerned, this is the first time SVM are used in this context in this particular way.

The approach is empirically applied to a leading household and personal care UK manufacturer with 229 weekly Stock Keeping Units (SKU) to forecast, with a horizon of 1 to 4 weeks ahead. Findings suggest that: i) Exponential Smoothing techniques are very good in this context both in terms of forecast accuracy and minimization of bias, maybe a reason why this is the technique most used in industry and business; ii) simple combination of forecasts (like mean and median) do not help much in this regard; iii) SVM classification techniques certainly manage to improve the forecasting results, both in terms of errors and bias.

6. Acknowledgements

This work was supported by the European Regional Development Fund and Spanish Government (MINECO/FEDER, UE) under the project with reference DPI2015-64133-R and by the Vicerrectorado de Investigación y Política Científica from UCLM by DOCM 31/07/2014 [2014/10340].

7. References

Adya, M., Collopy, F., Armstrong, J., and Kennedy, M. (2001). Automatic identification of time series features for rule-based forecasting. *International Journal of Forecasting*, 17(2):143 – 157.

Barone, D., Mylopoulos, J., Jiang, L., and Amyot, D. (2010a). The business intelligence model: Strategic modelling.

Barone, D., Yu, E., Won, J., Jiang, L., and Mylopoulos, J. (2010b). Enterprise modeling for business intelligence. pages 31–45.

Billah, B., King, M. L., Snyder, R. D., and Koehler, A. B. (2006). Exponential smoothing model selection for forecasting. *International Journal of Forecasting*, 22(2):239 – 247.

Cao, L., Chua, K. S., Chong, W., Lee, H., and Gu, Q. (2003). A comparison of pca, kpca and ica for dimensionality reduction in support vector machine. *Neurocomputing*, 55(1):321–336.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559 – 583.

Collopy, F. and Armstrong, J. S. (1992). Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations. *Management Science*, 38(10):1394–1414.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Davenport, T. H. (2006). Competing on analytics. *harvard business review*, 84(1):98.

Davenport, T. H. and Harris, J. G. (2007). *Competing on analytics: the new science of winning*. Harvard Business Press.

Duin, R. and Loog, M. (2004). Linear dimensionality reduction via a heteroscedastic extension of lda: the chernoff criterion. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):732–739.

Fildes, R. (1989). Evaluation of aggregate and individual forecast method selection rules. *Management Science*, 35(9):1056–1065.

Fildes, R. and Petropoulos, F. (2015). Simple versus complex selection rules for forecasting many time series. *Journal of Business Research*, 68(8):1692 – 1701. Special Issue on Simple Versus Complex Forecasting.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914.

Garcia, F. T., Villalba, L. J. G., and Portela, J. (2012). Intelligent system for time series classification using support vector machines applied to supply-chain. *Expert Systems with Applications*, 39(12):10590 – 10599.

Hira, Z. M. and Gillies, D. F. (2015). A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015.

Homaie-Shandizi, A.-H., Nia, V. P., Gamache, M., and Agard, B. (2016). Flight deck crew reserve: From data to forecasting. *Engineering Applications of Artificial Intelligence*, 50:106–114.

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

Jeston, J. and Nelis, J. (2008). *Business Process Management: Practical Guidelines to Successful Implementations*. Routledge.

Joachims, T. (2002). *Learning to classify text using support vector machines: Methods, theory and algorithms*. Kluwer Academic Publishers.

Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134.

Kourentzes, N., Barrow, D. K., and Crone, S. F. (2014). Neural network ensemble operators for time series forecasting. *Expert Systems with Applications*, 41(9):4235 – 4244.

Lenzerini, M., Vassiliou, Y., Vassiliadis, P., and Jarke, M. (2003). *Fundamentals of data warehouses*. Springer.

Li, C. and Hu, J.-W. (2012). A new arima-based neuro-fuzzy approach and swarm intelligence for time series forecasting. *Engineering Applications of Artificial Intelligence*, 25(2):295–308.

Ljung, G. and Box, G. (1978). On a measure of a lack of fit in time series models. *Biometrika*, 65(2):297–303.

Lu, C.-J. and Kao, L.-J. (2016). A clustering-based sales forecasting scheme by using extreme learning machine and ensembling linkage methods with applications to computer server. *Engineering Applications of Artificial Intelligence*, 55:231–238.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7.

Müller, K.-R., Smola, A. J., Rätsch, G., Schölkopf, B., Kohlmorgen, J., and Vapnik, V. (1997). Predicting time series with support vector machines. In *International Conference on Artificial Neural Networks*, pages 999–1004. Springer.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Pedregal, D. and Taylor, C. (2012). A flexible and general state space toolbox for matlab. In *System Identification, Environmental Modelling, and Control System Design*, pages 615–636. Springer Verlag.

Pedregal, D. and Trapero, J. (2012). The power of ecotool matlab toolbox. In *Industrial engineering: innovative networks*, pages 319–328. Springer.

Pegels, C. C. (1969). Exponential forecasting: Some new variations. *Management Science*, 15(5):311–315.

Petropoulos, F., Makridakis, S., Assimakopoulos, V., and Nikolopoulos, K. (2014). 'horses for courses' in demand forecasting. *European Journal of Operational Research*, 237(1):152 – 163.

Plattner, H. (2009). A common database approach for oltp and olap using an in-memory column database. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 1–2. ACM.

Poler, R. and Mula, J. (2011). Forecasting model selection through out-of-sample rolling horizon weighted errors. *Expert Systems with Applications*, 38(12):14778 – 14785.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*.

Sheikh, N. (2013). *Implementing Analytics: A Blueprint for Design, Development, and Adoption*. Newnes.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International Journal of Forecasting*, 16(4):437 – 450. The M3- Competition.

Taylor, J. (2011). *Decision Management Systems: A Practical Guide to Using Business Rules and Predictive Analytics*. Pearson Education.

Trapero, J. R., Kourentzes, N., and Fildes, R. (2012). Impact of information exchange on supplier forecasting performance. *Omega*, 40(6):738 – 747. Special Issue on Forecasting in Management Science.

Trapero, J. R., Kourentzes, N., and Fildes, R. (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society*, 66(2):299–307.

Wang, W., Pedrycz, W., and Liu, X. (2015). Time series long-term forecasting model based on information granules and fuzzy clustering. *Engineering Applications of Artificial Intelligence*, 41:17–24.

Wang, X.-Y., Zhang, B.-B., and Yang, H.-Y. (2013). Active svm-based relevance feedback using multiple classifiers ensemble and features reweighting. *Engineering Applications of Artificial Intelligence*, 26(1):368–381.

Yu, L., Dai, W., and Tang, L. (2016). A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. *Engineering Applications of Artificial Intelligence*, 47:110–121.