Classification of sags according to their origin based on the waveform similarity

J. Meléndez[1], X. Berjaga[1], S. Herraiz[1], V. Barrera[1], J.Sánchez[2] and M. Castro[2]

Abstract— **A statistical method for classification of sags according to their origin downstream or upstream from the recording point is proposed in this work. The goal is to obtain a statistical model using the sag waveforms useful to characterise one type of sags and to discriminate them from the other type. This model is built on the basis of multi-way principal component analysis an later used to project the available registers in a new space with lower dimension. Thus, a case base of diagnosed sags is built in the projection space. Finally classification is done by comparing new sags against the existing in the case base. Similarity is defined in the projection space using a combination of distances to recover the nearest neigbours to the new sag. Finally the method assigns the origin of the new sag according to the origin of their neighbours.**

Index Terms— **Fault location, voltage sag (dip), pattern classification, Power quality monitoring, Principal Component Analysis.**

I. Introduction

UTILITY companies have increased the number of power quality monitors installed in the distribution substations and are very interested in developing reliable methods to efficiently exploit the information contained in these registers. The goal in this work has been focused on the discrimination between sags originating in the transmission (HV) and distribution (MV) networks. With this aim, sags registered in three 25kV distribution substations have been used as case base. Additionally, the utility has provided information related to the origin, upstream (HV) or downstream (MV) from the transformer, of them.

Data mining principles can be applied to obtain the desired information and manage the huge volume of data contained in these registers efficiently. The basic principles of these strategies involve automatic classification, clustering, or pattern matching to recognize disturbances according to similarity criteria and associate them with the most plausible causes and origins. Researchers have classified sags according to their origins to assist utilities in locating faults. Determining whether sags have occurred in the distribution or transmission networks precedes the localization and mitigation stages [1]. Typical classification according to the origin consists in discriminating between transmission (or high voltage) and distribution (or medium voltage) origins. For this purpose, phase analysis and an unsupervised method were compared in [2] by extracting some temporal descriptors from the RMS representation of sags and using a Learning Algorithm for Multivariate Data Analysis (LAMDA). Recent research has also identified similarities among sags using the variability in the information contained in the waveform in statistical analyses based on Principal Component Analysis (PCA), which allows dimensionality reduction before similarity criteria are applied to sags, assigning them to different classes. In [2] sags are categorized into three classes using certain features run through a fuzzy system. A more recent method for locating the origin of a voltage sags in a power distribution system using the polarity of the real current component relative to the monitoring point has been introduced in [1].

Other approaches proposed for classifying voltage sags are related to defining and describing sag types with regard to their general three-phase nature. With these approaches, sags can be divided up according to the number of sagged phases and the presence of asymmetries using either the magnitude or the angle between phasors to identify sag typologies. Other strategies are related to evaluate both the minimum magnitude and the total duration of sags. This group of classifiers eliminates any possibility of classifying sags using their three-phase nature. With this approach, the sags are reduced to one simple square shape sag, which is represented by the minimum of all RMS phase voltages during the sag and the total duration of the sag in all sagged phases. Other sag, classification strategies take advantage of attributes extracted from the RMS waveform to represent sags in a feature space where classification algorithms are applied ([2]-[4]). In this paper we present new results obtained with a classification method based on the definition of similarity criterion in the projection space obtained when the Principal Component Analysis (PCA) is applied to sags waveforms [14]. The method proposes the exploitation of the whole information contained in the voltage and current waveforms instead of

[1] This research has been made possible by the interest and participation of ENDESA DISTRIBUCION and its Power Quality Department. It has also been supported by the research project DPI2006-09370 funded by the Spanish Government.

The authors (1) are with the eXiT Group in the Institute of Informatics and Applications of the University of Girona, Spain, Girona, Campus Montilivi, 17071, e-mail: quimmel@eia.udg.es, xberjaga@eia.udg.edu, sherraiz@eia.udg.es, vbarrera@eia.udg.edu

Authors (2) are with the PQ Department of Endesa Distribución, Barcelona, (Spain), e-mail jslosada@fecsa.es, MCastro@enher.es

The eXiT is part of Automation Engineering and Distributed Systems (AEDS) research group, awarded with a consolidated distinction (SGR-00296) for the 2005-2008 period in the Consolidated Research Group (SGR) project of the Generalitat de Catalunya.

978-1-4244-2218-0/08/$25.00 ©2008 IEEE.

obtaining features from them. With this goal PCA is used to cope with the dimensionality problem at the same time that it provides statistical indices to assess the quality of projected data in terms of adequacy to the projection model.

The paper is organised in four additional sections. In the next one we introduce the classification method. Next, in section three the validation procedure is explained. The fourth section is devoted to analyse the results obtained with registers gathered in three distribution substations. And finally, section five presents the conclusions extracted from this work and further work that can be performed.

II. Methodology

In this section we present the methodology used in this work that can be divided in two main tasks: model creation and model exploitation. The model creation consists of the data pre-treatment step and basic concepts on Principal Component Analysis (PCA) and its extension Multi-way PCA (MPCA) to deal with the waveform registers. On the other hand the model exploitation step implies the use of this model for the classification purposes ad involves the projection of data and definition of the similarity criteria.

A. Principal Component Analysis (PCA)

PCA is based on a Singular Value Decomposition (SVD) of the covariance matrix of a dataset, $X \in R^{m \times n}$ [13]. Rows (m) and columns (n) of X correspond to samples and measurements respectively. That is, each row contains the six auto scaled (zero mean and unit variance) RMS waveforms of voltages and currents. Thus, each row contains the whole information of a sag waveform.

According to PCA basis the dataset, X, can be expressed as a linear combination of r new variables, t_i assuming an error E [17]:

$$X = \sum_{i=1}^{r} t_i \times p_i^T + E \qquad (1)$$

Where t_i and p_i are named *scores* and *loading* vectors respectively and are computed to reflect relevant relation amongst observations (sags) (t_i) and variables (voltages and currents at every time instant) (p_i).

PCA assumes that the *loadings* with bigger eigenvalues are the best ones for expressing the data upon based on the maximum variance criteria. According to this condition, we keep those *loadings* that capture the majority of the variation and throw away others as meaningless variation caused by noise, E (error, also known as residual matrix). Thus, the first r principal components, instead the n original variables, build up a new space/model with a lower dimensionality than the original one. Projection of the data to the i-th axis in this new space can be done using the following linear transformation:

$$t_i = Xp_i, i = 1,..,r \qquad (2)$$

B. Multi-way Principal Component Analysis (MPCA)

Since we are dealing with the sag waveforms, the data structure we are working is a 3 dimensional matrix as shown in Fig. 1.

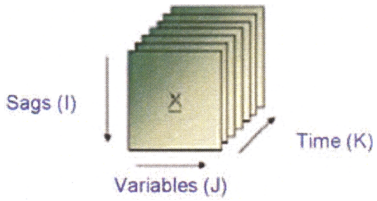

Fig. 1. Data matrix

As expressed in the previous section, PCA is a 2 dimensional procedure, so it is needed to unfold the matrix. From the six feasible unfolding directions, only 2 are meaningful for monitoring: unfold in the sag direction and unfold in the variable direction. When the monitoring procedure is used once the process is finished (the entire sag register) the best unfolding direction is the batch-wise [16]-[18]. The resulting data matrix is shown in Fig. 2 where variables x time represent the sequence of samples constituted by the three voltage and three current waveforms of a sag. In Fig. 3 is shown an example of a voltage sag unfolded this way.

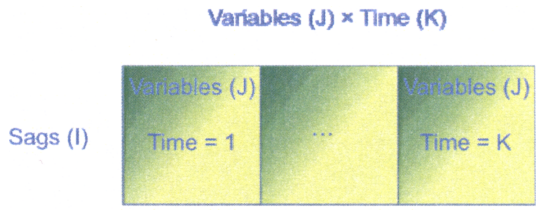

Fig. 2. Unfolding in the sag direction

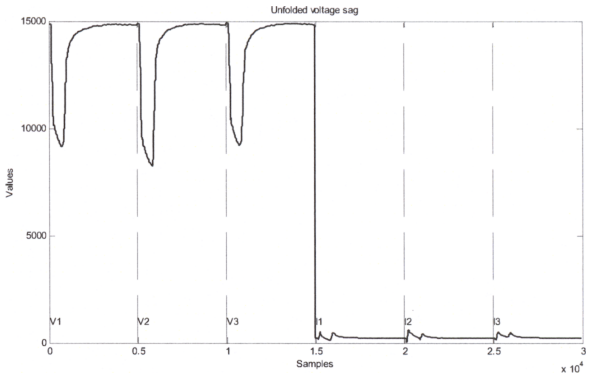

Fig. 3. Example of an unfolded voltage sag

Since voltages and currents have different range of values, a normalization step is needed. From the several possible approaches [16], the one we used is the so-called auto scaling, which formulation is as follows:

$$X_n = \frac{X_u - \bar{x}}{\sigma} \qquad (3)$$

Where X_n will be referred as the normalized or auto scaled data, X_u is the sag-wise unfolded data, $\bar{x}$ is the mean of the data (obtained for each column) and σ is the standard deviation of the data (obtained for each column). In **Fig. 4** is

shown the results of applying the auto scaling procedure on the voltage sag of **Fig. 3**.

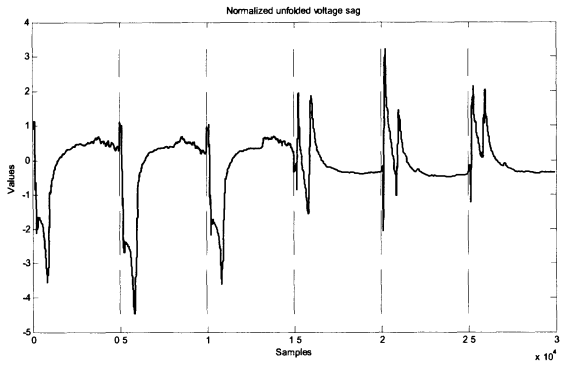

Fig. 4. Unfolded voltage sag after auto scaling

Once the data is unfolded and normalized, PCA can be applied directly. Since each row now contains an auto scaled sag, it can be decided the adequacy of them to the model using two statistics: Hotelling's T^2 and Q-residual [17]. T^2 is used to detect variations on the projection space, that is the hyper plane defined by the first r principal components that do not fit the model. This statistic is computed as follows:

$$T^2 = \sum_{i=1}^{r} \frac{t_i^2}{S_{t_i}^2} \tag{4}$$

Where $S_{t_i}^2$ is the estimated variance of score t_i(the eigenvalue associated to the p_i loading or eigenvector).

On the other hand, Q-residual or Squared Prediction Error (SPE) is the projection error obtained due to the use of only the first r principal components. This statistic is defined as:

$$Q_x = \sum_{i=1}^{n} \left(x_j - \hat{x}_j\right)^2 \tag{5}$$

Where $\hat{x}_j$ is computed with the PCA model using (1) and represents the reconstruction of the original variable using the model, n is the number of variables (in our case the total amount of samples contained in the three voltage and three current waveforms).

C. *Waveform similarity in the PCA space*

The voltage sags will be grouped in a single data structure in order to compare them, which will be called *Case Base*. Each individual (*case*) is the representation of a voltage sag and is formed by:

- Voltage and current RMS values of the sag.
- The r first principal components $(t_1,..,t_r)$.
- The Q-residual and T^2 control statistics.
- The name that contained the original sag values.
- Date and time when the sag was originated.
- Substation and transformer where was registered.

Once the structure to store sags has been defined, a similarity criterion is needed to compare sags. Since PCA is an optimal representation in terms of captured variance and it defines an orthogonal space we propose to define a distance in the projection space. Thus, it is easier to compute the distance in this space with reduced dimension and at same time the most similar cases in the original space are the most similar cases in the reduced space.

A two steps distance over the principal component space has been used based on this concept:

- First, the k_1 nearest neighbours of a sag are retrieved base on the absolute difference of the Q-residual, as shown in Eq. (6) in order to retrieve a small subset of cases fitting the a similar model structure.

$$d_Q = |Q_a - Q_b| \tag{6}$$

- In a second step, a non-weighted Euclidean distance -Eq. (7)- is used to select the most similar cases among the reduced subset from the previous step. Similarity among cases is measured using the scores of cases as attributes:

$$d_t = \sqrt{\sum_{i=1}^{r} \left(t_{a,i} - t_{b,i}\right)^2} \tag{7}$$

Where $t_{a,i}$ and $t_{b,i}$ represent scores of the new case, a, and a stored one, b, in the case base.

As an example of this similarity criterion, **Fig. 5** presents the nearest neighbour (thin line and empty point) of a new sag (fat line and filled point) with a distance value of 0.015466. **Fig. 6** shows its relative situation on the PCA space.

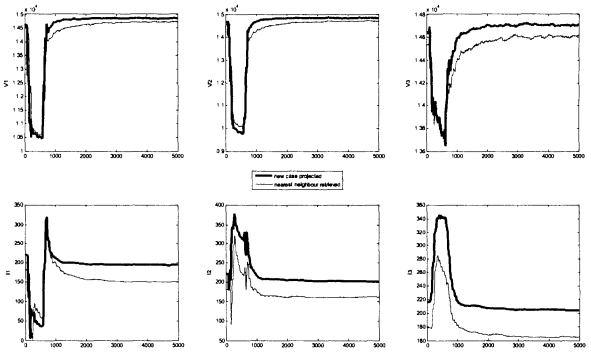

Fig. 5. Waveform comparison between a new case and its nearest neighbour

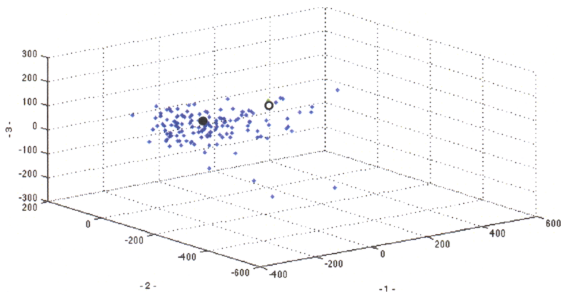

Fig. 6. Relative situation of the Nearest Neighbours in the PCA space

Finally, to determine the class of the new case, a decision threshold is used from the k_2 retrieved cases:

$$\frac{\sum_{o=1}^{nm} dist\_class\_ref_o}{\sum_{q=1}^{k_2} dist\_retr_q} > Th \qquad (8)$$

Where nm are all the cases retrieved from a reference class, $dist\_class\_ref_o$ refers to the distance value of the new case to its o-th nearest neighbour from the reference class, $dist\_retr_q$ is the retrieved distance of the new case to its q-th nearest neighbour and Th is the decision threshold to accept a new case of a reference class. Varying this threshold value, the Receiver Operating Characteristic (ROC) curve can be computed.

III. Results Evaluation

In this section we will present the methodology to evaluate the performance of the classification in an objective way.

A. Stratified l-Fold Cross Validation

In l-Fold Cross Validation, the available data is divided into l folders containing approximately the same number of examples. The stratified version of this technique takes into account the several ratios among classes present in the original set. Once the data is divided, one of the l folds of samples is retained for validation of the model formed by the remaining l-1 data fold. This process is repeated l times (once for each fold) [8]. **Fig. 7** presents this methodology in a graphical way. In this work, it has been applied a 4-Cross Validation (l=4).

B. Confusion Matrix and performance indices

In order to test the correct classification of the n-Fold Cross Validation, the *confusion matrix* is used. A confusion matrix is a form of contingency table showing the differences between the true and predicted classes for a set of labelled examples, as is shown in TABLE I [9].

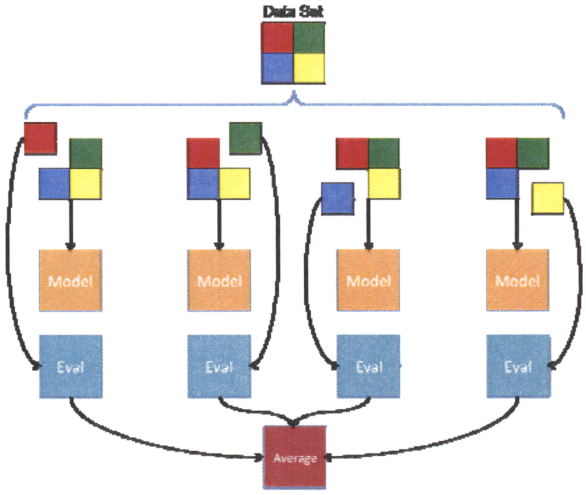

Fig. 7: l-Fold Cross Validation procedure

TABLE I
CONFUSION MATRIX

		Real Class	
		Ref	**No Ref.**
Predicted Class	**Ref**	*TP*	*FP*
	No ref.	*FN*	*TN*

Where *TP* stands for true positive (cases correctly predicted as the reference class), *TN* stands for true negative (cases correctly classified as non reference class), *FP* for false positive (cases classified as the reference class with its real class being of the non reference class) and *FN* for false negative (cases classified as a non reference class with its real class being of the reference class). The evaluation of these indices allows computing several performance parameters of the classifier, such as:

$$Sensitivity(SEN) = \frac{TP}{TP+FN} \qquad (9)$$

$$Specificity(SPC) = \frac{TN}{TN+FP} \qquad (10)$$

$$Accuracy(ACC) = \frac{TN+TP}{TN+TP+FN+FP} \qquad (11)$$

$$Precion(PRE) = \frac{TP}{TP+TN+FP+FN} \qquad (12)$$

In this work special attention is put on Sensitivity and Specificity to compute the ROC curve as it is explained in the next subsection.

C. ROC curves and Area under ROC curves

The ROC curves represent in a single figure the measure of the classifier's performance based on the relation of $\Pr(TP)$ (sensitivity) and $\Pr(FP)$ (specificity) as a decision threshold is varied [9].

The ROC curve representation is a two-dimensional graph where the y axis represents sensitivity and x axis represents de False Positive Rate (FPR), or what is the same, 1 – specificity. Observe that the lower left point (0,0) represents a classifier that never classifies correctly the cases of the model. The upper left point (0,1) represents the perfect classifier (it never

misses to classify the cases of the model, and also determine correctly the cases that are not represented by the model) and the upper right point (1,1) represents a classifiers that always classifies correctly cases fitting the model, but always classifies incorrectly cases different from the model [12].

Because in some operating points sensitivity can be increased with a minor loses in specificity and in others this is not possible, a non-ambiguous possible comparison of performance can be achieved by computing the Area Under the ROC Curve (AUC). A simple way of computing this value is using the trapezoidal integration method described in [9].

IV. Results

In this section we present the data with which we have worked on (upstream and downstream voltage sag), data pre-process (RMS value computation, auto scaling), first analysis of the data (k-means with several values of k over the projection space), experimental setup (number of parts in which the case base will be split, pair of values for the retrieval distance) and numerical results (ROC curve, confusion matrix, AUC).

From the original set of voltage sags registered in 3 substations (140 upstream sags and 81 downstream sags) we have removed those voltage sags that no presented a pre-fault stage (Fig. 8) because the performance of MPCA method decreases when data is misaligned. In the work all sags starts after the second period registered.

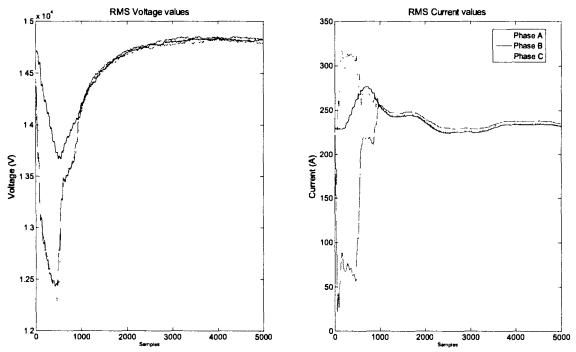

Fig. 8. Excluded sag example

For the RMS value computation, a Short Fourier Transform (SFT) in one cycle with a sliding window of one sample has been used. After that, the data is normalized using the auto scaling methodology. To build the model, only the upstream voltage sags have been used, what will cause that the downstream voltage sags projected onto them will have a large Q-residual statistic. This is exploited in the classification step. The number of principal components to retain has been fixed to 10 in order to explain at least a 95 % of the original variance of data. A 4-Stratified Cross Validation procedure has been used to build several Case Bases to test with all remaining voltage sags (upstream and downstream). Finally, to test the relation between neighbourhood size and performance, several pairs of k_1 and k_2 values has been tested. Numerical results obtained with these considerations are shown in TABLE II. The performance when the decision threshold is changed is represented in the Fig. 9 as a ROC curve of the tested classifiers zooming in the nearest area to the point (0,1).

TABLE II
CLASSIFICATION RESULTS

(k_1,k_2)	TP	FN	FP	TN	ACC	SEN	PRE	SPC	AUC
(15,1)	99	1	1	72	0.988	0.99	0.990	0.986	0.988
(15,3)	100	3	0	70	0.982	1	0.971	0.958	0.999
(15,5)	100	4	0	69	0.976	1	0.962	0.944	0.992
(10,1)	99	0	1	73	0.994	0.99	1	1	0.995
(10,3)	99	3	1	70	0.976	0.99	0.971	0.958	0.994
(10,5)	99	4	1	69	0.971	0.99	0.962	0.944	0.991

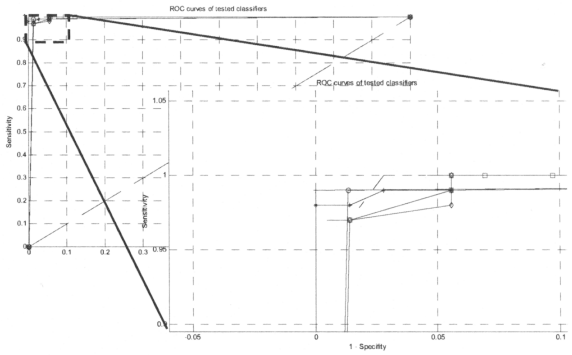

Fig. 9. ROC curves of the tested classifiers

As can be seen, the best classifier, what is the one with the highest AUC value (k_1=15,k_2=3), does not have the highest values of specificity with the default decision threshold (*Th* = 0.55), but it is the most regular classifier. Taking this classifier as a reference, if we analyze the four wrong classified sags (Fig. 10), it can be seen that 2 of then can be considered to be incorrectly labelled because its shape is the same as upstream voltage. But an accurate analysis concluded that those 2 voltage sags where an example of transformer energizing.

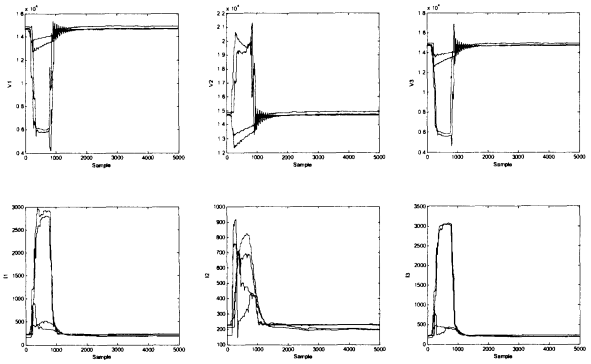

Fig. 10. The four incorrectly classified sags

V. CONCLUSIONS

Results obtained with the available data set have proved the utility of the presented methodology for relative location of voltage sags. The main advantage of this methodology is that similarity between voltage sags is performed using the voltage sag register without comparing directly its waveforms. As was stated in previously, a validation procedure of the original data sets has to be conducted in order to avoid wrong classifications due to the labelling. Another interesting point to work on is to extend to the classification to other available substation registers. Other distances are being studied, but early results state that performance of the classifier depends on the neighbourhood size than the distance criterion used.

VI. ACKNOWLEDGMENTS

The authors would like to thanks Magda Ruiz from the University of Girona for her assessment on the MPCA model and her suggestions.

VII. REFERENCES

[1] N. Hamzah, A. Mohamed and A. Hussain, "A new approach to locate the voltage sag source using real current component," Journal of Electric Power Systems Research, Vol. 72, pp. 113-123, 2004.

[2] J. Mora, D. Llanos, J. Melendez, J. Colomer, J. Sanchez and X. Corbella, "Classification of Sags Measured in a Distribution Substation based on Qualitative and Temporal Descriptors", in 17th Int. Conf. On Electricity Distribution, Barcelona, Spain, May 2003.

[3] M. Kezunovic and Y. Liao, "A new method for classification and characterization of voltage sags," Journal of Electric Power Systems Research, Vol. 52, No 1, pp 27-35, 2001.

[4] A. D. Gordon, *Classification*, Boca Raton, 1999.

[5] R. L. de Mantaras and E. Plaza, "Case-based reasoning: An overview," AI Communications, Vol. 10, No 1, pp 21-29.

[6] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," AI Communications, Vol. 7, No. 1, pp. 39-59, 1994.

[7] D. B. Leake, "Case-Based Reasoning: experiences, lessons and future direction," Press, 1996.

[8] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," Proceedings of 14th International Joint Conference on Artificial Intelligence, Vol. 2, No. 12, pp. 1137-1143.

[9] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," Pattern Recognition, Vol. 30, No. 7, pp. 1145-1159.

[10] K. Fukunaga, "Introduction to Statistical Pattern Recongition," San Diego, CA, Academic Press, 1990.

[11] C. W. Therrien, "Decision Estimation and Classification: An introduction to pattern recognition and related topics," Wiley, 1989.

[12] T. Fawcett, *Pattern Recognition*, Letters, pp. 861-874, 2006.

[13] B. M. Wise, N. B. Gallagher, S. Watts, D. D. White JR and G. G. Barna, "A comparison of PCA, Multiway PCA, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process," Journal of Chemometrics, Vol. 13, pp 379-396, 1999.

[14] J. Melendez, X. Berjaga, S. Herraiz, J. Sánchez and M. Castro, "Classification of Voltage Sags based on k-NN in the Principal Component Space," International Congress on Renewal Energies and Power Quality (ICREPQ), 2008.

[15] J. Camacho and J. Picó, "Multi-phase principal component analysis for batch processes modelling," Chemometrics and Intelligent Laboratory System, Vol. 81, pp. 127-136, 2006.

[16] M. Ruiz, K. Villez, G. Sin, J. Colomer and P. A. Vanrolleghem, "Influence of scaling and unfolding in PCA based monitoring of nutrient removing batch process," 6th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes, Sept., 2006.

[17] P. Nomikos and J. F. MacGregor, "Monitoring batch processes using multiway principal component analysis," AIChE, Vol. 40(3), pp. 1361-1375, 1994.

[18] D. Aguado, A. Ferrer, A. Seco and J. Ferrer, "Comparison of different predictive models for nutrient estimation in sequencing batch reactor for wastewater treatment," Chemometrics and Intelligent Laboratory System, Vol. 84, pp. 75-81, 2006.