Expert Systems with Applications 160 (2020) 113729

Contents lists available at ScienceDirect

Expert Systems with Applications

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

SVR-FFS: A novel forward feature selection approach for high-frequency time series forecasting using support vector regression

José Manuel Valente [a], Sebastián Maldonado [b,c,*]

[a] Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Mons. Álvaro del Portillo 12455 Las Condes, Santiago, Chile
[b] Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Santiago, Chile
[c] Instituto Sistemas Complejos de Ingeniería (ISCI), Chile

ARTICLE INFO

Article history:
Received 12 August 2019
Revised 5 July 2020
Accepted 5 July 2020
Available online 13 July 2020

Keywords:
Support vector regression
Feature selection
Forecasting
Energy load forecasting
Automatic model specification

ABSTRACT

In this paper, we propose a novel support vector regression (SVR) approach for time series analysis. An efficient forward feature selection strategy has been designed for dealing with high-frequency time series with multiple seasonal periods. Inspired by the literature on feature selection for support vector classification, we designed a technique for assessing the contribution of additional covariates to the SVR solution, including them in a forward fashion. Our strategy extends the reasoning behind Auto-ARIMA, a well-known approach for automatic model specification for traditional time series analysis, to kernel machines. Experiments on well-known high-frequency datasets demonstrate the virtues of the proposed method in terms of predictive performance, confirming the virtues of an automatic model specification strategy and the use of nonlinear predictors in time series forecasting. Our empirical analysis focus on the energy load forecasting task, which is arguably the most popular application for high-frequency, multi-seasonal time series forecasting.

1. Introduction

Feature selection is a very important element in designing expert systems, especially in supervised learning and forecasting (Jiang, Chin, Wang, Qu, & Tsui, 2017; Sanz-García, Fernández-Ceniceros, Antonanzas-Torres, Pernia-Espinoza, & de Pison, 2015; Zbikowski, 2015). Identifying and selecting the relevant information in machine learning has several advantages, such as a better generalization of new objects, a better understanding of the process that generated the data, and the reduction of data collection costs (Maldonado, Pérez, & Bravo, 2017).

Support vector regression (SVR) (Smola & Schólkopfmola & Schólkopf, 1998) is a well-known kernel method that has been successfully used for time series forecasting, thanks to its ability to construct nonlinear regressors (Karmy & Maldonado, 2019; Sanz-García et al., 2015; Wu, Tzeng, & Lin, 2009). This method, however, is not able to derive the feature importance automatically (Maldonado & López, 2018). In this sense, SVR and ARIMA are similar, and a search strategy can be useful in defining an automatic model specification approach for kernel-based regression.

This paper provides a multi-purpose methodological development designed for high-frequency, multi-seasonal time series data, with a special focus on energy load forecasting. This task is arguably the most popular application in the literature for this type of data (Anderson & Torriti, 2018; Taylor, 2010; Son & Kim, 2015). Feature selection via machine learning can be extremely useful in this domain for the following reasons:

- A small gain in predictive performance can be extremely profitable. Knowing as accurately as possible how much electricity will be consumed each hour is a must for providing a certain level of service and reducing operative/production costs. It was estimated that an extra 1% on the forecasting error increases these costs up to £10 million per year (Gross & Galiana, 1987).
- Energy load forecasting problem is intrinsically nonlinear (Henley & Peirson, 1997), and therefore kernel machines could be very useful for boosting prediction. The main disadvantage of kernel methods is that they are prone to overfitting in applications with few data samples, which is usually the case in time series forecasting. Although some studies argue that complex models tend not to perform as well as simpler ones in this domain (see e.g. Makridakis & Hibon, 2000), feature selection

* Corresponding author at: Department of Management Control and Information Systems, School of Economics and Business, University of Chile, Santiago, Chile.
E-mail addresses: jmvalente@miuandes.cl (J.M. Valente), sebastianm@fen.uchile.cl (S. Maldonado).

https://doi.org/10.1016/j.eswa.2020.113729
0957-4174/

has shown to be extremely useful at alleviating the risk of overfitting in machine learning (Guyon, Gunn, Nikravesh, and Zadeh, 2006).

- Understanding which seasonal patterns are relevant for a given forecasting task can be useful for making better managerial decisions, especially medium and long term decisions (Martínez, Frías, Pérez-Godoy, & Rivera, 2018). A model that is able to identify these patterns automatically can facilitate gaining important insights into the application.
- In the case of multivariate time series, exogenous data is often collected from external sources (Nicholson, Matteson, & Bien, 2017). Variable acquisition costs can be reduced through an adequate feature selection process that discards variables that do not improve performance (Maldonado et al., 2017).

Our contribution is twofold. First, we establish a parallel between feature selection in machine learning and the automatic model specification process for time series analysis; theory proposed by Hyndman and Khandakar (2008) with their well-known approach Auto-ARIMA. Variables in time series analysis are essentially lagged versions of the target variable over successive intervals, and the choice of these lags is usually arbitrary. Auto-ARIMA, however, finds the relevant lags in traditional time series analysis automatically by performing a search strategy, thus evaluating various ARIMA models in terms of Akaike information criterion (AIC) (Hyndman & Khandakar, 2008). We extend this idea to forecasting via machine learning.

Our second contribution is the development of an efficient and automated approach, called SVR-FFS (SVR – Forward Feature Selection), for kernel-based forecasting and model specification. Our strategy extends the reasoning behind the well-known SVM-RFE (Support Vector Machine – Recursive Feature Elimination) approach for binary classification, in which those variables whose removal has less impact on the objective function of the SVM problem are removed iteratively. Our proposal, in contrast, selects variables in a forward fashion, avoiding solving high-dimensional SVR problems with a high degree of redundancy.

Our proposal was applied to seven energy load forecasting datasets of different nature. This task is very challenging since it faces high-frequency data, which translates into large running times, especially for machine learning algorithms. Additionally, effective models must consider multiple seasonal patterns, such as daily, weekly, and yearly seasonality. Our experiments have permitted us to conclude that our proposal is the most suitable approach for dealing with such complex data.

The remainder of this paper is organized as follows: Section 2 provides a review of prior work on energy load forecasting, SVR for time series analysis, and forward feature selection for SVMs. The proposed SVR-FFS method for automatic model specification is presented in Section 3. Experimental results on energy load forecasting datasets are discussed in Section 4. Finally, in Section 5 the key conclusions are summarized, and future developments are proposed.

2. Previous work

There are many approaches designed to deliver a solution to the energy load forecasting problem, ranging from more traditional techniques, such as exponential smoothing with seasonality and ARIMA models (Hyndman & Khandakar, 2008; Taylor, 2010, 2003), to more sophisticated machine learning models (Ng, 2017; Sapankevych & Sankar, 2009). They all contribute to the development and understanding of algorithms which bring insights that contribute to reaching better forecasting in the different applications and variants of electrical consumption.

With this given, and due to the enormous sustainability and economic benefits that forecasting within the energy load context imply, and the challenging nature of such a problem, there has been a huge amount of research focused on finding the best approach to this matter depending on the planning horizon, such as short (STLF), medium (MTLF), and long term load forecasting (LTLF).

Prediction of electric demand in residential areas was studied in STLF by Son and Kim (2015), who combined SVR and fuzzy-rough feature selection with particle swarm optimization. Several studies used artificial neural networks (ANN) for this task with mixed results, mostly due to the known overfitting issue (Hippert, Pedreira, & Souza, 2001; Lee, Cha, & Park, 1992; Srinivasan, 1998). Traditional methods such as autoregressive moving average (Huang & Shih, 2003) and exponential smoothing (Christiaanse, 1971) are still used instead to face this challenge thanks to their simplicity and good predictive performance.

Regarding the MTLF problem, Hu, Bao, Chiong, and Xiong (2015) and Hu, Bao, and Xiong (2013) used multi-output SVR and memetic algorithms for feature selection. In Chikobvu and Sigauke (2013), the authors incorporated the influence of the temperature on the daily electric peak using segmented regressions. In Srinivasan (2008), mid term load forecasting was presented through GMDH networks for a monthly based energy demand prediction, leading to effective and accurate models. A hybrid method proposed by Amjady and Keynia (2008) uses a combination of different techniques to build up an evolutionary algorithm.

For the LTLF application, a combination of k-nearest neighbors, mutual information, and nonparametric noise estimation was proposed in Sorjamaa, Hao, Reyhani, Ji, and Lendasse (2007) for time series analysis and input selection. In another study, particle swarm optimization (PSO) was used, leading to superior performance (AlRashidi & El-Naggar, 2010).

SVR for time series analysis is formalized next in Section 2.1, while some feature selection methods for SVMs that are relevant to our proposal are discussed in Section 2.2.

2.1. SVR for time series

SVR is a machine learning model that has the flexibility of balancing the trade-off between minimizing the empirical error and the complexity of the resulting fitted function, reducing the risk of overfitting (Smola & Schólkopfmola & Schólkopf, 1998). Its well-deserved popularity is due to this property, which usually leads to the best predictive results in time series analysis (Sapankevych & Sankar, 2009).

The raw input for this task consists of a series of time-dependent values Y. A matrix of covariates X is constructed by computing lagged versions of the Y vector. Then, the m inputs for the supervised learning model can be seen as tuples $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \ldots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x}_i$ is the vector containing all the lags considered for period i, and with m being the total number of periods considered for training. The ε-SVR method solves the following problem (Smola & Schólkopfmola & Schólkopf, 1998):

$$\min_{\mathbf{w},b,\xi,\xi^*} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\mathbf{e}^\top(\xi+\xi^*) \quad \text{s.t. } \mathbf{y} - (A\mathbf{w}+b\mathbf{e}) \leqslant \varepsilon\mathbf{e}+\xi,\ \xi \geq 0,$$
$$(A\mathbf{w}+b\mathbf{e}) - \mathbf{y} \leqslant \varepsilon\mathbf{e}+\xi^*,\ \xi^* \geq 0, \tag{1}$$

where $A = [\mathbf{x}_1\,\mathbf{x}_2\ldots\mathbf{x}_m]^\top \in \Re^{m\times n}, \mathbf{y} = (y_1, y_2, \ldots, y_m) \in \Re^m, \mathbf{e} \in \Re^m$ is a vector of ones, $\xi, \xi^* \in \Re^m$ are slack vectors that indicate whether or not the samples are inside the ε-insensitive tube, and $C > 0$ is the hyper-parameter that controls the trade-off between complexity and empirical error minimization (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997).

The introduction of kernel functions lead to a nonlinear approach through an implicit mapping of the input data to a higher-dimensional space. In this study, the linear kernel (Eq. (2)) and the Radial Basis Function (RBF) were considered. With the inclusion of kernel functions, the dual form of ε-SVR (Eq. (1)) results in the following problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & \mathbf{y}^\top(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) - \varepsilon\mathbf{e}^\top(\boldsymbol{\alpha}+\boldsymbol{\alpha}^*) - \frac{1}{2}(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*)^\top K(A,A^\top)(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) \\ \text{s.t. } & \mathbf{e}^\top(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*) = 0,\ 0 \leqslant \boldsymbol{\alpha} \leqslant C\mathbf{e},\ 0 \leqslant \boldsymbol{\alpha}^* \leqslant C\mathbf{e}, \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ are the dual variables related to the constraints of Eq. (1), while $K(A,A^\top) \in \Re^{m\times m}$ is the kernel matrix whose elements are $k_{is} = \mathcal{K}(\mathbf{x}_i,\mathbf{x}_s)$. The RBF kernel has the following form:

$$\mathcal{K}(\mathbf{x}_i,\mathbf{x}_s) = \exp\left(-\gamma||\mathbf{x}_i-\mathbf{x}_s||^2\right) \tag{3}$$

where $\gamma = \frac{1}{2\sigma^2}$, and $\sigma > 0$ is the hyper-parameter that controls the shape of the kernel.

In addition to its application in energy load forecasting, SVR has been used in time series analysis in a wide variety of domains, such as finance, transportation systems, wind speed prediction, and sales forecasting, among others (S.Dhiman et al., 2019; Karmy & Maldonado, 2019; Kazem, Sharifi, Hussain, Saberi, & Hussain, 2013; Xu, Chan, & Zhang, 2019; Zbikowski, 2015).

2.2. Forward feature selection for SVMs

As discussed above in the introductory section, feature selection is a relevant topic in machine learning since it leads to several advantages in terms of predictive performance, interpretation, training efficiency, and cost reduction. Feature selection approaches for SVMs can be classified into three groups: filter, wrapper, and embedded methods (Guyon et al., 2006).

Filter methods assess the variable contribution before the learning process, filtering out irrelevant covariates, for example by using statistical measures (Guyon et al., 2006). The main issue with filter approaches in time series analysis is that the correlation between lags is not taking into account, and this is an important issue since lags are usually highly correlated. Another disadvantage is that the interaction between covariates and the model is also not considered. However, filter methods are fast strategies that provide good results in classification tasks (Fleuret, 2004; Pal & Foody, 2010; Song, Smola, Gretton, Bedo, & Borgwardt, 2012).

Wrapper methods evaluate a given predictive method using various subsets of input variables, computing their performance and selecting the one with the best predictive capabilities. Since exhaustive evaluation of possible subsets is intractable, even for medium-size problems, efficient heuristics and meta-heuristics are usually used (Guyon et al., 2006). For example, Gheyas and Smith (2010) proposed a wrapper method that combines simulated annealing with genetic algorithms to develop a method (SAGA) in order to improve convergence.

The main issue with the wrapper strategies is that they usually lead to very large training times due to the size of the SVM models. In order to remedy this problem, hybrid filter-wrapper approaches have been proposed. For example, Peng, Long, and Ding (2005) used the minimum redundancy-maximum relevance (mRMR) as a filter method for the first stage, and subsequently performed backward/forward feature selection to create a compact set of final selected features.

Embedded methods define model-based evaluation metrics for assessing the contribution of the variables, or penalizing the use of features during model training (Ghaddar & Naoum-Sawaya, 2018; Zhao, Chen, Pedrycz, & Wang, 2019). The SVM-RFE algorithm fits in the first category, which is relevant for our proposal. There are also several approaches that perform feature penalization have been reported in the literature (Lal, Chapelle, Weston, & Elisseeff, 2006; Maldonado & López, 2018; Maldonado & Weber, 2010). SVM-RFE uses a backward elimination strategy, removing those covariates whose elimination has less impact on the objective function of the SVM model (Guyon, Weston, Barnhill, & Vapnik, 2002).

Forward selection can be used for developing embedded strategies while avoiding the issue of computing large SVM models (Ambroise & McLachlan, 2002; Lin & Wei, 2005; Liu et al., 2007). For example, Weston et al. (2001) used forward selection based on SVM metrics for the bounds on the leave-one-out error (Span-bound and RW-bound). The approach called grafting (Perkins, Lacker, & Theiler, 2003) computes the gradient of a loss function that includes three different regularizers for feature penalization.

Although embedded feature selection has not been discussed in the context of time series analysis to the best of our knowledge, some feature selection methods have been proposed using other machine learning techniques. For example, Crone and Kourentzes (2010) proposed an hybrid filter-wrapper for neural networks with positive results.

3. Proposed SVR and forward feature selection method

In this section we present a novel strategy for automatic model specification of high-frequency time series using SVR. The main idea is to perform an embedded forward selection, adapting the SVM-RFE algorithm proposed for backward feature elimination in binary classification tasks. Starting with the first lagged variable, which is usually the most important one in most application, the goal is to include those lags and seasonal patterns that have a larger impact in the objective function of the SVR method in a forward manner. The reasoning behind this strategy is that the time series model can be specified automatically without performing expensive search strategies that require training models with large data matrices, which would be the case in exhaustive search (Auto-ARIMA approach) or in backward feature elimination (SVM-RFE algorithm).

The main advantages of our proposal are the following:

- It allows the use of kernel functions for capturing the inherent nonlinear patterns that are usually present in high-frequency time series, for example, in energy load forecasting.
- It can be extended naturally to multivariate time series analysis. Exogenous variables can be considered as additional covariates, and the model will identify which ones are relevant for prediction using the same selection process.
- It has all the theoretical advantages of embedded feature selection: it takes both the correlations between covariates and the interaction between the variables and the model into account. The first point is of utmost importance in high-frequency time series analysis since adjacent lags tend to be highly correlated, and filter methods, such as using the partial autocorrelation function (PACF), tend to be ineffective. In order to illustrate this issue, the PACF plot for one of the high-frequency data sets used in the experimental section (the GB data set) is presented in Fig. 1, considering only the first 352 lags. It can be seen in this plot that most partial autocorrelations are significantly different from zero, and, therefore, few attributes could be discarded if this measure were used as a filter strategy. Notice that the thresholds for statistical significance are represented in the PACF plot by blue dashed lines, and the lags that fall outside the boundaries defined by these lines can be considered to be statistically relevant.

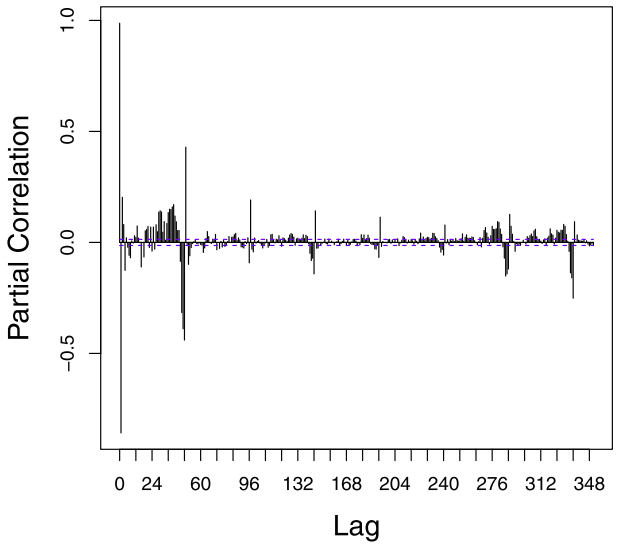


Fig. 1. PACF plot for the first 352 lags.

- The forward selection strategy has the advantage of being more efficient than any other wrapper/embedded strategy. Starting with a single covariate and including only relevant features iteratively avoid the construction of large SVR models with a high level of redundancy. In agreement with the previous example, a yearly seasonality can be identified in the lag $48 * 365 = 17,520$ (48 half hours = one day). A year of training samples leads to a dense data matrix X of more than 17,000 rows and columns, which is large for an SVR problem. Furthermore, a wrapper method can evaluate only few subsets of these 17,000 features in a reasonable training time. In particular, a backward approach such as SVM-RFE has to train first a model with all variables, which is almost intractable computationally.

The proposed method is formalized next. First, it can be noticed in Eq. (2) that the Euclidean norm of the weight vector that is minimized in Eq. (1) can be rewritten as follows, using the dual variables and the kernel functions:

$$W^2(\boldsymbol{\alpha}, A) = \frac{1}{2}\|\mathbf{w}\|^2 = -\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top K(A, A^\top)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*). \quad (4)$$

The SVM-RFE method for binary classification (Guyon et al., 2006) suggests that the variable in A that has less impact in $W^2(\boldsymbol{\alpha}, A)$ should be removed. Then, for all the variables in A, A_{-p} is computed as the data matrix with the current variables but p. The attribute to be discarded in the backward elimination strategy is the one with lower value for the following measure:

$$\|W^2(\boldsymbol{\alpha}, A) - W^2(\boldsymbol{\alpha}, A_{-p})\| \quad (5)$$

We propose a simplified contribution metric (CM) which is equivalent to $W^2(\boldsymbol{\alpha}, A)$ for feature ranking:

$$CM(\boldsymbol{\alpha}, A) = (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top K(A, A^\top)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) \quad (6)$$

As a forward selection method, the proposed approach adds those lags whose inclusion leads to a minimum value of this metric to a subset of relevant covariates. Following the notation used by Song et al. (2012), the full set of available lags is denoted by $\mathcal{L}$, while $\mathcal{L}^\dagger$ represents an ordered subset of $\mathcal{L}$ with only the relevant lags to be included in the SVR model. At each iteration, the subset of lags to be included in $\mathcal{L}^\dagger$ is represented by $\mathcal{I}$.

The proposed algorithm computes $CM(\boldsymbol{\alpha}, A)$ using all lags in $\mathcal{L}^\dagger$, plus an additional lag p from $\mathcal{L} \setminus \mathcal{L}^\dagger$. Formally, the contribution metric $CM_p(\boldsymbol{\alpha}, A_p)$ has the following form:

$$CM_p(\boldsymbol{\alpha}, A_p) = (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^\top K(A_p, A_p^\top)(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*), \quad (7)$$

where $A_p = [\mathbf{x}_1^{(p)}\, \mathbf{x}_2^{(p)} \ldots \mathbf{x}_m^{(p)}]^\top \in \Re^{m \times (|\mathcal{L}^\dagger|+1)}$, while $|\mathcal{L}^\dagger|$ represents the cardinality of $\mathcal{L}^\dagger$ and $\mathbf{x}_i^{(p)}$ the vector of covariates that includes the lags in $\mathcal{L}^\dagger$ plus the additional lag p.

Notice that a natural adaptation of SVM-RFE algorithm for forward selection would be to compute the difference between $CM(\boldsymbol{\alpha}, A)$ and $CM_p(\boldsymbol{\alpha}, A_p)$ and select the attribute p that the value of

$$\|CM(\boldsymbol{\alpha}, A) - CM(\boldsymbol{\alpha}, A_p)\| \quad (8)$$

is the largest. However, $CM(\boldsymbol{\alpha}, A)$ does not vary for all p in $\mathcal{L} \setminus \mathcal{L}^\dagger$. Therefore, we can simply select the attribute with the minimum value for $CM_p(\boldsymbol{\alpha}, A_p)$.

We consider the following approximation proposed in Guyon et al. (2006) for reducing the computational complexity of the algorithm: instead of re-training the model for every A_p, we assume that the solution obtained when using the already selected lags $\mathcal{L}^\dagger$ is close to the one that would be obtained with the inclusion of p. Therefore, a single $\boldsymbol{\alpha}$ solution is considered for the feature evaluation process (the one obtained with the variables in $\mathcal{L}^\dagger$).

The proposed forward selection strategy for SVR and time series analysis (SVR-FFS) is described in Algorithm 1.

Algorithm 1 Forward feature selection for SVR and time series analysis (SVR-FFS)

Input: The original set of lags $\mathcal{L}$ (excluding the first one)
Output: An ordered subset of relevant lags $\mathcal{L}^\dagger$
1: $\mathcal{L}^\dagger \leftarrow \{1\}$.
2: **repeat**
3: $\quad \boldsymbol{\alpha} \leftarrow$ SVR Training on $\mathcal{L}^\dagger$
4: $\quad \mathcal{I} \leftarrow \text{argmin}_{\mathcal{I}} \sum_{p \in \mathcal{I}} CM_p(\boldsymbol{\alpha}, A_p),\ \mathcal{I} \subset \mathcal{L}$
5: $\quad \mathcal{L} \leftarrow \mathcal{L} \setminus \mathcal{I}$
6: $\quad \mathcal{L}^\dagger \leftarrow (\mathcal{L}^\dagger, \mathcal{I})$
7: **until** $|\mathcal{L}^\dagger| = r$.

The $\mathcal{L}^\dagger$ set is initialized with one lag that is expected to be relevant in the SVR training (see step 1 of Algorithm 1). The natural choice would be to consider the first lagged variable.

At each iteration, the SVR-FFS approach trains an ε-SVR model with the lags in $\mathcal{L}^\dagger$ (step 3), and evaluates all lags $p \in \mathcal{L}$ using $CM_p(\boldsymbol{\alpha}, A_p)$ (step 4). Then, the $|\mathcal{I}|$ lags that lead to the smallest values of CM_p are included in $\mathcal{L}^\dagger$ and excluded from $\mathcal{L}$ (step 5-6).

At each iteration, increasing the size of $\mathcal{I}$ is recommended in order to speed up the learning process. Following the suggestion made for the SVM-RFE strategy in Guyon et al. (2006), the cardinality of $\mathcal{I}$ is doubled at each iteration, starting with $|\mathcal{I}| = 1$. In this way our approach should not require more than 10–15 iterations to find an adequate model.

As stopping criterion we set a predefined number r of selected lags, i.e. $|\mathcal{L}^\dagger| = r$ (see step 7 of Algorithm 1).

As a machine learning method, SVR-FFS requires an appropriate validation strategy for parameter tuning in order to avoid overfitting. Apart from the traditional SVR hyperparameters $\{C, \sigma, \varepsilon\}$, SVR-FFS parameters r (the final number of selected lags) and $|\mathcal{I}|$ (the number of lags to be included in $\mathcal{L}^\dagger$ at each iteration) can be considered as additional tuning parameters.

We suggest applying the SVR-FFS algorithm on a training set under different $\{C, \sigma, \varepsilon, |\mathcal{I}|\}$ configurations as empirical

framework, and evaluating the performance in terms of a suitable evaluation metric on the validation set. This process can be repeated several times on different training/validation subsets, following the guidelines of the rolling forecast origin approach (Hyndman & Athanasopoulos, 2013). For each training/validation configuration, the average value of a suitable measure can be computed and monitored at each iteration in order to determine the value for r. Finally, the best $\{C, \sigma, \varepsilon, |\mathcal{I}|, r\}$ configuration, i.e. the one with the lowest average error, is used for testing. Training and validation sets are then combined to train a final SVR-FFS model with the best parameter configuration, and applied to the test set for comparison with alternative methods.

4. Experimental results on energy load forecasting datasets

The proposed SVR-FFS algorithm was applied to six energy load forecasting data sets and the results were compared with well-known alternative time series strategies. Among the six data sets, five are high-frequency load forecasting tasks, while the remaining data set considers daily electric load with additional exogenous variables. On the one hand, experiments on high-frequency load forecasting are very important since we evaluate more than 17,000 candidate lags in order to model all seasonal patterns adequately, leading to high-dimensional problems. On the other hand, although daily electric load forecasting is rather a low-dimensional task with less than 400 candidate lags, the inclusion of exogenous variables, such as the forecasted temperature for the given day, is extremely valuable for assessing the model's ability to capture nonlinear patterns. According to Chikobvu and Sigauke (2013), temperature and energy consumption have a nonlinear relationship, which could be identified adequately via kernel methods.

First, the data sets are presented in Section 4.1. Next, the experimental framework is discussed in Section 4.2. Finally, a summary of the results obtained with all data sets is reported in Section 4.3.

4.1. Description of data sets

The following data sets were included in this study:

- *GB*: Great Britain half-hour electric load data set (Taylor, 2010).
- *E&W*: England and Wales half-hour electric load data set (Anderson & Torriti, 2018).
- *IO14_DEM*: Great Britain half-hour electric load data set, which excludes the net demand from imports/exports and pump storage demand, and includes transmission losses and station transformer load (Hawkins, Eager, & Harrison, 2011).
- *IO14_TGSD*: Great Britain half-hour total gross system demand (Bejan, Gibbens, & Kelly, 2012).
- *France* Import(+)/Export(−): Net half-hour exported load from Great Britain towards France (Taylor, 2010). In order to avoid negative values, we transformed this data set by adding the minimum value of the series to each observation.
- *Chile*: Chilean hourly real systemic electric demand, managed by the autonomous entity, *Coordinador Eléctrico Nacional*, in charge of coordinating the operations for all the entities involved in the national electric system.
- *GB Daily*: The Great Britain half-hour electric load data set aggregated at a daily level. This aggregation is done by computing the sum of all half-hour load consumptions on a given day. This data set also includes the forecast for the average temperature of a given day, and a dummy variable that indicates whether or not the given day is a holiday.

All data sets consider the period between April, 2001 and October, 2008, except for the *Chile* data set, which includes the period from 5th of May, 2017 to the 24th of October, 2018.[1] Table 1 presents the relevant meta-data for all the time series before preprocessing, indicating the maximum, minimum, and average values for the series, and their respective standard deviations.

4.2. Experimental framework

In time series analysis, an out-of-time validation strategy should be used, such as the rolling forecast origin (Hong, 2013; Hyndman & Athanasopoulos, 2013). Model validation includes training, validation, and test stages as presented in Fig. 2.

First, the training ($\mathcal{TR}$), validation ($\mathcal{V}$), and test ($\mathcal{T}$) subsets are defined for each data set. The number of periods considered for each set and the number of candidate lags $|\mathcal{L}|$ are presented in Table 2. Notice that the final sample size for the covariate matrix corresponds to the number of periods included for each set minus the largest lag in $\mathcal{L}^{\dagger}$.

The proposed SVR-FFS method is compared with the following alternative forecasting methods:

1. The one-step non-seasonal and seasonal naïve forecast approaches, in which the forecast is computed as the observed value from the prior period and the observed value from the previous seasonal pattern, respectively. In case of multiple seasonality, only the daily seasonality is considered for the latter model.
2. Traditional Box-Jenkins (ARIMA) models that take seasonal factors into account, and double-seasonal Holt-Winters. These strategies are well-established approaches for the load forecasting task (Hong, 2013; Mbamalu & El-Hawary, 1993; Taylor, 2010). The performance of these approaches relies heavily on an adequate model specification. For this task, we used the automatic model specification process proposed by Hyndman and Khandakar (2008), in which the relevant lags can be identified automatically by using the Akaike's Information Criterion (AIC). For the Box-Jenkins approach, we used the model called Auto-ARIMA (Hyndman & Khandakar, 2008), and the TBATS function is used for the Holt-Winters method (De Livera, Hyndman, & Snyder, 2011).
3. Various SVR configurations are considered using naïve subsets of lags based on the various seasonal patterns that are present in these data sets. These models are presented in Table 3. The various experiments are designed in an incremental fashion, and *Lags Exp. j* denotes the set of lags included in Experiment *j*.

The index related to d on Table 3 represents the lag that is being considered. For example, d_1 is the observation just before the one that is being predicted. Index s_k, with $k = \{1, 2, 3\}$ refers to the seasonal pattern, which depends on the data set, as described in Table 4.

Notice that the eighth SVR configuration in Table 3 is not available for the *GB Daily* data set since only two seasonal patterns are present. For this data set, all experiments were performed both with and without considering exogenous variables, and the best performance is reported for all methods. For consistency, $\boldsymbol{s_2}$ corresponds to weekly seasonality for all data sets since experiments 5 to 7 consider seasonal autoregressive processes of order 2 to 4 for the weekly seasonality.

As mentioned in the previous section, grid search on a validation set is suggested for parameter tuning for the SVR methods. The hyper-parameters explored for our proposal and the various

[1] https://www.coordinador.cl/sistema-informacion-publica/portal-de-operaciones/operacion-real/demanda-real/.

Table 1
Descriptive information for each data set (MW).

Data Set	Max	Min	Average	Std. Dev.
GB	60,098	20,993	38,260	7,542.46
E&W	54,431	0	34,552	6,896.37
IO14_DEM	60,588	19,296	36,758	7,429.19
IO14_TGSD	60,672	20,427	37,376	7,207.98
France	1,999	−2,021	956.2	998.46
Chile	10,528.86	6.48	8,572.24	814.74
GB Daily	2,406,803	1,325,938	1,836,476	233,517.8

Fig. 2. Model validation strategy.

Table 2
Sample size for the training, validation, and test subsets, and number of candidate lags for each data set.

Data Set	$\lvert\mathcal{TR}\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{L}\rvert$
GB	18,972	1,440	1,440	17,532
E&W	18,972	1,440	1,440	17,532
IO14_DEM	18,972	1,440	1,440	17,532
IO14_TGSD	18,972	1,440	1,440	17,532
France	18,972	1,440	1,440	17,532
Chile	11,472	720	720	8,766
GB Daily	2,711	30	30	365

Table 3
Different naïve SVR configurations for benchmarking.

Experiment	Lags
1	d_1
2	Lags Exp. 1, d_2 , d_3
3	Lags Exp. 2, $d_{s_1}, d_{s_1+1}, d_{s_1+2}, d_{s_1+3}$
4	Lags Exp. 3, $d_{s_2}, d_{s_2+1}, d_{s_2+2}, d_{s_2+3}$
5	Lags Exp. 4, $d_{2s_2}, d_{2s_2+1}, d_{2s_2+2}, d_{2s_2+3}$
6	Lags Exp. 5, $d_{3s_2}, d_{3s_2+1}, d_{3s_2+2}, d_{3s_2+3}$
7	Lags Exp. 6, $d_{4s_2}, d_{4s_2+1}, d_{4s_2+2}, d_{4s_2+3}$
8	Lags Exp. 7, $d_{s_3}, d_{s_3+1}, d_{s_3+2}, d_{s_3+3}$

Table 4
Seasonal patterns present in the different data sets.

Data Set	s_1	s_2	s_3
GB	48 (Daily)	336 (Weekly)	17.532 (Yearly)
E&W	48 (Daily)	336 (Weekly)	17.532 (Yearly)
IO14_DEM	48 (Daily)	336 (Weekly)	17.532 (Yearly)
IO14_TGSD	48 (Daily)	336 (Weekly)	17.532 (Yearly)
France	48 (Daily)	336 (Weekly)	17.532 (Yearly)
Chile	24 (Daily)	168 (Weekly)	8.766 (Yearly)
GB Daily	365 (Yearly)	7 (Weekly)	–

naïve SVR configurations are presented in Table 5. Notice that these values vary for all data sets because the grids were adapted for each method and data set in order to maximize predictive performance.

The size of $\mathcal{I}$ for the SVR-FFS algorithm was changed at each iteration in order to achieve faster convergence. The cardinality of $\mathcal{I}$ doubled at each iteration, starting with $|\mathcal{I}| = 1$, leading to a maximum of twelve iterations for each high-frequency data set and of eight iterations for the *GB Daily* data set. Twelve iterations on a high-frequency data set lead to 2048 lags selected, while eight iterations for the *GB Daily* data set lead to 128 lags selected.

Regarding performance measures, the mean absolute percentage error (MAPE), the root mean squared error (RMSE), and the seasonal mean absolute scaled error (MASE_s) were considered in this study. Given a point $t = \{1 \dots T\}$ from the validation or the test set, where O_t is the observed value and F_t is the forecast value, these metrics are computed as follows:

$$MAPE = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{O_t - F_t}{O_t}\right| \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(O_t - F_t)^2}{T}} \tag{10}$$

$$MASE_s = \frac{\frac{\sum_{t=1}^{T}|O_t - F_t|}{T}}{\frac{\sum_{i=s+1}^{m}|Y_i - Y_{i-s}|}{m-s}} \tag{11}$$

For the MASE_s measure, the denominator represents the mean absolute error (MAE) of the one-step seasonal naïve forecast approach. This error is computed in the training set, with $i = 1, \dots, m$. The observed value from the previous seasonal period is used for the seasonal naïve strategy ($F_i = Y_{i-s}$). In our series with multiple seasonal patterns, we consider the daily seasonality for the computation of the latter metric, which is the one with the biggest influence in the predictive performance.

4.3. Result summary

Table 6 summarizes the results for all the methods in terms of MAPE (%) on the test set. For all the SVR approaches, the best hyper-parameter configuration using the RBF kernel is reported, i.e. the one that achieves the lowest MAPE on the validation set. For the proposed SVR-FFS method, results using the linear kernel are also reported for illustrative purposes. The best performance in terms of MAPE (%) is emphasized in bold type. The final row

Table 5
SVR hyper-parameters explored for each data set.

Data Set	ε	C	γ
GB	0.1	$\{100,000; 110,000; \ldots; 150,000\}$	$\{0.0001; 0.3501; \ldots; 1.7501\}$
E&W	0.1	$\{75,000; 85,000; \ldots; 125,000\}$	$\{0.0001; 0.3501; \ldots; 1.7501\}$
IO14_DEM	0.1	$\{120,000; 130,000; \ldots; 170,000\}$	$\{0.0001; 0.3501; \ldots; 1.7501\}$
IO14_TGSD	0.1	$\{85,000; 95,000; \ldots; 135,000\}$	$\{0.0001; 0.3501; \ldots; 1.7501\}$
France	0.1	$\{6000; 16,000; \ldots; 56,000\}$	$\{0.0001; 0.3501; \ldots; 1.7501\}$
Chile	0.1	$\{2500; 4500; \ldots; 10,500\}$	$\{1x10^{-7}; 6x10^{-7}; \ldots; 9.96x10^{-5}\}$
GB Daily	0.1	$\{1x10^{7}; 1.5x10^{7}; 2x10^{7}\}$	$\{0.1; 0.6; \ldots; 5.1\}$

Table 6
Result summary for data sets. Average MAPE (%) in test set.

Experiment	Data Set						
	GB	E&W	IO14_DEM	IO14_TGSD	France	Chile	GB Daily
Naïve	18.36	18.8	18.51	16.12	37.13	7.44	6.45
Naïve	11.44	11.7	11.5	11.44	37.04	4.63	5.18
ARIMA	11.51	11.47	10.99	11.41	37.13	4.40	6.13
ARIMA	5.40	5.52	5.67	5.87	37.45	5.68	4.37
ARIMA	16.78	17.20	16.92	14.88	37.12	7.30	-
TBATS	15.63	8.47	8.28	7.70	42.94	5.65	2.75
SVR Exp. 1	16.04	30.26	27.36	19.60	37.41	7.52	98.00
SVR Exp. 2[g]	19.96	16.95	17.13	16.09	106.36	7.38	67.97
SVR Exp. 3[g]	17.96	965.31	22.11	480.14	20.93	6.64	3.82
SVR Exp. 4[g]	12.17	12.70	16.97	40.18	23.49	4.13	2.81
SVR Exp. 5[g]	34.83	21.01	9.70	187.67	30.80	3.77	2.25
SVR Exp. 6[g]	11.29	10.50	10.67	10.39	25.91	3.48	2.65
SVR Exp. 7[g]	14.54	9.83	11.24	7.12	30.20	3.10	2.56
SVR Exp. 8[g]	8.83	15.61	15.17	14.35	23.09	3.70	-
SVR-FFS	6.58	6.39	6.32	6.41	35.97	5.21	5.14
SVR-FFS	**4.88**	**5.44**	**5.20**	**5.57**	**20.12**	**2.44**	**1.82**
SVR-FFS (FS ratio)	11.68	11.68	11.68	11.68	0.02	1.46	4.38

[a] Non-seasonal naïve approach.
[b] Seasonal naïve approach using seasonal pattern s_1 (Table 4).
[c] Auto-ARIMA function using seasonal pattern s_1 (Table 4).
[d] Auto-ARIMA function using seasonal pattern s_2 (Table 4).
[e] Auto-ARIMA function using seasonal pattern s_3 (Table 4).
[f] Holt-Winter's TBATS function using seasonal patterns s_1 and s_2.
[g] Naïve SVR configurations using the lags reported in Table 3.
[h] Best iteration based on test performance, Linear kernel.
[i] Best iteration based on test performance, RBF kernel.
[j] Dimensionality reduction ratio (%). RBF kernel.

in Table 6 presents the dimensionality reduction ratio for the best configuration of the proposed SVR-FFS method. This ratio is computed as the number of selected attributes by the method divided by the total number of attributes in the search space $|\mathcal{L}|$ (see Table 2).

The *GB Daily* data set leads to a special set of experiments since it includes two exogenous variables: a dummy variable related to the nature of the day (normal or holiday), and the forecast of the average temperature for a given day. Since only two seasonal patterns are present in this data set, the experiments SVR Exp. 8 and Auto-ARIMA using seasonal pattern s_3 are not included.

It can be observed in Table 6 that our SVR-FFS method outperforms all the other approaches, achieving the best MAPE on all the data sets. Additionally, it can be observed that results using the RBF kernel are always better than those with the linear kernel, demonstrating the importance of nonlinear regression for high-frequency time series. Notice that no experiments were reported using alternative feature selection approaches for SVR since such techniques have not been formalized for time series analysis, to the best of our knowledge.

Regarding the dimensionality reduction ratio presented in the last row of Table 6, we can conclude that our approach is very effective at performing feature elimination, selecting approximately 10% of the variables or less.

The entity in charge of the *Chile* data set publishes a forecast for its hourly scheduled demand.[2] Based on this forecast, we estimate its performance in our test set, which leads to a MAPE of 2.46%. Our proposal, therefore, achieves slightly better results when compared with the strategy considered by the national coordinator. Notice that forecast includes exogenous variables, such as weather conditions, which are not considered in our approach for this data set.

For completeness, Tables 7 and 8 presents the results using RMSE and $MASE_s$ as performance metrics, respectively. Although our approach is not able to outperform all other methods, these results confirm the good performance achieved by our proposal, achieving the best performance in average. Table 8 also illustrate the ability of nonlinear approach such as our proposal to perform better than a naïve seasonal method.

It is noteworthy the relatively large MAPE values for the France dataset (see Table 6) while the errors are rather low for the remaining metrics in Tables 7 and 8. This can be partially explained by the fact that the output variable is smaller in average for this dataset, leading to low RMSE values. For this reason, the denominator is smaller for the MAPE metric when compared to the other datasets.

[2] https://www.coordinador.cl/sistema-informacion-publica/portal-de-operaciones/operacion-programada/demanda-programada/.

Table 7
Result summary for data sets. Average RMSE in test set.

Experiment	Data Set						
	GB	E&W	IO14_DEM	IO14_TGSD	France	Chile	GB Daily
Naïve	8211	7626	7622	6721	1144	732	133,179
Naïve	4868	4519	4529	4532	1142	536	112,514
ARIMA	4871	4458	4388	4508	1144	466	126,429
ARIMA	2823	2496	**2539**	**2579**	1148	705	102,300
ARIMA	7114	6631	6613	5737	1143	701	-
TBATS	5737	3182	3186	3022	1264	570	51,981
SVR Exp. 1	6396	11,432	8724	6299	1145	984	292,758
SVR Exp. 2[g]	6879	5601	5637	5341	3051	979	3,961,496
SVR Exp. 3[g]	6317	427,501	7630	264,308	**664**	883	287,896
SVR Exp. 4[g]	4851	4414	5818	12,213	732	457	88,790
SVR Exp. 5[g]	11,846	6483	3703	107,675	945	371	80,913
SVR Exp. 6[g]	4333	3630	3714	3641	795	338	66,582
SVR Exp. 7[g]	5107	3543	3900	2929	930	328	107,170
SVR Exp. 8[g]	4051	5490	5451	5186	709	387	-
SVR-FFS	2966	2799	2814	2830	1235	534	122,196
SVR-FFS	**2813**	**2425**	2700	2818	967	**274**	**43257**

[a] Non-seasonal naïve approach.
[b] Seasonal naïve approach using seasonal pattern s_1 (Table 4).
[c] Auto-ARIMA function using seasonal pattern s_1 (Table 4).
[d] Auto-ARIMA function using seasonal pattern s_2 (Table 4).
[e] Auto-ARIMA function using seasonal pattern s_3 (Table 4).
[f] Holt-Winter's TBATS function using seasonal patterns s_1 and s_2.
[g] Naïve SVR configurations using the lags reported in Table 3.
[h] Best iteration based on test performance, Linear kernel.
[i] Best iteration based on test performance, RBF kernel.

Table 8
Result summary for data sets. Average $MASE_s$ (%) in test set.

Experiment	Data Set						
	GB	E&W	IO14_DEM	IO14_TGSD	France	Chile	GB Daily
Naïve	3.1	3.07	3.06	2.66	1.89	1.68	2.2
Naïve	1.61	1.58	1.58	1.58	1.88	1.05	1.81
ARIMA	1.62	1.54	1.50	1.57	1.89	1.04	2.07
ARIMA	0.76	0.79	0.78	0.81	1.91	1.38	1.56
ARIMA	2.74	2.72	2.71	2.35	1.89	1.68	-
TBATS	6.27	3.34	3.32	3.30	26.45	2.21	**0.52**
SVR Exp. 1	2.51	4.87	3.52	2.66	1.91	2.19	15.38
SVR Exp. 2[g]	2.72	2.41	2.44	2.32	6.07	2.18	228.49
SVR Exp. 3[g]	2.51	143.86	3.03	71.24	**1.17**	1.91	10.73
SVR Exp. 4[g]	1.67	1.67	2.31	5.71	1.23	0.84	4.49
SVR Exp. 5[g]	4.96	2.91	1.34	27.36	1.57	0.69	3.44
SVR Exp. 6[g]	1.58	1.46	1.45	1.47	1.34	0.60	2.74
SVR Exp. 7[g]	2.05	1.35	1.54	0.99	1.53	**0.58**	4.91
SVR Exp. 8[g]	1.36	1.98	1.96	1.91	1.26	0.73	-
SVR-FFS	0.94	0.87	0.88	0.90	2.03	1.22	1.80
SVR-FFS	**0.72**	**0.77**	**0.74**	**0.79**	1.23	**0.58**	0.63

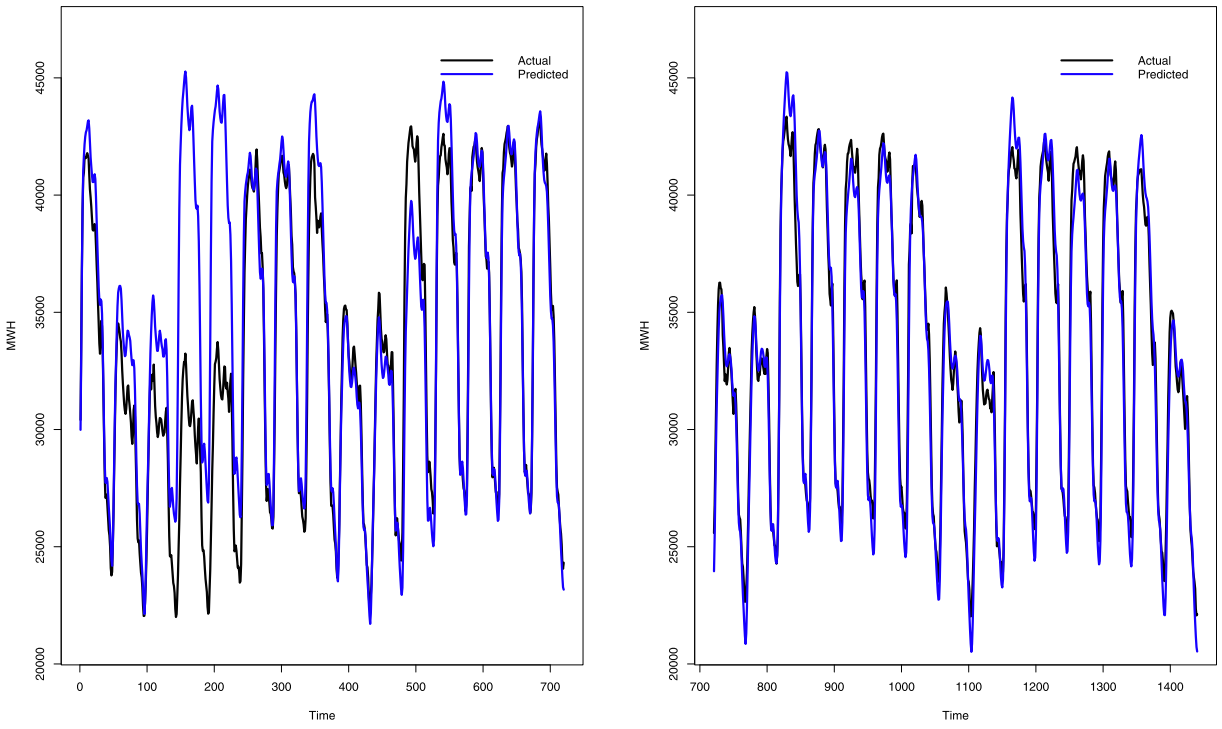
[a] Non-seasonal naïve approach.
[b] Seasonal naïve approach using seasonal pattern s_1 (Table 4).
[c] Auto-ARIMA function using seasonal pattern s_1 (Table 4).
[d] Auto-ARIMA function using seasonal pattern s_2 (Table 4).
[e] Auto-ARIMA function using seasonal pattern s_3 (Table 4).
[f] Holt-Winter's TBATS function using seasonal patterns s_1 and s_2.
[g] Naïve SVR configurations using the lags reported in Table 3.
[h] Best iteration based on test performance, Linear kernel.
[i] Best iteration based on test performance, RBF kernel.

The France dataset is probably more difficult to forecast accurately, leading to larger errors in percentage.

A large value for the metric may be a sign of overfitting for the machine learning methods (see Table 8). However, simple methods such as the naïve approaches may also incur a large error because the denominator of this metric is computed on the training set. In this case, the issue is not overfitting but a probable case of data fracture in which the training and test set distributions mismatch. For the France dataset, for example, the MASE values are relatively small, discarding a problem of dataset shift.

For the next set of experiments, the *GB* data set is used for illustrating some relevant aspects of our proposal. The first analysis illustrates the quality of the forecast provided by our proposal, comparing the output of the model (blue lines) with the actual time series for the month used for testing. This figure is divided in two parts: Fig. 3(a) presents the first two weeks of the month, while Fig. 3(b) reports the second half of the month.

From Fig. 3 we can conclude that our approach provides a very accurate forecast in relation with the real values. The forecast is sometimes over- or underpredicting the real outcome for some



(a) Weeks 1 and 2 (b) Weeks 3 and 4

Fig. 3. Comparison between SVR-FFS and the real time series for the month used for testing. GB data set.

periods, but this is somewhat expected in some real-world applications. In energy load forecasting, for example, the weather can be abnormally cold or hot in some days, causing this problem.

The time series literature acknowledges several issues that can affect predictive performance negatively. One of these issues is serial correlation (Hyndman & Khandakar, 2008). A regular regression approach assumes that the residuals are independent from one observation to the next. However, this assumption is not valid in auto-regressive processes, such as high-frequency data. Consequences of this issue include an inefficient estimation of the coefficients, under-estimation of the mean-squared error loss used as objective function, inaccurate estimation of the confidence intervals, and ultimately the inability to provide an accurate forecast (Hyndman & Khandakar, 2008).

Serial correlation is addressed via the inclusion of autoregressive processes in all our methods. The proposed and alternative

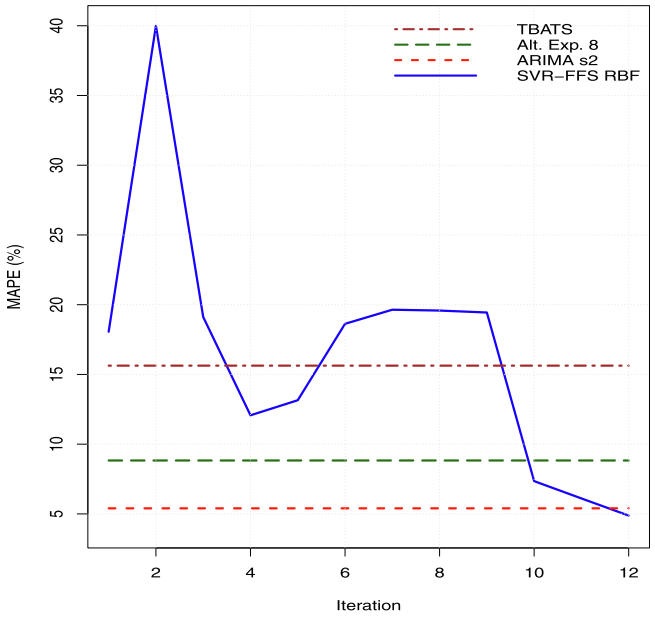


Fig. 4. Comparison between SVR-FFS and the best benchmark algorithms for an increasing number of selected lags. GB data set.

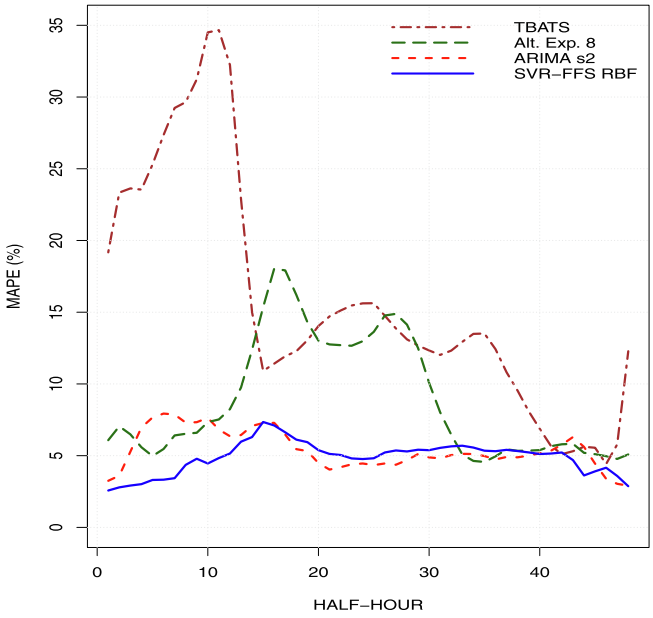


Fig. 5. Comparison between SVR-FFS and the best benchmark algorithms for an average day. GB data set.

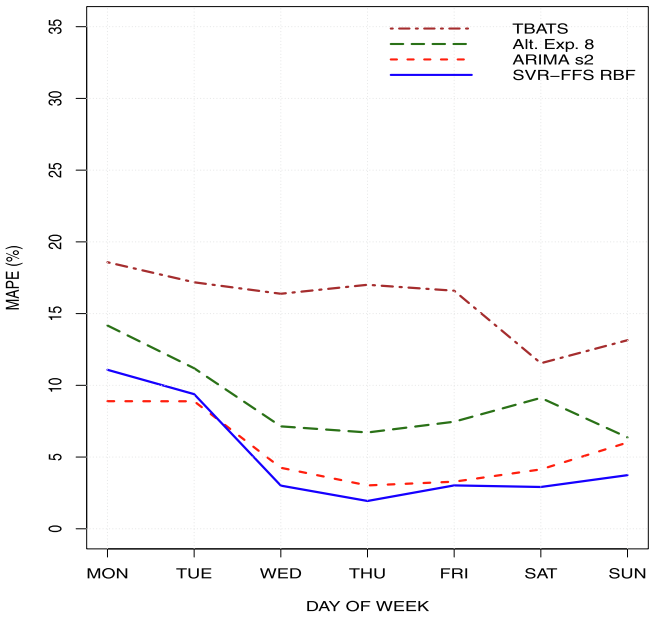


Fig. 6. Comparison between SVR-FFS and the best benchmark algorithms for an average week. GB data set.

methods include all the lags that are relevant for both predicting well and alleviating the issue of serial correlation. We support our analysis in the literature of energy load forecasting and high-frequency data in general. Our method is designed precisely for alleviating the serial correlation issue by incorporating the lags that are relevant for the problem with introducing additional noise, which is the case in most machine learning approaches for forecasting. The inclusion of an excessive number of lags leads to overfitting. Furthermore, the first lag AR(1) is included as covariate in our method and all the benchmark approaches.

Although the use of the MAPE metric may introduce some bias in the model selection process, the proposed method performed best in terms of predictive performance for the three metrics considered in this study. The three-stage model validation procedure reduces the risk of selecting a biased model or one affected negatively by serial correlation since it would have a poor out-of-sample performance in the validation set and therefore would not be selected.

The following experiment presents the performance of the SVR-FFS method for an increasing number of selected lags. For the following values of $|\mathcal{L}^{\dagger}| : |\mathcal{L}^{\dagger}| = \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024, 2048\}$, the MAPE is reported for each iteration of the algorithm. The result of this experiment is presented in Fig. 4. This figure also includes the performance of the best configurations for the ARIMA, Holt-Winter (TBATS), and SVR approaches, represented by lines of horizontal dashes.

It can be observed in Fig. 4 that our method is able to perform better than Holt-Winter, Naïve SVR, and ARIMA on the iteration 4, iteration 10, and iteration 12, respectively. A high variance in terms of performance can be observed for the different iterations of our proposal and, therefore, performing a similar strategy for determining the final set of lags is recommended.

Fig. 5 illustrates the daily performance in terms of average MAPE (in percentage, the Y-axis) for the best model configurations on the *GB* data set. The X-axis presents every half-hour of the day, from 1 (00 : 00) to 48 (23 : 30). Similarly, Fig. 6 shows the same metric aggregated by the day of the week.

It can be seen in Fig. 5 that the proposed SVR-FFS (blue line) clearly achieves the best performance during the first 14 half hours of the day, which are the times that exhibit the largest errors in terms of MAPE. The performance is rather similar when compared with the best alternative approach (SARIMA with weekly seasonality). The remaining methods are outperformed by these two approaches for this data set. A similar analysis can be done for the weekly aggregation (Fig. 6), in which SVR-FFS also has the best performances in general, followed by SARIMA. For all the methods, better performance is achieved from Wednesday to Saturday, decreasing from Sunday to Tuesday.

Our final set of experiments consists in the residual plots, which provides a visual inspection method for assessing the underlying assumptions of statistical methods. It must be noticed that SVR and SVR-FFS do not make any assumption on the distribution of the residuals. Fig. 7 presents the residual plots for the four methods discussed in the GB dataset analysis.

It can be concluded from this figure that no clear pattern is observed that could invalidate the results obtained with the various methods. The only interesting pattern that can be noticed is that the residuals seem bounded for the SVR-FFS approach (Fig. 7 (d)). This is due to the shape of the ε-tube constructed by the SVM classifier and the large value of parameter C obtained for the best hyperparameter configuration. A large value of C tend to force all predicted samples to lay within the tube.

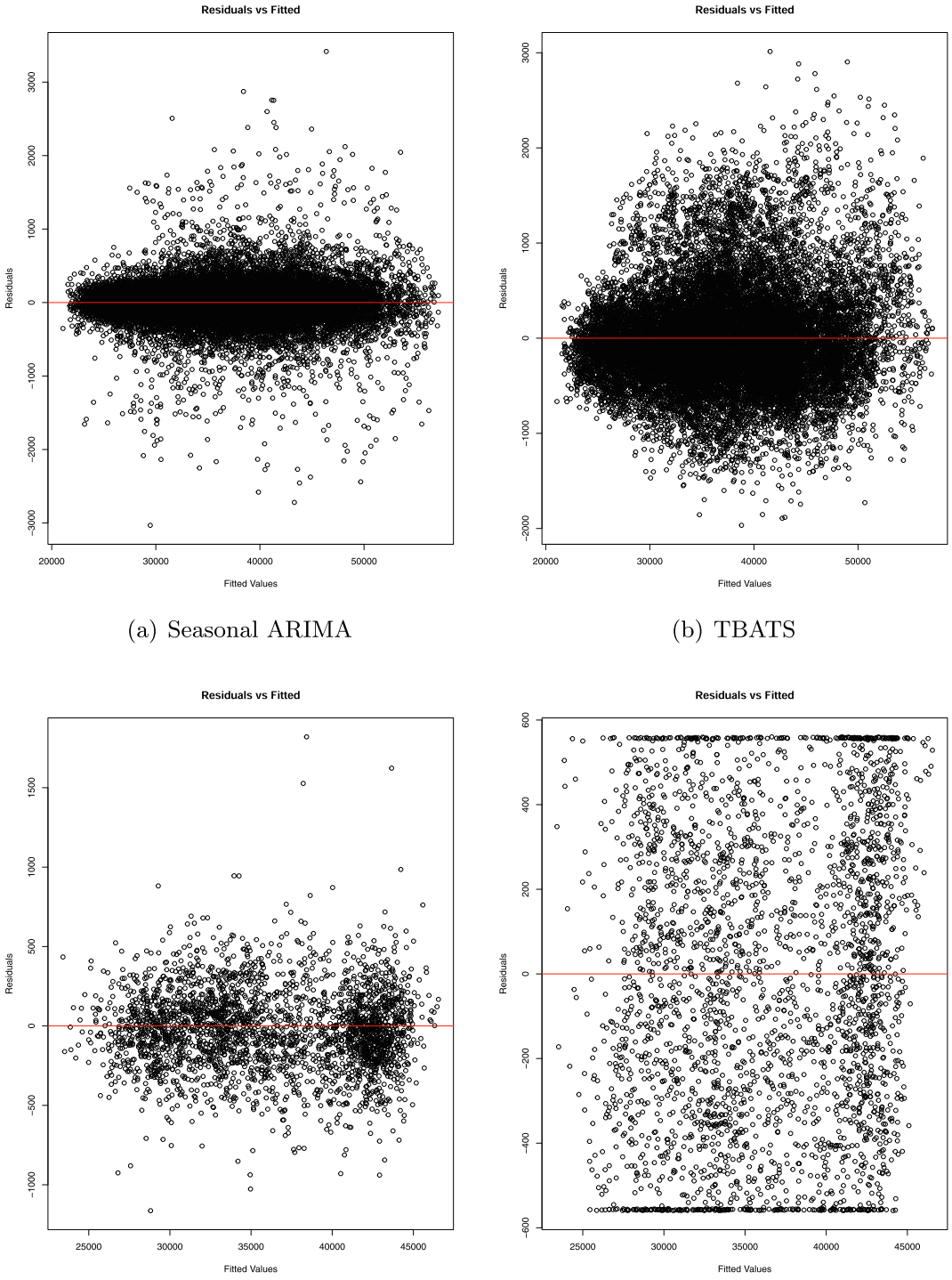
5. Conclusions

A novel approach for automatic lag selection using SVR is presented in this paper. A forward feature selection algorithm was designed in order to identify which lags are relevant by using the decision function. The proposed SVR-FFS corresponds to an embedded method since it determines the optimal set of lags simultaneously with the regression model. The main methodological advantages of SVR-FFS are its ability to construct nonlinear decision functions thanks to the use of kernel functions, the capacity of taking the interaction between covariates and the model into account, a faster training thanks to the use of forward variable selection, and the natural inclusion of exogenous variables.

The proposed method was applied to several short-term energy load forecasting data sets, which is a very challenging in forecasting given its high-frequency nature. In such data sets, more than 20,000 lags are required for identifying yearly seasonality, leading to dense, high-dimensional matrices. Our experiments demonstrated the virtues of our approach, which performed best on all the data sets in comparison with fourteen other forecasting strategies. Our experiments also confirm the importance of using nonlinear regression approaches, such as kernel methods, in time series analysis with high-frequency data.

Regarding the experiments on the *Chile* data set, the proposed SVR-FFS method has slightly better performance than the approach reported by the national electric coordinator (*Coordinador Eléctrico Nacional*) in decision-making, as can be seen on Table 6, even without including the exogenous variables that this forecast considers. This result shows the potential contribution of applying our proposal for improving short-term load decisions.

As for future developments, there are multiple research opportunities that could stem from this work. First, since the traditional ε-SVR implementation used in this study (LIBSVM) could be very time-consuming in high-dimensional problems, using more efficient base regression models can be considered to reduce training times. For example, linear approaches using highly-optimized learning strategies have been developed for classification (see e.g. Djuric, Lan, Vucetic, & Wang, 2013), and they can be adapted to SVR and time series analysis. Secondly,



(a) Seasonal ARIMA (b) TBATS

(c) Triple-seasonal SVR (d) SVR-FFS

Fig. 7. Residual plots for SVR-FFS and the best benchmark algorithms. GB data set.

the forward process is bottlenecked by the computation of kernel matrices using all training samples, which is a time consuming process that is avoided by SVR via incremental optimization methods, such as the sequential minimization optimization (SMO) (Hornik, Meyer, & Karatzoglou, 2006). Sampling strategies could be useful for speeding up our algorithm when each attribute is assessed in terms of its contribution. Thirdly, the issue of serial correlation in machine learning can be studied further in future studies. Although this issue is of utmost importance for the proper estimation of linear methods due to its parametric nature, we believe it can also be a relevant topic in the machine learning. This issue has not been discussed in the machine learning literature, to the best of our knowledge. Finally, cost-sensitive strategies could be considered instead of using symmetric error measures, such as MAPE, in the sense that underestimating the demand is equivalent to overestimating it. Symmetric error metrics may not be the right approach in applications such as energy load forecasting.

Declaration of Competing Interest

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

Acknowledgements

The authors gratefully acknowledge financial support from CONICYT PIA/ BASAL AFB180003 and FONDECYT-Chile, grants 1181809 and 1200221. The authors are grateful to the anonymous reviewers who contributed to improving the quality of the original paper.

References

AlRashidi, M. R., & El-Naggar, K. M. (2010). Long term electric load forecasting based on particle swarm optimization. *Applied Energy, 87*(1), 320–326.
Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences, 99*(10), 6562–6566.
Amjady, N., & Keynia, F. (2008). Mid-term load forecasting of power systems by a new prediction method. *Energy Conversion and Management, 49*(10), 2678–2687.
Anderson, B., & Torriti, J. (2018). Explaining shifts in uk electricity demand using time use data from 1974 to 2014. *Energy Policy, 123*, 544–557.
Bejan, A. I., Gibbens, R. J., & Kelly, F. P. (2012). Statistical aspects of storage systems modelling in energy networks. In *2012 46th annual conference on information sciences and systems (ciss)* (pp. 1–6).
Chikobvu, D., & Sigauke, C. (2013). Modelling in influence of temperature on daily peak electricity demand in south africa. Journal of Energy in Southern Africa, 24.
Christiaanse, W. R. (1971). Short-term load forecasting using general exponential smoothing. *IEEE Transactions on Power Apparatus and Systems, 2*, 900–911.
Crone, S. F., & Kourentzes, N. (2010). Feature selection for time series prediction – a combined filter and wrapper approach for neural networks. *Neurocomputing, 73*, 1923–1936.
De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association, 106*(496), 1513–1527.
Djuric, N., Lan, L., Vucetic, S., & Wang, Z. (2013). Budgetedsvm: a toolbox for scalable svm approximations. *Journal of Machine Learning Research, 14*, 3813–3817.
Drucker, H., Burges, C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In Advances in neural information processing systems (Vol. 9, p. 155–161). MIT Press.
Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research, 5*(Nov), 1531–1555.
Ghaddar, B., & Naoum-Sawaya, J. (2018). High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research, 265*(3), 993–1004.
Gheyas, I. A., & Smith, L. S. (2010). Feature subset selection in large dimensionality domains. *Pattern Recognition, 43*(1), 5–13.
Gross, G., & Galiana, F. D. (1987). Short term load forecasting. *Proceedings of the IEEE, 75*, 1558–1573.
Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). *Feature extraction, foundations and applications*. Berlin: Springer.
Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning, 46*(1–3), 389–422.
Hawkins, S., Eager, D., & Harrison, G. P. (2011). Characterising the reliability of production from future british offshore wind fleets. In *IET conference on renewable power generation (RPG 2011), Edinburgh, UK*.
Henley, A., & Peirson, J. (1997). Non-linearities in electricity demand and temperature: parametric versus non-parametric methods. *Oxford Bulletin of Economics & Statistics, 59*, 149–162.
Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems, 16*(1), 44–55.
Hong, W. (2013). *Modeling for energy demand forecasting*. London: Springer-Verlag.
Hornik, K., Meyer, D., & Karatzoglou, A. (2006). Support vector machines in r. *Journal of Statistical Software, 15*(9), 1–28.
Huang, S. J., & Shih, K. R. (2003). Short-term load forecasting via arma model identification including non-gaussian process considerations. *IEEE Transactions on Power Systems, 18*(2), 673–679.
Hu, Z., Bao, Y., Chiong, R., & Xiong, T. (2015). Mid-term interval load forecasting using multi-output support vector regression with a memetic algorithm for feature selection. *Energy, 84*, 419–431.
Hu, Z., Bao, Y., & Xiong, T. (2013). Electricity load forecasting using support vector regression with memetic algorithms. *The Scientific World Journal, 2013*, 10.
Hyndman, R.J., & Athanasopoulos, G. (2013). Forecasting: principles and practice. O Texts.
Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software, 27*(3).
Jiang, S., Chin, K.-S., Wang, L., Qu, G., & Tsui, K. L. (2017). Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department. *Expert Systems with Applications, 82*, 216–230.
Karmy, J., & Maldonado, S. (2019). Hierarchical time series forecasting via support vector regression in the european travel retail industry. *Expert Systems and Applications, 137*, 59–73.
Kazem, A., Sharifi, E., Hussain, F. K., Saberi, M., & Hussain, O. K. (2013). Support vector regression with chaos-based fire fly algorithm for stock market price forecasting. *Applied Soft Computing, 13*(2), 947–958.
Lal, T. N., Chapelle, O., Weston, J., & Elisseeff, A. (2006). Embedded methods. In *Feature extraction* (pp. 137–165). Springer.
Lee, K. Y., Cha, Y. T., & Park, J. H. (1992). Short-term load forecasting using an artificial neural network. *IEEE Transactions on Power Systems, 7*(1), 124–132.
Lin, Y. L., & Wei, G. (2005). Speech emotion recognition based on hmm and svm. *Machine learning and cybernetics, 2005. Proceedings of 2005 international conference on machine learning and cybernetics* (Vol. 8, pp. 4898–4901).
Liu, B., Li, S., Wang, Y., Lu, L., Li, Y., & Cai, Y. (2007). Predicting the protein sumo modification sites based on properties sequential forward selection (psfs). *Biochemical and Biophysical Research Communications, 358*(1), 136–139.
Makridakis, S., & Hibon, M. (2000). The m3-competition: results, conclusions and implications. *International Journal of Forecasting*.
Maldonado, S., & López, J. (2018). Dealing with high-dimensional classimbalanced datasets: embedded feature selection for svm classification. *Applied Soft Computing, 67*, 94–105.
Maldonado, S., Pérez, J., & Bravo, C. (2017). Cost-based feature selection for support vector machines: An application in credit scoring. *European Journal of Operational Research, 261*(2), 656–665.
Maldonado, S., & Weber, R. (2010). Feature selection for support vector regression via kernel penalization. In *Proceedings of the 2010 international joint conference on neural networks, barcelona, spain* (pp. 1973–1979).
Martínez, F., Frías, M. P., Pérez-Godoy, M. D., & Rivera, A. J. (2018). Dealing with seasonality by narrowing the training set in time series forecasting with knn. *Expert Systems with Applications, 103*, 38–48.
Mbamalu, G., & El-Hawary, M. (1993). Load forecasting via suboptimal seasonal autoregressive models and iteratively reweighted least squares estimation. *IEEE Transactions on Power Systems, 8*, 343–348.
Ng, A. (2017). Machine learning yearning.
Nicholson, W. B., Matteson, D. S., & Bien, J. (2017). Varx-l: Structured regularization for large vector autoregressions with exogenous variables. *International Journal of Forecasting, 33*(3), 627–651.
Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by svm. *IEEE Transactions on Geoscience and Remote Sensing, 48*(5), 2297–2307.
Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and minredundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238.
Perkins, S., Lacker, K., & Theiler, J. (2003). Grafting: Fast, incremental feature selection by gradient descent in function space. *Journal of Machine Learning Research, 3*(Mar), 1333–1356.
Sanz-García, A., Fernández-Ceniceros, J., Antonanzas-Torres, F., Pernia-Espinoza, A. V., & de Pison, F. J. M. (2015). Ga-parsimony: A ga-svr approach with feature selection and parameter optimization to obtain parsimonious solutions for predicting temperature settings in a continuous annealing furnace. *Applied Soft Computing, 35*, 13–28.
Sapankevych, N., & Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine, 4*, 24–38.
S.Dhiman, H., Deb, D., & Guerrero, J.M. (2019). Hybrid machine intelligent svr variants for wind forecasting and ramp events. Renewable and Sustainable Energy Reviews, 108, 369-379.
Smola, A. J., & Schölkopf, B. (1998). *A tutorial on support vector regression (Tech. Rep.). NeuroCOLT Technical Report NC-TR-98-030*. London, UK: Royal Holloway College, University of.
Song, L., Smola, A., Gretton, A., Bedo, J., & Borgwardt, K. (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research, 13*, 1393–1434.
Son, H., & Kim, C. (2015). Forecasting short-term electricity demand in residential sector based on support vector regression and fuzzyrough feature selection with particle swarm optimization. *Procedia Engineering, 118*, 1162–1168.
Sorjamaa, A., Hao, J., Reyhani, N., Ji, Y., & Lendasse, A. (2007). Methodology for long-term prediction of time series. *Neurocomputing, 70*(16–18), 2861–2869.
Srinivasan, D. (1998). Evolving artificial neural networks for short term load forecasting. *Neurocomputing, 23*(1–3), 265–276.
Srinivasan, D. (2008). Energy demand prediction using gmdh networks. *Neurocomputing, 72*(1–3), 625–629.
Taylor, J. W. (2003). Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society, 54*(8), 799–805.
Taylor, J. (2010). Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research, 204*, 139–152.
Weston, J., Mukherjee, S., Chapelle, O., Ponntil, M., Poggio, T., & Vapnik, V. (2001). *Feature selection for SVMs. Advances in neural information processing systems 13* (Vol. 13). MIT Press.

Wu, C., Tzeng, G., & Lin, R. (2009). A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression. *Expert Systems with Applications, 36*, 4725–4735.

Xu, S., Chan, H. K., & Zhang, T. (2019). Forecasting the demand of the aviation industry using hybrid time series sarima-svr approach. *Transportation Research Part E: Logistics and Transportation Review, 122*, 169–180.

Zbikowski, K. (2015). Using volume weighted support vector machines with walk forward testing and feature selection for the purpose of creating stock trading strategy. *Expert Systems with Applications, 42*(4), 1797–1805.

Zhao, J., Chen, L., Pedrycz, W., & Wang, W. (2019). Variational inferencebased automatic relevance determination kernel for embedded feature selection of noisy industrial data. *IEEE Transactions on Industrial Electronics, 66*(1), 416–428.