IChemE

www.icheme.org/journals
doi: 10.1205/cherd.04246

0263–8762/05/$30.00+0.00

Trans IChemE, Part A, August 2005
Chemical Engineering Research and Design, 83(A8): 1009–1018

CUSTOMER DEMAND FORECASTING VIA SUPPORT VECTOR REGRESSION ANALYSIS

A. A. LEVIS and L. G. PAPAGEORGIOU*

Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), London, UK

This paper presents a systematic optimization-based approach for customer demand forecasting through support vector regression (SVR) analysis. The proposed methodology is based on the recently developed statistical learning theory (Vapnik, 1998) and its applications on SVR. The proposed three-step algorithm comprises both nonlinear programming (NLP) and linear programming (LP) mathematical model formulations to determine the regression function while the final step employs a recursive methodology to perform customer demand forecasting. Based on historical sales data, the algorithm features an adaptive and flexible regression function able to identify the underlying customer demand patterns from the available training points so as to capture customer behaviour and derive an accurate forecast. The applicability of our proposed methodology is demonstrated by a number of illustrative examples.

Keywords: customer demand forecasting; support vector regression; statistical learning theory; nonlinear optimization.

INTRODUCTION

Recent studies have clearly identified customer demand as the ultimate driver of business management in process industries ranging from the traditional oil and gas industry (Lasschuit and Thijssen, 2004) to the high-risk agrochemical and pharmaceutical industries (Maravelias and Grossmann, 2001; Levis and Papageorgiou, 2004). Given the importance of customer demand, one can easily realize the potential benefits of an accurate customer demand forecasting tool in process industries.

Forecasting has gained widespread acceptance as an integral part of business planning and decision-making in areas such as sales planning, marketing research, pricing, production planning and scheduling (Makridakis and Wheelwright, 1978). From a historical perspective, exponential smoothing methods and decomposition methods were the first forecasting approaches to be developed back in the mid-1950s. During the 1960s, as computer power became more available and cheaper, more sophisticated forecasting methods appeared. Box–Jenckins (Box and Jenckins, 1969) methodology gave rise to the autoregressive integrated moving average (ARIMA) models. Later on during the 1970s and 1980s, even more sophisticated forecasting approaches were developed including econometric methods and Bayesian methods (Makridakis and Wheelwright, 1982). The consolidation and improvement of the aforementioned approaches provided forecasting tools of ever increasing complexity, before artificial neural networks (ANN) emerged as a novel and promising forecasting approach in the 1990s taking full advantage of the number-crunching capabilities of super-computers (Foster *et al.*, 1992).

However, it is very interesting to notice that the increasing complexity of forecasting approaches is not always accompanied by the desired increased predictive accuracy as pointed out by Makridakis and Hibon (1979). Their remark is consistent with recent criticism towards the excessive use of parameters in unnecessarily complex ANN applications in chemical engineering problems (Bhat and McAvoy, 1992). There exists a clearly identified need for a new generation of forecasting tools that share all the benefits of ANN while at the same time maintain the underlying formulation as simple as possible.

Support vector machines (SVM) constitute such a novel learning paradigm that provides an inherently simple formulation and yet offers the promise of increased flexibility. The growing popularity of the SVM is mainly attributed to the solid theoretical foundations and the practical applications in a broad range of the scientific spectrum. Based on the statistical learning theory recently developed by Vapnik (1998), SVM applications have been proposed for a number of classification and regression problems ranging from discrete manufacturing (Prakasvudhisarn *et al.*, 2003) to bioinformatics

**Correspondence to*: Dr L. G. Papageorgiou, Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), London WC1E 7JE, UK.
E-mail: l.papageorgiou@ucl.ac.uk

(Myasnikova *et al.*, 2002). Agrawal *et al.* (2003) portray SVM as a useful tool for process engineering applications while Kulkarni *et al.* (2004) and Chiang *et al.* (2004) provide support vector classification applications in process engineering problems.

However, in order to present a balanced perspective, we must also mention that so far the applicability of support vector regression (SVR) is hindered by the notorious problem of parameter selection. Although, the number of parameters to be tuned is not prohibitively large, parameter values affect significantly the predictive capabilities. Exhaustive grid-search (Chang and Lin, 2001) and heuristic-based rules for parameter selection (Cherkassky and Ma, 2004) are currently used for SVR while further research in this area is in progress. As a typical case of any emerging forecasting research field, those heuristic rules can be regarded an initial step towards the identification of a more formal way for parameter selection in the near future.

Despite the parameter selection problem, SVR still enjoys numerous advantages when compared with other forecasting methodologies. Similar to ANN, SVR employs an adaptive basis regression function without postulating any pre-determined family of basis functions (e.g., high-order polynomial parametric regression). Support vectors provide a completely new way of parameterization of the adaptive function (Cherkassky and Mulier, 1998) leading to increased flexibility while avoiding the trap of over-complexity. Unlike ANN, SVR employs only a handful of parameters while its unique mathematical formulation guarantees a computationally tractable global optimal solution. This is a very attractive feature for applications in the process industries where repeatability and consistency are of paramount importance. Furthermore, SVR requires no *a priori* fundamental understanding of the process being studied since it is a training data-driven methodology and therefore is very well-suited for process industry forecasting applications where historical data is abundant.

Overall, support vector regression is identified as a novel emerging forecasting technique and the aim of this paper is to validate the applicability of SVR analysis for forecasting customer demand in process industries. The rest of the paper is organized as follows. In the next section, the main characteristics of the customer demand forecasting problem are described. The SVR section provides a brief mathematical description of support vector regression analysis. A three-step algorithm is then described in the Proposed Forecasting Algorithm section before it is validated through a number of illustrative examples in the following section. Finally, some concluding remarks are drawn in the last section of the paper.

PROBLEM DESCRIPTION

We assume that customer demand at time period t (y_t) depends on a number of z independent variables (x_{1t}, x_{2t}, ..., x_{zt}) that are called attributes and form the associated input vector x_t.[1] Therefore, the dependant variable y_t is a function of the input vector $\boldsymbol{x}_t$ which in turns contains the multiple independent variables as shown in the following equations.

$$\boldsymbol{x}_t = \begin{pmatrix} x_{1t} \\ x_{2t} \\ . \\ . \\ . \\ . \\ . \\ x_{zt} \end{pmatrix} \tag{1}$$

$$y_t = F(\boldsymbol{x}_t) \tag{2}$$

The problem of customer demand forecasting via SVR analysis can be formally stated as follows. Given a set of training data (time series) in the form of N training points $(\boldsymbol{x}_1, y_1)$, $(\boldsymbol{x}_2, y_2)$, ..., $(\boldsymbol{x}_N, y_N)$ where $\boldsymbol{x}_t$ is the input vector and y_t is the associated customer demand for every $\boldsymbol{x}_t$, as well as a forecasting horizon of size M, we would like to *determine* the output values $\hat{y}_{N+1}$, $\hat{y}_{N=2}$, ..., $\hat{y}_{N+M}$. The mean absolute percent error (MAPE) is a commonly used forecasting error metric for quantifying and assessing the accuracy of the predicted output values. Mathematically, it is given by the following formula (Makridakis and Wheelwright, 1978):

$$\text{MAPE} = 100 \cdot \frac{\sum_{t=N+1}^{N+M} |(y_t - \hat{y}_t)/y_t|}{M} \tag{3}$$

where y_t is the actual customer demand and $\hat{y}_i$ is the predicted demand at time period t. Clearly, accurate predictions would result in low MAPE values, which implies small absolute deviations between the actual and predicted output values.

Based on the available training points ($\boldsymbol{x}_t$, y_t), the ultimate goal of support vector regression analysis is to extract as much information as possible from the historical data so as comprehend the complicating relationships between customer demand and all the different attributes before identifying an appropriate regression function F able to accurately predict future unknown output values from a given input vector of attributes.

In our time series forecasting problem, customer demand attributes can be classified into a number of different main categories such as:

- *Past demand attributes*: those attributes represent customer demand for a predetermined number of previous time periods. Employing past demand attributes can be extremely helpful to relate present customer demand with historical customer demand values. According to our experience, past demand attributes prove to be very efficient when dealing with periodical customer demand patterns.
- *Calendar attributes*: those attributes illustrate a specific characteristic of the time period under investigation and are usually treated as binary parameters representing true or false statements with one and zero values respectively. For example, calendar attributes could be

[1]Symbols in bold fonts represents vectors.

employed to represent the day of the week, the month or week of the year and so on. Moreover, calendar attributes could also be used to identify customer demand patterns on national holidays or weekends. Therefore, calendar attributes prove to be a very critical source of information when trying to predict time-sensitive output values such as electricity load demand or seasonality-dominated customer demand patterns such as swimware sales.

Although there may exist more categories of attributes other than the two mentioned above, we restrict ourselves to only those two main categories since any other attribute is viewed as problem specific. For example, in an ice-cream demand forecasting case, it would be very beneficial to incorporate a temperature attribute or any other weather attribute that can reflect the dependancy between ice-cream consumption and environmental conditions.

Based on our ability to know their future values or not, attributes can further be classified into two main categories, namely deterministic and stochastic attributes. Deterministic attributes are those whose future values are known (or can be predicted with a very high-accuracy). Calendar attributes fall under the deterministic category since anyone can accurately predict the date of next weekend or next Monday. However, there exist a number of attributes that affect our output values considerably, whose values unfortunately can not be predicted or accurately estimated. Such attributes are called stochastic and include for example future temperature profiles, oil prices, dollar-to-pound exchange rates and so on.

In our proposed methodology, customer demand forecasting is based entirely on past demand attributes.[2] Past demand attributes belong to a very special case of attributes that can be regarded as semi-deterministic as explained in detail in the section entitled Proposed Forecasting Algorithm. According to equations (1) and (2), knowing the attributes of the input vectors is only the first step towards a valid prediction. What is foremost needed is to establish a solid relationship between the input vector attributes and the target value. In our case, customer demand and past demand attributes are related through a SVR function F. Such a regression function is needed in order to translate past demand attributes into accucate demand forects. In the next section, the derivation of the SVR function F is explained in full detail.

SUPPORT VECTOR REGRESSION

In this section, we briefly describe the SVR based on the statistical learning theory developed by Vapnik (1998). Given training data

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\} \subset R^Z \times R \tag{4}$$

where $\boldsymbol{x}_t$ is the input vector at time period t and y_t is the associated customer demand for every $\boldsymbol{x}_t$, the goal is to find regression function $F(\boldsymbol{x}_t)$:

$$F(\boldsymbol{x}_t) = \boldsymbol{w}^T \boldsymbol{x}_t + \beta \quad \boldsymbol{w}, \boldsymbol{x}_t \in R^Z, \ \beta \in R \tag{5}$$

The main insight of the statistical learning theory is that in order to obtain a regression function with high generalisation behaviour, one needs to control both model complexity and training error tolerance (Chalimourda *et al.*, 2004). Model complexity is illustrated by the flatness of the function F which in turns means small $\boldsymbol{w}$ values. One way to ensure this is to minimise the Euclidean norm $\|\boldsymbol{w}\|$. On the other hand, the regression function should not be too flat but rather complicated enough so as to fit closely with the demand training points. In order to control training error tolerance, the ε-insentitive loss-function $|\xi|_{\varepsilon;}$ can be employed:

$$|\xi|_\varepsilon = \max(0, |F(\boldsymbol{x}_t) - y_t| - \varepsilon) \tag{6}$$

The ε-insensitive loss function ensures that errors less than ε are not taken into consideration. However, we penalize any deviations larger than ε, meaning all training points that lie outside the ε-insensitive tube as shown graphically in Figure 1.

Overall, the SVR analysis takes the form of the following constrained optimization problem (Vapnik, 1998):

[Problem ε-SVR]

$$\min_{\boldsymbol{w},\beta,\xi_t,\xi_t^*} \frac{1}{2}\|\boldsymbol{w}\|^2 + C \cdot \sum_{t=1}^{N} (\xi_t + \xi_t^*)$$

subject to

$$\begin{aligned}
&y_t - \boldsymbol{w}^T \boldsymbol{x}_t - \beta \le \varepsilon + \xi_t \quad \forall t = 1, \ldots, N \\
&\boldsymbol{w}^T \boldsymbol{x}_t + \beta - y_t \le \varepsilon + \xi_t^* \quad \forall t = 1, \ldots, N \\
&\xi_t \ge 0 \quad \forall t = 1, \ldots, N \\
&\xi_t^* \ge 0 \quad \forall t = 1, \ldots, N
\end{aligned}$$

The first term in the objective function represents model complexity (flatness) while the second term represents the model accuracy (error tolerance). The parameter C is a positive scalar determining the trade-off between flatness and error tolerance (regularization parameter), while ξ_t and ξ_t^* represent the absolute deviations above and below the ε-insensitive tube.

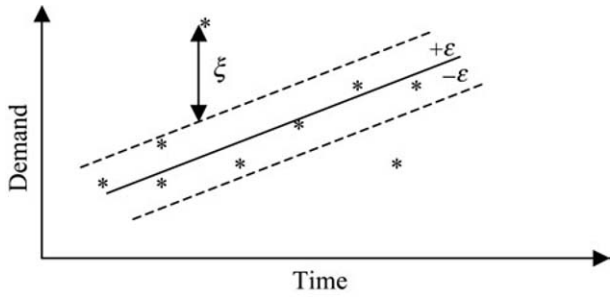

Figure 1. Graphical representation of support vector regression (* depicts training points).

[2]In this work, past customer demand attributes (x_{zt}) are defined according to the following formula:

$$x_{zt} = \begin{cases} y_{t-z} & \text{if } t > z \\ y_t & \text{if } t \le z \end{cases}$$

From a mathematical point of view, the aforementioned nonlinear optimization problem features a number of very interesting properties. Problem ε-SVR constitutes a convex nonlinear programming NLP optimization problem since it involves the minimization of a quadratic function subject to a linear set of constraints, meaning that every local solution to the problem is also a global solution (Bertsekas, 1995; Floudas, 1995). Furthermore, ε-SVR is a convex primal problem satisfying strong duality conditions. Therefore, instead of solving primal problem ε-SVR, we can obtain the exact same global minimum solution by solving its dual counterpart. Thanks to its reduced size both in terms of constraints and variables, the dual model formulation requires significantly less computational effort to solve. Without compromising the quality of the obtained solution, the dual problem formulation can also easily be extended to accommodate the general case of nonlinear regression through appropriately defined kernel functions as it is demonstrated later on in this section.

We can easily construct the Lagrangean function of the primal problem by bringing all constraints into the objective function with the use of appropriately defined Lagrange multipliers λ_t, λ_t^*, μ_t, μ_t^* as follows:

$$\begin{aligned} L = {} & \frac{1}{2}\|\boldsymbol{w}\|^2 + C \cdot \sum_{t=1}^{N} (\xi_t + \xi_t^*) \\ & - \sum_{t=1}^{N} \lambda_t \cdot (\varepsilon + \xi_t - y_t + (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + \beta)) \\ & - \sum_{t=1}^{N} \lambda_t^* \cdot (\varepsilon + \xi_t^* + y_t - (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + \beta)) \\ & - \sum_{t=1}^{N} \mu_t \cdot \xi_t - \sum_{t=1}^{N} \mu_t^* \cdot \xi_t^* \end{aligned} \tag{7}$$

At the optimal point Karush–Kuhn–Tucker (KKT) conditions impose that the partial derivatives of L with respect to the primal variables ($\boldsymbol{w}$, β, ξ_t, ξ_t^*) equal zero:

$$\begin{aligned} & \frac{\partial L}{\partial w} = 0 \Rightarrow \boldsymbol{w} - \sum_{t=1}^{N} (\lambda_t - \lambda_t^*) \cdot \boldsymbol{x}_t = 0 \\ & \frac{\partial L}{\partial \beta} = 0 \Rightarrow \sum_{t=1}^{N} (\lambda_t - \lambda_t^*) = 0 \\ & \frac{\partial L}{\partial \xi_t} = 0 \Rightarrow C - \lambda_t - \mu_t = 0 \quad \forall\, t = 1, \ldots, N \\ & \frac{\partial L}{\partial \xi_t^*} = 0 \Rightarrow C - \lambda_t^* - \mu_t^* = 0 \quad \forall\, t = 1, \ldots, N \end{aligned} \tag{8}$$

By substituting equations (8) into (7), we obtain the dual optimization problem:

[Problem D]

$$\begin{aligned} & \max_{\lambda_t, \lambda_t^*} \frac{1}{2} \sum_{t'=1}^{N} \sum_{t=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot (\lambda_t - \lambda_t^*) \cdot \boldsymbol{x}_{t'}^{\mathrm{T}} \boldsymbol{x}_t \\ & - \varepsilon \sum_{t=1}^{N} (\lambda_t + \lambda_t^*) + \sum_{t=1}^{N} y_t \cdot (\lambda_t - \lambda_t^*) \end{aligned}$$

Subject to

$$\begin{aligned} & \sum_{t=1}^{N} (\lambda_t - \lambda_t^*) = 0 \\ & 0 \le \lambda_t \le C \quad \forall\, t = 1, \ldots, N \\ & 0 \le \lambda_t^* \le C \quad \forall\, t = 1, \ldots, N \end{aligned}$$

The dual problem optimization problem maximizes a quadratic objective function with respect to Lagrange multipliers λ_t and λ_t^* which now play the role of dual variables. The solution of the dual problem derives the optimal vector $\boldsymbol{w}$ as well as the regression function $F(\boldsymbol{x}_\tau)$ as follows:

$$\boldsymbol{w} = \sum_{t=1}^{N} (\lambda_t - \lambda_t^*) \cdot \boldsymbol{x}_t \tag{9}$$

$$F(\boldsymbol{x}_t) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + \beta \tag{10}$$

Or by using equation (9), we obtain the following expression for the regression function:

$$F(\boldsymbol{x}_t) = \sum_{t'=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot \boldsymbol{x}_{t'}^{\mathrm{T}} \boldsymbol{x}_t + \beta \tag{11}$$

Parameter β can be calculated from the KKT complementarity conditions which state that the product between dual variables and constraints has to vanish as follows:

$$\lambda_t \cdot (\varepsilon + \xi_t - y_t + (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + \beta)) = 0 \quad \forall\, t = 1, \ldots, N \tag{12}$$

$$\lambda_t^* \cdot (\varepsilon + \xi_t^* + y_t - (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_t + \beta)) = 0 \quad \forall\, t = 1, \ldots, N \tag{13}$$

$$(C - \lambda_t) \cdot \xi_t = 0 \quad \forall\, t = 1, \ldots, N \tag{14}$$

$$(C - \lambda_t^*) \cdot \xi_t^* = 0 \quad \forall\, t = 1, \ldots, N \tag{15}$$

According to the aforementioned KKT complementarity conditions, training points lying outside the ε-insensitive tube have $\lambda_t = C$ (or $\lambda_t^* = C$) and $\xi_t \neq 0$ (or $\xi_t^* \neq 0$). Those points are called support vectors. Furthermore, there exists no set of dual vaiariables λ_t and λ_t^* which are both nonzero simultaneously as this would require nonzero slack variables in both directions. Finally, training points within the ε-insensitive tube have $\lambda_t \in (0, C)$ (or $\lambda_t^* \in (0, C)$) and also $\xi_t = 0$ (or $\xi_t^* = 0$) (Smola and Scholkopf, 1998).

Alternatively, a practical way of calculating β and slack variables ξ_t and ξ_t^* is by solving a slightly differentiated version of the primal problem:

[Problem P]

$$\min_{\beta,\, \xi_t,\, \xi_t^*} \sum_{t=1}^{N} (\xi_t + \xi_t^*)$$

subject to

$$y_t - \sum_{t'=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot \boldsymbol{x}_{t'}^T \cdot \boldsymbol{x}_t - \beta \le \varepsilon + \xi_t \quad \forall t = 1, \ldots, N$$

$$\sum_{t'=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot \boldsymbol{x}_{t'}^T \cdot \boldsymbol{x}_t + \beta - y_t \le \varepsilon + \xi_t^* \quad \forall t = 1, \ldots, N$$

$$\xi_t \ge 0 \quad \forall t = 1, \ldots, N$$

$$\xi_t^* \ge 0 \quad \forall t = 1, \ldots, N$$

The solution of Problem P derives simultaneously the values for both β and variables ξ_t and ξ_t^*. It is worth mentioning that λ_t and λ_t^* are now treated as parameters whose values are given from the solution of the dual problem solved earlier on. Notice also that Problem P is a simple linear programming (LP) model and therefore it can be solved with great computational efficiency even for large number of training points.

As shown in equation (11), function F is used to perform a linear regression in input space R^Z based on input vectors $\boldsymbol{x}_{t'}$ and $\boldsymbol{x}_t$. For nonlinear regression however, we need to exploit the way training data appears in our problem. More specifically, according to equation (11) regression function F depends only on the inner product of input vectors ($\boldsymbol{x}_{t'}^T \boldsymbol{x}_t$) and therefore we can employ the following kernel trick (Aizerman *et al.*, 1964) as described by Burges (1998). We first map input vectors into a high-dimensional feature space via mapping function Φ as follows:

$$\Phi: R^Z \to R^{Z'} \tag{16}$$

Regression function F then takes the following form:

$$F(\boldsymbol{x}_t) = \sum_{t'=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot \Phi(\boldsymbol{x}_{t'}^T) \cdot \Phi(\boldsymbol{x}_t) + \beta \tag{17}$$

The difference between equations (11) and (17) is that function F is used to perform a linear regression in different spaces, the input space R^Z and the feature space $R^{Z'}$, respectively.

However, there is no particular need to define mapping function Φ explicitly, since the inner product of vectors in the feature space can be represented with a kernel function K as follows:

$$K(\boldsymbol{x}_{t'}, \boldsymbol{x}_{t'}) = \Phi(\boldsymbol{x}_{t'}^T) \cdot \Phi(\boldsymbol{x}_t) \tag{18}$$

It is worth mentioning that kernel function K is defined as a function of vectors in the original input space R^Z. In that sense, the expression of regression function F can now easily be extended to accommodate the case of nonlinear regression in input space R^Z by performing a linear regression in feature space $R^{Z'}$ via the kernel function transformation as follows:

$$F(\boldsymbol{x}_t) = \sum_{t'=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot K(\boldsymbol{x}_{t'}, \boldsymbol{x}_t) + \beta \tag{19}$$

Table 1. Different types of kernel functions.

No.	Name of kernel	Expression
1	Polynomial	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = ((\boldsymbol{x}_i \cdot \boldsymbol{x}_j) + 1)^p \; p = 1, 2, \ldots$
2	Gaussian radial basis function	$K(x_i, x_j) = \exp\left(\frac{\|\|x_i - x_j\|\|^2}{-2p^2}\right)$
3	Exponential radial basis function	$K(x_i\, ax_j) = \exp\left(\frac{\|x_i - x_j\|}{-2p^2}\right)$
4	Multi-layer perception	$K(x_i, x_j) = \tanh(b(x_i \cdot x_j) - c)$
5	Fourier series	$K(x_i, x_j) = \frac{\sin(N + (1/2))(x_i - x_j)}{\sin(1/2(x_i - x_j))}$
6	Tensor product splines	$K(x_i, x_j) = \prod_{m=1}^{n} K_m(x_{im}, x_{jm})$

Any function satisfying Mercer's theorem (Mercer, 1909) may be employed as a kernel function. The types of functions most commonly used in SVM literature as kernel functions are summarized in Table 1 (see for example, Kulkarni *et al.*, 2004).

Finally, it is very interesting to notice that the introduction of kernel functions in the expression of the regression function as shown in equation (19) does not affect any of the previous analysis on linear SVR. All previous considerations hold intact with the only difference being that linear regression in now performed in a high-dimensional feature space $R^{Z'}$ in order to create a nonlinear regression function in the original input space R^Z (Gunn, 1998). Based on the equation (19), dual and primal model formulations can now easily be extended for nonlinear support vector regression as follows:

[Problem D1]

$$\max_{\lambda_t, \lambda_t^*} -\frac{1}{2} \sum_{t=1}^{N} \sum_{t'=1}^{N} (\lambda_t - \lambda_t^*) \cdot (\lambda_{t'} - \lambda_{t'}^*) \cdot K(\boldsymbol{x}_t, \boldsymbol{x}_{t'}) - \varepsilon \sum_{t=1}^{N} (\lambda_t + \lambda_t^*) + \sum_{t=1}^{N} y_t \cdot (\lambda_t - \lambda_t^*)$$

subject to

$$\sum_{t=1}^{N} (\lambda_t - \lambda_t^*) = 0$$

$$0 \le \lambda_t \le C \qquad \forall t = 1, \ldots, N$$

$$0 \le \lambda_t^* \le C \qquad \forall t = 1, \ldots, N$$

and

[Problem P1]

$$\min_{\beta, \xi_t, \xi_t^*} \sum_{t=1}^{N} (\xi_t + \xi_t^*)$$

subject to

$$y_t - \sum_{t'=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot K(\boldsymbol{x}_t, \boldsymbol{x}_{t'}) - \beta \leq \varepsilon + \xi_t \quad \forall t = 1, \ldots, N$$

$$\sum_{t'=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot K(\boldsymbol{x}_t, \boldsymbol{x}_{t'}) + \beta - y_t \leq \varepsilon + \xi_t^* \quad \forall t = 1, \ldots, N$$

$$\xi_t \geq 0 \quad \forall t = 1, \ldots, N$$

$$\xi_t^* \geq 0 \quad \forall t = 1, \ldots, N$$

The aforementioned mathematical models are used as part of our proposed forecasting algorithm presented in the next section.

PROPOSED FORECASTING ALGORITHM

Based on the SVR analysis presented in the previous section, we proposed the following three-step algorithm.

[Algorithm A1]

Step 1: Solve Problem D1 to determine variables λ_t and λ_t^*.
Step 2: Fix dual variables λ_t, λ_t^* and solve Problem P1 to determine β.
Step 3: For $t := N+1$ to $N+M$, do:

i. Calculate customer demand prediction $y_t = \sum_{t'=1}^{N} (\lambda_{t'} - \lambda_{t'}^*) \cdot K(\boldsymbol{x}_{t'}, \boldsymbol{x}_t) + \beta$.
ii. Update input vectors $\boldsymbol{x}_t$.

The first two steps of Algorithm A1 are used for determining regression function F from the available N training points. In particular, the first step determines optimal values for the λ_t and λ_t^* from the solution of the dual NLP problem (Problem D1). In Step 2, λ_t and λ_t^* are fixed to their optimal levels while we determine parameter β by solving a LP model (Problem P1). The final step of the algorithm constitutes a post-processing recursive forecasting methodology. Having identified regression function F from the training data in steps 1 and 2, a customer demand prediction can be made for the next M time periods based on the semi-deterministic past demand attributes.

Given a time series in the form of N training points $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)$ predicting value y_{N+1} is done by considering the last Δ elements of the time series as deterministic attributes of input vector $\boldsymbol{x}_{N+1}$. The very next prediction for demand y_{N+2} would normally require the *actual* past demand attribute of point $N+1$. Although such a deterministic past demand attribute is not available to us, it can be substituted with the predicted demand value we estimated previously. In other words, the newly predicted demand value is used as a stochastic attribute (since it is not an actual demand value but merely a prediction) in the input vector $\boldsymbol{x}_{N+2}$ for predicting value y_{N+2}. In that fashion, by sequentially adding newly obtained data as attributes in the input vector (and removing the earliest elements), we construct a recursive forecasting algorithm based on semi-deterministic past demand data. More specifically, we employ an iterative moving forecasting horizon approach where in each iteration we calculate customer demand for only the following one time period. The customer demand prediction is then used in a recursive fashion to update the input vector information before calculating customer demand for the very next time period. The algorithm terminates when the entire forecasting horizon M is scanned.

Needless of course to mention that the predictive capabilities of the proposed algorithm are restricted by the inevitable error propagation and accuracy deterioration due to inaccurate forecasted data entering as attributes in our calculations. However, for the purposes of short to medium-term forecast (forecasting/training points ratio between 5–15%) the proposed algorithm performs with great accuracy, as it is demonstrated by the illustrative examples presented in the next section.

Parameter Tuning

Before Algorithm A1 is implemented, a number of parameters should be determined *a priori.* Parameter selection is a notorious problem that usually hinders the applicability of forecasting techniques. However, SRV only relies on a handful of parameters as listed below:

(1) regularization parameter C;
(2) width of ε-insensitive tube;
(3) kernel function K;
(4) number of attributes Δ included in the input vector.

As mentioned earlier, parameter C controls the trade-off between model complexity and model accuracy. Model underfitting occurs when C is too small since model does not have enough detail to describe the training data. On the other hand, overfitting occurs when C is too high (Chiang *et al.*, 2004). The optimal value of parameter C is usually determined by employing a grid-search in either a n-fold cross validation or leave-one-out error estimate approach (Kulkarni *et al.*, 2004). However, an exhaustive grid search is a time-consuming and computionally-expensive way for parameter selection. Alternative ways for determining SVR parameters is an ongoing research area. In our methodology, we determine parameter C based on a heuristic rule recently proposed by Cherkassky and Ma (2004) as follows:

$$C = \max(\bar{y}_t + 3 \cdot \sigma_{y_t}, \bar{y}_t - 3 \cdot \sigma_{y_t}) \tag{20}$$

where $\bar{y}_t$ is the mean average and σ_{y_t} is the standard deviation of the customer demand training points. The proposed formula for determining C has been validated for a number of different cases.

The size of ε influences the number of support vectors (training points lying outside the ε-insensitive tube) and therefore allows direct control over the complexity of the model. Therefore in practice, parameter ε is chosen so as to reflect our relative view towards error and noise through the implementation of different ε-insensitive loss functions. In our experience, values of ε equal to approximately one order of magnitude less than the mean average of the training points target values provide good performance for various data sets. Mathematically, we have:

$$\varepsilon = \frac{\bar{y}_t}{k} \tag{21}$$

where k is a constant scalar taking values in the range $[10, 30]$.

The Gaussian radial basis function (RBF) is chosen as the kernel function in our proposed algorithm since it is the most commonly used kernel for SVM.

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{-2p^2}\right) = \exp\left(-\gamma \cdot \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right) \quad (22)$$

Cherkassky and Ma (2004) proposed an empirical rule for determining RBF kernel width parameter p. According to their formula for z-dimensional problems, where all z input attributes are scaled to [0, 1], p is determined approximately by the following formula:

$$p^Z \approx (0.2 - 0.5) \quad (23)$$

In our methodology, input vector attributes are scaled between zero and one and therefore we calculate parameter p by employing the mean value of the proposed formula, resulting in the following mathematical expression:

$$p = 0.35^{1/Z} \Rightarrow \gamma = \frac{1}{2} \cdot 0.35^{-2/Z} \quad (24)$$

The number of attributes included in the input vector should reflect the seasonality pattern underlying customer demand training points. A heuristic rule for accurate predictions is to employ a number of past attributes equal to an integer multiple of the period points. Say for example one wishes to predict a daily customer demand pattern that repeats itself more or less every week, then it would be advisable to use seven or 14 past demand attributes.

Forecasting Assessment

The assessment of the proposed forecasting algorithm is performed though the employment of the following accuracy criteria:

(1) Prediction Accuracy (P.A.): derives from the MAPE criterion and is used to compare actual demand and predicted demand values over the forecasting horizon time periods. It is defined as follows:

$$\text{P.A.} = 100 - \text{MAPE} = 100 - 100 \cdot \frac{\sum_{t=N+1}^{N+M} |(y_t - \hat{y}_t)/y_t|}{M} \quad (25)$$

(2) Fitting Accuracy (F.A.): it is used to compare actual demand and predicted demand values over all the training points time periods. It is defined as follows:

$$\text{F.A.} = 100 - 100 \cdot \frac{\sum_{t=1}^{N} |(y_t - \hat{y}_t)/y_t|}{N} \quad (26)$$

(3) Overall Accuracy (O.A.): it is used to compare actual demand and predicted demand values over all time periods (both training and forecasting). It is defined as follows:

$$\text{O.A.} = 100 - 100 \cdot \frac{\sum_{t=1}^{N+M} |(y_t - \hat{y}_t)/y_t|}{N + M} \quad (27)$$

Table 2. Electrical appliances daily sales (in hundreds of pieces).

Week	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	7.80	7.70	6.00	6.80	9.20	12.00	3.20
2	5.30	6.30	5.80	5.80	6.00	10.00	5.00
3	5.50	6.40	5.80	5.90	6.10	10.20	5.00
4	6.20	7.80	6.40	6.40	7.20	12.00	6.00
5	7.00	9.00	7.80	7.90	8.30	14.00	7.00
6	9.00	11.00	9.80	9.90	10.40	17.80	9.00
7	8.58	8.47	6.60	7.48	10.12	13.20	3.52
8	5.83	6.93	6.38	6.38	6.60	11.00	5.50
9	6.05	7.04	6.38	6.49	6.71	11.22	5.50
10	6.82	8.58	7.04	7.04	7.92	13.20	6.60
11	7.70	9.90	8.58	8.69	9.13	15.40	7.70
12	9.90	12.10	10.78	10.89	11.44	19.58	9.90
13	10.14	10.01	7.80	8.84	11.96	15.60	4.16
14	6.89	8.19	7.54	7.54	7.80	13.00	6.50
15	7.15	8.32	7.54	7.67	7.93	13.26	6.50
16	*8.06*	*10.14*	*8.32*	*8.32*	*9.36*	*15.60*	*7.80*
17	*9.10*	*11.70*	*10.14*	*10.27*	*10.79*	*18.20*	*9.10*

The applicability of the proposed forecasting algorithm along with the heuristic-based rules for parameter estimation and assessment criteria is demonstrated by a number of illustrative examples presented in the next section.

ILLUSTRATIVE EXAMPLES

This section presents customer demand forecasting results for three illustrative examples. All runs were implemented in GAMS (General Algebraic Modelling System) (Brooke *et al.*, 1998) and solved with commercially available solvers CONOPT for the nonlinear models and CPLEX for the linear models using an IBM RS/6000 workstation.

Example 1

Illustrative example 1 is concerned with an electrical appliances distribution company. The company has collected customer sales daily data from all points-of-sale for the last 15 weeks as shown in Table 2. Based on this historical data only, the company wishes to make an accurate customer demand forecast for the following 2 weeks. The additional 2 weeks data (week 16 and week 17) shown in Table 2 in italics are not used as training points but they are employed only for validating the accuracy of the proposed methodology.

The SVR parameter values used for example 1 can be found in Table 3. Notice that the number of past demand attributes (Δ) equals 14 days, meaning that for each prediction point we base our calculations on the previous 2 weeks data. This is done so as to reflect the periodic behaviour of customer demand. The values for SVR parameters C, ε and

Table 3. SVR parameter values for the illustrative examples.

Parameter	Example 1	Example 2	Example 3
$\bar{y}_t$	8.39	7879.241	4.638
σ_{y_t}	2.825	1831.659	2.472
k	20	20	30
Δ	14	24	12
C	16.864	13374.218	12.054
ε	0.419	393.962	0.155
γ	0.581	0.546	0.596

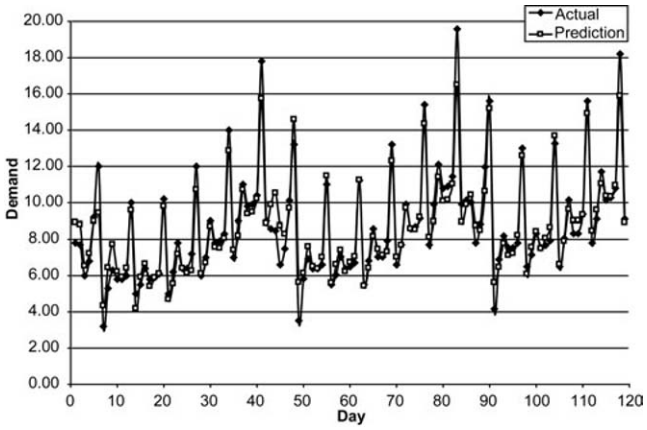

Figure 2. Customer demand forecast for example 1.

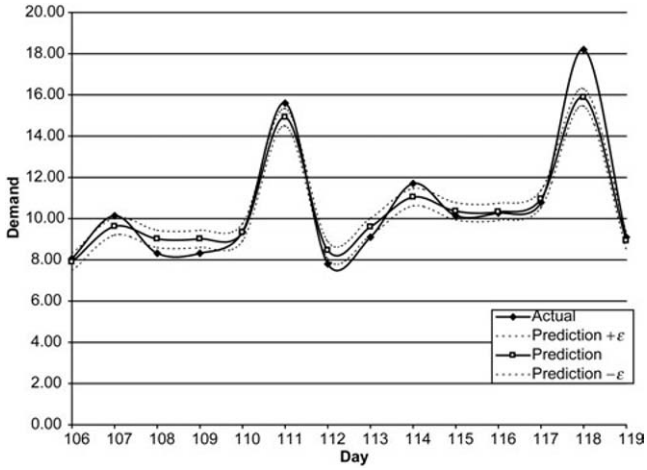

Figure 3. Focus on customer demand forecast for example 1.

Table 4. Forecasting assessment criteria for the illustrative examples.

Assessment criterion	Example 1	Example 2	Example 3
P.A.	95.22%	95.24%	93.29%
F.A.	92.48%	95.38%	91.92%
O.A.	92.80%	95.36%	92.08%

γ are derived from the mean average and standard deviation of customer demand training points by employing the heuristic rules presented in the previous section.

Customer demand forecasting results are shown graphically in Figures 2 and 3. According to the results, the proposed regression function is able to capture the basic customer demand pattern and derive accurate forecast for the 2 weeks forecasting horizon. It is very interesting to notice that the regression function can easily follow the actual customer demand even in extreme time periods where demand fluctuates much above average. For instance, increased customer demand during Saturdays is captured efficiently for both weeks in the forecasting horizon (see Figure 3, time periods 111 and 118).

Moreover, based on the training points, SVR can also capture the positive trend underlying customer demand and produce increased customer sales expectations for the forecasting horizon under investigation. Overall, the quality of the proposed forecast is determined based on the previously defined assessment criteria. Table 4 presents the forecasting assessment criteria for all illustrative examples.

Example 2

Example 2 presents a customer demand forecasting example for a chemical process industry. Monthly sales data is provided for the last 9 years (see Table 5) and the question is to accurately forecast customer demand for the next 12 months. Sales data points for the last year, shown in italics in Table 5, are not used as training points but merely employed for assessing the predictive capabilities of the SVR.

The parameter values used for this example are given in Table 3 while customer demand forecasting results are shown graphically in Figures 4 and 5. Again, just as in example 1, SVR manages to capture successfully both the positive trend and the periodic pattern of demand.

Moreover in this example, every year contains an internal pattern resulting in spiky M-shape time periods. By employing an increased number of attributes ($\Delta = 24$ months $= 2$ periods), SVR manages to overcome this difficulty in demand pattern recognition and provides accurate forecasts that follow closely the actual customer demand for both high and low demand time periods (see Figure 4). Furthermore, the proposed methodology is able to capture the effect of decreased customer demand occurring every July and derives a precise forecast for the seventh month of the forecasting horizon (see Figure 5, time period 115). According to the forecasting assessment

Table 5. Monthly chemical sales (in tonnes).

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	5055	5904	6180	7073	7545	8037	6650	6947	7222	6132	6073	3724
2	3994	4326	5176	5900	7485	8034	6431	7215	7744	8017	7685	4950
3	5971	6816	7179	8338	8549	8773	7727	8295	8679	7275	5434	4270
4	5649	7188	7837	8631	9805	9760	7921	9094	9413	8755	8069	5174
5	6066	7084	7416	8487	9054	9644	7980	8336	8666	7358	7287	4468
6	4792	5191	6211	7080	8982	9640	7717	8658	9292	9620	9222	5940
7	7165	8179	8614	10005	10258	10527	9272	9954	10414	8730	6520	5124
8	6778	8625	9404	10357	11766	11712	9505	10912	11295	10506	9682	6208
9	7077	8265	8652	9902	10563	11251	9310	9725	10110	8584	8502	5213
10	*5591*	*6056*	*7246*	*8260*	*10479*	*11247*	*9003*	*10101*	*10841*	*11223*	*10759*	*6930*

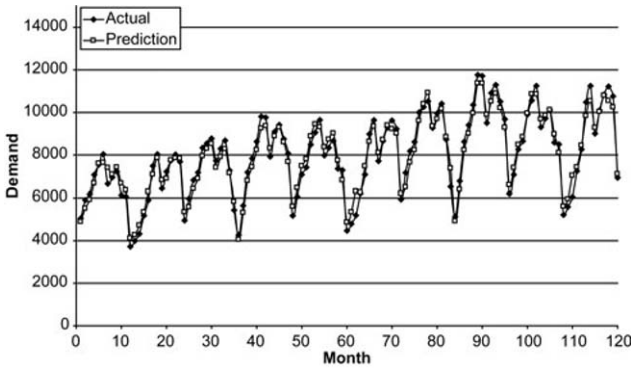

Figure 4. Customer demand forecast for example 2.

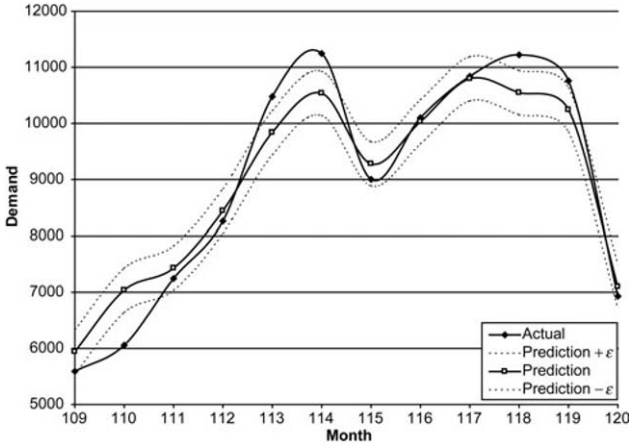

Figure 5. Focus on customer demand forecast for example 2.

criteria for example 2 given in Table 4, SVR scores well above 95% in all accuracy measures.

Example 3

Example 3 is a customer demand forecasting example in the food and drinks process industry. The monthly champagne sales data for the period between January 1964 and September 1972 has been collected from the Association of French Champagne Firms and published by Makridakis and Wheelwright (1978) as shown in Table 6. Based on the monthly champagne sales data only for the period between January 1964 and September 1971 (training points), we would like to forecast champagne sales for the following 12 months.

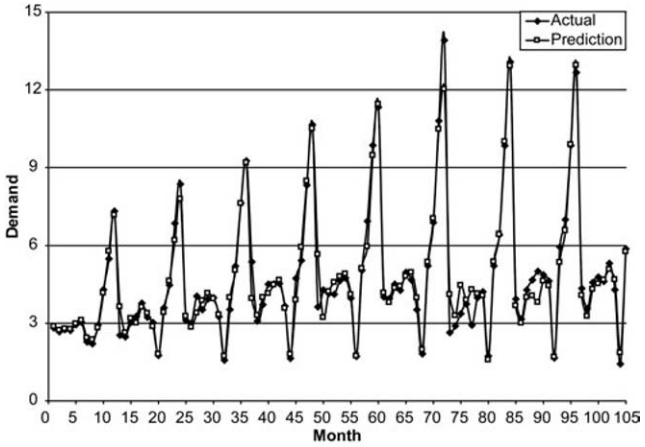

Figure 6. Customer demand forecast for example 3.

From a preliminary study of the given historical sales data, we can clearly identify a customer demand pattern that consistently repeats itself each year. Customer demand for champagne is relatively low during summer months (low-peaks every August) while it steadily builds up during autumn before reaching its high-peak every December mainly due to holidays celebrations. Based on this observation, we choose to employ 12 past demand attributes in the SVR regression input vectors. All SVR parameters used for example 3 are given in Table 3. Forecasting results for example 3 are shown graphically in Figures 6 and 7. According to the results, the proposed methodology derives a very precise prediction for the entire forecasting horizon including the December high-peak and the August low-peak (see Figure 6, time periods 96 and 104, respectively). It is also very interesting to notice that the regression function keeps very good track of the irregular customer demand and successfully reproduces the rising trend of December demands (see Figure 5). However, SVR does not naively mimic the training points but rather learns from them. The proposed forecasting algorithm manages to distinguish noisy data points from structural data points (see Figure 5, time period 72) and therefore derives a regression function that not only avoids *overfitting* but also interprets correctly the underlying customer demand pattern into a forecast of over 93% accuracy as shown in Table 4.

CONCLUDING REMARKS

In this paper, a systematic optimization-based approach for customer demand forecasting was presented based on

Table 6. Monthly champagne sales—France (millions of bottles).

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1964	2.815	2.672	2.755	2.721	2.946	3.036	2.282	2.212	2.922	4.301	5.494	7.312
1965	2.541	2.475	3.031	3.266	3.776	3.230	3.028	1.759	3.595	4.474	6.838	8.357
1966	3.113	3.006	4.047	3.523	3.937	3.986	3.260	1.573	3.528	5.211	7.614	9.254
1967	5.375	3.088	3.718	4.514	4.520	4.539	3.663	1.643	4.739	5.428	8.314	10.651
1968	3.633	4.292	4.154	4.121	4.647	4.753	3.965	1.723	5.048	6.922	9.858	11.331
1969	4.016	3.957	4.510	4.276	4.968	4.677	3.523	1.821	5.222	6.872	10.803	13.916
1970	2.639	2.899	3.370	3.740	2.927	3.986	4.217	1.738	5.221	6.424	9.842	13.076
1971	3.934	3.162	4.286	4.676	5.010	4.874	4.633	1.659	5.951	*6.981*	*9.851*	*12.670*
1972	*4.348*	*3.564*	*4.577*	*4.788*	*4.618*	*5.321*	*4.298*	*1.431*	*5.877*	—	—	—

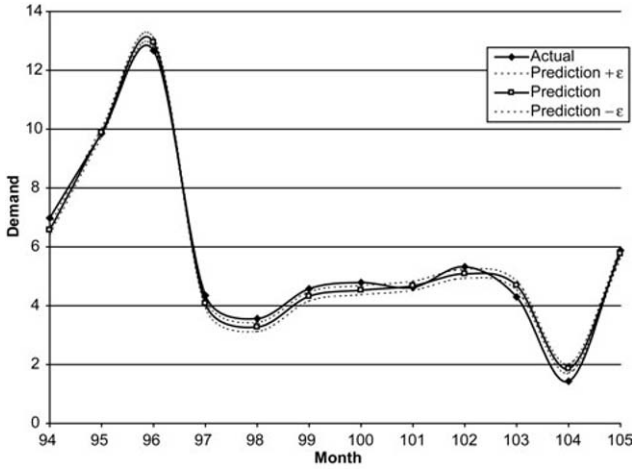

Figure 7. Focus on customer demand forecast for example 3.

SVR analysis. Historical customer demand patterns were used as training points attributes for the SVR. The proposed approach employed a three-step algorithm in order to extract information from the training points and identify an adaptive basis regression function before it performs a recursive methodology for customer demand forecasting. The applicability of the proposed forecasting approach was then validated through a number of illustrative customer demand forecasting examples. In all three examples, the proposed methodology handled successfully complex nonlinear customer demand patterns and derived forecasts with prediction accuracy of more than 93% in all cases. Although, future work should consider a formal way of determining SVR parameters, support vector regression can still be regarded as a parsimonious alternative to complex artificial neural networks forecasting.

NOMENCLATURE

C	regularization parameter
F	regression function
F.A.	fitting accuracy
k	scalar for error tolerance
K	kernel function
L	Lagrangean function
M	number of training and prediction points
N	number of training points
O.A.	overall accuracy
P.A.	prediction accuracy
p,γ	kernel function parameters
t	time periods
$\boldsymbol{w}$	slope vector
$\boldsymbol{x}_t$	input vector at time period t
x_{zt}	zth input vector attribute at time t
y_t	actual customer demand
$\bar{y}_t$	mean average customer demand
$\hat{y}_t$	predicted customer demand
z	input vector attributes

Greek Symbols

β	constant for regression function
Δ	number of past demand attributes
ε	error tolerance
$\lambda_t,\lambda_t^*,\mu_t,\mu_t^*$	Lagrange multipliers
$\lvert\xi\rvert_\varepsilon$	ε-insentitive loss function
ξ_t,ξ_t^*	slack variables at time period $t\sigma_{y_t}$standard deviation of customer demand
Φ	mapping function

REFERENCES

Agrawal, M., Jade, A.M., Jayaraman, V.K. and Kulkarni, B.D., 2003, Support vector machines: a useful tool for process engineering applications, *Chem Eng Prog*, **99**: 57–62.

Aizerman, M.A., Braverman, E.M. and Rozoner, L.I., 1964, Theoretical foundations of the potential function method in pattern recognition learning, *Automation and Remote Control*, **25**: 821–837.

Bhat N.V. and McAvoy, T.J., 1992, Determining model structure for neural models by network stripping, *Comput Chem Eng*, **16**: 271–281.

Bertsekas, D.P., 1995, *Nonlinear Programming* (Athena Scientific, Massachusetts, USA).

Box, G.E.P. and Jenckins, G.M., 1969, *Time Series Analysis, Forecasting and Control* (Holden-Day, San Francisco, USA).

Brooke, A., Kendrick, D., Meeraus, A. and Raman, R., 1998, *GAMS: A User's Guide* (GAMS Development Corporation, Washington, USA).

Burges, C.J.C., 1998, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, **2**: 121–167.

Chalimourda, A., Scholkopf B. and Smola A.J., 2004, Experimetally optimal v in support vector regression for different noise models and parameters settings, *Neural Networks*, **17**: 127–141.

Chang, C.-C. and Lin C.-J., 2001, *LIBSVM: a library for support vector machines*, software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

Cherkassky, V. and Ma, Y.Q., 2004, Practical selection of SVM parameters and noise estimation for SVM regression, *Neural Networks*, **17**: 113–126.

Cherkassky, V. and Mulier, F., 1998, *Learning From Data: Concepts, Theory and Methods* (John Wiley and Sons, New York, USA).

Chiang, L.H., Kotanchek, M.E. and Kordon, A.K., 2004, Fault diagnosis based on Fisher discriminant analysis and support vector machines, *Comput Chem Eng*, **28**: 1389–1401.

Foster, W.R., Collopy, F. and Ungar, L.H., 1992, Neural network forecasting of short noisy time series, *Comput Chem Eng*, **16**: 293–297.

Floudas, C.A., 1995, *Nonlinear and Mixed-Integer Optimization: Fundamentals and Applications* (Oxford University Press Inc., New York, USA).

Gunn, S.R., 1998, *Support Vector Machines for Classification and Regression*, Technical Report, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton, UK.

Kulkarni, A., Jayaraman, V.K. and Kulkarni, B.D., 2004, Support vector classification with parameter tuning assisted by agent-based technique, *Comput Chem Eng*, **28**: 311–318.

Lasschuit, W. and Thijssen, N., 2004, Supporting supply chain planning and scheduling decision in the oil and chemistry industry, *Comput Chem Eng*, **28**: 863–870.

Levis, A.A. and Papageorgiou, L.G., 2004, A hierarchical solution approach for multi-site capacity planning under uncertainty in the pharmaceutical industry, *Comput Chem Eng*, **28**: 707–725.

Makridakis, S. and Hibon, M., 1979, Accuracy of forecasting: an empirical investigation, *J Roy Statistical Society A*, **142**: 97–125.

Makridakis, S. and Wheelwright, S.C., 1978; *Interactive Forecasting: Univariate and Multivariate Methods* (Holden-Day Inc., San Francisco, USA).

Makridakis, S. and Wheelwright, S.C., 1982, *The Handbook of Forecasting: A Manager's Guide* (John Wiley and Sons, New York, USA).

Maravelias, C.T. and Grossmann, I.E., 2001, Simultaneous planning for new product development and batch manufacturing facilities, *Ind Eng Chem Res*, **40**: 6147–6164.

Mercer, J., 1909, Functions of positive and negative type and their connection with the theory of integral equations, *Philos Trans Roy Soc London A*, **209**: 415–446.

Myasnikova, E., Samsonova, A., Samsonova, M. and Reinitz, J., 2002, Support vector regression applied to the determination of the developmental age of a Drosophila embryo from its segmentation gene expression patterns, *Bioinformatics*, **18**: S87–S95.

Prakasvudhisarn, C., Trafalis, T.B. and Raman, S., 2003, Support vector regression for determination of minimum zone, *J Manuf Sci Eng*, **125**: 736–739.

Smola, A.J. and Scholkopf, B., 1998, *A Tutorial on Support Vector Regression*, NeuroCOLT2 Technical Report Series, available from website: http://www.neurocolt.com.

Vapnik, V.N., 1998, *Statistical Learning Theory* (John Wiley and Sons, New York, USA).

The manuscript was received 26 August 2004 and accepted for publication after revision 6 June 2005.