International Journal of Computational Intelligence Systems, Vol. 7, No. 2 (April 2014), 225-241

Ensemble of Kernel Regression Models for Assessing the Health State of Choke Valves in Offshore Oil Platforms

Piero Baraldi[a, *], Enrico Zio[b,a], Francesca Mangili[a]
[a]*Dipartimento di Energia, Politecnico di Milano, Italy*
[b]*Chair on Systems Science and the Energetic Challenge, European Foundation for New Energy-Electricité de France, Ecole Centrale Paris-Supelec,France*

Giulio Gola[c,d], Bent H. Nystad[c],
[c]*Institutt for energiteknikk, Halden, Norway*
[d]*IO Center for Integrated Operations, Trondheim, Norway*

Received 14 September 2011

Accepted 27 November 2013

Abstract

This paper considers the problem of erosion in choke valves used on offshore oil platforms. A parameter commonly used to assess the valve erosion state is the flow coefficient, which can be analytically calculated as a function of both measured and allocated parameters. Since the allocated parameter estimation is unreliable, the obtained evaluation of the valve erosion level becomes inaccurate and undermines the possibility of achieving good prognostic results. In this work, cluster analysis is used to verify the allocated parameter values and an ensemble of Kernel Regression models is used to correct the valve flow coefficient estimates.

Keywords: AHP, Choke Valve Erosion, Ensemble, Kernel Regression, Prognostics and Health Management.

* Dipartimento di Energia, Politecnico di Milano, via Ponzio 34/3, 20133 Milano, Italy. E-mail: piero.baraldi@polimi.it.

1 Introduction

Predicting the evolution of equipment degradation allows efficient planning of maintenance operations.[1-3] In general, a prognostic model can be developed based on information directly or indirectly related to equipment degradation.[4] In practice, however, field data are affected by noise, sensor faults and extrapolation errors and need to be verified and possibly corrected before they are used for developing the prognostic model. Thus, the necessity of pre-treating degradation-related data arises in real industrial applications. Reducing the uncertainty on the data used by prognostic models can lead to the reduction of the uncertainty on the model output, i.e. the remaining useful life, and, thus improve maintenance and operation scheduling. Providing a thorough analysis and general solutions for prognostic data pre-treatment is a very difficult task, since the solutions typically depend strongly on the specific application. In this work, some data pre-treatment methods have been developed and are presented with reference to a case study related to the erosion of choke valves located topside at wells on the Norwegian Continental Shelf.[5,6] The difference between the actual valve flow coefficient and its theoretical value is retained as the indicator of the choke valve health state and is used to assess the degree of erosion affecting the choke. While the theoretical value of the valve flow coefficient depends only on the choke opening, the actual valve flow coefficient is analytically calculated on a daily basis as a function of the pressure drop through the choke which is directly measured and oil, gas and water flow rates which are allocated based on the measured total production from a number of wells and on physical parameters (pressures and temperatures) related to the single well. Such flow rates are actually measured only during a number of well tests carried out throughout the valve life. In practice, the resulting indicator of the choke valve state is very noisy and lacks the physical monotonicity of the erosion process; the allocated values of oil, gas and water flow rates are conjectured to be the cause of the large inaccuracies and uncertainties in the calculation of the actual valve flow coefficient.

To verify this, data are processed by the Fuzzy C Means (FCM) clustering algorithm.[7,8] FCM is applied to the projections of the five-dimensional dataset into the subspace of the two measured parameters (pressure drop and choke opening) and the subspace of the three allocated parameters (oil, water and gas flow rates). The two partitions are compared to investigate the coherence of the information conveyed by the parameters. A supervised clustering algorithm based on Mahalanobis metrics[9] is used to obtain a partition of the entire five-dimensional dataset as close as possible to that obtained based only on the two measured parameters. A measure of the importance of the parameters in the clustering is calculated and used to verify the coherence of the information conveyed by the less reliable allocated parameters with that conveyed by the two reliable ones.

If found unreliable, the values of oil, gas and water flow rates are corrected based on the relations among all parameters. To this aim, an ensemble of Kernel Regression (KR) models is here devised. KR is a distance-based regression algorithm[10,11]; an ensemble of four KR models is used to avoid the need of selecting the optimal model and to increase the robustness and reduce the uncertainty of the estimate.[12,13] Diversity is injected in the ensemble by differentiating the training procedure for each KR model. The aggregation of the KR model outcomes is obtained through an original procedure based on the weighted average of the single model outcomes with weights calculated using the Analytic Hierarchy Process (AHP).[14]

Since a validation dataset is not available for the choke valve case study, the ensemble-based reconstruction approach is verified on an artificial dataset which does not attempt to reproduce the physical behaviors of the choke valve system and only shares some of the main characteristics of the choke valve dataset. The artificial dataset contains five-dimensional patterns randomly sampled from as many multivariate Gaussian distributions as the number of clusters found in the real dataset; a white Gaussian noise is added to three

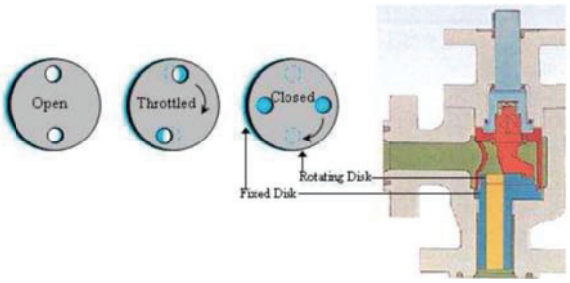

Fig. 1. Typical choke valve of rotating disk type (http://www.vonkchokes.nl/).

parameters in order to simulate the uncertainty in their values, in analogy to what is observed in the three allocated parameters of the choke valve case study.

The main contributions of this work to the field of prognostic concern the pre-treatment of noisy and unreliable data. In this context we have developed an original procedure which allows evaluating the quality of the prognostic data available and, eventually, improving it. In particular we have: (i) proposed a monotonicity-based index for the evaluation of the quality of a degradation indicator; (ii) developed a clustering-based procedure for establishing whether allocated parameter estimates are reliable; (iii) developed a method for improving the estimates of unreliable parameters based on an original strategy for the aggregation of multiple model outcomes.

The application of these methods to the problem of choke valve erosion assessment can potentially improve the accuracy in the estimation of the choke valve flow coefficient, which is extensively used in the oil & gas industry for wells condition monitoring. Furthermore, the methods can be applied in many other situations where some unreliable parameter estimates are used.

The paper is framed as follows. The traditional procedure for the construction of a health indicator assessing the choke valve erosion state is presented in Section 2. Section 3 illustrates the clustering procedures introduced to verify the reliability of the allocated parameters; based on the results of the cluster analysis, an artificial dataset is built for validating the effectiveness of the proposed clustering method (Section 4); to improve the accuracy of the allocated flow rates, a KR ensemble is developed, verified on the artificial case study and then applied to the real case study; finally, the estimated flow rates are used to calculate the health indicator (Section 5). Conclusions and potential perspectives for future work are drawn in the last Section.

2 Choke Valve Erosion Assessment

In oil and gas industries, choke valves are normally located on top of each well and are used to balance the pressure on several wells into a common manifold to control flow rates and protect the equipment from unusual pressure fluctuations.

In Fig. 1, a choke valve is sketched. The throttle mechanism consists of two circular disks, each with a pair of circular openings to create variable flow areas. One of the disks is fixed in the valve body, whereas the other is rotated either by manual operation or by actuator, to vary or close the opening. For large pressure drops, the well streams which contains gas, liquid and sand particles can reach 400-500 m/s and produce heavy metal loss mainly due to solids, liquid droplets, cavitation and combined mechanisms of erosion-corrosion, resulting in choke lifetimes of less than a year. Erosion management is vital to avoid failures that may result in loss of containment, production being held back, and increased maintenance costs. Moreover, several chokes are located subsea, where the replacement cost is high. Then, the need has increased for reliable models to estimate erosion and lifetime of choke valves, in order to allow implementing effective maintenance strategies.[15-17]

2.1 Choke valve health state indicator

A common indicator of the valve flow capacity is the flow coefficient C_V, which is related to the effective flow cross-section of the valve. Given a differential pressure ΔP, the flow rate q across the valve is proportional to the flow coefficient C_V [18]:

$$q = C_V \sqrt{\frac{\Delta P}{\rho / \rho_w}} \tag{1}$$

where ρ/ρ_w is the relative density of the substance across the valve, i.e. the ratio of the substance density to the water density. Tests are performed by manufacturers on new valves to evaluate the theoretical valve flow coefficient $C_V^{th}(\theta)$ for different values of the valve opening θ. In practice, $C_V^{th}(\theta) \propto \theta^{\alpha}$, where α is close to 1 and depends on the type of choke considered.

Erosion is a slow process. For a specific valve opening, erosion produces a gradual increase of the valve area available for the flow transit. Given θ and ΔP, erosion determines an increase in q modeled by a corresponding increase in C_V (eq. 1). For this reason, the difference δC_V between the actual (C_V) and the nominal (C_V^{th}) values of the valve flow coefficient is retained as the health indicator for the choke[6]:

$$\delta C_V(\theta) = C_V(\theta) - C_V^{th}(\theta) \tag{2}$$

During operation, C_V is not directly measured but computed for a two-phase flow as[18]:

$$C_V = \frac{\dot{m}}{27.3 \cdot F_p \sqrt{\Delta P \left(\frac{f_g}{\rho_g \cdot J^2} + \frac{f_w}{\rho_w} + \frac{f_o}{\rho_o} \right)}} \quad (3)$$

where $\dot{m} = \dot{m}_o + \dot{m}_w + \dot{m}_g$ is the total mass flow rate of the oil-water-gas mixture, $f_{o,w,g} = \dot{m}_{o,w,g} / \dot{m}$ is the fraction of the oil, water and gas mass flow rates, respectively, $\rho_{o,w,g}$ are the corresponding densities, J is the gas expansion factor, $F_p(\theta)$ is the piping geometry factor accounting for the geometry of the valve/pipe reducer assembly and ΔP is the pressure drop through the choke. Eq. (3) and the values of $\rho_{o,w,g}$, J, $F_p(\theta)$ and N_6 are derived from fluid dynamics; parameters ΔP, θ, $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$ are measured or allocated during operation.

2.2 Choke valve dataset

For a correct assessment of the choke erosion state and the prediction of its remaining useful life, it is fundamental to obtain frequent and reliable measurements or estimates of the parameters ΔP, θ, $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$ used to compute the health indicator δC_V. Nevertheless, only the pressure drop ΔP and the valve opening θ are measured during standard daily inspections (SI), whereas measures of water, oil and gas flows rates are taken downstream of the choke only during well tests (WT) with a multiphase flow separator. On a daily basis, the values of $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$ are allocated for a single well by a software based on the measured total production from a number of wells and on physical parameters (pressures and temperatures) related to the specific well. The available information consists in 259 ΔP and θ measurements performed every operational day, in 7 $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$ measurements performed at times t=0.4, 18.4, 61.5, 135.8, 180.3, 250.6, 276,7 during well tests and in the 259 daily allocated values of these latter three parameters (Table 1). Fig. 2 shows the parameters trends during standard inspections (continuous line) and well tests (stars). Fig. 3 shows the values of the health indicator δC_V computed using daily standard inspections data (continuous line) and well test measurements (stars).

Table 1. Available information

	Number of patterns	**ΔP and θ**	**$\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$**
Standard Inspections (SI)	N_{SI}=259	Measured	Allocated
Well Test Inspections (WT)	N_{WT}=7	Measured	Measured

In general, δC_V is expected to be monotonic since erosion cannot decrease in time unless maintenance actions are performed. A quantitative index of monotonicity is the Spearman's rank correlation used in statistics to assess how well the relationship between two variables can be described using a monotonic function.[19] The curve of δC_V computed using the SI data, is highly noisy and presents remarkable oscillations. The Spearman's rank correlation coefficient r_S between δC_V and time t_k at which the measurements are taken is computed as:

$$r_S = 1 - \frac{6 \sum_{k=1}^{N} \left(R_{\delta C_V}(\mathbf{x}_k) - k \right)^2}{N(N^2 - 1)} \quad (4)$$

where $\mathbf{x}_k$ is the five-dimensional vector containing the parameter values collected at time t_k, and $R_{\delta C_V}(\mathbf{x}_k)$ and k are the ranks (i.e., the relative positions) of pattern $\mathbf{x}_k$ when all patterns are ordered with respect to the values of δC_V and t_k, respectively. Values of r_S close to 1 are expected for a monotonic quantity.

Results show that δC_V behaves monotonically (r_S=0.9643) only when WT measurements are used to compute it. On the contrary, the lower monotonicity (r_S=0.7401) obtained when δC_V is calculated using SI data suggests that some of the allocated mass flow rates may be unreliable. A cluster analysis is performed in the next Section in order to verify this hypothesis.

3 Clustering

Let $\mathbf{X}$ be a generic set of N patterns $\mathbf{x}_k = [\mathbf{x}_k^r \quad \mathbf{x}_k^u]$, k=1,…,N, of P parameters which can be divided in a vector $\mathbf{x}_k^r$ of p^r reliable parameters $x_p^r(t_k)$, p=1,…, p^r and another vector $\mathbf{x}_k^u$ of p^u unreliable parameters $x_p^u(t_k)$, p=p^r+1,…,P.

In general, the distinction between reliable and unreliable parameters can be achieved considering expert judgment, data analysis or by resorting to data validations techniques which allow detecting anomalous behaviors in datasets. In the choke valve case study, the

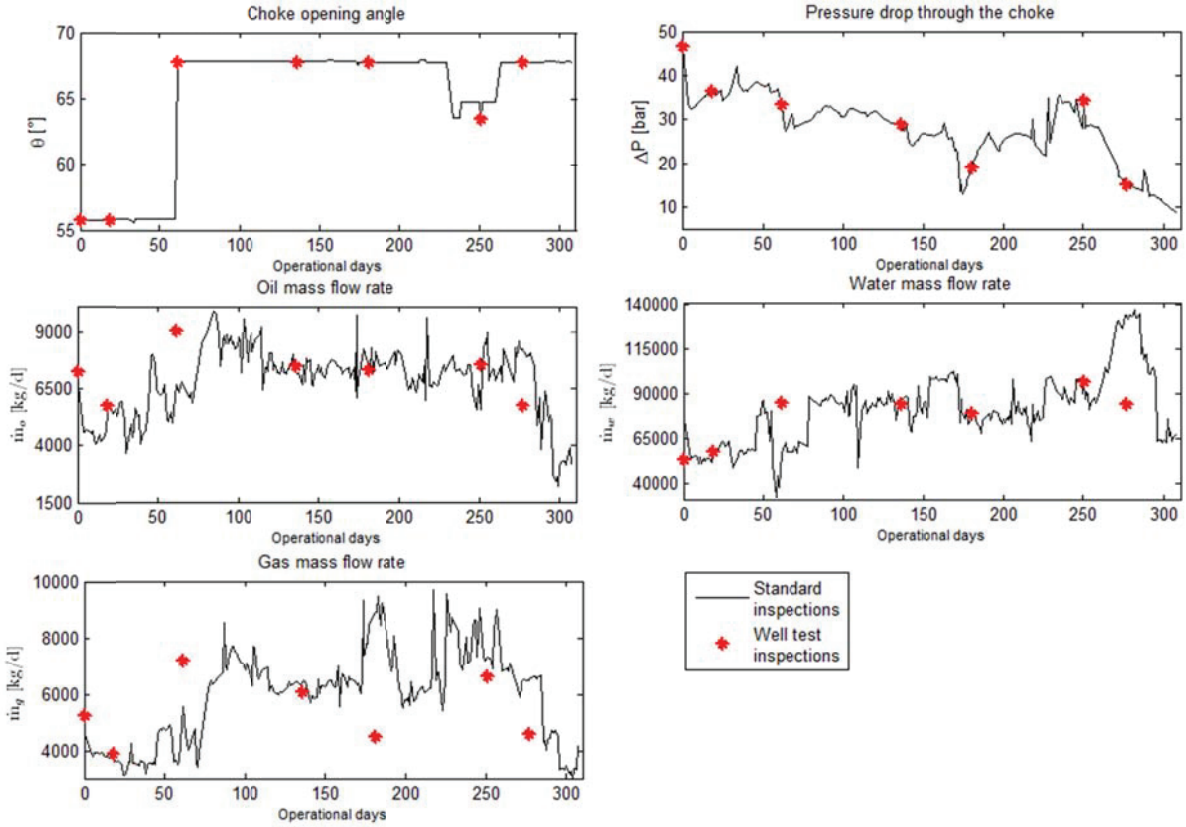

Fig. 2. Parameters trends (continuous line represents SI, stars indicate WT).

measured parameters ΔP and θ are classified as reliable according to expert judgment, whereas the allocated ones, $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$ are judged unreliable.

The aim is here to propose a procedure for verifying whether the information provided by the unreliable parameters in $\mathbf{x}_k^u$ is coherent with that of the reliable parameters in $\mathbf{x}_k^r$. This is done by considering the relative positions of the patterns in the p^r-dimensional subspace S^r of the reliable parameters, and in the p^u-dimensional subspace S^u of the unreliable parameters.

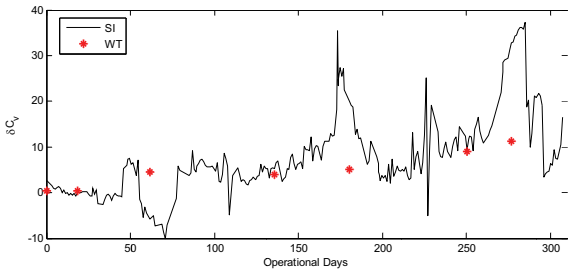

Fig. 3. Health indicator δC_V using SI (continuous line) and WT (stars).

An effective technique to find a structure in a collection of unlabeled objects is unsupervised clustering, consisting in the organization (partition) of the patterns into non-overlapping, non-empty groups (clusters) so that patterns of the same cluster are *similar* between them and *dissimilar* to the patterns belonging to other clusters.[20] The problem of clustering has been addressed in many contexts and disciplines. Several algorithms, such as the hard c-mean[32] and the evolving clustering methods,[31] have been proposed to identify clusters of objects. A limitation of these approaches is that they constrain each pattern to belong to one cluster only, when, in practice, the clusters may not be completely disjoint and patterns could be classified as belonging to one cluster almost as well as to another. Alternatively, fuzzy clustering algorithms which assign to each object a set of membership values, one for each cluster, have been proposed. The implication of this is that the class boundaries are not 'hard' but rather 'fuzzy'. The clustering technique employed in this work is the Fuzzy

C-Means (FCM)[21] which is based on the minimization of a weighed sum Y of the distances $d(x_k, v_c)$ between the patterns x_k and the cluster centers v_c.

$$Y = \sum_{c=1}^{C}\sum_{k=1}^{N}[\mu_c(\mathbf{x}_k)]^{\omega} d^2(\mathbf{x}_k, \mathbf{v}_c) \qquad (5)$$

where the weight $\mu_c(x_k)$ denote the membership of x_k to cluster c, and ω is a parameter which controls the degree of fuzziness of the clusters (often a value between 1 and 2 is found suitable in applications[9]). In the traditional algorithm[7], the distance is Euclidean. The membership values $\mu_c(x_k)$ and the cluster centers v_c are computed via an iterative procedure reported, for completeness, in Appendix A.

In this work, for the validation of the unreliable parameters, two different partitions (Γ^r and Γ^u) of the dataset X into C clusters are considered: Γ^r is obtained using the unsupervised Fuzzy C-Means (FCM) clustering technique in the parameters space S^r, whereas Γ^u is obtained by applying the same technique in the parameters space S^u.

In Section 3.1, the main steps of the procedure of cluster analysis proposed are presented; in Section 3.2, the results of its application to the choke valve erosion case study are discussed.

3.1 Cluster analysis

The information used to build the partition Γ^r is incomplete, since only p^r out of P parameters are used; on the other hand, the cluster structure thereby identified is assumed as reference in the comparison with the partition Γ^u, since it is built using only the p^r reliable parameters in $\mathbf{x}_k^r$.

Notice that, due to the incompleteness of the Γ^r information base, one could observe disagreement between Γ^r and Γ^u not only when the values of the unreliable parameters in $\mathbf{x}_k^u$ used to build Γ^u are incorrect, but also when they give information which, despite being correct, is uncorrelated with that given by the reliable parameters in $\mathbf{x}_k^r$. For example, two different clusters can coincide when projected on S^r and be well separated on S^u instead; in such a situation, one can obtain partitions Γ^r and Γ^u which are significantly different even if the uncertain parameters estimates are accurate. Since in the choke valve case study the unreliable parameters $\mathbf{x}_k^u = [\dot{m}_o \quad \dot{m}_w \quad \dot{m}_g]$ are somehow correlated to the reliable parameters $\mathbf{x}_k^r = [\Delta P \quad \theta]$ (see eq. (3)), situations where uncorrelated signals have to be handled are not considered in this work.

Operatively, the cluster analysis is performed as follows:

(i) The optimal number of clusters C to be used for the partitions Γ^r and Γ^u is identified. This is obtained by considering the minimum of the compactness and separation validity function $s(C)$:

$$s(C) = \frac{\frac{1}{N}\sum_{c=1}^{C}\sum_{k=1}^{N}\mu_c^{\omega}(\mathbf{x}_k)d(\mathbf{x}_k, \mathbf{v}_c)}{\min_{i,j} d(\mathbf{v}_i, \mathbf{v}_j)} \qquad (6)$$

which represents the ratio between the cluster compactness, measured by the average distance of the patterns from their cluster centers and the separation between the clusters, measured by the minimum distance between two cluster centers. Notice that the numerator tends to decrease when the compactness increases and the denominator tends to increase when the separation increases. Thus, in order to obtain a partition characterized by highly compact and well separated clusters, one has to find the optimal number of clusters which minimizes the validity function $s(C)$.

(ii) The fuzzy partitions Γ^r and Γ^u of the N data into C clusters are obtained using the FCM clustering algorithm (see Appendix A).

(iii) The clusters of Γ^r and Γ^u are bi-univocally associated $c_r \leftrightarrow c_u$ by minimizing the partition distance $D(\Gamma^r, \Gamma^u)$ between the partitions Γ^r and Γ^u. In this respect, the distance $D(\Gamma^r, \Gamma^u)$ defined in Ref. 22 has been used:

$$D(\Gamma^r, \Gamma^u) = \sum_{c=1}^{C}\sum_{k=1}^{N_{SI}} \frac{\left|\mu_c^r(\mathbf{x}_k) - \mu_c^u(\mathbf{x}_k)\right|}{2N} \qquad (7)$$

where $0 \le \mu_c^{r,u}(\mathbf{x}_k) \le 1$ is the membership of the k-th pattern to the c-th cluster of the partition Γ^r and Γ^u.

(iv) Crisp partitions Ω^r and Ω^u are obtained from the fuzzy partitions Γ^r and, respectively, Γ^u, by assigning a pattern $\mathbf{x}_k$ to a given cluster c if its degree of membership to the cluster, $\mu_c(\mathbf{x}_k)$, exceeds a predefined threshold $\gamma \in (0,1)$, which represents the required degree of confidence for the assignment. If the condition $\mu_c(\mathbf{x}_k) > \gamma$ is not fulfilled for any cluster or if it is verified for more than one cluster, the pattern is not associated to any cluster. The crisp partitions Ω^r and Ω^u are compared by considering the difference between the sets of patterns $\mathbf{X}_{c_r}$ and $\mathbf{X}_{c_u}$ assigned to the associated

clusters c_r and c_u. A large difference in the assignment of the patterns to the clusters is taken as a symptom that the information conveyed by the unreliable parameters may be misleading.

3.1.1 Results

According to this procedure, the dataset $\mathbf{X}^{SI}$ of the N_{SI} =259 SI available patterns $\mathbf{x}_k$, k=1,…, N_{SI} is projected into the subspaces $S^r=\Delta P\times\theta$ and $S^u=\dot{m}_o\times\dot{m}_w\times\dot{m}_g$ of the measured (reliable) and allocated (unreliable) parameters of the choke valve case study, respectively. Two partitions Γ^r and Γ^u of the dataset $\mathbf{X}^{SI}$ into C=5 clusters are obtained using the FCM algorithm with degree of fuzziness ω=2.

The clusters of Γ^r and Γ^u are then coupled by minimizing the partition distance $D(\Gamma^r, \Gamma^u)$ in eq. (7) and the same cluster index c=1,…,5 is assigned to each member of the pair of associated clusters. The minimal value found for the partition distance is 0.47 which is high considering that, by definition, the maximum partition distance is 1. With a degree of confidence γ=0.4, 255 patterns out of the total 259 patterns of $\mathbf{X}^{SI}$ are assigned without ambiguity to the clusters of Γ^r and 219 to the clusters of Γ^u. The remaining patterns are ambiguous. Ambiguous patterns in Γ^r, which differ from those in Γ^u, are located at the boundaries between clusters 1 and 3 and clusters 2 and 3, and for this reason they are assigned to both clusters.

Fig. 4 shows the partitions Γ^r and Γ^u of the 259 SI patterns in the space S^r. It can be seen that in Γ^r, the clusters are clearly separated, contrarily to what happens in Γ^u. Moreover, one can observe large differences in clusters' composition, e.g. many patterns that belong to cluster 1 in Γ^r are assigned to cluster 5 in Γ^u; patterns of clusters 2, 3 and 4, which are well separated in Γ^r, are, instead, mixed in Γ^u.

Table 2 compares the number of patterns assigned to the same cluster in Γ^r and Γ^u (4th column) to the total number of patterns assigned separately to each cluster of Γ^r and Γ^u (2nd and 3rd column, respectively). Notice that, globally, less than half of the patterns (47%) assigned to a cluster of Γ^r are assigned to the associated cluster of Γ^u (last row in the Table).

Table 2. Number of patterns assigned to each cluster in Γ^r (2nd column), in Γ^u (3rd column), in both Γ^r and Γ^u (4th column) and percentage of patterns assigned to the same cluster in both partitions with respect to the number of patterns assigned to that cluster in Γ^r

Cluster c	Γ^r	Γ^u	Γ^r & Γ^u	(Γ^r & Γ^u)/(Γ^r)
1	45	15	14	31.11%
2	56	49	15	26.79%
3	77	48	32	41.56%
4	25	47	15	60.00%
5	52	60	43	82.69%
Σ	255	219	119	46.67%

3.2 *Supervised evolutionary clustering*

To confirm the conclusions drawn in the previous Section, a further analysis based on a supervised clustering technique is here performed. Firstly, a partition Γ^s, as similar as possible to Γ^r, is obtained using a supervised evolutionary clustering technique based on Mahalanobis metrics in the space of all parameters.

A set $\mathbf{X}^{lab}$ of N_{lab} labeled training data is built by choosing, among the N patterns of $\mathbf{X}$, those belonging to one of the C clusters in Γ^r with a membership

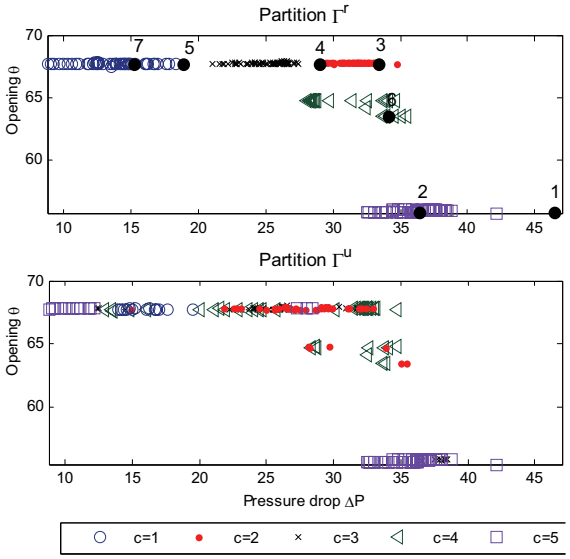

Fig. 4. Visualization on the space $S^r=\Delta P\times\theta$ of the patterns assigned to the five clusters in Γ^r (top) and Γ^u (bottom). In the top graph, the WT patterns are also shown (black dots, numbered in chronological order).

$\mu_{c_r}(\mathbf{x}_k) > 0.9$ and labeling them with the index c of the cluster they are assigned to. The evolutionary algorithm searches for the optimal metrics to be used by the FCM in order to achieve clusters as close as possible to the clusters of the labeled patterns.

In this view, each cluster c is defined by an individual distance through a dedicated Mahalanobis metric, defined by a definite positive matrix $\mathbf{M}_c$:

$$d^2_{\mathbf{M}_c}(\mathbf{x}_k, \mathbf{v}_c) = (\mathbf{x}_k - \mathbf{v}_c)^{\mathrm{T}} \mathbf{M}_c (\mathbf{x}_k - \mathbf{v}_c) \qquad (8)$$

The classification task amounts to an optimization problem in which the metrics, i.e., the geometric distance functions, become additional parameters to be determined besides the fuzzy partition. The supervised target of the optimization is that of minimizing the partition distance D(Γ,Γ*) between the a priori known partition Γ and the obtained partition Γ^* as defined in eq. (7).

For the optimization, we integrate an evolutionary algorithm for determining the C optimal geometric distance functions[21] with the FCM algorithm for determining the optimal fuzzy partition based on such distance. For more details on the algorithm one can refer to Ref. 9.

A measure of importance $I_{\mathbf{M}_c}(x_p)$ of a parameter x_p, p=1,…,P for the assignment of a pattern to a cluster c is:

$$I_{\mathbf{M}_c}(x_p) = \sum_{j=1}^{5} g^2_{c_{j,p}} \qquad (9)$$

where $g_{c_{j,p}}$, j,p=1,…,P are the coefficients of the lower triangular matrix $\mathbf{G}_c = [g_{c_{j,p}}]$ for cluster c obtained from the decomposition of the Mahalanobis matrix $\mathbf{M}_c$ into its Cholesky factors $\mathbf{G}_c$, i.e., $\mathbf{M}_c = \mathbf{G}_c{}^{\mathrm{T}} \mathbf{G}_c$.[9]

3.2.1 Results

The importance values $I_{\mathbf{M}_c}(\dot{m}_o)$, $I_{\mathbf{M}_c}(\dot{m}_w)$ and $I_{\mathbf{M}_c}(\dot{m}_g)$ associated to the allocated parameters are compared to those associated to the measured ones ($I_{\mathbf{M}_c}(\Delta P)$, $I_{\mathbf{M}_c}(\theta)$): if the importance of allocated and measured parameters is similar, one can conclude that they both convey useful information for defining the partition Γ^s; vice versa, if the importance of the allocated parameters is lower than that of the measured parameters, one should doubt about their reliability, since the information they convey appears to be incoherent with that of the measured parameters.

Table 3. Measures of importance $I_{\mathbf{M}_c}$ of the different parameters

	Measured parameters		**Allocated parameters**		
Cluster c	ΔP	θ	$\dot{m}_o$	$\dot{m}_w$	$\dot{m}_g$
1	2.221	1.770	0.095	0.048	0.105
2	2.410	5.933	0.000	0.001	0.002
3	2.175	4.443	0.050	0.009	0.011
4	0.362	7.847	0.013	0.696	0.008
5	0.288	3.802	0.044	0.097	0.199

Table 3 reports the measures of importance $I_{\mathbf{M}_c}$ obtained for the five parameters for each cluster. The allocated parameters have low importance compared to the measured ones, meaning that they do not significantly contribute to the assignment of the patterns to any of the clusters.

The analysis performed in this Section has shown that the information conveyed by the allocated parameters, $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$, i.e., the oil, water and gas mass flow rates, respectively, are unreliable and thus contribute to lower the quality of the choke valve health indicator δC_V. For this reason, a method for providing more accurate estimates of the mass flow rates has been developed. To test the performance of this method, an artificial dataset reproducing some of the main features of the choke valve dataset is built.

4 Artificial dataset

An artificial dataset X^A of N_A=250 five-dimensional patterns has been generated by sampling the values of the first three parameters, x_1^A, x_2^A and x_3^A, from C=5 multivariate Gaussian distributions representing the five clusters of the choke valve dataset (Fig. 4). Table 4 reports the mean and standard deviation values employed for sampling the patterns. The values of the remaining two parameters, x_4^A and x_5^A, are obtained by using the following deterministic functions of x_1^A, x_2^A and x_3^A.

$$x_4^A = x_2^A + \left| \frac{\left(x_1^A\right)^{2.5}}{\sqrt[3]{x_1^A}} \right|; \; x_5^A = \left| x_1^A \right| \sqrt[4]{x_2^A} - x_3^A \qquad (10)$$

In analogy with the choke valve case study, the parameters are divided into a vector $\mathbf{x}^r = [x_1^A, x_2^A]$ of two reliable parameters and another vector $\mathbf{x}^u = [x_3^A, x_4^A, x_5^A]$ of three unreliable parameters. In

order to realistically reproduce the uncertainties affecting the mass flow rates in the choke valve case study, a second dataset $X^{A,noise}$ has been built by adding to the unreliable parameters x_3^A, x_4^A and x_5^A of the patterns of X^A a white Gaussian noise with probability 0.5. To this purpose, the intensity of the noise affecting the allocated parameters of the choke valve case study has been roughly guessed by considering the root of the mean square difference (RMSD) between the seven WT mass flow rate measurements and the corresponding SI values. Table 4 shows that the values obtained for the noise are in a range between 0.8 and 1.25 times the standard deviations of the parameters computed using the SI data. The dataset $X^{A,noise}$ has been built by considering Gaussian noises with standard deviations equal to the parameter standard deviations.

Table 4. Mean and standard deviation of x_1^A, x_2^A and x_3^A

	Mean			**Standard deviation**		
Cluster *c*	x_1^A	x_2^A	x_3^A	x_1^A	x_2^A	x_3^A
1	**10.5**	**38**	**10**	**0.5**	**1.5**	**0.2**
2	**11.5**	**-7**	**-3**	**0.3**	**0.3**	**0.5**
3	**9**	**51**	**7**	**0.5**	**1.1**	**0.5**
4	**10**	**10**	**-5**	**0.2**	**1**	**0.4**
5	**10.3**	**22**	**0**	**0.07**	**2.5**	**0.2**

Nevertheless, since the intensity of the noise applied to x_3^A, x_4^A and x_5^A is large, when it is added to all patterns, the FCM algorithm is not able to find well separated clusters; on the contrary, in the choke valve case study the FCM algorithm is able to find separated clusters, despite the presence of noise on the mass flow rates $\dot{m}_o$, $\dot{m}_w$, and $\dot{m}_g$. For this reason, a smaller global amount of noise is inserted in the artificial case study by sampling the points to disturb with probability 0.5.

Table 5. Estimate of the standard deviations of the mass flow rate noises

	$\dot{m}_o$	$\dot{m}_w$	$\dot{m}_g$
RMSD	1.495	21.677	2582.769
σ (based on SI data)	1.793	19.064	2104.903
RMSD/σ	0.834	1.137	1.227

The cluster analysis procedure described in Section 3.1 has been applied to parameters x_1^A and x_2^A of the artificial case study, which are not affected by noise. In the partition Γ^r thus obtained, all the 250 patterns have been assigned to a cluster with a degree of membership higher than 0.4. Repeating the same cluster analysis on parameters x_3^A, x_4^A and x_5^A in case of both undisturbed and noisy data, we have obtained a partition Γ^u for the undisturbed dataset $\mathbf{X}^A$ characterized by 9 ambiguous patterns, i.e. patterns not assigned to any cluster with a degree of membership higher than 0.4, and a partition $\Gamma^{u,noise}$ for the disturbed dataset $\mathbf{X}^{A,noise}$ with 44 ambiguous patterns, thus demonstrating that, in case of noise, the identification of clearly separated clusters is more difficult.

Table 5 reports the number of patterns assigned to the same cluster in the partition Γ^r obtained by considering x_1^A and x_2^A and in the partitions Γ^u and $\Gamma^{u,noise}$ based on x_3^A, x_4^A and x_5^A, in both cases of undisturbed and noisy parameters, respectively. Notice that, in absence of noise, the two partitions almost coincide, whereas they are quite dissimilar in case of noise. These results confirm that, in absence of noise one should expect similarity of the partitions Γ^r and Γ^u. On the contrary, in case of noise on the allocated parameters, fewer patterns can be assigned to one cluster without ambiguity and many are assigned to different clusters.

Table 6. Number of patterns assigned to the same cluster in Γ^r and Γ^u in case of undisturbed and noisy data. Undisturbed data: number of patterns assigned to each cluster in Γ^r (column a), Γ^u (column b) and in both Γ^r and Γ^u (column c). Noisy data: number of patterns assigned to each clusters in Γ^u (column b) and in both Γ^r and Γ^u (column c)

Cluster *c*	Γ^r (a)	Γ^u (b)	Γ^r & Γ^u (c)	(c)/(a)
Undisturbed data $\mathbf{X}^A$				
1	48	50	48	1
2	52	50	50	0.96
3	50	50	50	1
4	43	50	43	1
5	48	50	48	1
Σ	241	250	239	0.99
Noisy data $\mathbf{X}^{A,noise}$				
1	48	36	25	0.54
2	52	42	40	0.63
3	50	49	49	0.86
4	43	43	25	0.72
5	48	36	26	0.63
Σ	241	206	163	0.68

Finally, seven patterns of the artificial dataset $\mathbf{X}^{A}$ are randomly sampled and left without noise in order to reproduce the situation of the seven WT patterns of the choke valve case study which have small uncertainties.

5 Improving the Quality of the Allocated Parameters

After having verified that the values of $\dot{m}_o$, $\dot{m}_w$, and $\dot{m}_g$ of the choke valve case study are noisy and unreliable, a procedure for improving the accuracy of those parameter estimates is here proposed. This is done by means of empirical models which learn from a training set the relationships between the parameters, and provide as output an estimate $\hat{\mathbf{x}}_k$ of the input parameters $\mathbf{x}_k$. Different regression techniques such as those based on the use of principal component analysis,[23] artificial neural networks,[24,25] support vector machines,[26] evolving clustering methods[27] have been applied to this purpose. In this work, Kernel Regression models[10,11] have been chosen.

Nonparametric Kernel Regression (KR) is used to build a model for improving the quality of the allocated values of oil, water and gas mass flow rates. Compared with parametric methods, which are defined by sets of parameters and predefined functional relationships, nonparametric methods have the advantage that they do not require any assumption about the mathematical structure of the regression model.[10]

KR models provide estimates by developing local models in the neighborhoods of the test patterns they are fed with. Estimates are obtained as weighted averages of the training patterns, with weights decreasing as the distance between the test and the training patterns increases. In this view, training patterns closer to the test pattern are conjectured to be more similar to it, thus giving the most relevant contribution to its estimate. Distances between test and training patterns are evaluated based on a subset of the available parameters belonging to the predictor group (PG). More details about the KR method are given in Appendix B.

In the choke valve case study, the choice of training dataset and predictor parameters is critical. In this respect, four different models can be devised by differentiating the training set as listed in Table 7.

Table 7. Model training procedures

Model	Training set	Predictor parameters
1	Well test data $\mathbf{X}^{WT}$	Measured $\mathbf{x}_k^r = [\Delta P, \theta]$
2	Standard inspections data $\mathbf{X}^{SI}$	Measured $\mathbf{x}_k^r = [\Delta P, \theta]$
3	Well test data $\mathbf{X}^{WT}$	Measured & allocated $\mathbf{x}_k = [\Delta P, \theta, \dot{m}_o, \dot{m}_w, \dot{m}_g]$
4	Standard inspections data $\mathbf{X}^{SI}$	Measured & allocated $\mathbf{x}_k = [\Delta P, \theta, \dot{m}_o, \dot{m}_w, \dot{m}_g]$

The KR models return in output the unreliable parameters that need to be estimated $\mathbf{x}_k^{RG} = \mathbf{x}_k^u = [\dot{m}_o, \dot{m}_w, \dot{m}_g]$.

Since the performances of the models depend on the characteristics of the parameter to be estimated and the intensity of the noise, as shown below in Section 5.1, it is difficult to identify a single best model.

Using an ensemble of models allows overcoming this dilemma. Indeed, the general idea underlying ensembles is to create many models and combine their outputs in order to achieve a performance which is better than that provided by each individual model in the ensemble.[12] Models' prediction diversity plays a fundamental role when ensemble approaches are devised. In fact, individual models committing diverse errors can be opportunely combined in such a way that the error of the aggregated prediction is smaller than the error of any of the individual models.

In Ref. 31, it is shown that in the case of very noisy parameters the reconstruction error can be reduced by iterating the reconstruction procedure: the reconstruction of the noisy parameters obtained at the previous iteration is repeatedly given in input to the reconstruction model. In this application, in order to obtain the estimate at one iteration, the values of the allocated parameters in $\mathbf{x}_k^u$ estimated by the ensemble at the previous iteration are given in input to the ensemble together with the original values of the measured parameters in $\mathbf{x}_k^r$.

5.1 Outcome aggregation with Analytic Hierarchy Process

Different techniques for the aggregation of the outcomes of individual models have been proposed in the literature, the most common being statistical methods like the simple mean, the median and the trimmed mean.[29,30] Other aggregation techniques, which allow improving the ensemble performance, consider weighted averages of the model outcomes with weights proportional to the performance of the individual models. In this respect, both global approaches (in which the performance is computed on all the available patterns) and local approaches (which measure the performance only on the patterns closed to the test pattern) have been proposed.[31] Note that these techniques, which require the availability of a complete input-output data set of patterns in order to compute the individual model performances, cannot be used in the choke valve case study considered in this work, since the available output values (allocated values of $\dot{m}_o, \dot{m}_w, \dot{m}_g$) are not reliable. For this reason, a new strategy for outcome aggregation in ensemble systems is here proposed. The strategy is based on the use of the Analytic Hierarchy Process (AHP).[14]

AHP is a multi-criteria decision method that uses hierarchic structures to represent a decision problem and provides ranking of different choices.[14] It has been extensively studied since its proposal by Saaty in the 1970s. Here, beyond its traditional purpose, this technique is used in an original way to assign performance weights to the models of the ensemble. The proposed procedure allows ranking different model outcomes using relative performance measurements, without resorting to an absolute measurement of the model performance. AHP consists of two main steps: 1) structuring a hierarchy; 2) assigning priorities to the elements of each hierarchy level by comparative judgments of the elements based on a pre-defined scale.

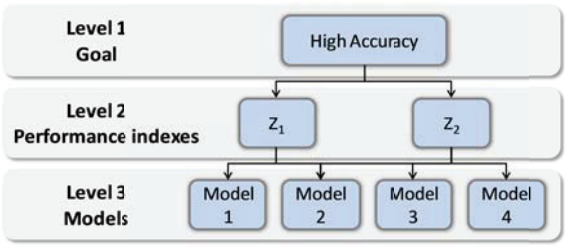

Fig. 5. Model weighting hierarchy structure.

In this application, the hierarchy structure sketched in Fig. 6 is used. The four models on level 3 are compared with respect to the two criteria Z_1 and Z_2 of the level 2 towards the goal (level 1) of obtaining high model accuracy.

The basic tools for assigning priorities to the elements of a level of the hierarchy are matrices of pairwise comparisons based on the criteria defined at the previous level. For the hierarchy of Fig. 5, two matrices of comparisons $\mathbf{A}_{Z_1}$ and $\mathbf{A}_{Z_2}$ have to be defined, each one containing elements a_{ij} representing the relative importance of model i when compared to model j based, respectively, on criteria Z_1 and Z_2.

Once a matrix of comparisons $\mathbf{A}_{Z_l}$ is defined, the vector of priorities $\boldsymbol{\pi}_{Z_l}$ of the models in level 3 of the hierarchy with respect to criterion Z_l is given by the eigenvector associated to the maximum eigenvalue of matrix $\mathbf{A}_{Z_l}$. The priority vectors obtained for each criterion are weighted with the priority assigned to the corresponding criterion and averaged to obtain the overall priority vector $\boldsymbol{\pi}=[\pi_1, \pi_2, \pi_3, \pi_4]$ assigning the priority π_m to model m.

In the proposed aggregation method, the priorities assigned to each model are used as weights to aggregate the models' outcomes through a weighted average:

$$\hat{\mathbf{x}}_{tst}^{u} = \frac{\sum_{m=1}^{4} \pi_m \hat{\mathbf{x}}_{tst}^{u,m}}{\sum_{m=1}^{4} \pi_m} \tag{13}$$

where $\hat{\mathbf{x}}_{tst}^{u,m}$ is the estimate provided by model m of the unreliable parameters in $\mathbf{x}_{tst}^{u}$.

In this application, the first criterion Z_1 chosen to evaluate the relative importance $a_{ij}(\mathbf{x}_{tst})$ of model i with respect to model j in the reconstruction of a test pattern $\mathbf{x}_{tst}$ is the relative similarity of the two models outcomes $\hat{\mathbf{x}}_{tst}^{u,i}$ and $\hat{\mathbf{x}}_{tst}^{u,j}$ to the remaining models outcome $\hat{\mathbf{x}}_{tst}^{u,m}$, $m \neq i, j$. Assuming that the outcomes of the models left out of the pair-wise comparison are distributed around the correct value, this criterion assigns larger weights to the model (i or j) whose outcome is more similar to that of the models left out.

The similarity of two patterns $\hat{\mathbf{x}}_{tst}^{u,i}$ and $\hat{\mathbf{x}}_{tst}^{u,m}$ has been estimated by the inverse of their Euclidean distance $d(\hat{\mathbf{x}}_{tst}^{u,i}, \hat{\mathbf{x}}_{tst}^{u,m})$; the relative importance $a_{ij}^{m}(\mathbf{x}_{tst})$ of a model i with respect to model j when model m is taken as reference is defined by:

$$a_{ij}^{m}(\mathbf{x}_{tst}) = d(\hat{\mathbf{x}}_{tst}^{u,j}, \hat{\mathbf{x}}_{tst}^{u,m}) / d(\hat{\mathbf{x}}_{tst}^{u,i}, \hat{\mathbf{x}}_{tst}^{u,m}) \qquad (14)$$

and the entry a_{ij} of the comparison matrix $\mathbf{A}$ is given by the product of the relative importance values $a_{ij}^{m}(\mathbf{x}_{tst})$ m=1,…,4, $m \neq i, j$:

$$a_{ij} = \prod_{m \neq i,j} a_{ij}^{m}(\mathbf{x}_{tst}) \qquad (15)$$

According to the AHP method, the quality of a matrix of comparison can be evaluated considering its consistency. Matrix $\mathbf{A}_{Z_1}$ is consistent if the following equation is satisfied for any i, j and k [14]:

$$a_{ij} a_{jk} = \frac{\pi_i}{\pi_j} \frac{\pi_j}{\pi_k} = \frac{\pi_i}{\pi_k} = a_{ik} \qquad (16)$$

In our case, to substitute eqs. (14) and (15) in eq. (16) gives:

$$\begin{aligned} a_{ij} a_{jk} &= \prod_{m \neq i,j} \frac{d^{jm}}{d^{im}} \prod_{m \neq j,k} \frac{d^{km}}{d^{jm}} \\ &= \left[\frac{d^{jk}}{d^{ik}} \prod_{m \neq i,j,k} \frac{d^{jm}}{d^{im}} \right] \left[\frac{d^{ki}}{d^{ji}} \prod_{m \neq j,k,i} \frac{d^{km}}{d^{jm}} \right] \\ &= \frac{d^{jk}}{d^{ji}} \prod_{m \neq i,j,k} \frac{d^{km}}{d^{im}} = \prod_{h \neq i,k} \frac{d^{km}}{d^{im}} = a_{ik} \end{aligned}$$

where $d^{ij} = d(\hat{\mathbf{x}}_{tst}^{u,i}, \hat{\mathbf{x}}_{tst}^{u,j})$ and, by definition, $d^{ij} = d^{ji}$. This shows that, in the proposed approach, matrix $\mathbf{A}_{Z_1}$ is consistent.

A second criterion Z_2 for evaluating the performance of a model takes into account the RMSE in reconstructing the reliable parameters in $\mathbf{x}_{tst}^{r}$, i.e. the root mean square difference between the reconstructed and measured values. This second criterion takes into account the fact that robust and reliable models should be able to correctly reconstruct the reliable parameters of $\mathbf{x}_{tst}^{r}$ despite the noise on the unreliable parameters of $\mathbf{x}_{tst}^{u}$.

Since all model performances are evaluated with respect to the same reference, i.e. the reliable measurements in $\mathbf{x}_{tst}^{r}$, the pair-wise comparison is not needed, and the vector of priorities $\boldsymbol{\pi}_{Z_2}$ is computed by taking for each model h=1,…,4, the inverse of its RMSE, i.e. $\pi_{Z_2}^{m} = 1 / RMSE^{m}$.

Finally, the two criterions Z_1 and Z_2 of level 2 of the hierarchy are given the same importance and thus the priority vector $\boldsymbol{\pi}$ is given by:

$$\boldsymbol{\pi} = [0.5 \quad 0.5] \cdot \begin{bmatrix} \boldsymbol{\pi}_{Z_1} \\ \boldsymbol{\pi}_{Z_2} \end{bmatrix} \qquad (17)$$

5.2 Results

Given the impossibility of verifying the correctness of the oil, water and gas mass flow rates estimates provided by the AHP aggregated ensemble of KR models in the choke valve case study, its performance is firstly verified with respect to the artificial case study introduced in Section 4.

5.2.1 Application to the artificial case study

In this Section the KR models and the ensemble approach are applied to estimate parameters x_3^A, x_4^A and x_5^A in the artificial case study of Section 4 for different values of the standard deviation σ_n of the noises applied to the unreliable parameters x_3^A, x_4^A and x_5^A. For each model, the bandwidth parameter h (eq. (12)) has been set through a trial and error procedure in order to minimize the root mean square error (RMSE) of the model in estimating the noisy parameters. Fig. 6 reports the reconstruction errors of the four KR models for different values of the noise standard deviation σ_n.

Notice that the performance of model 1 does not depend on the noise intensity, since the information used to develop the model (the training set of undisturbed patterns simulating the WT measurements) and the information fed to the model to estimate the unreliable parameters (predictor parameters $\mathbf{x}_k^r = [x_1^A, x_2^A]$) are not affected by noise. As expected, the other model performances tend to decrease as the noise intensity increases. In particular, model 4, which is built using training patterns affected by the noise and receives in input noisy parameters, is the most affected by the noise. Model 2 tends to outperform the other models for small noise intensities. This is due to the fact that model 2 is built using the largest training dataset and receives in input only the undisturbed parameters x_1^A and x_2^A; on the other side, large noise intensities tend to reduce the performance of this model since they affect the value of the response parameters x_3^A, x_4^A and x_5^A of the training patterns.

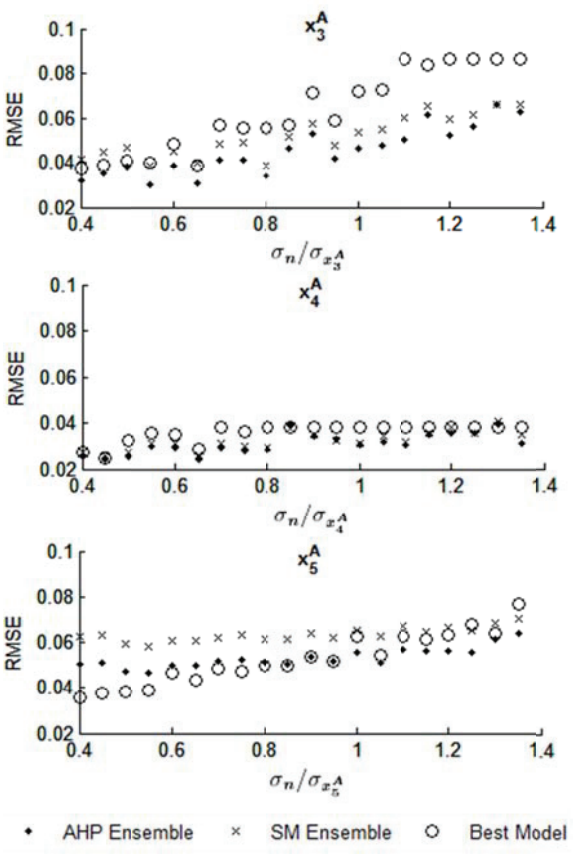

Fig. 7. Comparisons of the performance of the SM and AHP ensembles and of the best individual model for different noise intensities.

Then, the ensemble of models have been tested using the same values of standard deviation noises on the unreliable parameters x_3^A, x_4^A and x_5^A. Fig. 7 compares the performances of the ensemble aggregated using the AHP strategy with those of the ensemble aggregated using the simple mean (SM) of the model outcomes and those of the best performing model. Results show that the AHP ensemble outperforms all the four KR models in 77% of the cases, whereas the best model slightly outperforms the AHP ensemble only in the reconstruction of parameter x_5^A when low values of σ_n are considered. However, these noise values are out of the range [0.8, 1.25]. This confirms the higher robustness standards achievable with the AHP ensemble approach which also generally outperforms the SM aggregation.

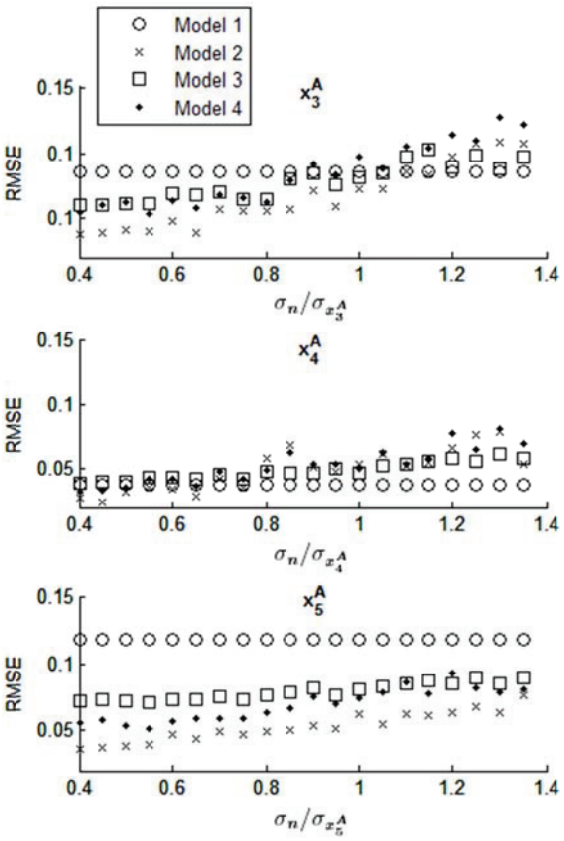

Fig. 6. Comparison of the reconstruction performance obtained by the four KR models for different noise intensities.

Table 7 compares the RMSE (averaged over different values of σ_n) obtained in estimating the noisy parameters x_3^A, x_4^A and x_5^A by the four individual models and by the SM and AHP ensembles. Results confirm that, in average, the AHP ensemble reconstruction outperforms the others.

The artificial case study represents a general situation characterized by the presence of reliable and unreliable parameters, considering different correlations between the parameters and different noise levels. Since the AHP ensemble has provided satisfactory performance in the artificial case study, we expect that it also will provide accurate reconstructions of the mass flow rates in the choke valve case study.

Table 8. RMSE of the KR models and ensembles in estimating parameters x_3^A, x_4^A and $\boldsymbol{x}_5^A$

Model 1	Model 2	Model 3	Model 4	SM ensemble	AHP ensemble
0.0810	0.0550	0.0822	0.0817	0.0582	0.0472

5.2.2 Application to the choke valve case study

The ensemble approach is finally applied to the choke valve case study to improve the quality of the mass flow rates $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$ allocations. The test set is made by the 259 patterns of X^{SI}. A leave-one-out cross validation procedure has been adopted[29]: according to this procedure, at each cross-validation a single pattern from the original dataset X^{SI} is used as test and the remaining N_{SI}-1 patterns as training. This is repeated N_{SI} times so that each pattern of the dataset is used once as test. The estimates are then used to calculate the choke valve health indicator δC_V (eqs. (2) and(3)). The procedure is iterated 10 times. Table 8 compares the value of the Spearman's rank correlation coefficient r_S of the health indicator obtained using the SI dataset, the estimates of the four individual models and those of the SM and the AHP ensembles.

Results in Table 8 show that estimating $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$ allows increasing the monotonicity of the health indicator δC_V with respect to that obtained by directly using the value computed during standard inspections. Furthermore, notice that in this case model 3 generates a health indicator slightly more monotone than that obtained by using the AHP ensemble. Nevertheless, since the performance of this model in the more data-rich and robust artificial case study are much worse than those obtained by the AHP ensemble (Table 7), the estimates obtained by the latter are used to calculate the choke valve health indicator.

Table 9. Monotonicity r_s of the health indicator calculated using the SI dataset, the individual models estimates and those of the SM and AHP ensembles

Method for mass flow rate estimation	r_s
SI data	0.740
Model 1	0.847
Model 2	0.903
Model 3	0.920
Model 4	0.843
SM ensemble	0.918
AHP ensemble	0.919

Fig. 8 shows the δC_V obtained using the SI allocated values of $\dot{m}_o$, $\dot{m}_w$ and $\dot{m}_g$ and those estimated by the AHP ensemble. Notice that the values of δC_V obtained using the estimated values are more monotonic and more similar to those obtained in correspondence of the WT inspections (dots). Nevertheless, neither the AHP ensemble nor any of the single models considered can produce a totally monotonic indicator and some anomalous behaviors remains (e.g., some peaks such as the one occurring between 150 and 200 operational days which corresponds to a decrease in the pressure drop not followed by a decrease of the allocated values of the mass flow rates).

6 Conclusions

In this paper, we have tackled the problem of providing a reliable health indicator of a choke valve used in offshore oil platforms which undergoes erosion. The health indicator is derived from the valve flow coefficient which is a valve parameter that regulates the analytical relationship between the pressure drop across

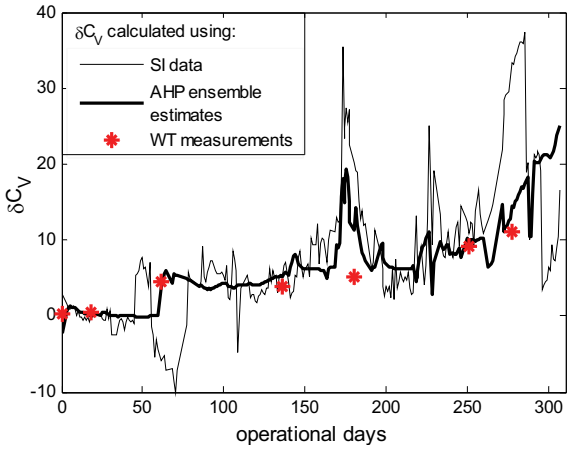

Fig. 8. Comparison of the health indicator obtained using the allocated values of the mass flow rates and those estimated by the AHP ensemble.

the choke and the flow of oil, water and gas through the choke. The difference between the theoretical and actual value of the valve coefficient highlight the contribution of the erosion. The theoretical value is given by the valve producer, while the actual value can be analytically calculated. A major problem is due to the inaccuracy of oil, water and gas mass flow rates which are used to calculate the actual valve flow coefficient. In

fact, such values are not directly measured, but allocated for a single well by a software based on the measured total production of a number of wells and on physical parameters (pressures and temperatures) related to the single well. They are therefore affected by large uncertainties which lead to highly inaccurate calculations of the erosion state of the choke valve.

The scope of this paper has been to devise a procedure to improve the quality of those allocated parameters based on the other available measurements (pressure drop and choke opening) which are conjectured to be reliable. Operatively, a number of well tests have been performed throughout the valve life and few reliable measurements are available also for the oil, water and gas flow rates.

In the paper, Fuzzy C Means clustering has been applied to verify the consistency of the measured and allocated parameter. A comparison of the FCM partitions obtained in the space of the measured and allocated parameters has been made and the importance of each parameter has been evaluated in the data partitioning by a supervised evolutionary clustering. The results of the analyses performed on the choke valve data have indicated the low reliability of the allocated values of the mass flow rates. This has led to the development of a method for improving their quality.

To this aim, Kernel Regression models have been devised. Different training procedures have been adopted to generate diverse models within an ensemble approach. To aggregate the outcomes of the individual models, an original technique based on the Analytic Hierarchy Process (AHP) method has been used. The results obtained in an artificial case study, reproducing the choke valve case study, have confirmed the improved performances of the ensemble with respect to any of the single KR models. The application of the proposed method to the choke valve case study has allowed significant improvement of the oil, water and gas mass flow rates calculation and, as a consequence, it has improved the quality of the health indicator.

Since a general application of the proposed approach is envisioned in situations in which unreliable parameter values need to be improved by resorting to a set of reliable parameters, future works will be devoted to demonstrate its applicability in different industrial contexts.

Appendix A: The Unsupervised Fuzzy C Means Technique

The Fuzzy C Means (FCM) technique is an unsupervised clustering technique, since it makes no use of a priori known information on the true classes of the data. The clustering is based on the minimization of a weighed sum Y of the distances $d(\mathbf{x}_k,\mathbf{v}_c)$ between the patterns $\mathbf{x}_k$ and the cluster centers $\mathbf{v}_c$,

$$Y = \sum_{c=1}^{C}\sum_{k=1}^{N}[\mu_c(\mathbf{x}_k)]^{\omega} d^2(\mathbf{x}_k,\mathbf{v}_c) \tag{A1}$$

where the weight $\mu_c(\mathbf{x}_k)$ denotes the membership of $\mathbf{x}_k$ to clusters c and ω is a parameter which controls the degree of fuzziness of the clusters (often a value of 2 has been found suitable as in Ref. 9). In the traditional algorithm[7] the distance is Euclidean:

$$\begin{aligned} d^2(\mathbf{x}_k,\mathbf{v}_c) &= d_I^2(\mathbf{x}_k,\mathbf{v}_c) \\ &= (\mathbf{x}_k-\mathbf{v}_c)^{\mathrm{T}}\mathbf{I}(\mathbf{x}_k-\mathbf{v}_c) \end{aligned} \tag{A2}$$

where $\mathbf{I}$ is the identity matrix.

The membership values $\mu_c(\mathbf{x}_k)$ which minimize Y (eq. (A1)) for a given a set of centers $\mathbf{v}_c$, c=1,…,C, are computed as in eq. (A3) and used in eq. (A4) to compute a new optimal set of clusters centers, which are in return used in eq. (A3) to update the membership values. The iterative procedure provides the optimal fuzzy partition of the dataset.

$$\mu_c(\mathbf{x}_k) = \frac{\left[\dfrac{1}{d_I^2(\mathbf{x}_k,\mathbf{v}_c)}\right]^{1/(\omega-1)}}{\displaystyle\sum_{i=1}^{C}\left[\frac{1}{d_I^2(\mathbf{x}_k,\mathbf{v}_i)}\right]^{1/(\omega-1)}} \tag{A3}$$

$$\mathbf{v}_c = \frac{\displaystyle\sum_{k=1}^{N}[\mu_c(\mathbf{x}_k)]^{\omega}\mathbf{x}_k}{\displaystyle\sum_{k=1}^{N}[\mu_c(\mathbf{x}_k)]^{\omega}} \tag{A4}$$

Based on the set of optimal centers $\mathbf{v}_c$, c=1,…,C, a generic pattern $\mathbf{x}_k$ is assigned to cluster c provided that its membership $\mu_c(\mathbf{x}_k)$ exceeds a threshold $\gamma \in (0,1)$ representing the degree of confidence that $\mathbf{x}_k$ belongs to c. If the condition $\mu_c(\mathbf{x}_k) > \gamma$ is never fulfilled or if it is verified for more than one value of c, the pattern is not associated to any cluster.

Appendix B: The Kernel Regression method

Let $\mathbf{X}^{trn}=\{\mathbf{x}_k\}$, $k=1,\ldots, N_{trn}$ be the training set used for the estimate of the test pattern $\mathbf{x}_{tst}$. To develop the KR model, parameters are divided into a predictor group (PG) and a response group (RG) (with the two groups possibly overlapping). For the estimate of $\mathbf{x}_{tst}$, the KR algorithm assigns to each training pattern $\mathbf{x}_k$ a weight $w_k=\mathrm{K}[d_{PG}(\mathbf{x}_{tst},\mathbf{x}_k)]$, where K is the kernel function which produces the weight for a given distance $d_{PG}(\mathbf{x}_{tst},\mathbf{x}_k)$, between the training and the test patterns, computed considering only the parameters of the predictor group. The estimate $\hat{\mathbf{x}}_{tst}^{RG}$ of the RG parameters of the test patterns is obtained as a weighted average of the RG parameters of the training patterns:

$$\hat{\mathbf{x}}_{tst}^{RG} = \frac{\sum_{k=1}^{N_{trn}} w_k \cdot \mathbf{x}_k^{RG}}{\sum_{k=1}^{N_{trn}} w_k} \qquad (11)$$

The kernel function K must be such that training patterns with small distances from the test pattern are assigned large weights and vice versa. Among the several functions which satisfy this criterion, the Gaussian kernel is commonly used[28]:

$$K(d_{PG}) = \frac{1}{\sqrt{2\pi h^2}} \exp\left(-\frac{{d_{PG}}^2}{2h^2}\right) \qquad (12)$$

where the parameter h defines the kernel bandwidth and is used to control how close training patterns must be to the test pattern to be assigned a large weight. In order to compute d_{PG}, the PG parameters are normalized to mean equal to 0 and standard deviation equal to 1.

Acknowledgments

The authors wish to thank Erling Lunde and Morten Løes at Statoil ASA for proving us with the operational choke valve data and the IO Center for Integrated Operations in the Petroleum Industry (www.ntnu.no/iocenter) for funding this research project.

References

1. Garvey, D.R., Baumann, J., Lehr, J., Hughes, B., and Hines, J.W. 2009. Pattern Recognition Based Remaining Useful Life Estimation of Bottom Hole Assembly Tools. SPE/IADC Drilling Conference and Exhibition, 17-19 March, Amsterdam, NL. doi: 10.2118/118769-MS.
2. Vachtsevanos, G., Lewis, F.L., Roemer, M., Hess, A., and Wu, B. 2006. Intelligent Fault Diagnosis and Prognosis for Engineering Systems, 1st edition, Hoboken, New Jersey, John Wiley & Sons.
3. Jarrell, D.B., Sisk, D.R., and Bond, L.J. 2004. Prognostics and Condition-Based Maintenance: A New Approach to Precursive Metrics, Nuclear Technology, Vol. 145, pp. 275-286.
4. Hines, J.W., and Usynin, A. 2008. Current Computational Trends in Equipment Prognostics. International Journal of Computational Intelligence Systems, Vol. 1, Issue 1, pp. 94–102. doi: 10.2991/ijcis.2008.1.1.7.
5. Nystad, B.H., Gola, G., Hulsund, J.E., and Roverso, D. 2010. Technical Condition Assessment and Remaining Useful Life Estimation of Choke Valves subject to Erosion. Proc. PHM Society 2010 Ann. Conf., October 11-13, Portland, OR.
6. Haugen, K., Kvernvold, O., Ronold, A., Sandberg, R. 1995. Sand Erosion of Wear Resistant Materials: Erosion in Choke Valves. Wear, Vol. 186-187, Part 1, pp. 179-188. doi:10.1016/0043-1648(95)07158-X.
7. Dunn, J.C. 1974. A fuzzy relative of the isodata process and its use in detecting compact, well separated clusters. Cybernetics and Systems, Vol 3, Issue 3, pp. 32–57.
8. Jain, A.K., Murty, M.N., and Flyn, P.J. 1999. Data Clustering: A Review. ACM Computing Surveys (CSUR), Vol. 31, Issue 3, pp. 264–323. doi: 10.1145/331499.331504.
9. Zio, E., and Baraldi, P. 2005. Identification of nuclear transients via optimized fuzzy clustering. Ann. of Nucl. Energy 32, pp.1068–1080. doi: 10.1016/j.anucene.2005.02.012.
10. Nadaraya, E.A. 1964. On Estimating Regression. Theory of Probability and Its Applications, Vol. 10, pp. 186–190. doi: 10.1137/1109020.
11. Atkeson, C.G., Moore, A.W., and Schaal, S. 1997. Locally Weighted Learning. Artificial Intelligence Review, Vol. 11, pp. 11–73. doi: 10.1.1.73.9932.
12. Perrone, M.P., and Cooper, L.N. 1992. When networks disagree: ensemble methods for hybrid neural networks, National Science Fundation, USA. doi: 10.1.1.32.3857.
13. Bonissone, P.P., 2010. Soft Computing: A Continuously Evolving Concept, International Journal of Computational Intelligence Systems, Vol.3, No. 2, pp. 237-24.
14. Saaty, T.L. 1980. The analytic Hierarchy Process, Planning, Priority Setting, Resource Allocation. McGraw-Hill, New York.
15. Wold, K., Hopkins, S., Jakobsen, T., Lilleland, S.E., Roxar, R.S., and Brandal, Ø. 2010. New Generation Software Integrates Intrusive and Non-intrusive Systems for Corrosion and Sand/erosion Monitoring. SPE Int. Conf. on Oilfield Corrosion, 24-25 May, Aberdeen, UK. doi: 10.2118/130569-MS.
16. Nøkleberg, L., and Søntvedt, T. 1995. Erosion in choke valves-oil and gas industry applications. Wear, Vol. 186-

187, Part 2, pp. 401-412. doi: 10.1016/0043-1648(95)07138-5.
17. Birchenough, P.M., Cornally, D., Dawson, S.G.B., McCarthy, P., and Susden, S. 1994. Assessment of Choke Valve Erosion in a High-Pressure, High-Temperature Gas Condensate Well Using TLA. European Petroleum Conference , 25-27 October , London, UK. doi: 10.2118/28887-MS.
18. Metso Automation. 2005. Flow Control Manual. 4th edition.
19. Myers, J.L., Well, A.D., and Lorch, R.F. 2010. Research Design and Statistical Analysis. Taylor & Francis, New York.
20. Devijver, P.A., and Kittler, J. 1982. Pattern Recognition: A Statistical Approach. Prentice/Hall, Englewood Cliffs, NJ.
21. Yuan, B., Klir, G., Swan-Stone, J. 1995. Evolutionary fuzzy c-means clustering algorithm. Proc. Fourth IEEE Int. Conf. Fuzzy Syst., Vol 4, pp. 2221–2226. doi: 10.1109/FUZZY.2007.4295668.
22. Zio, E., Baraldi, P., 2004. A fuzzy clustering approach for transients classification. In: Ruan, D., D_hondt, P., Cock, M.D., Nachtegael, M., Kerre, E.E. (Eds.), Applied Computational Intelligence. Word Scientific. doi: 10.1142/9789812702661_0103.
23. Baraldi, P., Zio, E., Gola, G., Roverso, D., and Hoffmann, M. 2009. A procedure for the reconstruction of faulty signals by means of an ensemble of regression models based on principal components analysis, NPIC-HMIT Topical Meeting, April 5-9, Knoxville, USA.
24. Fantoni, P.F., and Mazzola, A. 1996. Multiple-Failure Signal Validation in Nuclear Power Plants using Artificial Neural Networks, Nuclear technology, Vol. 113, Issue 3, pp. 368-374.
25. Marseguerra, M., Zio, E., and Marcucci, F. 2006. Continuous Monitoring and Calibration of UTSG Process Sensors by Autoassociative Artificial Neural Network. Nuclear Technology, Vol. 154, Issue 2, pp. 224-236.
26. Sun, B.Y., Huang, D.S., and Fang, H.T. 2005. Lidar Signal Denoising Using Least-Squares Support Vector Machine. IEEE Signal Processing Letters, Vol. 12, Issue 2, pp.101-104. doi: 10.1109/LSP.2004.836938(410) 12.
27. Chevalier, R., Provost, D., and Seraoui, R. 2009. Assessment of Statistical and Classification Models For Monitoring EDF's Assets, Sixth American Nuclear Society Int. Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies, Knoxville, USA, 2009.
28. Wand, M.P., and Schucany, W.R. 1990. Gaussian-based kernels for curve estimation and window width selection. Canadian Journal of Statistics, Vol. 18(3), pp. 197–204.
29. Polikar, R. 2007. Bootstrap-inspired techniques in computational intelligence, IEEE Signal Processing Magazine, Vol. 59, pp. 59–72. doi: 10.1109/MSP.2007.4286565.
30. Baraldi, P., Cammi, A., Mangili, F., and Zio, E. 2010. Local Fusion of an Ensemble of Models for the Reconstruction of Faulty Signals, IEEE Trans. on Nucl. Sci., Vol.57, Issue 2, pp. 793 - 806. doi: 10.1109/TNS.2010.2042968.
31. Hruschka, E., Campello, S., Freitas, A. De Carvalho, A. IEEE Transactions on A Survey of Evolutionary Algorithms for Clustering, Systems, Man, and Cybernetics, Part C: Applications and Reviews,
32. J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," Applied Statistics, Vol. 28, pp.100-108, 1979.