Nuclear power plant components condition monitoring by probabilistic support vector machine

Jie Liu, Redouane Seraoui, Valeria Vitelli, Enrico Zio

Nuclear Power Plant Components Condition Monitoring by Probabilistic Support Vector Machine

Jie Liu[a], Redouane Seraoui[b], Valeria Vitelli[a], Enrico Zio[a,c,*]

[a]Chair on Systems Science and the Energetic challenge, European Foundation for New Energy - Électricité de France, École Centrale Paris, Chatenay-Malabry, France, and Supelec (École Supérieure d'Électricité), Plateau de Moulon, Gif-sur-Yvette, France.

[c]Energy Department, Politecnico di Milano, Milano, Italy

[b]EDF R&D, Simulation and information TEchnologies for Power generation System (STEPS) Department, Chatou, France

[*]Corresponding author: Phone number: +33 1 41 13 19 14

Email: jie.liu@ecp.fr, redouane.seraoui@edf.fr, valeria.vitelli@ecp.fr, enrico.zio@ecp.fr

Abstract

In this paper, an approach for the prediction of the condition of Nuclear Power Plant (NPP) components is proposed, for the purposes of condition monitoring. It builds on a modified version of the Probabilistic Support Vector Regression (PSVR) method, which is based on the Bayesian probabilistic paradigm with a Gaussian prior. Specific techniques are introduced for the tuning of the PSVR hyerparameters, the model identification and the uncertainty analysis. A real case study is considered, regarding the prediction of a drifting process parameter of a NPP component.

Key words: Probabilistic support vector machine, Condition monitoring, Nuclear power plant, Point prediction

1. Introduction

Production systems are becoming increasingly complex and demand sophisticated methods to anticipate, diagnose and control abnormal events in a timely manner, as the consequences of unexpected faults can bring high economic losses for a company (Venkatasubramanian, 2005).

For an optimized operation, the conditions of NPP components and systems are usually monitored at regular intervals (Condition Monitoring), and a warning is triggered when the monitored signals exceed predefined thresholds (Fault Detection) (Zio et al., 2010).The plant operators must identify the plant state and the components out of control (Diagnostics), and predict the future development of the scenarios (Prognostics) to decide the actions to take to regain safe control of the plant (Zio, 2012). Then, while diagnostics aims at identifying the cause of the deviation from normal behavior and at determining the state of the parameters critical for the plant operation and safety, prognostics aims at the prediction of the Residual Useful Life (RUL) of the components (Zio, 2012).

In general, two strategies for condition monitoring, detection, diagnostics and prognostics are possible: either based on physical models, or based on data-driven approaches (Zio, 2012; Ma and Jiang, 2011). In the case of complex systems, physical models can be built only after simplification of the physical relations. Then, in most cases, they cannot timely provide the plant operators with a sufficiently precise diagnostics of the plant situation (Zio, 2012). On the contrary, data-driven approaches are attractive for NPPs, also considering that most components are monitored since the commissioning of plants, and, hence, a large amount of measured data is available to drive the tuning of the models (Ma and Jiang, 2011).

A substantial amount of research has concerned the development of data-driven approaches for condition monitoring, detection, diagnostics and prognostics. Artificial Neural Network (ANN), Support Vector Machine (SVM), Genetic Algorithm (GA) and Auto-Associative Kernel Regression (AAKR) are among some of the most studied and applied (Chevalier et al., 2009; Baraldi et al., 2010; Baradi et al., 2011; Santosh et al., 2009; Li et al., 2012; Yazikov et al., 2012; Rand et al., 2012a; Rand et al., 2012b; Muralidharan and Sugumaran, 2012; Ekici, 2012; Zio and Gola, 2006; Lu and Upadhyaya, 2005; Jeong et al., 2003; Zio et al., 2009). These approaches are already mature, especially for detection and diagnostics. On the contrary, the amount of research dealing with prognostics is limited, especially in the context of NPP components. Some recent references, referring to prognostics for engineering systems, are Li and Nilkitsaranont (2009), Niu and Yang (2010) and Wang et al. (2004). Support Vector Regression (SVR) is used in Trontl et al. (2007) and Bae et al. (2008) to fulfill the point estimation with satisfactory results. In Elnokity et al. (2012), a hybrid modeling combined with the Industrial Source Complex (ISC) model and an Adaptive Neuro-Fuzzy Inference System (ANFIS) has been used to improve the modeling ability of predicting tracer concentrations. SVR method is used in Cai (2012) to predict the critical heat flux, while Fuzzy Neural

Networks are used in Na et al. (2006) to estimate the collapse moment due to the wall-thinned defects of bends and elbows in piping systems. However, uncertainty quantification is not included in the previously described data-driven models. In Zio et al. (2010) and Zio and Di Miao (2010), a fuzzy similarity analysis is introduced to compare the evolving failure scenario with a library of reference patterns describing the multidimensional evolution of monitored process variables. The aim is to find a combination of the reference patterns, weighed by their similarity to the observed failure pattern, to determine the future evolution of the scenario and to derive the corresponding RUL. However, failure patterns in NPP components are rare and thus a "solid" library of references cannot be easily formed. SVR has also been used in Kim et al. (2012) to build Prediction Intervals (PIs) for the same problem. However, since the method has been trained on a relatively small amount of data, its generalization power is not assured.

It is well recognized that there exists no prognostic method that is ideal for every situation (Jardine et al., 2006; Y-C and Pepyne, 2001). A variety of methods have been developed for specific situations or specific classes of systems. In the present work, we propose a method for prediction with uncertainty quantification, in the context of NPP components condition monitoring and prognostics. We address the problem of predicting process variables under conditions of fault of a NPP component. A modified Probabilistic Support Vector Regression (PSVR) is developed and used to provide in output the PIs of a process variable. To the author's knowledge, this is the first time that such technique is applied in the specific application context of interest. A real case study is considered, related to the condition monitoring of a component of a NPP of Électricité De France (EDF). A main challenge arises from the need of building a model based on only one scenario, which is a realistic situation given the rarity of faults in NPP components.

The paper is structured as follows. Section 2 provides a description of the PSVR method for prognostics. Section 3 presents the characteristics of the data of the real case study, and the pre-treatment techniques used to remove the outliers, to reconstruct the missing data points, and to identify the most proper model. In Section 4, the results of the application of PSVR for prognostics are presented, and comparisons with the standard SVR method and other empirical apporaches are also given in this Section. Some conclusions are drawn in Section 5.

2. Probabilistic Support Vector Regression (PSVR)

Standard Support Vector Machines (SVMs) (Cortes and Vapnik, 1995; Vapnik et al.; 1996; Boser et al., 1992; Drucker et al., 1997; Cristianini and Taylor, 2000) are learning machines implementing the Structural Risk Minimization (SRM) inductive principle to obtain good generalization performance on a limited number of learning patterns (Gao et al., 2002; Jardine et al., 2006; Poggio and Girosi, 1998; Girosi, 1998). However, the parameters need to be specifically tuned for the

problem at hand and this may be difficult. Another problem related to SVMs is that the classification and regression results are provided as point estimates only, while it would be more informative to obtain a Prediction Interval (PI) with an associated probability that the true value lies in the interval. Also, the distribution of the predicted value is a constructive indicator for practical purposes.

To overcome these limitations, the Bayesian probabilistic paradigm has been considered in combination with SVM (Mackay, 1997; Neal, 1996; Williams, 1997). Recently, it has been shown that SVMs can be interpreted as a Maximum A Posteriori (MAP) solution to a Bayesian inference problem with Gaussian priors and an appropriate likelihood function. This probabilistic interpretation enables Bayesian methods to be employed to determine the regularization parameters in the SVM framework (Kim et al., 2012; Sollich, 1999). The method using MAP for SVM estimation is called Probabilistic Support Vector Regression (PSVR). Bayesian approaches for SVM can estimate the parameter and feature spaces simultaneously by maximizing the evidence function, and they allow obtaining an error bar for the prediction (Lin and Weng, 2004).

2.1 PSVR Using Ɛ-Insensitive Loss Function

Let us assume that the input data is a n-dimensional set of vectors $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$, independently drawn in $\boldsymbol{R}^p$, and that we also have an independent sample from the target value $\boldsymbol{Y} = \{y_1, y_2, \dots, y_n\}$, where $y_i \in \boldsymbol{R}, i = 1, 2, \dots, n$.

In regression methods, the final aim is to find an underlying function $a(\boldsymbol{x})$: $\boldsymbol{R}^p \rightarrow \boldsymbol{R}$ describing the relation between the input data and the target. We will now briefly state the PSVR approach to the estimation of $a(\boldsymbol{x})$; further mathematical details on the derivation of the method can be found in the Appendix, and in the references therein.

We make the following assumptions:

(1) Training data set $\boldsymbol{\Gamma} = \{\boldsymbol{X}, \boldsymbol{Y}\}$ follows an identical and independent distribution (i.i.d).

(2) The *a priori* probability distribution is $P[\boldsymbol{a}(\boldsymbol{X})] \propto \exp(-\frac{1}{2} \| \hat{P} a \|^2)$, where $\| \hat{P} a \|^2$ is a positive semi-definite operator and $\boldsymbol{a}(\boldsymbol{X}) = (a(\boldsymbol{x}_1), a(\boldsymbol{x}_2), \dots, a(\boldsymbol{x}_n))^T$.

(3) The Ɛ -insensitive loss function is chosen as the loss function.

(4) The covariance function is $K(\boldsymbol{x}, \boldsymbol{x}')$, and $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2}{2\gamma^2})$, where $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ are the input data points in $\boldsymbol{X}$.

The a posteriori probability of $\boldsymbol{a}(\boldsymbol{X})$ can be written as

$$P[\boldsymbol{a}(\boldsymbol{X})|\boldsymbol{\Gamma}] = \frac{[G(C,\varepsilon)]^N}{\sqrt{det 2\pi K_{X,X}} P[\boldsymbol{\Gamma}]} \exp\{-C \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} L_\varepsilon\big(y_i - a(\boldsymbol{x}_i)\big) - \frac{1}{2}\boldsymbol{a}(\boldsymbol{X})^T K_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{a}(X)\}, \quad (1)$$

where $G(C,\varepsilon) = \frac{1}{2}\frac{C}{C\varepsilon+1}$, and $K_{\boldsymbol{X},\boldsymbol{X}} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]$ is the covariance matrix of the data points of $\boldsymbol{X}$.

We find the maximum of Equation (1) using the so-called MAP. This requires finding the minimum of the following function

$$R_{GSVM}(a) = C \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} L_\varepsilon\big(y_i - a(\boldsymbol{x}_i)\big) + \frac{1}{2}\boldsymbol{a}(\boldsymbol{X})^T K_{\boldsymbol{X},\boldsymbol{X}}^{-1}\boldsymbol{a}(\boldsymbol{X}) \quad (2)$$

We can see that the risk of Gaussian SVMs is equivalent to the standard SVM. Following the discussion in Mackay (1997), Tikhonov and Arsenin (1997), Girosi (1998) and Burges (1998), we can write the solution of the minimization problem associated to Equation (2) in the following form

$$a^*(\boldsymbol{x}) = \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \beta_i K(\boldsymbol{x}_i, \boldsymbol{x}) \quad (3)$$

where $\beta_i = a_i - a_i^*$ is a combination of the Lagrange Multipliers associated to the optimization problem (Smola and Scholköpf, 2004). The a_i and a_i^* can be determined by a Quadratic Programming approach. According to Smola and Scholköpf (2004), $\forall\, i = 1, \ldots, n$, a_i and a_i^* lie in the interval $[0, C]$, and β_i consequently lies in the interval $[-C, C]$, which is the domain of the optimization problem. See Na et al. (2006) for more details on the implementation.

2.2 Hyperparameters

According to the description of the PSVR method given in the previous Section, we shall now detail a strategy to determine the three hyperparameters C, ε, γ, before the optimization algorithm is initialized.

Parameter C is the penalty factor. It controls the trade-off between complexity and the proportion of non-separable samples, and must be selected by the user (Vladimir et al., 1998). If it is too large, it will induce a high penalty for non-separable points, hence we may store too many support vectors and go towards over fitting. If it is too small, it may result in underfitting (Alpaydin, 2004). For what concerns the optimization process, C influences the computational burden of the regression: the bigger C is, the heavier the computational burden is.

Parameter ε controls the sparsity of the data. It has an effect on the smoothness of the SVM response and it affects the number of support vectors, so both the complexity and the generalization power of the network depend on its value (Horváth, 2001). By inspecting the ε-insensitive loss function (see the details in the Appendix), we see that data points inside a tube of radius ε surrounding

the predicted values, are not considered in training the regression model. This is graphically exemplified in Figure 1.

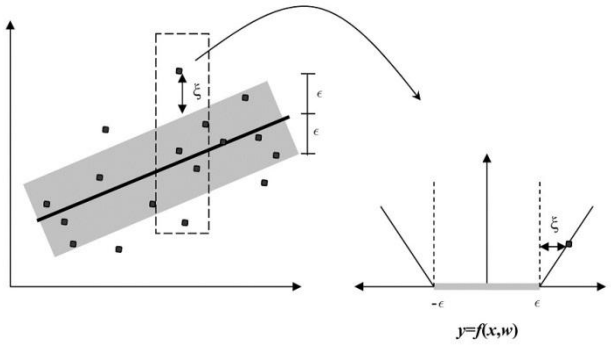

Fig.1 A picture of the ε-insensitive loss-function behavior.

Finally, parameter γ influences the width of the kernel, and hence the accuracy of the prediction and its variability.

There are already some methods in the literature to determine these hyperparameters, e.g. VC-theory in Vapnik (1995), Bayesian method in Mackay (1991), AIC in Akaike (1974), NIC in Murata et al. (1994) and Maximizing Evidence Function in Kim et al. (2012). In this paper, an interpolation method based on an innovative criterion is used to obtain the best values of these three parameters. The details are illustrated in Section 4, directly in relation to the case study.

2.3 Error Bar Estimation

In a Bayesian treatment of the prediction problem, error bars arise naturally from the predictive distribution. They are made up of two terms, one due to the *a posteriori* uncertainty (the uncertainty of $a(\boldsymbol{x})$), and the other due to the intrinsic noise in the data (Kim et al., 2012). Suppose that $\boldsymbol{x}$ is a test input vector, and that the corresponding value of the target is the random variable y, obtained adding to $a(\boldsymbol{x})$ an unknown noise δ with zero mean; then

$$P[\boldsymbol{\Gamma}|\boldsymbol{a}(\boldsymbol{X})] \propto \exp(-C\sum_{i=1}^{n} l(\delta_i)). \tag{4}$$

We can also obtain the density of the noise δ

$$P[\delta] = \frac{C}{2(C\varepsilon+1)}\exp(-Cl_{\varepsilon}(\delta)), \tag{5}$$

and the noise variance

$$\sigma_\delta^2 = \frac{2}{C^2} + \frac{\varepsilon^2(C\varepsilon+3)}{3(C\varepsilon+1)}. \quad (6)$$

The conditional probability distribution of $a(\boldsymbol{x})$ given $\boldsymbol{\Gamma}$ can instead be written as

$$P[a(\boldsymbol{x})|\boldsymbol{\Gamma}] = \frac{1}{\sqrt{2\pi}\sigma_t}\exp\{-\frac{(a(\boldsymbol{x}) - a^*(\boldsymbol{x}))^2}{2\sigma_t^2}\}, \quad (7)$$

with

$$\sigma_t^2(\boldsymbol{x}) = K(\boldsymbol{x},\boldsymbol{x}) - K_{\boldsymbol{X}_M,\boldsymbol{x}}^T K_{\boldsymbol{X}_M,\boldsymbol{X}_M}^{-1} K_{\boldsymbol{X}_M,\boldsymbol{x}}. \quad (8)$$

Consequently, the error bar width of the prediction corresponding to the test input point $\boldsymbol{x}$ is

$$\sigma^2(\boldsymbol{x}) = \sigma_\delta^2 + \sigma_t^2(\boldsymbol{x}) = \frac{2}{C^2} + \frac{\varepsilon^2(C\varepsilon+3)}{3(C\varepsilon+1)} + K(\boldsymbol{x},\boldsymbol{x}) - K_{\boldsymbol{X}_M,\boldsymbol{x}}^T K_{\boldsymbol{X}_M,\boldsymbol{X}_M}^{-1} K_{\boldsymbol{X}_M,\boldsymbol{x}}. \quad (9)$$

The conditional probability distribution and the error bar are given in Equations (7) and (9). See Na et al. (2006) for more details on the calculations.

3. Case Study Description

A set of data from the Reactor Coolant Pump (RCP) of one of EDF's NPPs is used to test the efficiency and the accuracy of the PSVR modeling approach developed in our work. In the following, we describe the data and illustrate the pre-processing steps.

3.1 Data Description

The dataset includes the measurements of the RCP of a NPP, with increasing leak flow in the first seal (a variable denoted with IntVar 9). The dataset contains the values of seventeen different variables recorded by seventeen different sensors along a period of 406 days. The variables whose measurements concern sensors inside the RCP are hereafter called internal variables; the others are called external variables. The description of all the internal and external variables and their physical meanings are given in Table 1.

There are nine internal variables and eight external variables. Each of the variables is observed hourly for a period of more than 13 months, about 9200 observation points. The evolution of four of the variables is shown in Figure 2: from left to right, and from top to bottom, ExtVar 2, ExtVar 6, ExtVar 7 and IntVar 9. At the 5700th observation instance, we observe the fault, manifested by the variable IntVar 9 going out of control. We note that: all the variables are time-dependent, and there are seventeen variables in total, hence leading to a multivariate problem; each variable is measured hourly, giving 9205 measurements for each variable, and hence making computations challenging; all the variables show a nonlinear behavior, hence requiring a nonlinear model; the data need pre-processing,

because there are many outliers and missing observations. Missing data are due to the absence of sensors recording during some time instances, while outliers correspond to bad (extremely high or low) sensor recordings. Concerning internal variables, the total number of missing data is 377 in IntVar 1, 415 in IntVar 2, 512 in IntVar 3, IntVar 4, IntVar 5 and IntVar 6, 493 in IntVar 7, 462 in IntVar 8 and 409 in IntVar 9. In the time series of external variables, there are 434 missing data in ExtVar 1, 409 in ExtVar 2, 422 in ExtVar 3, 428 in ExtVar 4, 372 in ExtVar 5, 453 in ExtVar 6, 512 in ExtVar 7, and 500 in ExtVar 8.

Tab.1 Physical meaning of each internal and external variable.

Internal variables		**External variables**	
Name	**Physical meaning**	**Name**	**Physical meaning**
IntVar 1	T cold leg loop 1 [WR]	ExtVar 1	T by-pass hot leg loop 3
IntVar 2	T water seal #1 051PO	ExtVar 2	T seal injection line
IntVar 3	T stator winding motor 051PO	ExtVar 3	P primary amount file B [GL]
IntVar 4	T motor lower bearing 051PO	ExtVar 4	Debit general file A
IntVar 5	T lower thrust bearing 051PO	ExtVar 5	Debit general file B
IntVar 6	T motor upper bearing 051PO	ExtVar 6	T aval exchanges file A
IntVar 7	T motor upper thrust bearing 051PO	ExtVar 7	T aval exchanges file B
IntVar 8	Flow seal injection supply RCP051PO	ExtVar 8	Debit refrigeration GMPP 051PO
IntVar 9	Seal leak flow #1 RCP051PO		

3.2 Data Pre-processing

Since the dataset we are going to analyze contains both missing data and outliers, we have to deal with both these issues. First of all, we will remove anomalous data, since their extreme values would affect the results of the analysis. Outliers can be easily detected by deciding some constraints, e.g. the limits $\overline{x} \pm 3 * \sigma_x$ where $\overline{x}$ is the mean of all the data points and σ_x is their standard deviation. These limits are needed to detect the outliers, selected as those data points bigger than $\overline{x} + 3 * \sigma_x$ or smaller than$\overline{x} - 3 * \sigma_x$, and subsequently removed. Note that we used such constraints, rather than the usual ones based on the median and the InterQuartile Range (IQR), to be more conservative in the outlier selection, due to the dependence among data.

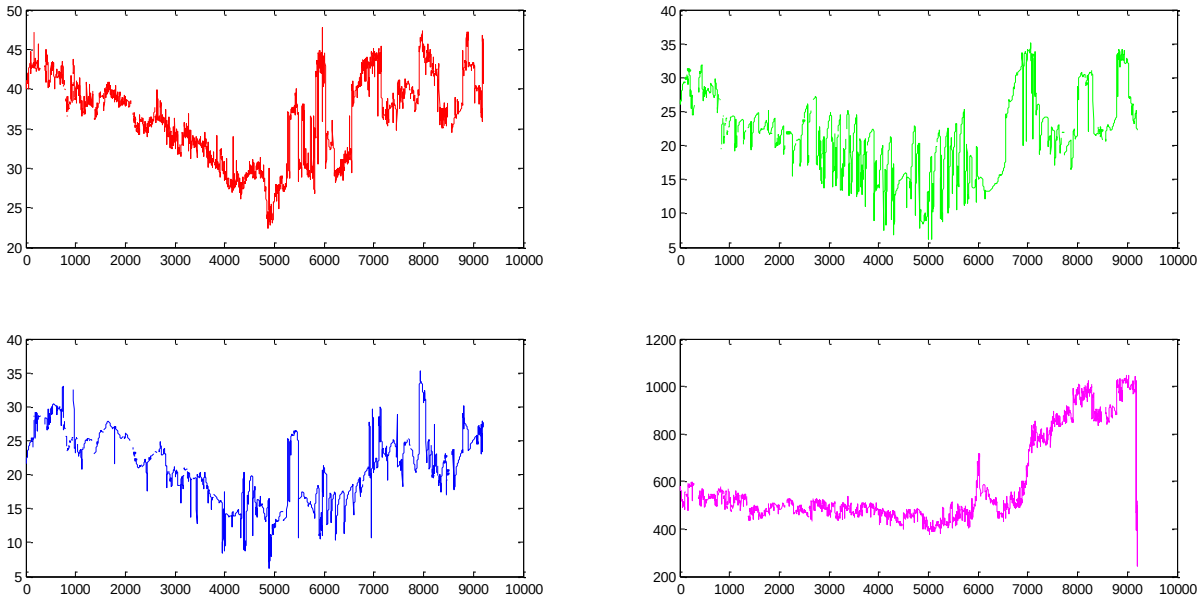

Fig.2 The evolution of four of the variables included in the dataset: from left to right, and from top to bottom, the measurements of ExtVar 2, ExtVar 6, ExtVar 7 and IntVar 9. On the axis, time measured in hours.

Secondly, we want to reconstruct missing data. Note that, after the outlier selection and elimination procedure, the number of missing data has increased. A possible way to deal with the reconstruction of missing data is the local polynomial regression fitting (Masry and Mielniczuk, 1999). This local least squares regression technique estimates effectively the values of the internal and external variables when there are missing data points. Moreover, it can also be used to perform the smoothing of the available observations, in order to reduce noise. We will thus use this technique both to reconstruct data where missing, and to obtain a smoother and less noisy time series in all remaining time instances.

Precisely, if we denote by t_0 a generic time instance, we execute the following steps to perform local polynomial regression:

(1) Find the k-nearest neighbors of t_0, which constitute a neighborhood $N(t_0)$: this means finding the k time instances in the time series which are closest to t_0. The number k is determined by setting it equal to a selected percentage (called *span*) of the data; note that the *span* can be eventually different for each variable to allow flexibility. In the case of our application, three different values of the span have been selected, according to a trial-and-error procedure: 0.5% (high), 0.2% (medium) and 0.08% (low). For each of the variables, the most proper value of the *span* (high, medium or low) has been selected to be the most suited to the noise level of the variable.

(2) Calculate $D(t_0) = \max d(t, t_0)$ over $t \in N(t_0)$, where d is the Euclidean distance between the data at time t and t_0.

(3) For each point $t \in N(t_0)$, calculate its weight $W(t) = (1 - |\frac{t - t_0}{D(t_0)}|^3)^3$ with a tri-cube weight function.

(4) Calculate the weighted least square fit of t_0 on the neighborhood $N(t_0)$.

By repeating these steps for all time instances, all the variables are smoothed and reconstructed. Some examples of the so obtained time series are shown in Figure 3: they are the smoothed and reconstructed data corresponding to the variables in Figure 2. For the variables shown in Figure 3, the *span* parameter has been fixed to 0.5%.

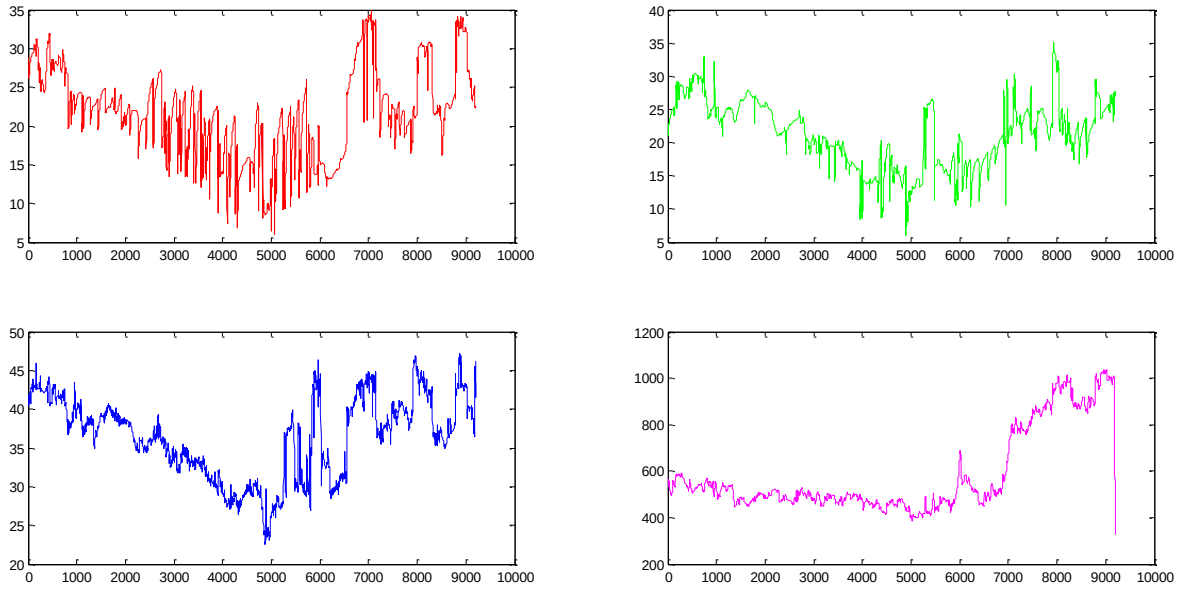

Fig.3 The smoothing and reconstruction of the evolution of the four variables whose raw observations are shown in Figure 2.

3.3 Model Identification

In order to select the most proper variables to be included as inputs in the PSVR model for improved prediction accuracy and reduction of the computational burden, a correlation analysis is carried out between the target variable IntVar 9 and the other internal and external variables. The inputs are chosen to be the variables maximizing their correlations with the target IntVar 9. Correlations are measured by the classical Pearson correlation coefficient (Rodgers and Nicewander, 1988). Table 2 shows the results of the analysis.

Tab.2 Correlations of the target variable with other internal and external variables.

Correlations	**Internal variables**							
	IntVar 1	IntVar 2	IntVar 3	IntVar 4	IntVar 5	IntVar 6	IntVar 7	IntVar 8
IntVar 9	0.03128	0.48797	0.55268	0.50926	0.50701	0.58884	0.48164	0.19193
Correlations	**External variables**							
	ExtVar 1	ExtVar 2	ExtVar 3	ExtVar 4	ExtVar 5	ExtVar 6	ExtVar 7	ExtVar 8
IntVar 9	-0.44992	0.50569	0.12352	-0.24375	0.24569	0.43695	0.37322	-0.03361

Three external variables are the most related to the target: ExtVar 2, ExtVar 6 and ExtVar 7, corresponding to a correlation of 50.5%, 43.7% and 37.3%, respectively (see Table 2). Some of the internal variables have also a strong correlation with the target, with a correlation of more than 48%: IntVar 2, IntVar 3, IntVar 4, IntVar 5, IntVar 6 and IntVar 7. Hence, these six most related internal variables, and the three most related external variables, are included as inputs in the prediction model. IntVar 8 is also chosen as input, as suggested by expert judgment. The results are given in the next Section.

Historical values of the target can also be exploited as inputs to improve the accuracy of the prediction. In order to determine the most proper temporal horizon of the target for prediction purposes, i.e. the number of previous values to be used in the model, an autocorrelation analysis is carried out on the time series of the target values. The results of this analysis are reported in Figure 4, where the empirical partial autocorrelation function is plotted against the corresponding temporal lag (a multiple of one hour). It is evident that the correlations decrease with lag, and after a lag of three time steps (i.e. three hours) they are no longer significantly different from zero. Indeed, the dashed horizontal lines in the plot are the limits of the region of acceptance for a statistical test with null hypothesis being zero partial autocorrelation. Hence, only the first three historical values of the target, i.e. three hours before, are added as inputs to the three most correlated external variables.

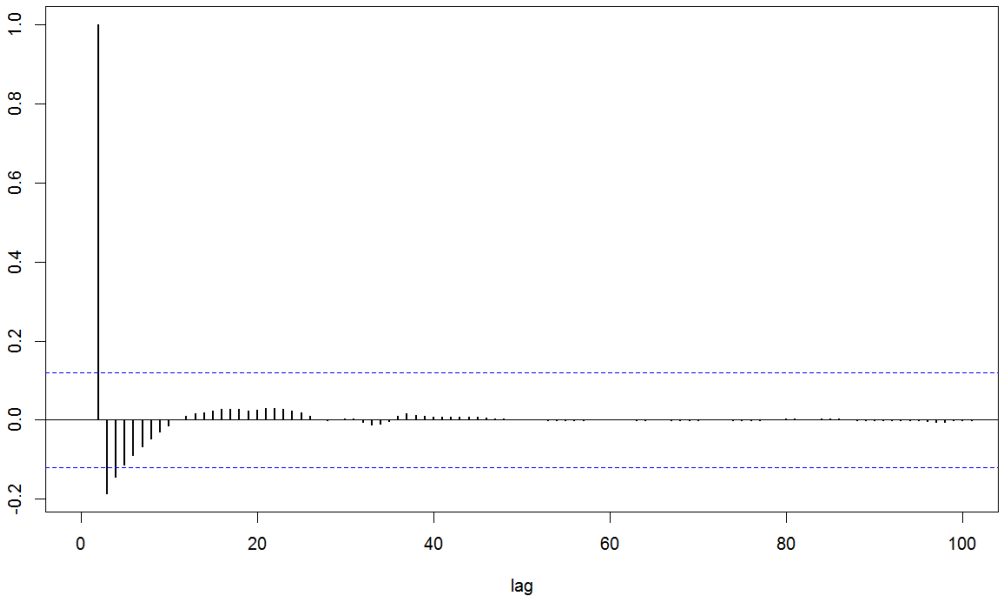

Fig.4 Empirical partial autocorrelation function of the time series of the target values (IntVar 9) with respect to time laps (multiples of hours)

4. Case Study Results

In this Section, we describe the results obtained using the PSVR method to give the prediction intervals for the target of interest in the context of condition monitoring of NPP components. The target is the variable IntVar 9, observed in its "out-of-control" regime after a fault occurred, and we focus on short-term (1-hour ahead) prediction. Assuming that we are at time t and we want to predict the target value at time $t+1$, we use as inputs the historical values of the target itself till three time steps before t, the values of the three most correlated external variables at time t and the values of the six most correlated internal variables at time t. We select a portion of the scenario under fault to apply PSVR for prediction: from the 5600th to the 8000th observed data. In this Section, 200 data points (5600th-5800th observations) are used for training and the rest for testing.

4.1 Tuning of the Hyperparameters

In order to achieve good prediction performance, we need to select the values of the hyperparameters C, ε and γ. The values of the hyperparameters influence the results of PSVR but a unifying method to determine their values has not yet been established. We propose a novel method which gives promising results. A comparison with two alternative methods is also conducted.

The method proposed in the present paper to determine the best values for the three hyperparameters is a simple but effective iterative search based on interpolation. Each parameter is initially selected within a given interval. The best values are to be found by minimizing the following criterion

$$C_1 \sum_{i=1}^{n} \sigma_i + C_2 \sum_{i=1}^{n} |y_i^* - y_i| \quad (10)$$

where σ_i is the error bar width, $y_i^* = a^*(\boldsymbol{x}_i)$ the predicted value and y_i the target value of the i^{th} input data point. C_1 and C_2 are the two weights of the two parts of the objective function (Equation (10)), the error bar width and the bias of the prediction. If C_1 is smaller than C_2, it means that we pay more attention to the variance of the prediction (error bar width) than to the accuracy in the prediction (distance between target and predicted values), and vice versa for C_1 bigger than C_2. We fix $C \in [10, 10^5]$, $\gamma \in [10^{-7}, 10^3]$, $\varepsilon \in [10^{-3}, 10^{-1}]$, $C_1 = 4$ and $C_2 = 5$ by a trial-and-error process. For each parameter, a geometric sequence included in the corresponding interval is considered. In this applicative context, geometric sequences are better than arithmetic ones, since the parameter's influence on the objective function (Equation (10)) is highly non-linear. For C, ε and γ, geometric sequences of size 4, 10 and 4 are formed respectively. Note that for different training data sets, the best values of the parameters can change: hence, the tuning of the parameters in a feasible computational time is a relevant issue. In this case, the optimization of the objective function (Equation (10)) leads to the following choice for C, ε and γ: (6309.6, 0.0032, 7).

The results obtained via PSVR where the tuning of the hyperparameters is conducted according to the method proposed by the authors are compared with two alternative methods based on, respectively: the minimization of the objective function of the PSVR (Equation (2)) and the widely used minimization of the Mean Square Error (MSE) between the predicted value and the target of the training data set. The best combinations of C, ε and γ determined by these last two approaches are (398.1072, 0.3162, 2.5119) and (6309.6, 0.001, 3), respectively. A comparison of the results obtained with each of these strategies will be shown in the next Section.

4.2 PI for the Target and Conditional Predictive Distribution

The results of the application of PSVR are shown below. Figure 5 depicts the prediction of the target, with the corresponding Prediction Interval (PI) with a confidence level of 95%, obtained by tuning the hyperparmeters according to the novel strategy proposed by the authors. The solid line is the target, the dash-dot line is the point prediction, while the two dashed lines are the upper and lower bounds of the 95% PI computed according to the predictive distribution. Hence, for each test point $\boldsymbol{x}$, the PI bounds are the values $L(\boldsymbol{x})$ and $U(\boldsymbol{x})$ corresponding to a 95% confidence that $y(\boldsymbol{x})$ lies in the interval $[L(\boldsymbol{x}), U(\boldsymbol{x})]$. In particular, the PI corresponding to the test point $\boldsymbol{x}$ is $[a^*(\boldsymbol{x}) - 2\sigma(\boldsymbol{x}), a^*(\boldsymbol{x}) + 2\sigma(\boldsymbol{x})]$, where $a^*(\boldsymbol{x})$ is the predicted value according to Equation (3) and $\sigma(\boldsymbol{x})$ is the variance associated to the prediction (error bar) and given by Equation (9). We remark that the predictive distribution in $\boldsymbol{x}$ is a Gaussian with mean $a^*(\boldsymbol{x})$ and variance $\sigma(\boldsymbol{x})$. Figure 6 shows the predictive distribution associated to the 7500th target data point according to Equation (7). The circle in Figure 6 is drawn in correspondence to the target value.

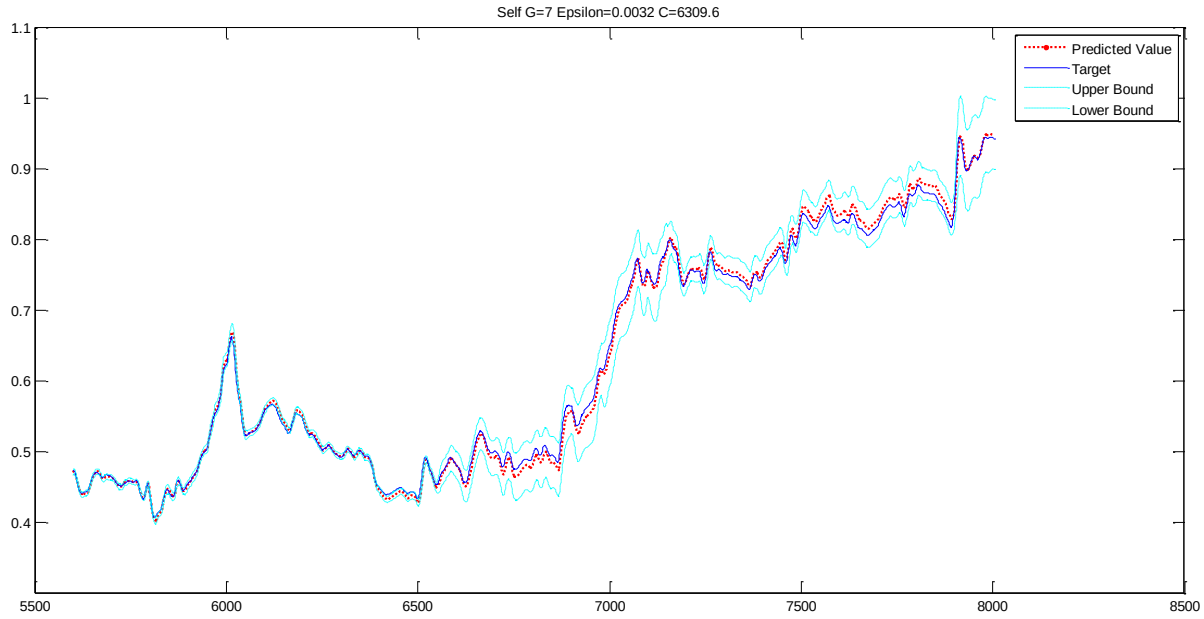

Fig.5 Point prediction and associated PIs for the target of interest (both the training and testing data points) using PSVR with hyperparameters tuning according to the proposed method.

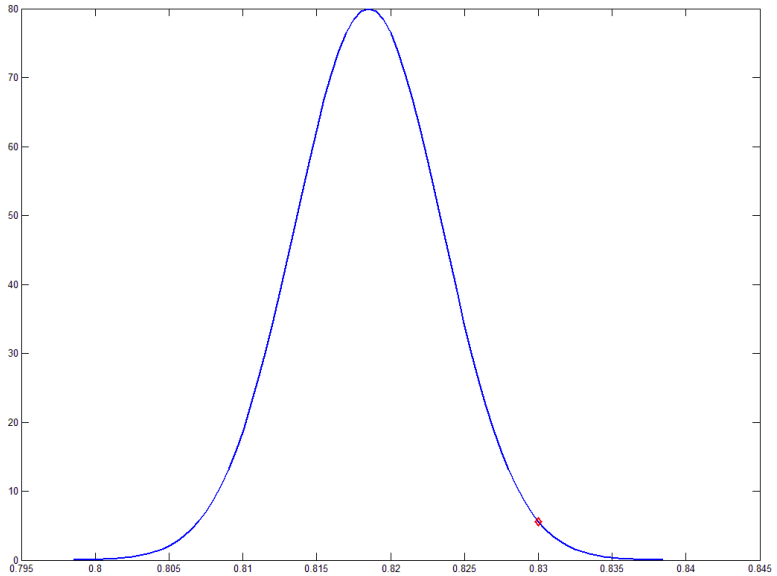

Fig.6 Predictive distribution associated to the 7500th target data point (circle), and obtained by using the PSVR with hyperparameters tuning according to the proposed method.

The prediction interval empirical coverage estimated on the whole testing set is 91.50%. The MSE is $5.4332*10^{-5}$. The relative error is smaller than 4%. If the model is trained using a bigger training data set, the relative error and absolute error will decrease.

The results of the application of PSVR (prediction of the target and 95% confidence PIs) with tuning of the hyperparameters according to the objective function (Equation (2)) and the MSE are shown in Figure 7 and Figure 8, respectively. Moreover a comparison of the three methods used for determining the values of the hyperparameters in terms of average width of PIs, mean relative error and mean absolute error is offered in Table 3. It is obvious that the proposed method is the best both in terms of prediction accuracy and precision. It is reasonable that the objective function of PSVR gives the worst results, because the objective function (Equation (2)) is used as a criterion to determine the weights of the support vectors in PSVR, and thus it is not expected to be suited also for determining the values of the hyperparameters. The results obtained via MSE are a little worse than the ones obtained by using the method proposed by the authors. This is mainly caused by the fact that MSE looks only to prediction accuracy and not at PIs width.

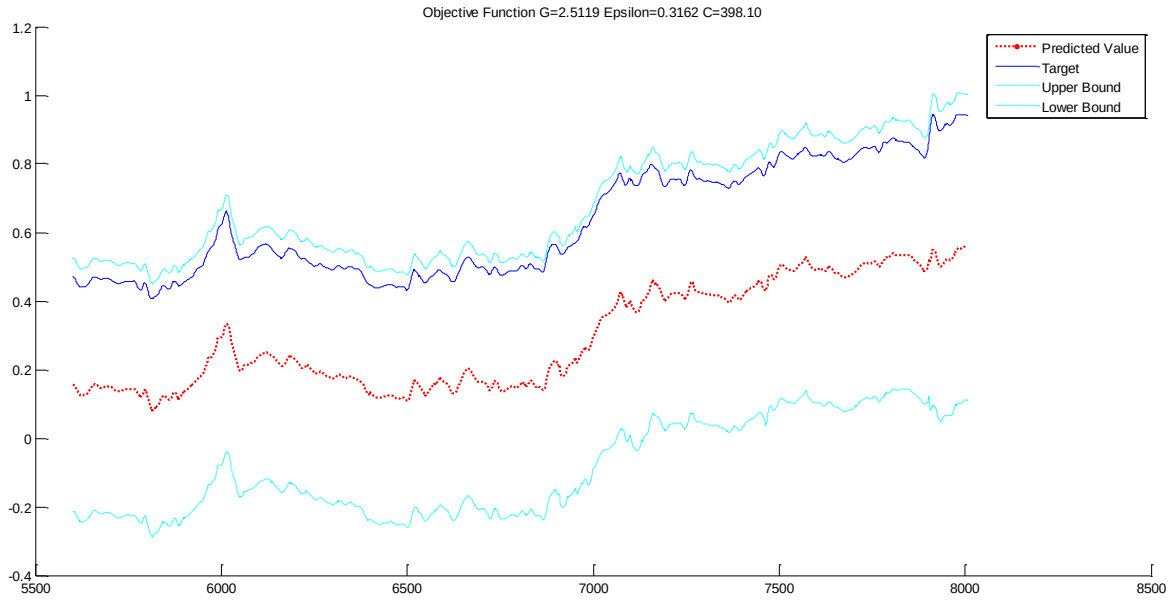

Fig.7 Point prediction and associated PIs for the target of interest (both the training and testing data points) using PSVR with hyperparameters tuning according to the objective function.

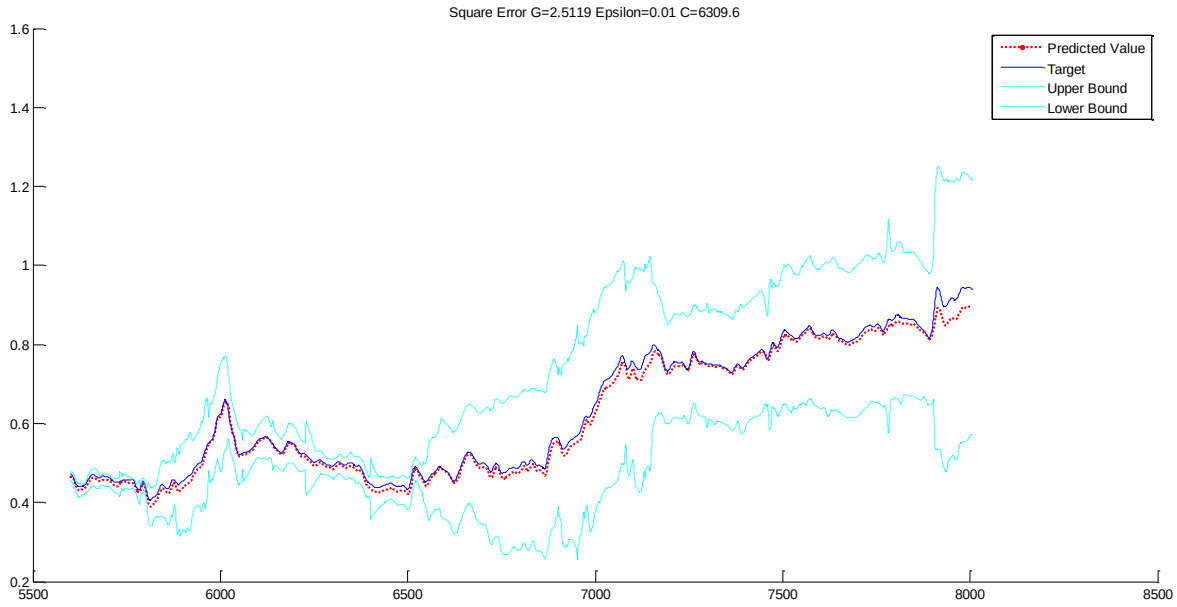

Fig.8 Point prediction and associated PIs for the target of interest (both the training and testing data points) using PSVR with hyperparameters tuning according to MSE.

Tab.3 Comparison of the results of different methods for determining the values of hyperparameters

Methods	Average Weights of PIs	Mean Relative Error	Mean Absolute Error
Proposed Method	0.0099	0.0093	0.0059
Objective Function of PSVR	0.1903	0.5613	0.3318
Mean Square Error	0.0671	0.0183	0.0114

4.3 Comparisons with other empirical approaches

In this Section, a comparison of the results obtained by PSVR and by other standard empirical approaches to short term prediction is illustrated. The empirical approaches we are going to consider are Auto-Associative Kernel Regression (AAKR) method, a well-established benchmark empirical approach to condition monitoring and prognostics, and Standard SVR, which corresponds to PSVR in a non-Bayesian framework.

AAKR is a well-known and established method suited both for reconstruction and for prediction purposes. It is an empirical modeling technique in which the prediction is found as a weighted sum of the previous values of the target variable. In order to determine the weights, AAKR makes use of historical observations of all signals to compute a global similarity measure, typically based on a Gaussian kernel, at each time. Further details on the method can be found in Baraldi *et al.* (2010). The main difference of this approach from both PSVR and SVR is the lack of a model for prediction:

internal and external variables are used by AAKR just to compute the weights, but their patterns are not further exploited in the prediction process. On the contrary, both PSVR and SVR aim at finding the best non-linear empirical model relating the input variables (internal and external variables, and historical values of the target) to the future value of the target.

The comparison with AAKR has been carried out for three different training datasets, which correspond to the measurements intervals [6800th, 6950th], [6900th, 7050th] and [7000th, 7150th]. The bandwidth of the Gaussian kernel used in AAKR has been tuned for each dataset by a trial-and-error process, and the resulting best values are 2, 2 and 1, respectively. We show in Figure 9 the results obtained by AAKR on the second training dataset, [6900th, 7050th], where we trained AAKR on the same signals (internal and external variables) used as inputs in both PSVR and SVR models: the solid line in the Figure is the target and the dashed line is the prediction given by AAKR. Note that the data normalization strategy used in the AAKR procedure is different from the one used in PSVR and SVR: in the former case, data have been normalized to have zero mean and standard deviation equal to 1, while in the latter case they have been forced to lie in the interval [0,1].

It is evident from Figure 9 that AAKR method does not give satisfactory results. Actually these poor results should be expected, since AAKR is in general proficiently used for condition monitoring and fault detection, but it is not a proper method for prognostics: it is capable of effectively reconstructing the operational behavior of a signal, but since it computes only a weighted average of the signals in the training set, its generalization power is low in the case of out-of-control signals.

For what concerns the comparison of PSVR with SVR, the SVM-Toolbox of Matlab is used. The comparison is carried out for the same three training datasets considered for the comparison with AAKR. Using the same values for the hyperparameters selected for PSVR, the result of SVR on the training dataset [6900th, 7050th] is shown in Figure 10, where the solid line is the target and the dashed line is the predicted value.

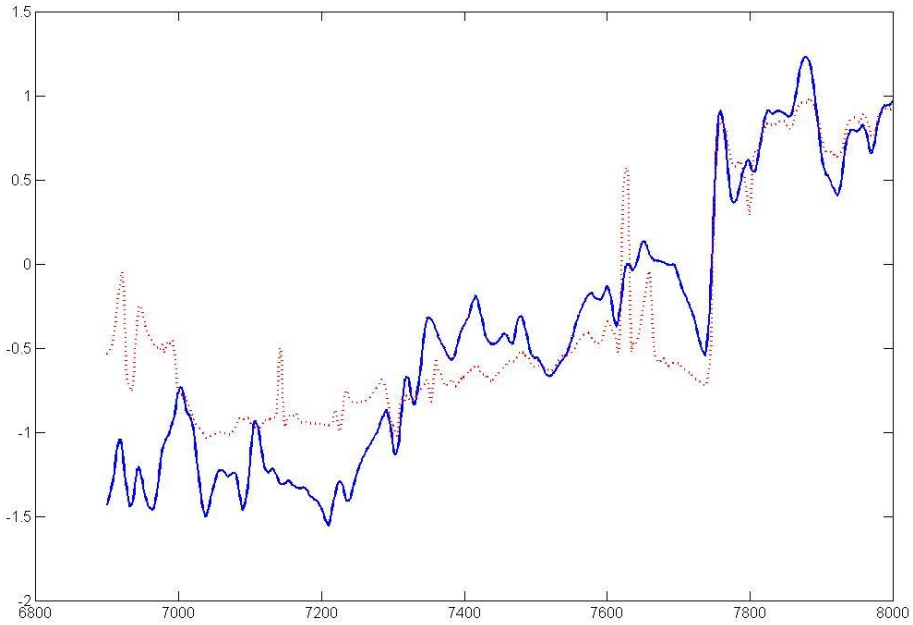

Fig.9 Point prediction for the target of interest (for both the training and testing data points) using AAKR (the bandwidth of the Gaussian kernel is set equal to 2).

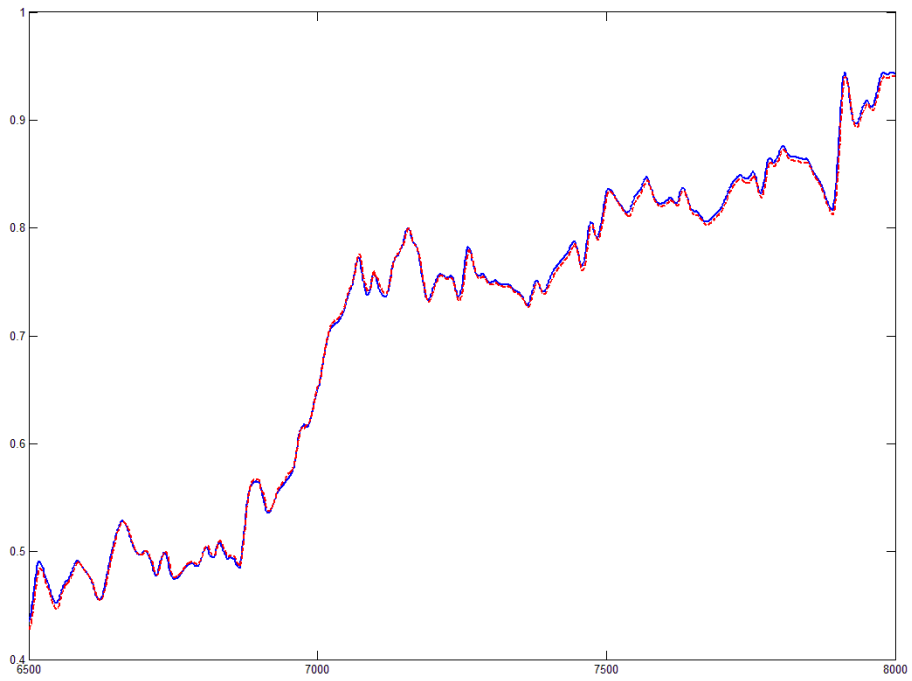

Fig.10 Point prediction for the target of interest (for both the training and testing data points) using standard SVR (with Matlab Toolbox).

Differently from the case of AAKR, the prediction obtained with SVR is quite accurate. Hence, to obtain a more precise comparison, in Table 4 the values of the Mean Square Errors obtained for the three training data sets with SVR and PSVR are reported. From inspection of the Table, it can be noticed that PSVR and standard SVR give comparable results, since the result of PSVR is slightly better for the first and third training data sets, but it is slightly worst for the second one. This is probably due to the empirical nature of the nonlinear regression methods we are using. On the other hand, standard SVR can only give a point estimate, while PSVR can also provide the uncertainty quantification, e.g. the PIs for the target and the predictive distribution.

Tab.4 Comparison of the results of PSVR and standard SVR

	SVR	PSVR
[6800th, 6950th]	$3.3353*10^{-4}$	$2.2267*10^{-5}$
[6900th, 7050th]	$1.4105*10^{-5}$	$5.3534*10^{-5}$
[7000th, 7150th]	$1.9787*10^{-4}$	$2.8867*10^{-5}$

5. **Conclusion**

In this paper, an approach is proposed for prediction of parameters of NPP components under fault conditions. It includes pre-processing for data reconstruction and model selection, and PSVR for estimation of the prediction interval and conditional predictive distribution of the target of interest. The results of the application to a real case study of leak flow in the first seal of a RCP are satisfactory. The coverage of the prediction interval is 91.50% with a confidence level of 95%. The conditional predictive distribution provides the probability distribution of the values of the target. These two indicators, the PI and the predictive distribution, are very informative for the NPP operators in case of accident.

The future work will focus on the development of a method to extend condition monitoring to prognostics, by computing the NPP components RUL on the basis of the prediction of its evolving parameters. This entails propagating the uncertainties in the prediction, due to both the observed data and the model itself.

References

H. Akaike, 1974. “A new look at the statistical model identification”. IEEE Trans. Auto. Control. 19(6), PP.716–723.

E. Alpaydin, 2004. “Introduction to Machine Learning”. The MIT Press. Cambridge, Massashusetss, London, England.

I.H. Bae, M.G. Na, Y.J. Lee, G.C. Park, 2008. “Calculation of the power peaking factor in a nuclear reactor using support vector regression models”. Ann. Nucl. Energy. 35, PP.2200-2205.

P. Baraldi, R. Canesi, E. Zio, R. Seraoui and R. Chevalier, 2010. “Signal grouping for condition monitoring of nuclear power plant components”. NPIC&HMIT 2010, Las Vegas, Nevada, USA, November.

P. Baraldi, R. Canesi, E. Zio, R. Seraoui and R. Chevalier, 2011.“Genetic algorithm-based wrapper approach for grouping condition monitoring signals of nuclear power plant components”. Integr. Comput. Eng. 18(3), PP. 221-234.

B.E. Boser, I.M. Guyon, V.N. Vapnik, 1992. “A training algorithm for optimal margin classifiers”. 5th Annual ACM Workshop on COLT, Pittsburgh, PA, ACM Press, pp. 144-152.

C.J.C. Burges, 1998. “A Tutorial on Support Vector Machines for Pattern Recognition”. Data Min. & Knowl. Discov. 2(2), PP.121-167.

J.J. Cai, 2012. “Applying support vector machine to predict the critical heat flux in concentric-tube open thermosiphon”. Ann. Nucl. Energy. 43, PP.114-122.

R. Chevalier, D. Provost and R. Seraoui, 2009. “Assessment of statistical and classification models for monitoring EDF's assets”. Sixth American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control, and Human-Machine Interface Technologies, Knoxville, Tennessee, USA.

C. Cortes, V.N. Vapnik, 1995. “Support-Vector Networks". Machine Learning, vol. 20.

N. Cristianini and G. S. Taylor, 2000. “An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods”, Cambridge University Press.

H. Drucker, C. J. C. B urges, L. Kaufman, A.J. Smola, V.N. Vapnik, 1997. “Support vector regression machines”. Advances in Neural Information Processing Systems (NIPS 1996; MIT Press), 9, PP. 155-161.

S. Ekici, 2012. “Support Vector Machines for classification and locating faults on transmission lines”. Appl. Soft Comp. 12, PP.1650-1658.

O. Elnokity, I.I. Mahmoud, M.K. Refai, H.M. Farahat, 2012. “ANN based Sensor Faults Detection, Isolation, and Reading Estimates –SFDIRE: Applied in a nuclear process”. Ann. Nucl. Energy. 49, PP.131-142.

J.B. Gao, S.R. Gunn and C.J. HarrisetM. Brown, 2002. “A Probabilistic Framework for SVM Regression and Error Bar Estimation”. Mach. Learn. 46(1-3), PP.71-89.

F. Girosi, 1998. “An equivalence between sparse approximation and support vector machines”. Neural Comp. 10(6), PP.1455–1480.

G. Horváth. “Neural Networks in System Identification”. In: V. Piuri (Ed.) Neural Networks in Measurement Systems NATO ASI NIMIA, Crema, Italy 2001. Oct. IOS Press, in print.

A.K.S. Jardine, D. Lin and D. Banjevic, 2006. “A review on machinery diagnostics and prognostics implementing condition-based maintenance”. Int. J. Adv. Manuf. 28(9-10), PP.1012-1024.

H. Jeong, W. Hwang, E. Kim and M. Han, 2003. “Hybrid modeling approach to improve the forecasting capability for the gaseous radionuclide in a nuclear site”. Ann. Nucl. Energy. 30, PP.1365-138.

D.S. Kim, J.H. Kim, M. Gyunna and J.W. Kim, 2012. “Uncertainty analysis of data-based models for estimating collapse moments of wall-thinned pipe bends and elbows”. Nucl. Eng. & Tech. 44(3), PP. 323-330.

F. Li, U.B.R. Perillo and S.R.P, 2012. “Fault Diagnosis of Helical Coil Steam Generator Systems of an Integral Pressurized Water Reactor Using Optimal Sensor Selection”. IEEE Trans. Nucl. Sci. 59(2), PP. 403-410.

Y. G. Li and P. Nilkitsaranont, 2009. “Gas turbine performance prognostic for condition-based maintenance”, Applied Energy, 86(10), PP. 2152–2161.

C.J. Lin and R.C. Weng, 2004. “Simple probabilistic predictions for support vector regression”. National Taiwan University, Taipei.

B. Lu, B.R. Upadhyaya, 2005. “Monitoring and fault diagnosis of the steam generator system of a nuclear power plant using data-driven modeling and residual space analysis”. Ann. Nucl. Energy. 32, PP.897-912.

J.P. Ma and J. Jiang, 2011. “Applications of fault detection and diagnosis methods in nuclear power plants: A review”. Prog. Nucl. Energy, 53, PP.255-266.

D.J. MacKay, 1991. “Bayesian Modelling and Neural Networks”. PhD thesis, California Institute of Technology, Pasadena, CA.

D.J. MacKay, 1997. “Gaussian processes, a replacement for neural networks”. NIPS tutorial 1997, Cambridge University.

E. Masry and J. Mielniczuk, 1999. “Local linear regression estimation for time series with long-range dependence”. Stoch. Process. & Appl. 82 (2), PP.173–193.

V. Muralidharan and V. Sugumaran, 2012. “A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis”. Appl. Soft Comp. 12, PP.2023-2029.

N. Murata, S. Yoshizawa and S. Amari, 1994. “Network information criterion–determining the number of hidden units for artificial neural network models”. IEEE Trans. Netw. 5, PP.865–872.

M. G. Na, J. W. Kim, and D. N. Moreton, 2006. “Estimation of Collapse Moment for the Wall-Thinned Pipe Bends Using Fuzzy Model Identification”. Nucl. Eng. & Des. 236, PP. 1335–1343.

R. Neal, 1996. “Bayesian Learning for Neural Networks”. Lecture Notes in Statistics. Springer, New York.

G. Niu and B. S. Yang, 2010. “Intelligent condition monitoring and prognostics system based on data-fusion strategy”, Expert Systems with Applications, 37(12), PP. 8831–8840.

T. Poggio and F. Girosi, 1998. “A sparse representation for function approximation”. Neural Comp. 10, PP.1445–1454.

C.P. du Rand and G. van Schoor, 2012a. “Fault diagnosis of generation IV nuclear HTGR components – Part I: The error enthalpy–entropy graph approach”. Ann. Nucl. Energy. 40, PP.14-24.

C.P. du Rand and G. van Schoor, 2012b. “Fault diagnosis of generation IV nuclear HTGR components – Part II: The error enthalpy–entropy graph approach”. Ann. Nucl. Energy. 41, PP.79-86.

J.L. Rodgers and W.A. Nicewander, 1988. “Thirteen ways to look at the correlation coefficient”. Am. Stat. 42 (1), PP. 59-66.

A.J. Smola and B. Schölkopf, 2004. “A tutorial on support vector regression”. Stat. & Comp. 14(3), PP.199-222.

P. Sollich, 1999. “Probabilistic interpretations and Bayesian methods for support vector machines”. Technical report, King’s College London, London, UK.

A.N. Tikhonov and V.Y. Arsenin, 1977. “Solution of Ill-posed Problems”. W.H. Winston, Washington, D.C..

K. Trontl, T. Smuc, D. Pevec, 2007. “Support vector regression model for the estimation of c-ray buildup factors for multi-layer shields”. Ann. Nucl. Energy. 34, PP.939-952 (2007).

T. V. Santosh, A.Srivastava, V.V.S.SanyasiRao, A.K.Ghosh and H.S.Kushwaha, 2009. “Diagnostic system for identification of accident scenarios in nuclear power plants using artificial neural networks”. Reliab. Eng. Syst. Saf. 94, PP. 759-762.

V.N. Vapnik, 1995. “The Nature of Statistical Learning Theory”. Springer. New York.

V. N. Vapnik, S. E. Golowich and A. Smola, 1996. “Support vector method for function approximation, regression estimation, and signal processing”, Proceedings of the 10th Neural Information Processing Systems (NIPS) Conference, Denver, Colorado.

V. Venkatasubramanian, 2005. “Prognostic and Diagnostic Monitoring of Complex Systems for Product Lifecycle Management: Challenges and opportunities”. Comput. Chem. Eng. 29(6), PP. 1253-1263.

W. Q. Wang, M. F. Golnaraghi and F. Ismail, 2004. “Prognosis of machine health condition using neuro-fuzzy systems”, Mechanical Systems and Signal Processing, 18(4), PP. 813–831.

C. Vladimir and F. Mulier, 1998. “Learning from Data: Concepts, Theory, and Methods”. John Wiley camp; Sons, Inc. New York, N.Y., USA.

C.K. Williams, 1997. “Computing with infinite networks”. In M.C. Mozer, M.I. Jordan, and T. Petsche, editors, Info. Process. Sys. volume 9, PP.295–301. MIT Press.

M. Yazikov., G. Gola , O. Berg, J. Porsmyr, H. Valseth, D. Roverso and M. Hoffmann, 2012. “On-Line Fault Recognition System for the Analogic Channels of VVER 1000/400 Nuclear Reactors”. IEEE Trans. Nucl. Sci. 59(2), PP. 411-418.

H. Y-C and DL Pepyne, 2001. “Simple Explanation of the No Free Lunch Theorem of Optimization”. In Proceedings of the 40th IEEE Conference on Decision and Control, 5 PP. 4409-4414. December 4-7, Orlando, Florida. IEEE, Piscataway, New Jersey.

E. Zio, 2012. “Diagnostics and Prognostics of Engineering Systems: Methods and Techniques”, Chapter 17. Engineering Science Reference. USA.

E. Zio, F. Di Maio and M. Stasi, 2010. “A data-driven approach for predicting failure scenarios in nuclear systems”, Ann. Nucl. Energy. 37(4), PP. 482-491.

E. Zio and F. Di Maio, 2010. “A data-driven fuzzy approach for predicting the remaining useful life in dynamic failure scenarios of a nuclear system”. Reliab. Eng. Syst. Saf. 95, PP. 45-57.

E. Zio, G. Gola, 2006. “Neuro-fuzzy pattern classification for fault diagnosis in nuclear components”. Ann. Nucl. Energy. 33, PP.415-426.

E. Zio, P. Baraldi, I. C. Popescu, 2009. “A fuzzy decision tree method for fault classification in the steam generator of a pressurized water reactor”. Ann. Nucl. Energy. 36, PP.1159-1169.

Appendix A

Let us assume that the input data is a n-dimensional set of vectors $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$, independently drawn in$\boldsymbol{R}^p$, and that we also have an independent sample from the target value $\boldsymbol{Y} = \{y_1, y_2, \dots, y_n\}$, where $y_i \in \boldsymbol{R}, i = 1, 2, \dots, n$.

In regression methods, the final aim is to find an underlying function describing the relation between the input data and the target. Here, this function will be indicated as an element of the generic space

$$F = \{a(\boldsymbol{x}): \boldsymbol{R}^p \to \boldsymbol{R}\}. \tag{1}$$

Moreover, we assume that the training set $\boldsymbol{\Gamma} = \{\boldsymbol{X}, \boldsymbol{Y}\}$ has been drawn from the probability distribution $P(\boldsymbol{x}, \boldsymbol{y})$: $\boldsymbol{R}^{p+1} \to \boldsymbol{R}$, which is not known. The Maximum A Posteriori (MAP) method consists in finding the $a(\boldsymbol{x})$ which minimizes the risk

$$R_{ERM}(\boldsymbol{x}) = \int l(a(\boldsymbol{x}) - y) dP(\boldsymbol{x}, y), \tag{2}$$

where $l(\boldsymbol{x}, y)$ is the loss function. There are many possible choices for the loss functions, e.g. square loss function, 1-norm loss function, Huber's loss function, etc. In this paper, we adopt one of the most common choices, the ε-insensitive loss function

$$l(x) = \begin{cases} 0 & if\ |\boldsymbol{x}| < \varepsilon \\ |x| - \varepsilon & if |x| \geq \varepsilon \end{cases}. \tag{3}$$

The ε-insensitive loss function has a good sparseness property, because all the data points whose margin between the predicted and target values is smaller than ε, are not used in the estimation process.

In SVM, the Empirical Risk Minimization (ERM) is used to solve the optimization problem in Equation (2), where $dP(\boldsymbol{x}, y)$ is replaced by $\frac{1}{n}$, recalling that n is the number of input data points. This means assuming that all the data points follow an identical and independent distribution (i.i.d), and using their empirical sample distribution. However, according to Tikhonov and Arsenin (1977), this is an ill-posed approach, whose generalization property is not good.

The Structural Risk Minimization (SRM) is formulated to solve the problem. A positive semi-definite operator $\| \hat{P} a \|^2$ is added to the ERM, and the so obtained new risk functional, called SRM, is given by

$$R_{SRM}(\boldsymbol{x}) = C \sum_{\boldsymbol{x_i}} l(a(\boldsymbol{x_i}) - y_i) + \frac{1}{2} \| \hat{P} a \|^2. \tag{4}$$

The operator $\widehat{\mathrm{P}}$ maps the space F into a dot-product space, and the kernel kernel $K = (\hat{P}^T \hat{P})^{-1}$ is derived after the Gaussian Process (GP) is introduced as a prior into the regression problem.

Indicating with $\boldsymbol{a}(\boldsymbol{X}) = (a(\boldsymbol{x}_1), a(\boldsymbol{x}_2), \dots, a(\boldsymbol{x}_n))^T$ the vector of function values, $P[\boldsymbol{a}(\boldsymbol{X})|\boldsymbol{\Gamma}]$ is the conditional probability of $\boldsymbol{a}(\boldsymbol{x})$ given the training set $\boldsymbol{\Gamma}$. $P[\boldsymbol{\Gamma}|\boldsymbol{a}(\boldsymbol{X})]$ is the likelihood of $\boldsymbol{X}$ having been originated by the corresponding target $\boldsymbol{Y}$ given the underlying function $a(\boldsymbol{x})$. $P[\boldsymbol{a}(\boldsymbol{X})]$ is the a priori probability of the underlying function $a(\boldsymbol{x})$. $P[\boldsymbol{\Gamma}]$ is the evidence.

Applying the Bayesian Rule, we can derive the relation

$$P[\boldsymbol{a}(\boldsymbol{X})|\boldsymbol{\Gamma}] = \frac{P[\boldsymbol{\Gamma}|\boldsymbol{a}(\boldsymbol{X})]P[\boldsymbol{a}(\boldsymbol{X})]}{P[\boldsymbol{\Gamma}]} \tag{5}$$

We make the following assumptions:

(1) Training data are i.i.d.

(2) The *a priori* probability distribution is $P[\boldsymbol{a}(\boldsymbol{X})] \propto \exp(-\frac{1}{2} \| \hat{P}a \|^2)$.

(3) The ε-insensitive loss function is chosen as the loss function.

(4) The covariance function is $K(\boldsymbol{x}, \boldsymbol{x}')$, and $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\frac{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2}{2\gamma^2})$, where $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ are the input data points in $\boldsymbol{X}$.

Following Equation (6), the a posteriori probability of $\boldsymbol{a}(\boldsymbol{X})$ can be written as

$$P[\boldsymbol{a}(\boldsymbol{X})|\boldsymbol{\Gamma}] = \frac{[G(C,\varepsilon)]^N}{\sqrt{det 2\pi K_{\boldsymbol{X},\boldsymbol{X}}} P[\boldsymbol{\Gamma}]} \exp\{-C \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} L_\varepsilon\big(y_i - a(\boldsymbol{x}_i)\big) - \frac{1}{2} \boldsymbol{a}(\boldsymbol{X})^T K_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{a}(X)\} \tag{6}$$

where $G(C, \varepsilon) = \frac{1}{2} \frac{C}{C\varepsilon + 1}$, and $K_{\boldsymbol{X},\boldsymbol{X}} = [K(\boldsymbol{x}_i, \boldsymbol{x}_j)]$ is the covariance matrix of the data points of $\boldsymbol{X}$.

We find the maximum of Equation (7) using the so-called MAP. This requires finding the minimum of the following function

$$R_{GSVM}(a) = C \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} L_\varepsilon\big(y_i - a(\boldsymbol{x}_i)\big) + \frac{1}{2} \boldsymbol{a}(\boldsymbol{X})^T K_{\boldsymbol{X},\boldsymbol{X}}^{-1} \boldsymbol{a}(X) \tag{7}$$

We can see that the risk of Gaussian SVMs is equivalent to the standard SVM. Following the discussion in Mackay (1997), Tikhonov and Arsenin (1997), Girosi (1998) and Burges (1998), we can write the solution of the minimization problem associated to Equation (8) in the following form

$$a^*(\boldsymbol{x}) = \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \beta_i K(\boldsymbol{x}_i, \boldsymbol{x}) \tag{8}$$

where $\beta_i = a_i - a_i^*$ is a combination of the Lagrange Multipliers associated to the optimization problem (Smola and Scholköpf, 2004). The a_i and a_i^* can be determined by a Quadratic Programming approach. According to Smola and Scholköpf (2004), $\forall\, i = 1, \ldots, n$, a_i and a_i^* lie in the interval $[0, C]$, and β_i consequently lies in the interval $[-C, C]$, which is the domain of the optimization problem.

There are different medium-scale and large-scale algorithms that can be used for optimizing under constraints the objective function in Equation (8). An active set algorithm which focuses on the solution of the Karush-Kuhn-Tucker (KKT) equations is used in this paper. The KKT equations are both necessary and sufficient conditions for obtaining a global solution point when the problem is a convex programming problem. Sequential Quadratic Programming (SQP) method is used to compute directly the Lagrange multipliers which balance the deviations in magnitude of the objective function and constraint gradients.