Vol. 18, No. 1 (2019) 83-98

Simulación y Control

Revista Mexicana de Ingeniería Química

A FAULT DIAGNOSIS PROPOSAL WITH ONLINE IMPUTATION TO INCOMPLETE OBSERVATIONS IN INDUSTRIAL PLANTS

UNA PROPUESTA DE DIAGNÓSTICO DE FALLOS CON IMPUTACIÓN EN LÍNEA PARA OBSERVACIONES INCOMPLETAS EN PLANTAS INDUSTRIALES

O. Llanes-Santiago[1]*, B.C. Rivero-Benedico[1], S.C. Gálvez-Viera[1], E.F. Rodríguez-Morant[1], R. Torres-Cabeza[1], A.J. Silva-Neto[2]

[1]*Universidad Tecnológica de La Habana José Antonio Echeverría, CUJAE. Calle 114, No 11901, CUJAE, Marianao, La Habana, Cuba, CP 19390.*

[2]*Instituto Politécnico da Universidade do Estado do Rio de Janeiro (IPRJ-UERJ). Rua Bonfim 25-Camous UERJ, Vila Amelia – CEP 28625-670, Nova Friburgo, RJ. Brasil.*

Received: February 21, 2018; Accepted: April 24, 2018

Abstract

In this paper, the problem of fault diagnosis in complex industrial systems in the presence of missing data is addressed. Firstly, how to perform online imputation when there are missing values in the observations obtained by the data acquisition system is presented. Later, the possibility to apply advanced statistical techniques as Sequential Regression Multiple Imputation, Singular Value Decomposition, Local Least Squares Imputation and k- Nearest Neighbors as examples of possible tools to be used in the online imputation is displayed. In addition, the effects on the fault diagnosis process, when using these statistics tools to estimate the missing data are analyzed. A Neural Network Multi-layer Perceptron for the fault diagnosis system was used. The study was done using the Tennessee Eastman benchmark process. The results show the viability of the proposal. *Keywords*: missing data, online imputation, statistical techniques, fault diagnosis, industrial process.

Resumen

En este trabajo se presenta una propuesta de diagnóstico de fallos en sistemas industriales complejos cuando en el proceso de adquisición de datos se produce pérdida de información. Primeramente se presenta cómo realizar la imputación en línea cuando hay valores de variables perdidas en cada observación que es obtenida por el sistema de adquisición de datos. Posteriormente se presenta como aplicar las técnicas estadísticas de Imputación Múltiple con Regresión Secuencial, Descomposición en Valores Singulares, Mínimos Cuadrados Locales y k-Vecinos más Cercanos como ejemplos de posibles herramientas a utilizar en la imputación en línea. También se analizan las afectaciones que se producen en el proceso de diagnóstico de fallos cuando se utilizan estos métodos para estimar las variables perdidas. Como herramienta para el diagnóstico se utilizó una Red Neuronal Perceptrón Multicapa. El estudio se realizó utilizando el proceso de prueba Tennessee Eastman y sus resultados muestran la viabilidad de la propuesta.

Palabras clave: pérdida de información, imputación en línea, herramientas estadísticas, diagnóstico de fallos, procesos industriales.

1 Introduction

As industries develop and the processes taking place in them become more complex, the necessity to automatically detect the faults affecting the industrial systems have increased. In addition, it is very important the location and identification of them to counteract their negative consequences on the processes.

From the point of view of plant safety, the fast and efficient diagnosis of faults has become a very important task, since their occurrence in processes of medium or large magnitude can cause great economic, environmental and human losses (MacGregor and Cinar 2012, Prieto-Moreno *et al*. 2013, Zhang *et al*. 2013, García-Morales *et al*. 2015, Téllez-Anguiano *et al*. 2016). For that reason, it is fundamental the training of technical personnel in this regard (Castello *et al*. 2015).

* *Corresponding author. E-mail*: orestes@tesla.cujae.edu.cu, oresteslls@gmail.com ; *Tel. +53 72663280*

doi: https://doi.org/10.24275/uam/izt/dcbi/revmexingquim/2019v18n1/Llanes
issn-e: 2395-8472

From the knowledge of the process in its normal operation mode, the objective of the fault diagnosis methods is to analyze the behavior of the process in order to determine if it corresponds to its normal operation state or another known state characterized by the presence of a fault. (Zhang *et al*. 2013, Téllez-Anguiano *et al*. 2016). The development of digital instrumentation, industrial networks and Supervisory Control and Data Acquisition (SCADA) systems, allow storing a large volume of data from industrial processes, which permit the use of fault diagnosis methods based on historical data. These methods are very advantageous in very complex processes where it is very difficult to obtain models that represent their operation satisfactorily (Venkatasubramanian *et al*. 2002).

It is well known that methods based on historical data are affected by missing information (Walczak and Massart, 2001a; Walczak and Massart, 2001b; Nelson *et al*., 2006). In the case of the chemical industries, there are some aspects that may cause incomplete data sets (Askarian *et al*. 2016). Therefore, it is necessary to address this problem with the aim of having robust fault diagnosis in order to avoid false alarms and obtain reliable fault diagnosis systems (Askarian *et al*., 2016; Severson *et al*., 2017). The treatment of missing data has become a fundamental requirement when monitoring the process state. An incorrect treatment of them can cause great errors or false results in the classification process (Nelson *et al*., 2006; Zhang *et al*., 2013). In some articles as García-Laencina *et al.* (2010), Luengo *et al*. (2012a), Askarian *et al*. (2016), and Severson *et al*. (2017) this subject is analyzed, and it is considered as a very common current problem. Then, in order to achieve a satisfactory performance in fault diagnosis systems, it is necessary to select the techniques to be used to deal with the missing information.

Several methods for the treatment of missing data have been proposed in the scientific literature (Raghunathan *et al*. 2001, Little and Rubin 2002, Li *et al*. 2004, García-Laencina *et al*. 2010, Jerez *et al.* 2010, Luengo *et al*. 2012a, b, Askarian *et al*. 2016, Sovilj *et al*. 2016 and Severson *et al*. 2017). The main methods according to García-Laencina *et al.* (2010) are:

- Ignoring and eliminating incomplete data: only data that is complete is used.
- Imputing or estimating missing data using statistical or computational intelligence tools.
- Model based on the distribution of the data.
- Learning Machines: procedures where the missing values are incorporated into the classifier.

After studying these techniques, it was concluded that the elimination of the observation with missing variables and the imputation of the missing values are the most used techniques to treat the missing data in the fault diagnosis systems. From these two variants, the most recommended in the literature consulted is the imputation, which estimates the missing values using the information available and does not eliminate observations that may contain important information for the diagnostic work (García-Laencina *et al*.2010, Askarian *et al*. 2016, Severson *et al*. 2017).

In the large number of studied papers that address this issue, a common characteristic of the methods used to carry out the imputation is the necessity of having a set of observations of the process (data matrix), some of them with missing data. With the data matrix formed, the imputation is made to estimate the missing values in the data matrix using fundamentally, the relationship among the variables. When the imputation process is finished, the complete data matrix is obtained. This result enables to begin the classification of the observations that form the data matrix for the fault diagnosis system. The latter implies the need to store a set of observations that cannot be classified by the diagnostic system until the imputation process is developed.

The previous approach establishes a key question about its effectiveness as follows:

If it is necessary to wait until having a set of observations stored in order to have the minimum data matrix that allows applying the imputation methods, then it is possible that in the time used to accumulate these observations, a fault could occur and the fault diagnosis system might not early detect it.

In the case of the modern industrial systems where the demands of efficiency, quality and safety are very high, the aforementioned procedure implies a hard limitation. The imputation of the missing variables of an observation obtained from the process must be done online and with strict time requirements determined by the sampling time that has been established in the data acquisition system. Taking into account the above-mentioned, the main objective of this paper is to propose a procedure to carry out the imputation process online for each observation obtained by the data acquisition system which will permits its immediate classification by the fault diagnosis system. In the paper, it is also analyzed how to apply a group of effective statistical tools widely used for imputation in large databases (Sequential Regression Multiple Imputation, Singular Value Decomposition, Local Least Squares Imputation and k- Nearest Neighbors). A comparative analysis of the results obtained, the influence of the imputation process on the subsequent classification process, and the time requirements for the online imputation process is also presented in this paper. The main contributions of this paper are the following:

1. A proposal of procedure to estimate online the missing data for each observation obtained from the data acquisition system.
2. The implementation of the highly effective statistical methods to carry out the on-line imputation for each observation obtained with missing variables in the fault diagnosis system. In addition, a comparative analysis of the performance of these imputation tools is made.

It was not an objective of this paper to compare some classification tools. A Multilayer Perceptron Neuronal Network (MLP) was used as a classifier because it has been identified as one of the computational intelligence tools with better results in classification processes in the scientific literature (Patan *et al*. 2008). The structure of this paper is as follows, in the Materials and Methods section, the proposed procedure for the estimation online of the missing data for each observation obtained, the general characteristics and operation of the imputation techniques used in this paper (SRMI, SVD, LLSI, k-NN), the Neural Network Multilayer Perceptron as a classification tool and the Tennessee Eastman Process benchmark to evaluate the proposed procedure are presented. The Experiments and Results section shows the experiments developed and the analysis of the obtained results. Finally, conclusions and recommendations for future researches are exhibited.

2 Materials and methods

2.1 *Proposed scheme for the online imputation process*

When the data acquisition system obtains an observation from the process, the first step in the diagnostic system is to analyze if it has the complete information or not.

In case that the observation is complete, the diagnostic system classifies it otherwise: the imputation process is performed to estimate the missing values. Afterwards, the completed observation is classified. This procedure is carried out for each observation that is obtained from the process.

In this way, it is ensured that each observation received is classified either immediately or after imputation process if it has missing variables.

2.2 *Database for training supervised classification tools*

It has been reported in the scientific literature the need of performing a previous offline training when supervised diagnostic tools in fault diagnosis systems are used. For this training, a database containing the necessary number of observations representative of the normal operating state and each one of the fault states is needed (Watanabe *et al*. 1989, Venkatasubramanian and Chan 1989). Several methods have been developed in order to build a training database. The selection of a specific method depends on the nature of the process. In general, these training databases can be divided into two major groups according to the way they are generated (Leonhardt and Ayoubi 1997):

- Databases generated analytically
- Databases generated in a heuristic form

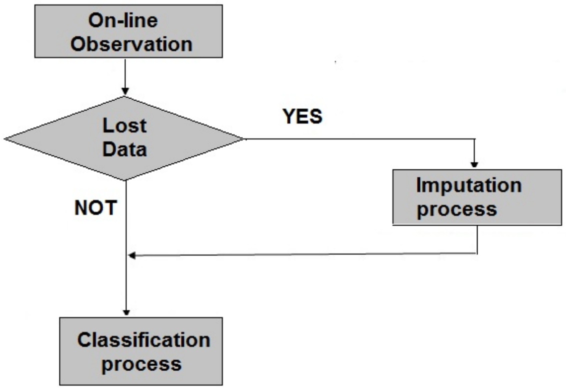

Fig. 1. Flowchart of the on-line imputation and diagnosis.

The databases generated analytically are based on the measurements and those generated in a heuristic form are based on the observations of human operators (Leonhardt and Ayoubi 1997).

In this paper, it is assumed the existence of a training database $E \in \mathbb{R}^{i \times p}$ composed of vectors $Ob_i = \{v_1, v_2, ..., v_p\}$ of p variables where $i = 1, ..., f$. The f observations are needed to characterize adequately, taking into account the type of process, the normal operating state and each one of the fault states.

This training database is used in the offline training of the supervised classification tools and in the online imputation process to estimate the missing variables. The last constitutes one of the key proposals in the main contribution of this paper.

2.3 Sequential regression multiple imputation

Sequential Regression Multiple Imputation (SRMI) has emerged as a popular approach for handling incomplete data with complex features. In this approach, imputations for each missing variable are produced cyclically based on a regression model using other variables as predictors. As an unsupervised tool, it does not need previous off-line training and is being widely used for its satisfactory results (Raghunathan *et al*. 2001).

The basic strategy relies on creating imputations through a sequence of regressions. The type of regression depends on the variable that will be imputed, and the objective of the algorithm is to find the correlation between the variables (Raghunathan *et al*. 2001).

Next, the general characteristics of the algorithm are presented.

Let $M \in \mathbb{R}^{m \times n}$ be a data matrix of m observations of n variables each one. The columns of the matrix M are reordered such that the variables with missing information are grouped in a sub-matrix $Y \in \mathbb{R}^{m \times k}$ of k column vectors Z_j where $j = 1, ..., k$, and the variables with the complete information are grouped in a sub-matrix $X \in \mathbb{R}^{m \times (n-k)}$ of $n - k$ columns. The reordered matrix M can be represented as $(Z_1, Z_2, \cdots, Z_k, X)$. From here, the correlation coefficients of each variable Z_j with missing data and the variables with full information are calculated. Taking into account these correlation coefficients, the variables with missing data are ordered from the most correlated to the least correlated, being ordered in a sub-matrix $Y = [Y_1, Y_2, \cdots, Y_k]$.

In the initial iteration, the imputation is developed using a regression model, according to the following conditioned distributions:

$$\begin{aligned} &Y_1|_X \\ &Y_2|_{X,Y_1} \\ &Y_3|_{X,Y_1,Y_2} \\ &\vdots \\ &Y_k|_{X,Y_1,Y_2,\cdots,Y_{k-1}} \end{aligned}$$

First, the regression of the most correlated variable, Y_1, is performed on X. Once a prediction of Y_1 has been obtained, this variable is incorporated into the X matrix of complete variables and the matrix $[X; Y_1]$ is obtained. Later, the regression of Y_2 (second best correlated) on $[X; Y_1]$ is performed, and so on until the missing values of the variable Y_k are imputed. At the end of the first iteration, there are not missing values in any variable.

In the following iterations, this initial iteration is repeated, but all variables are included because they have no missing values:

$$\begin{aligned} &Y_1|_{X,Y_2,\cdots,Y_k} \\ &Y_2|_{X,Y_1,Y_3,\cdots,Y_k} \\ &Y_3|_{X,Y_1,Y_2,Y_4,\cdots,Y_k} \\ &\vdots \\ &Y_k|_{X,Y_1,Y_2,\cdots,Y_{k-1}} \end{aligned}$$

Since the SRMI algorithm is iterative, it is necessary to establish the stop criteria. For this paper, two stopping conditions were established:

- Error $\epsilon < \beta$ where the value of β is set by the experts, and $\epsilon = M_i^{iter} - M_i^{iter-1}$ is the error calculated from the difference between the observation obtained from the imputation in the current iteration and the observation obtained in the previous iteration.

- The maximum number of iterations $Iter_{max}$ is reached.

Note that this imputation method requires the existence of a data matrix to carry out the imputation. The training database previously established and mentioned in subsection 2.1 will be used to apply this imputation method offline.

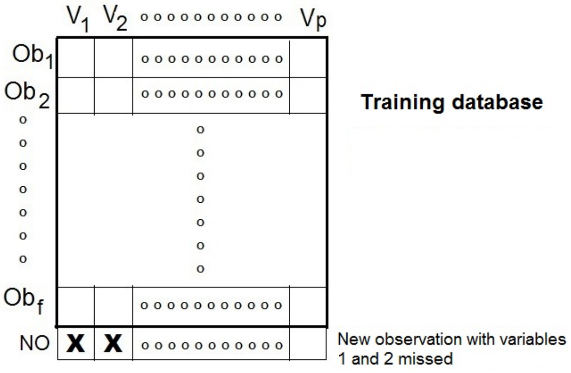

Fig. 2. Example of database to perform the imputation using SRMI algorithm to an observation with two missing variables.

The use of the training database guarantees the best conditions considering the number of observations representing each state, and the quality of these observations which is fundamental to obtain high performances of the imputation algorithms.

When a new observation is received, it is analyzed to determine if there are missing data. If the observation does not have missing data, it is classified by the diagnostic system. If the observation has missing data, it is added as the last row to the training data matrix to perform the imputation. Fig. 2 shows an example of the $M \in \mathbb{R}^{m\times n}$ when it is built with the training database and an observation that arrives missing the values of the variables 1 and 2. In this case $m = f + 1$, $n = p$ and $k = 2$.

To develop the imputation online, it is proposed the following procedure:

1. An observation $Ob_i \in \mathbb{R}^{1\times n}$ with k missing data is received online. The missing variables are eliminated in the new observation. Using the variables available, the distance of the new observation to the center of each class represented in the training database is calculated. The minimum distance is determined, and with this, the class to which the new observation will be associated. With the r observations belonging to that class in the training data matrix, a new sub-matrix $N \in \mathbb{R}^{r\times n}$ of a small dimension is formed, and it will be used to perform the imputation process. This simplifies the computational complexity of that process.

2. The new observation with missing data is added as the last row of the matrix N by forming the matrix $N' \in \mathbb{R}^{(r+1)\times n}$.

3. As it was previously presented, the imputation process is performed by using matrix N'.

4. Once all the missing values have been estimated, the observation is classified using the diagnostic system classification tool.

Remark: It is possible that the diagnostic tool classifies the observation with estimated data in a class different from the one used for the imputation. This could be explained by the degree of overlapping among the classes in the observation space.

2.4 Singular value decomposition

The imputation method based on the Singular Value Decomposition (SVD) was proposed in Troyanskaya *et al*. (2001), and it has been used effectively to impute missing data (Wang *et al*. 2014). As supervised tool, it needs previous offline training using the training database that has been built. SVD is a matrix factorization which allows determining the singular values of any matrix.

Consider the complex matrix $A \in \mathbb{C}^{m\times n}$ and its decomposition into singular values:

$$A = U\Sigma V^T \tag{1}$$

where $U \in \mathbb{C}^{m\times m}$ is an orthogonal matrix, $V \in \mathbb{C}^{n\times n}$ is an orthonormal matrix and $\Sigma \in \mathbb{C}^{m\times n}$ is a diagonal matrix whose non-zero elements in the main diagonal are the singular values σ_i ordered decreasingly. The column vectors of U are known as singular vectors by the left of A, and the column vectors of V as singular vectors by the right of A.

The imputation based on singular value decomposition has been used successfully to estimate the values in DNA microarrays (Troyanskaya *et al*. 2001) and in genetic interaction data (Wang *et al*. 2014). It is characterized by the speed of execution (Smith *et al*. 2015).

Like the SRMI algorithm, the SVD algorithm needs a data matrix to perform the imputation and for this, the training database will be used.

The procedure to make the online imputation using SVD algorithm for observations with missing data is the following:

1. Idem to step 1 using the SRMI algorithm. In this step, the matrix $N \in \mathbb{R}^{r\times n}$ is obtained.

2. Idem to step 2 using SRMI algorithm. In this step, the matrix $N' \in \mathbb{R}^{(r+1)\times n}$ is obtained.

3. The values of the missing variables in the last row of the matrix N' are filled using the average of the values of the respective column in order to complete the initial matrix.

4. The Expectation-Maximization (EM) algorithm (Dempster *et al.* 1977) is used to find the estimated values and the Singular Value Decomposition is applied for the new obtained matrix. This procedure is also iterative and it is repeated until the total change of the matrix is less than a threshold of 0.01 established empirically (Troyanskaya *et al.* 2001).

The matrix Σ contains in its main diagonal the singular values of the matrix $N' \in \mathbb{R}^{(r+1)\times n}$. Taking a certain number of significant singular values is sufficient to estimate the missing data (Troyanskaya *et al.* 2001). Therefore, how many and which singular values should be used to achieve the best estimation is determined by the experts with experiments. The latter will depend on the type of process, and the relationship between its variables.

2.5 Local least squares imputation

The local least squares imputation (LLSI) is a method proposed by Kim in 2004. It has had a great application in the estimation of data in DNA microarrays (Kim *et al.*, 2004a, b).

LLSI algorithm assumes $C \in \mathbb{R}^{m\times n}$ to denote the expression of the data associated to each one of the variables through an array with m observations and n variables.

In the matrix G presented in (2), a column $\boldsymbol{g}_i \in \mathbb{R}^{m\times 1}$ contains the values of the variable i$-th$ in m experiments.

$$G = (\boldsymbol{g}_1 \cdots \boldsymbol{g}_n) \in \mathbb{R}^{m\times 1} \tag{2}$$

A missing value in the $l-th$ localization of the variable $i-th$ is denoted by α (see Eq. 3)

$$G(l,i) = g_i(l) = \alpha \tag{3}$$

To simplify the description of the algorithm, the estimation of all missing values is described assuming a missing value in the first position of the first variable as it is presented in (4).

$$G(1,1) = g_1(1) = \alpha \tag{4}$$

The method of Local Least Squares Imputation has two steps (Kim *et al.* 2004b):

1. Selection of the k genes by using the Pearson correlation coefficients.

2. Regression and estimation.

2.5.1 Selection of genes or variables

To estimate a missing value in the first position $g_1(1)$ of $\boldsymbol{g}_1$ in $\in \mathbb{R}^{m\times n}$ by using the Pearson correlation coefficients, the k closest variables to the missing value are selected (Lei *et al.* 2017).

When there is a missing value in the first position of $\boldsymbol{g}_1$, the Pearson correlation coefficient between two vectors $\boldsymbol{g}'_1 = (g_{12}, \cdots, g_{1n})$ and $\boldsymbol{g}'_j = (g_{j2}, \cdots, g_{jn})$ is defined as Eq. (5):

$$r_{ij} = \frac{1}{n-1}\left(\frac{g_{1k} - \bar{g_1}}{\sigma_1}\right)\left(\frac{g_{jk} - \bar{g}_j}{\sigma_j}\right) \tag{5}$$

where $(\bar{\boldsymbol{g}}_1)$, $(\bar{\boldsymbol{g}}_j)$ represent the average values in $(\bar{\boldsymbol{g}}'_1)$, $(\bar{\boldsymbol{g}}'_j)$ and σ_1, σ_j represent the standard deviation of those values respectively.

2.5.2 Imputation by using LLSI

Based on the k genes closest to the missing variable, which have been selected using the Pearson correlation coefficients, the matrix $N \in \mathbb{R}^{k\times(n-1)}$ and the vectors $\boldsymbol{b} \in \mathbb{R}^{k\times 1}$ and $\boldsymbol{w} \in \mathbb{R}^{(n-1)\times 1}$ are formed. The k rows in the matrix N consist of the nearest k genes $\boldsymbol{g}_i^T \in \mathbb{R}^{1\times n}$ $1 \leq i \leq k$ with its first value removed. The elements of the vector $\boldsymbol{b}$ are formed by the first components of the k vectors $\boldsymbol{g}_i^T$ and the elements of the vector $\boldsymbol{w}$ are the $n-1$ elements of the vector $\boldsymbol{g}_1$. After forming the matrix N and the vectors $\boldsymbol{b}$ and $\boldsymbol{w}$, the least squares problem is formulated as:

$$\min_x \|N^T \boldsymbol{x} - \boldsymbol{w}\|_2 \tag{6}$$

The missing value α is estimated by the following linear combination:

$$\alpha = \boldsymbol{b}^T \boldsymbol{x} = \boldsymbol{b}^T (N^T)^+ \boldsymbol{w} \tag{7}$$

where $(N^T)^+$ represents the pseudo-inverse of N^T.

Procedure to impute online with LLSI:

1. Idem to step 1 using the SRMI and SVD algorithms. In this step, the matrix $N \in \mathbb{R}^{r\times n}$ is obtained.

2. Idem to step 2 using SRMI and SVD algorithms. In this step, the matrix $N' \in \mathbb{R}^{(r+1)\times n}$ is obtained.

3. The $k-$ genes of the matrix $N' \in \mathbb{R}^{(r+1)\times n}$ closest to the missing variables in the observation are selected using the Pearson correlation coefficients.

4. Based on the $k-$ genes closest to the missing variables selected in the previous step, the matrixes $N'' \in \mathbb{R}^{(n-q)\times(r+1)}$, $b \in \mathbb{R}^{(r+1)\times q}$ and the vector $w \in \mathbb{R}^{(n-q)\times 1}$ are formed.

5. The least squares problem is formulated by estimating the missing values using the following linear combination:

$$\alpha_{est} = \boldsymbol{b}^T (N''^T)^+ \boldsymbol{w} \tag{8}$$

2.6 k-Nearest neighbors

The algorithm k-Nearest Neighbors (k-NN) is a tool widely used in imputation tasks (Chen and Chao 2000, Eskelson *et al*. 2009) and classification (Bathia and Ashev 2010) due to its simplicity and good results. It is an unsupervised tool so, it is not necessary to perform a previous offline training.

In general form, the k-NN algorithm for imputation operates as follow: given m observations x_i and an observation y_1 with missing variables, the distance between y_1 and each one of the m observations x_i is calculated by using a measure of similarity, by eliminating from both observations used to calculate the distance, the missing variables in y_1. Later, the k observations x_i that are closer to y_1 are chosen. The values of each missing variable in y_1 are imputed by using the average of the sum of the same variable in the k observations x_i selected (García-Laencina *et al*. 2010, Luengo *et al*. 2012a, Askarian *et al*. 2016) as it is presented in (9):

$$\hat{y}_{lj} = \frac{\sum_{i=1}^{k} x_{ij}}{k} \tag{9}$$

where $\hat{y}_{lj}$ is the estimated value of the variable j in the observation y_1 using the k observations x_i selected.

The distance measure used in this paper was the Euclidean distance (García-Laencina *et al*., 2010; Askarian *et al*., 2016) which is presented in (10):

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{10}$$

There is no single criterion established to choose the value of k. For this reason, it is usually determined empirically. The algorithm k-NN is not iterative. It ends when all the distances are calculated, the nearest k observations are found and the values of the missing variables are imputed.

In order to use this algorithm online, the training database (set of observations x_i) is used. Also, the value of k should be defined by the experts. The new observation that arrives with missing data is identified as y_1.

Steps to impute online with k-NN:

1. From the training database $M \in \mathbb{R}^{m\times n}$ and the received observation with missing variables y_1, the information corresponding with the missing variables are removed.

2. The k observations of the training database closest to y_1 are determined using the Euclidean distance (See Eq. (10)).

3. The variables missing in y_1 are estimated using the Eq. (9).

2.7 Multilayer Perceptron Artificial Neural Network

For this paper, the Multi-layer Perceptron Artificial Neural Network (MLP) was chosen because it is one of the most used architectures due to its simplicity and excellent performance (Patan *et al*., 2008; Portillo *et al*., 2009; Ramírez *et al*., 2015).

Let denoting by a_i, $i = 1,2,3,...,n$ the inputs to the MLS artificial neural network; b_j, $j = 1,2,3,...,n$ the outputs of the hidden layer; c_k, $k = 1,2,3,...,n$ the outputs of the final layer, and t_k the target outputs. In addition, w_{ij} and θ_j are the weights and thresholds of the hidden layer, and w'_{kj} and θ'_k are the weights and thresholds of the output layer, respectively. The activation functions $f_1(t)$ and $f_2(t)$ correspond to the hidden layer and the output layer respectively. The operation of an MLP with a hidden layer is expressed as (del Brio and Molina, 2006):

$$\begin{aligned} c_k &= f_2\Big(\sum_j w'_{kj} b_j - \theta'_k\Big) \\ &= f_2\Big(\sum_j m w'_{kj} f_1(w_{ij} a_i - \theta_j) - \theta'_k\Big) \end{aligned} \tag{11}$$

The MLP Artificial Neural Network uses a supervised learning by error backpropagation (BP). Since the goal is to obtain an output of the neural network as close as possible to the desired output, the learning of the network is formulated as a problem of error minimization (del Brio and Molina, 2006).

$$E = \frac{1}{2}\sum_{\mu=1}^{p}\sum_{i=1}^{n}\left(t_i^{\mu} - c_k\right)^2 \tag{12}$$

The minimization is carried out by means of a descendent gradient algorithm, but there will be a gradient with respect to the weights of the output layer and another one with respect to those of the hidden layer (del Brio and Molina 2006).

$$w'_{kj}(n) = w'_{kj}(n-1) - \alpha\frac{\partial E}{\partial w'_{kj}} \tag{13}$$

$$w_{ij}(n) = w_{ij}(n-1) - \alpha\frac{\partial E}{\partial w_{ij}} \tag{14}$$

2.8 Tennessee Eastman (TE) process

The Tennessee Eastman (TE) test process is based on an industrial chemical process of the Eastman Chemical Company that was published in Downs and Vogel (1993), with the purpose of making available to the scientific community a reference problem to develop and evaluate different techniques of process control, optimization and monitoring and diagnostic methods.

The process consists of five major units: a reactor, a condenser, a recycle compressor, a separator, and a stripper; all interconnected as shown the flow diagram in Figure 3. The control objectives, suggested potential applications and features of the process simulation are described in more detail in Downs and Vogel (1993), and Chiang *et al.* (2001). The TE process contains 21 preprogrammed faults and one normal operating condition data set.

The data sets from the TE are generated along 48 hours of operation with the inclusion of faults after 8 simulation hours. Each historical data set contains 52 variables (41 measured variables, plus 11 manipulated variables) with a sampling time of 3 min. and Gaussian noise incorporated in all measurements. Concerning the study performed in this paper, only the 33 variables available online were considered, as shown in Table 1. A description of simulated faults is shown in Table 2. All data sets used in this paper can be downloaded from http://web.mit.edu/braatzgroup/TE_process.zip.

Table 1. Monitored variables in the Tennessee Eastman process.

No.	Variable	No.	Variable
1	A feed	18	Stripper temperature
2	D feed	19	Stripper steam flow
3	E feed	20	Compressor work
4	Total feed	21	Reactor cooling water outlet temperature
5	Recycle flow	22	Separator cooling water outlet temperature
6	Reactor feed rate	23	D feed flow valve
7	Reactor pressure	24	E feed flow valve
8	Reactor level	25	A feed flow valve
9	Reactor temperature	26	Total feed flow valve
10	Purge rate	27	Compressor recycle valve
11	Product separator temperature	28	Purge valve
12	Product separator level	29	Separator pot liquid product flow rate
13	Product separator pressure	30	Stripper liquid product flow valve
14	Product separator underflow	31	Stripper steam valve
15	Stripper level	32	Reactor cooling water flow
16	Stripper pressure	33	Condenser cooling water flow
17	Stripper underflow		

Table 2. Monitored variables in the Tennessee Eastman process.

Fault	Description	Type
1	A/C feed ratio, B composition constant	Step
4	Reactor cooling water inlet temperature	Step
5	Condenser cooling water inlet temperature	Step
11	Reactor cooling water inlet temperature	Random variation

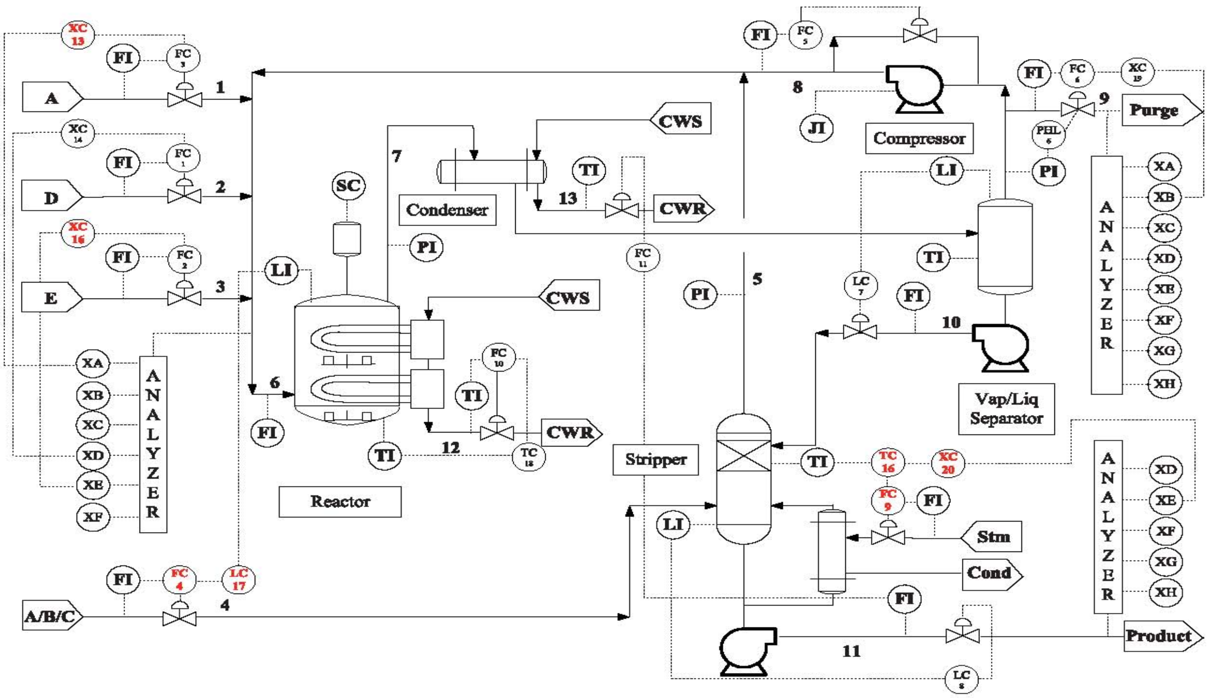

Fig. 3. Flow diagram of the Tennessee Eastman benchmark process.

To show the principal aim of this paper five states of the TEP were selected. Four of them represent fault states and the other represents the normal operation condition (NOC). The selected faults are shown in Table 2. These faults were chosen because they affect several parts of the process. (Quiñones-Grueiro *et al*. 2014).

3 Experiments and results

In this paper four experiments were designed:

1. Detection and classification without missing data.
2. Imputation of missing data randomly for 1, 2, 3, 4, 5 and 6 missing variables per observation which represents up to 18% of missing variables per observation. It was considered that a greater percent of missing variables represent a fault in the data acquisition system.
3. Detection and classification with imputed data.
4. Imputation, detection and classification of the data with 10, 20 and 30 percent of missing information with the mechanism missing completely at random (MCAR) in the database.

All experiments were carried out in a computer with the following characteristics: Intel® Core^{TM} i7-3537U 2.00 GHz processor, 8 Gb of RAM.

The training database used in the experiments had 300 observations by state to be diagnosed (normal operation condition and four fault states), i.e. a total of 1500 observations of 33 variables each one.

The datasets used to prove the procedure proposed in this paper had 1000 observations each one by state to be diagnosed for a total of 5000 observations. For the second and third experiments, six datasets were created, i.e. one for each number of missing variables per observation.

For the fourth experiment, were created three datasets of 5000 observations each one with the 10%, 20% and 30% of missing variables respectively, which satisfies the MCAR pattern.

The parameters used in the SRMI algorithm were $\beta = 0.0001$, and $Iter_{max} = 30$. The parameters used in the case of the k-NN algorithm, were the Euclidean distance and $k = 3$. In the case of the Singular Values Decomposition algorithm, after some experiments was determined to use 30 singular values to develop the experiments. A greater number of singular values do not improve the results.

In the case of the MLP neural network, it was used the following configuration: 33 neurons in the input layer (one per variable of the TEP), ten neurons in the hidden layer and four neurons in the output layer. As activation function in the hidden layer, it was used the sigmoidal function and the lineal function was selected to the activation function in the output layer.

3.1 Results of experiment 1: Detection and classification without missing data

The objective of this experiment is to obtain the performance in detection and classification of the fault diagnosis system without missing variables in the observations. Later, these results are used to analyze the impact caused in the performance of the fault diagnosis system when the imputation of missing variables is made using the different tools used in the paper.

Table 3 presents the results of detection and classification of the observations without missing variables using the MLP Neural Network.

Table 3. Results in detection and classification of the MLP without missing data.

	Detection Rate (%)	Classification Rate (%)
MLP	96.25	94.83

3.2 Results of experiment 2: Analysis of the imputation errors

The experiment 2 had as main objective the analysis of the performance of the algorithms in the imputation process. The imputation error was the indicator to evaluate the performance of each algorithm.

In addition, it was also analyzed the performance of the algorithms for 1, 2, 3, 4, 5 and 6 variables missed per observation with the aim of evaluating the affectation in the performance in the diagnostic system due to the imputation of different numbers of missing values by observation.

The data eliminated in each observation in order to create the missed variables were selected randomly but as the number of missing variables per observation is established, the pattern could be classified inside the type Missing at Random (MAR).

The errors were calculated taking into account the difference between the original values and the estimated values for the missed variables. The performances obtained by the different imputation algorithms are shown in Tables 4, 5, 6 and 7 where $\overline{Error}$ represents the average error in %, σ represents the standard deviation and $\bar{t_1}$ represents the average time in seconds used in the imputation process.

The Friedman's nonparametric statistical test (Luengo *et al.* 2009) was applied to determine if the number of missing variables by observation affected the performance of each imputation algorithm taking into account the estimation error.

It was found that there was no significant difference in the performance obtained from each algorithm regardless of the number of missing variables. Then, the imputation error of each algorithm is not affected by the number of missing variables. Furthermore, the Friedman nonparametric statistical test was applied to compare the results of the performances of each algorithm. It was determined that at least one of them had a behavior significantly different from the others.

The Wilcoxon non-parametric statistical test (Luengo *et al.* 2009) was applied and the result showed that the LLSI algorithm had the best behavior in the imputation process.

The results show that in all cases, the imputation times are small compared with the time constants and the sampling time of the SCADA systems of the most processes in the industrial plants such as chemical, pharmaceutical, food processing, just to mention some examples. The algorithm with the smallest imputation times was the k-NN.

3.3 Results of experiment 3: Detection and classification with imputed values

With the missing values imputed by each algorithm, the detection and classification process was developed with the aim of evaluating the affectations that the imputation process produces in the detection and correct classification of the faults and the normal operating state.

Table 4. Imputation results using SRMI algorithm in experiment 2.

Missing Variables	$\overline{Error}(\%)$	σ	$\bar{t_i}seg$
1	3	0,1373	0,053
2	3,35	0,1859	0,057
3	3,33	0,1906	0,058
4	3,59	0,2355	0,060
5	3,61	0,2490	0,067
6	3,55	0,2378	0,076

Table 5. Imputation results using SVD algorithm in experiment 2.

Missing Variables	$\overline{Error}(\%)$	σ	$\bar{t}_i seg$
1	3,24	0,0969	0,1271
2	2,60	0,1162	0,1345
3	2,75	0,0810	0,1317
4	2,62	0,0799	0,1787
5	2,48	0,0785	0,2310
6	2,51	0,0786	0,2393

Table 6. Imputation results using LLSI algorithm in experiment 2.

Missing Variables	$\overline{Error}(\%)$	σ	$\bar{t}_i seg$
1	1,85	0,0621	0,4483
2	1,28	0,0352	0,4314
3	1,16	0,0422	0,4813
4	1,52	0,0455	0,5070
5	1,22	0,0402	0,5291
6	1,46	0,0426	0,5383

Table 7. Imputation results using k-NN algorithm in experiment 2.

Missing Variables	$\overline{Error}(\%)$	σ	$\bar{t}_i seg$
1	2,78	0,0545	0,0033
2	2,51	0,0488	0,0035
3	2,80	0,0950	0,0034
4	2,75	0,0862	0,0033
5	2,84	0,0889	0,0032
6	2,95	0,1035	0,0033

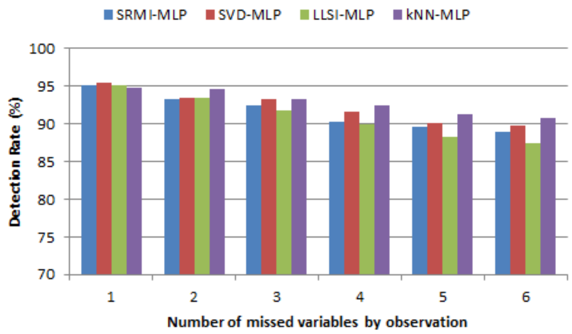

Fig. 4. Results of the detection rate with data imputed in experiment 3.

It is necessary to remark that detection can be seen as a binary classification where the observation is classified into two major groups: fault of non-fault (normal operation condition).

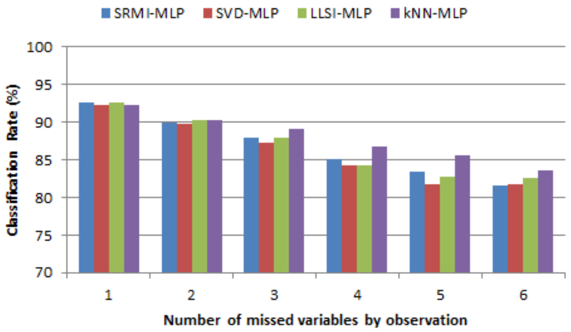

Fig. 5. Results of the classification rate with data imputed in experiment 3.

3.3.1 Detection stage

In Figure 4 are presented the percentage of detection rates (DR) by using the artificial neural network MLP, for 1, 2, 3, 4, 5 and 6 missed variables per observation.

The main conclusion obtained from these results is that the detection rate decreases when the number of missing variables in the observations increases.

However, the percentages of success in the detection for 6 missing variables per observation guarantee an acceptable level of effectiveness for the diagnostic system of more than 87% in the detection rate for the four imputation algorithms.

3.3.2 Classification stage

The influence of the number of missing variables per observation in the classification error after the online imputation is presented in Figure 5.

The classification process was developed by using the MLP artificial neural network. As it was expected, the Fig. 5 shows a downward trend behavior in the classification rate of the observations with imputed variables when the number of them increases.

However, similar to the detection stage, the success classification rates reflect the possibility of maintaining an acceptable performance of the diagnostic system in the classification process higher than 81%.

The Friedman nonparametric statistical test was applied to compare the results between the performances of each algorithm in detection and classification. No significant differences were found in the performance obtained by each algorithm

3.4 Results of experiment 4: Imputation, detection and classification of data for 10, 20, 30 percent of the random missing data

The main objective of the experiment is to analyze the effects of the imputation process on the detection and classification process when there are high percentages of missing variables with the pattern Missing Completely At Random (MCAR) globally.

In this experiment, datasets with 10, 20 and 30 percent of missing information globally were used, and the missing variables occur randomly (MCAR).

The imputation and classification processes were developed with the proposed procedure, and by using the imputation tools presented in section 2.

3.4.1 Imputation stage

The performance of the imputation algorithms for 10, 20, and 30 percent of the missing data globally are shown in Tables 8, 9, 10 and 11 where $\overline{Error}$ represents the average error in %, σ represents the standard deviation and $\bar{t_1}$ represents the average time in seconds used in the imputation process.

3.4.2 Detection stage

In Figure 6 are presented the results of the detection rates (DR) using the MLP neural network for the different percentages of global missing information for each imputation algorithm.

Table 8. Imputation results using SRMI algorithm in experiment 4.

Missing data (%)	$\overline{Error}(\%)$	σ	$\bar{t}_i seg$
10	2,59	0,1256	0,0491
20	2,63	0,1225	0,0525
30	2,04	0,0789	0,0541

Table 9. Imputation results using SVD algorithm in experiment 4.

Missing data (%)	$\overline{Error}(\%)$	σ	$\bar{t}_i seg$
10	3,29	0,0519	0,1328
20	3,08	0,0471	0,1437
30	3,25	0,1047	0,1826

Table 10. Imputation results using LLSI algorithm in experiment 4.

Missing data (%)	$\overline{Error}(\%)$	σ	$\bar{t}_i seg$
10	2,05	0,0847	0,1893
20	2,23	0,0808	0,2244
30	1,17	0,0588	0,2526

Table 11. Imputation results using kNN algorithm in experiment 4.

Missing data (%)	$\overline{Error}(\%)$	σ	$\bar{t}_i seg$
10	2,87	0,0554	0,0035
20	2,90	0,0737	0,0035
30	3,08	0,1167	0,0034

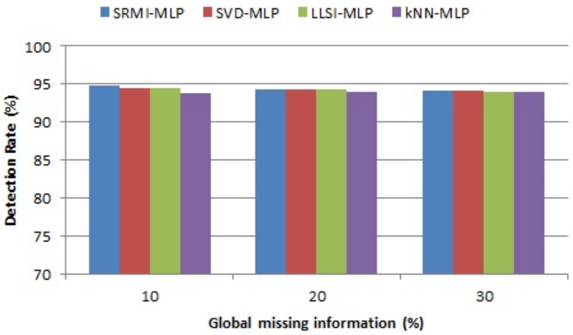

Fig. 6. Results of the detection rate with data imputed in experiment 4.

Figure 6 shows a small decrease on the detection rate when the percentage of global missing information increases. However, this decrease is not really significant.

The Friedman nonparametric statistical test was applied and no significant differences were found in the performance of the four variants imputation algorithm – MLP neural network.

Table 12 shows the average percent of the detection rate for the datasets obtained with each imputation algorithm taking into account the different percentages of the global missing information.

It is very interesting to note the small difference between these performances in the detection process and the result displayed in Table 1 when the same process was developed with no missing data.

Table 12. Average percent of the detection rate for each variant Imputation Algorithm – MLP neural network.

Imp.Alg-MLP	$\overline{DR}$
SMRI-MLP	94.19
SVD-MLP	94.08
LLSI-MLP	93.92
kNN-MLP	93.61

In addition, it is necessary to address the significant differences between these results and those presented in Figure 3 for the worst case (six missing variables per observation).

3.4.3 Classification stage

Figure 7 shows the results for the classification rates (CR) using the MLP neural network after the imputation process for different percentages of global missing information for each imputation algorithm.

In Fig. 7, it can be seen that the performances of each algorithm are not significantly affected by the increase in the percentage of global missing information. The application of nonparametric Friedman statistical test to compare the results between the four variants imputation algorithm – MLP neural network for classification also showed that there was no significant differences between these variants.

Table 13 shows the average percent of the classification rate for the datasets obtained with each imputation algorithm taking into account the different percentages of the global missing information. Analyzing the results displayed in Table 13, it can be said that there are not show high differences compared with respect the results presented in Table 1 when the classification process was developed with no missing data.

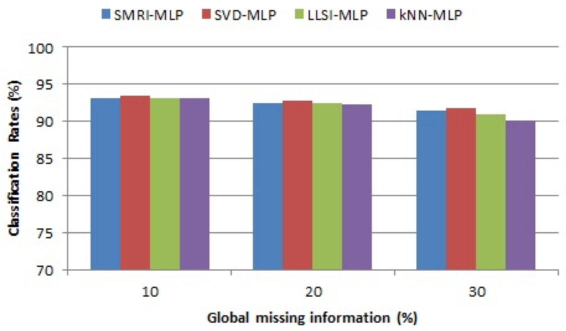

Fig. 7. Results of the classification stage with data imputed in experiment 4.

Table 13. Average percent of the classification rate for each variant Imputation Algorithm – MLP neural network.

Imp.Alg-MLP	$\overline{CR}$
SMRI-MLP	92.37
SVD-MLP	92.70
LLSI-MLP	92.17
kNN-MLP	91.88

However, there are significant differences between these results and those presented in Figure 4 for the worst case (six missing variables per observation) that which represents 18% of the missing variables.

The principal difference between the established conditions for the experiments 3 and 4 was the pattern of the missing variables. This evidence that a Missing at Random pattern (MAR) affects the prediction capacity of the algorithms analyzed when there are fix number of missing variables per observation. In general, from de practical point of view, however, it can be assumed the Missing Completely At Random (MCAR) pattern for the missing data in the industry. The Friedman's non-parametric statistical test was applied with two objectives. First, no significant differences were found in the performance of each algorithm with respect to the missing information percent globally. Second, no significant differences were found between the performances of the imputation algorithms. The imputation times had a similar behavior to experiment 2.

Conclusions

In this paper a new procedure has been presented for the online imputation of missing variables in the observations obtained by the data acquisition systems in industrial processes. In addition, the procedure ensures the immediate classification by the fault diagnosis system of all observations received which is a fundamental element to obtain a fast detection of faults.

The proposals found in the scientific and technical literature need the accumulation of a group of data to be able to make the imputation which decreases the effectiveness of the diagnostic systems in the fast detection of faults that appear in this period of time.

The paper also presents the results of the proposed procedure using the SRMI, SVD, LLSI and k-NN algorithms, which show very well results in the scientific literature in the imputation of large databases. The very satisfactory results obtained show the feasibility of the proposal.

In all experiments, the time required to carry out the online imputation of each observation obtained by the SCADA system is very small compared with the typical sampling times of the large chemical plants. Therefore, the online imputation works in favour of a fast detection and identification of a fault by the fault diagnostic system. For future work, it is recommended to analyze other computational intelligence tools that can improve the results in the imputation-detection-classification system.

Acknowledgements

The authors acknowledge the financial support provided by FAPERJ, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro; CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico; CAPES, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, research supporting agencies from Brazil; UERJ, Universidade do Estado do Rio de Janeiro and CUJAE, Universidad Tecnológica de La Habana José Antonio Echeverría.

References

Askarian, M., Escudero, G., Graellsc, M., Zarghamia, R., Jalali-Farahania, F. and School, N.M. (2016). Fault diagnosis of chemical processes with incomplete observations: a comparative study. *Computers and Chemical Engineering 84*, 104-116.

Bathia, N. and Ashev, V. (2010). Survey of nearest neighbor techniques. *Journal of Computer Science and Information Security 8*, 302-305.

Castelló, R.C., Puig, V. and Blesa, J. (2016). Introduction to model based fault diagnosis using project based learning. *RIAI – Revista Iberoamericana de Automática e Informática Industrial 13*, 186-195.

Chen, J. and Chao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics 16*, 113-131.

Chiang, L.H., Braatz, R.D. and Rusell, E.L. (2001). *Fault Detection and Diagnosis in Industrial Systems*. Springer Verlag.

Del Brío, B.M. and Molina, A.S. (2006). *Redes Neuronales y Sistemas Borrosos*. Ra-Ma S.S. Editorial y Publicaciones.

Dempster, A. P., Laird, M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological) 39*, 1-38.

Downs, J. and Vogel, E. (1993). A plant wide industrial process problem control. *Computers and Chemical Engineering 17*, 245-255.

Eskelson, B. Temesgen, H. Lemay, V., Barret, T., Crookston, N. and Hudak, A. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. *Scandinavian Journal of Forest Research 24*, 235-246.

García- Laencina, P.J., Sancho-Gómez, J.L. and Figueiras-Vidal, A.R. (2010). Pattern classification with missing data: a review. *Neural Computing and Applications 19*, 263-282.

García-Morales, J., Adam-Medina, M., Escobar, R.F., Astorga-Zaragoza, C.M. and García-Beltran, C.D. (2015). Multiple-sensor Fault Diagnosis in a heat exchanger using sliding-mode observers based on super-twisting algorithm. *Revista Mexicana de Ingeniería Química 14*, 553-565.

Jerez, J., Molina, I., García-Laencina, P., Alba, E., Ribelles, N. and Martín, M. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine 50*, 105-115.

Kim, H., Golub, G.H. and Park., H. (2004a). Imputation of missing values in DNA microarray gene expression data. In *Computational Systems Bioinformatics Conference*, 2004. CSB 2004. Proceedings IEEE.

Kim, H., Golub, G.H. and Park, H. (2004b). Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics 21*, 187-198.

Lei, B., van der Heijden, F., Xu, G., Feng, M., Zou, Y. and de Ridder, D. (2007). *Classification, Parameter Estimation and State Estimation: and Engineering Approach Using MATLAB*. John Wiley & Sons.

Leonhardt, S. and Ayoubi, M. (1997). Methods of fault diagnosis. *Control Engineering Practice 5*, 683-692.

Li, D., Deogun, J., Spaulding, W. and Shuart, B. (2004). Towards missing data imputation: A study of fuzzy K-means clustering method. Volume 3066 of *Lecture Notes in Computer Science*, Springer, 573-579.

Little, J.A. and Rubin, B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, U.S.A., 2nd Edition.

Luengo, J., García, S. and Herrera, F. (2009). A study on the use of statistical tests for experimentation with neural networks: Analysis of parametric test conditions and non-parametric tests. *Expert Systems and Applications 36*, 7798-7808.

Luengo, J., García, S. and Herrera, F. (2012a). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems 32*, 77-108.

Luengo, J., García, S. and Herrera, F. (2012b). Missing data imputation for fuzzy rule-based classification systems. *Softcomputing 16*, 863-881.

Lyman, P. and Georgak, C. (1995). Plant wide control of the Tennessee Eastman Problem. *Computers and Chemical Engineering 19*, 321-331.

MacGregor, J. and Cinar A. (2012). Monitoring, fault diagnosis, fault-tolerant control and optimization: data driven methods. *Computers & Chemical Engineering 47*, 111-120.

Nelson, P.R., MacGregor, J.F. and Taylor, P.A. (2006). The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemometrics and Intelligent Laboratory Systems 80*, 1-12.

Prieto-Moreno, A., Llanes-Santiago, O., Bernal de Lázaro, J.M. and García Moreno, E. (2013). Evaluation and classification methods used in the fault diagnosis of industrial process. *IEEE Latin America Transactions 11*, 682-689.

Patan, K., Witczak, M. and Korbicz, J. (2008). Towards robustness in neural network based fault diagnosis. *International Journal of Applied Mathematics and Computer Science 18*, 443-454.

Portillo, E., Cabanes, I., Marcos, M. and Zubizarreta, A. (2009). Aplicación de redes neuronales en la detección de regímenes degradados en el proceso wedm. *RIAI- Revista Iberoamericana de Automática e Informática Industrial 6*, 39-50.

Quiñones-Grueiro, M., Prieto-Moreno, A. and Llanes-Santiago, O. (2014). A proposal to configure the Fast-ICA algorithm in the fault diagnosis of industrial systems. *Revista Ingeniería Electrónica Automática y Comunicaciones, RIELAC 35*, 73-89.

Raghunathan, T., Lepkowski, J., Hoewyk, J.V. and Olenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models: Survey Methodology. 27 Statistics Canada. *Catalogue No. 12A001 27*, 85-95.

Ramírez, L., Mejías, R. and Coello, M. (2015). Single imputation with multilayer perceptron and multilayer imputation combining multilayer perceptron and k-nearest neighbors for monotone patterns. *Applied Sof Computing 29*, 65-74.

Smith, P., Mandel, J. and Guedj, M. (2015). A comparison of six methods for missing data imputation. *Journal of Biometrics and Biostatistic 6*, 1-6.

Sovilj, D., Eirola, E., Miche, Y., Björk, K., Nian, R. and Akuso, A. (2016). Extreme learning for missing data using multiple imputations. *Neurocomputing 174, Part A*, 220-231.

Téllez-Anguiano, A. C., Astorga-Zaragoza, C.M., Escobar, R.F., Alcorta-García, E. and Juárez-Romero, D. (2016). Continuous-Discrete observer-based fault detection and isolation system for distillation columns using a binary mixture. *Revista Mexicana de Ingeniería Química 15*, 275-290.

Tronskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T. and Tibshirani, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics 7*, 520-525.

Venkatasubramanian, V. and Chan, K. (1989). A neural network methodology for process fault diagnosis. *AIChE Journal 35*, 1993-2002.

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S.N. and Yin, K. (2002). A review of process fault detection and diagnosis, Part III: Process history based methods. *Computers and Chemical Engineering 27*, 327-346.

Walczak, B. and Massart, D. (2001a). Dealing with missing data: part I. *Chemometrics and Intelligent Laboratory Systems 58*, 15-27.

Walczak, B. and Massart, D. (2001b). Dealing with missing data: part II. *Chemometrics and Intelligent Laboratory Systems 58*, 29-42.

Wang, Y., Wang, L., Yang, D. and Deng, M. (2014). Imputing missing values for genetic interaction data. *Methods 21*, 187-198.

Watanabe, K., Matsuura, I., Abe, M., Kubota, M. and Himmelblau, D. M. (1989). Incipient fault diagnosis processes via artificial neural networks. *AIChE Journal 35*, 1803-1812.

Zhang, Z., Zhu, J. and Pan, F. (2013). Fault detection and diagnosis for data incomplete industrial systems with new Bayesian network approach. *Systems Engineering and Electronics 24*, 500-511.