Classification of Voltage Sags based on k-NN in the Principal Component Space

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

[1] Institut d'Informàtica i Aplicacions
eXiT., Universitat de Girona
Campus Montilivi, 17071 Girona (Spain)
Phone/Fax number:+0034 972 418391, e-mail: **quimmel@eia.udg.edu**, **xberjaga@eia.udg.edu**, **sherraiz@eia.udg.edu**

[2] Power Quality Department of Endesa Distribución, Barcelona (Spain)
Mailing address
Phone, fax, **jslosada@fecsa.es**, **MCastro@enher.es**

Abstract

A new method for the classification of sags is proposed. The goal is to deduce the origin of sags (upstream or downstream of the transformer) registered in distribution substations. The method is based on the existence of a case base of previous registers which origin is well known. This case base is used to infer the origin of new sags based on the retrieval of similar sags using a distance criterion. This distance computed in the principal component space is also used in the decision step to decide the origin of the new sag.

Key words

Fault location, voltage sag (dip), pattern classification, Power quality monitoring.

1. Introduction

THE utility companies have increased the number of power quality monitors installed in distribution substations and are very interested in developing reliable methods to efficiently exploit the information contained in these registers in order to automatically discriminate between sags originating in the transmission (HV) and distribution (MV) networks and to assess and diagnose them. With this aim, this work is focused on monitoring sags registered in 25kV distribution substations in order to assign their origin to the MV or the HV, i.e. upstream or downstream of the transformer. Data mining principles can be applied to obtain the desired information and manage the huge volume of data contained in these registers more 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 is reduced to one a 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]).

The paper is organised in five additional sections. In the next one the proposed method for classification in the principal component space is outlined. Next, in section

three and four the theoretical fundamentals of Multiway Principal Component Analysis and Case Based Reasoning are introduced. Section five describes the validation procedure used in this work. The sixth section is devoted to analyse the results obtained with registers gathered in three distribution substations. And finally, section seven presents the conclusions extracted from this work.

2. The proposed method

We introduce a new method to assist power quality monitoring for classification of electrical sags gathered in distribution substation in order to determine their origin. PCA is proposed to model datasets of sags which origin, in MV or HV, are known. A PCA model built with existing sags originated in HV is built and later used to identify new sags by projecting them against this model and evaluating similarity of new sags in this new space. The goal is to capture the relevant information of sags originated in HV useful to discriminate them from those originated in MV. Next we present the ways for preparing the data as well as the whole procedure in details.

The procedure steps could be broken down in two general stages: (i) case base preparing and construction in the principal component space, and (ii) model exploitation.

The tasks to perform the first stage include the creation of the PCA model and the projection of the existing registers (voltage and current waveforms) into the principal component space. This is performed with a training set of data which origin is known. Here, they are described in more details following the execution order:

- *RMS Value Computations:* Instantaneous RMS value for each variable (three voltages and three currents) is computed.
- *Autoscaling:* Since RMS magnitudes of voltages and currents are completely different autoscaling is a preprocessing stage needed to avoid overweighting of voltage variables towards currents. It results in zero-mean centered data with a unit variance.
- *Model Creation:* HV-PCA model is created using training subsets.
- *Projection:* HV and MV sags are projected into HV-MPCA model resulting in a case base in the principal components space.

The case base obtained in the new space presents an important reduction on the number of variables that will be taken into account. Moreover, this new variables are independent, which means that the Euclidean distance will reflect the differences between variables considering the correlation among them, as exposed in [4]. At the same time, the use of a projection operator that models one class of data produce a better separation of cases in the new space, also allowing the use of well known statistics as Q and T^2 in this space to evaluate similarity between new sags and those in the case base.

The second stage, the exploitation of this case base built in the projection space, is based on the similarity criteria between new sags and those in the case base previously diagnosed. Thus, the methodology to classify sags follows these steps:

- *Projection of new sags:* The same HV-PCA model is used to project new sags in the principal component space.
- *Retrieval of the k-NN:* A subset of *k* nearest neighbors is identified based on two criteria: Q and T^2 statistics. Firstly, a larger selection of candidates is made based on its adequacy to the model using the Q statistic. From it, the *k* best candidates are selected by comparing the T^2 statistic (distance to the center of the model).
- *Class Determination:* The class of the sag is determined by the comparison between a threshold and the relation of the distances of all the retrieved cases that belong to the model (its class is HV) and all the distances that have been obtained.

Fundamentals of these tasks are based on the principles of Case Based Reasoning briefly introduced in section 4.

3. PCA

PCA has been developed based on Singular Value Decomposition (SVD) of the covariance matrix of a dataset, $X \in R^{m \times n}$. Rows (m) and columns (n) of X correspond to samples and measurements respectively. That is, each row contains the six autoscaled RMS waveforms of voltages and currents. Thus, each row contains the whole information of a sag waveform.

The sample-covariance matrix is then computed as follows:

$$C = \frac{X^T X}{m-1} \tag{1}$$

Then, the dataset, *X*, can be expressed as a linear combination of *r* new variables, t_i assuming an error *E*:

$$X = \sum_{i=1}^{r} t_i . p_i^t + E \tag{2}$$

Where t_i and p_i are named scores and loading vectors respectively and are computed to reflect relevant relation amongst samples (t_i). While p_i highlights the correlation among variables and they correspond to eigenvectors of covariance matrix (C).

$$Cp_i = \lambda_i p_i \tag{3}$$

PCA assumes that the eigenvectors 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 eigenvectors which capture the majority of the variation and throw away others as meaningless variation caused by noise, E (error or residual matrix). Thus, the first r principal components 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:

$$t_i = Xp_i, i = 1,..,r \tag{4}$$

The lack of model fit can be measured using two statistical criteria named Q-residual and T^2. Thus, when PCA model is built using a dataset gathered during the same process conditions those indices are used to detect abnormal situations not included in the initial dataset, i.e. fault detection or alarm generation in process monitoring tasks. Multivariate control charts based on T^2 can be plotted as follows:

$$T^2 = \sum_{j=1}^{r} \frac{t_j^2}{S_{t_j}^2} \tag{5}$$

Where S_t^2 is the estimated variance of t_j. This control chart is used to detect variations in the hyperplane defined by the first r principal components not fitting the model. Other types of abnormalities not represented by the original dataset used to build the PCA model are those that project the new observation out of this hyperplane. Those types of events can be detected by computing the Q-statistic or Squared Prediction Error (SPE) of the residual for new observations. It is defined as

$$Q_x = \sum_{j=1}^{r} \left(x_j - \hat{x}_{j,new} \right) \tag{6}$$

Where $\hat{x}_{j,new}$ is computed from the reference PCA model. Normally, Q-statistic is much more sensitive than T^2 to data not following the same structure (relationship among variables in terms of covariance) as the original data set. It is due to the fact that Q is typically small and orthogonal to the hyperplane defined by t_j and consequently any minor change in the data structure will be observable. T^2 has a great variance and therefore requires a great change in the system characteristic for it to be detectable. Variation in T^2 does not imply a change in the correlation structure.

4. Case Based Reasoning fundamentals

Case-Based Reasoning (CBR) is a reasoning approach to problem solving capable of using the knowledge acquired by previous experiences [5]. The basic functions that all CBR present are known as the "4-Rs" [6], and can be organised in a cycle as it is depicted in Fig. 1:

1. RETRIEVE the most similar cases of the new case.
2. REUSE the information in these cases to solve the new problem.
3. REVISE the proposed solution.
4. RETAIN the new information of the new experience in order to solve new similar problems.

To solve a new problem, the most similar cases are retrieved from the experiences previously stored. The information contained in these retrieved cases is then reused to propose a possible solution. Once the solution is evaluated, the case is retained, if necessary, for further classifications.

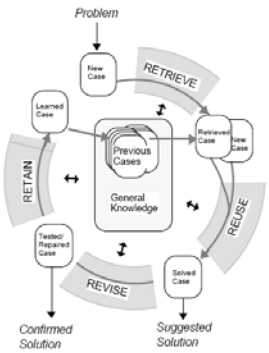

Fig. 1. The CBR cycle [6]

The past experiences are stored in a data structure representing problem (sags) and solutions (origin) called cases and the set of cases constitutes the Case Base. Those cases are the minimal representation of the problem and the functional solution [7].

In order to determine the most similar cases, it is needed to find the similarities between cases. This is done by applying a distance criterion. Local distances between attributes of a case are combined to define a global distance between cases. In this work a two steps Euclidean distance has been used in the Principal Component Space:

- First the k_1 nearest cases are retrieved using the Euclidean distance measured in the direction of Q statistic (See Eq. (7)) in order to select a reduced subset of similar cases fitting the data structure.

$$\sqrt{(Q_a - Q_b)^2} \tag{7}$$

- And in a second step a weighted Euclidean distance – Eq. (8) - is used to select the best k_2 from the previous retrieved subset (k1) of cases. In this case scores are used as attributes to compute this similarity:

$$\sqrt{\sum_{i=1}^{10} \left(t_{a,i} - t_{b,i} \right)^2} \tag{8}$$

Where $t_{a,i}$ and $t_{b,i}$ represent the components of the new and stored cases respectively in the principal component space (10 component have been used in this work) after the projection using the PCA model. Finally a decision threshold is used to determine the class using the k_2 retrieved cases:

$$\frac{\sum_{i=1}^{nm} d_i}{\sum_{j=1}^{k_2} d_j} > Th \tag{9}$$

Where nm are those k_2 retrieved cases from the model

class and Th is the decision threshold to determine if the new case is close enough to the model class. Several values of this threshold have been explored in the validation step. The ROC (Receiver Operation Characteristic) curve represents the performance of classification for different values of that threshold.

5. Validation methodology

Validation of the method has been done with data from distribution substations using *n-fold* cross validation and computing the sensitivity and specificity for each experiment. Several decision thresholds have been used in the test stage. Results have been used to represent the Receiver Operation Characteristic (ROC) curve.

A. n-Fold Cross Validation

In *n*-fold cross validation, the available data is divided into *n* folders containing approximately the same number of examples. Once the data is divided, one of the *n* folds of samples is retained for validation of the model formed by the remaining *n* – 1 data fold. This process is repeated *n* times (once for each fold) [8]. Then the average performance of these *n* experiments is computed. The application of this simple method allows using all data for testing and training (in different executions) and consequently a more representative performance index is obtained. In this paper we applied a 5-fold cross validation (*n* = 5).

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 labeled examples, as is shown in Table 1[9].

Table 1: Confusion Matrix elements

		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:

$$Sensitivity\left(SEN\right)=\frac{TP}{TP+FN} \tag{10}$$

$$Specificity\left(SPC\right)=\frac{TN}{TN+FP} \tag{11}$$

$$Accuracy\left(ACC\right)=\frac{TP+TN}{TP+TN+FN+FP} \tag{12}$$

$$Precision\left(PRE\right)=\frac{TP}{TP+TN+FP+FN} \tag{13}$$

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

The Receiver Operating Characteristic (ROC) curve is used, in conjunction with the Neyman-Pearson method, in signal detection theory [10]-[11] and also in medicine for measuring the performance of diagnostics. It is a good way of visualizing the performance of a classifier in order to select a suitable operating point, or decision threshold.

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

The ROC curve 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 classify 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]. Example of a ROC curve obtained following theses steps can be observed in Fig. 5.

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],

$$AUC=\sum_i\left\{\left(1-\beta_i\times\Delta\alpha\right)+\frac{1}{2}\left[\Delta\left(1-\beta\right)\times\Delta\alpha\right]\right\} \tag{14}$$

Where:

$$\Delta\left(1-\beta\right)=\left(1-\beta_i\right)-\left(1-\beta_{i-1}\right) \tag{15}$$

$$\Delta\alpha=\alpha_i-\alpha_{i-1} \tag{16}$$

Observe that α and β are related with the False and True Positive ratios of the classifier:

$$\alpha=\Pr\left(FP\right);1-\beta=\Pr\left(TP\right) \tag{17}$$

6. Results

This methodology has been tested with sags registered in several substations resulting in a very good performance.

A. Origin of Data

The data used are 212 registers of sags, which contain phase voltage and current waveforms registered during 0.8 seconds in three 25kV substations. From each registered waveform, 4993 samples have been obtained (128 samples / period).

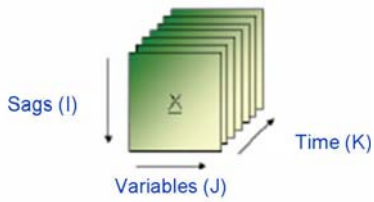

Fig. 2. Data Matriz

The data has been separated in two classes: HV (141 sags produced upstream) and MV (81 sags produced downstream). In order to simplify its management, the data has been group in a 3D matrix, as shown in Fig. 2.

B. Training and Testing Subsets

The original subset of data has been split in 5 folders using the *n*-Fold Cross Validation methodology in order to have a more realistic classifier's performance evaluation. This division has been done by preserving the proportion of both types of sags: HV origin and MV origin in each fold.

C. Preprocessing

Fast Fourier Transformation (FFT) in one cycle (20 ms) with a sliding window has been used to estimate the magnitude of nominal frequency (50 Hz) during the sag has been used to compute the RMS values. Due to the fact that voltage and current have a different nominal value, the data will be scaled using the autoscaling procedure. First, a mean and standard deviation are computed for each variable at each time instant. Then the data is scaled using the following equation:

$$X_S = \frac{X - \overline{x}}{s^*} \qquad \textbf{(18)}$$

Where X_S is the scaled data, X is the original data, $\overline{x}$ is the mean of the data and s^* is the standard deviation of the data computed from the available data base.

D. PCA

Once the data is scaled, the matrix is unfolded in the sag wise to apply standard PCA calculation. The precision of the model generated is desired to be above 95%, it results in the use of 10 principal components. In Fig. 3, a representation of the first 3 principal components of one of the folders in which the data was divided is depicted.

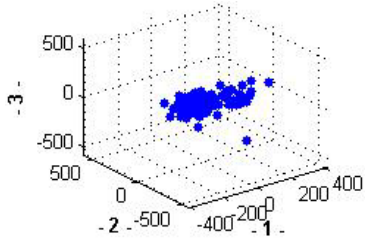

Fig. 3: PCA space representation

E. Case representation

The information stored for every sag constitutes a case. That is:

- Voltage and Current waveforms for future computation.
- The 10 first principal components $(t_1, t_2, ..., t_{10})$.
- The Q statistic.
- The origin of the sag (HV, MV).
- The name of the original file where the information was extracted.
- Date and Time
- Substation and transformer where was registered

F. Determining the k-NN cases

The k-Nearest Neighbors have been obtained following the two steps distance criterion described in the previous section. Fig. 4 shows an example of two sags with a distance of 0.0248 measured in the principal component space. Observe that both waveforms (three voltages and three currents each) present similar shapes.

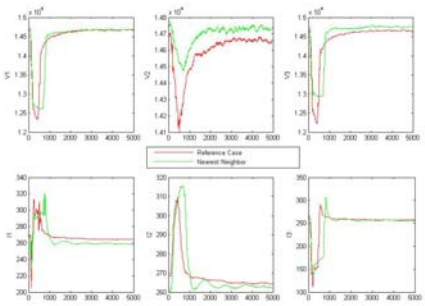

Fig. 4. New case a the nearest one retrieved from the case base.

An additional benefit of using the distance criterion in the principal component space is that it avoids comparing waveforms in the time domain.

G. Results

The methodology has been tested using different pairs of values for the k_1 and k_2 retrieved cases and using

different thresholds to compute the ROC curve depicted in Fig. 5. It can be observed that all the tests had a very good performance: AUC near 1. The red dashed line represents a random classifier. According to AUC (Table 2) criteria, the best classifier is the one using $k_1 = 20$ and $k_2 = 3$ because it has the greatest value, also having the nearest operating point to (0,1) in the curve.

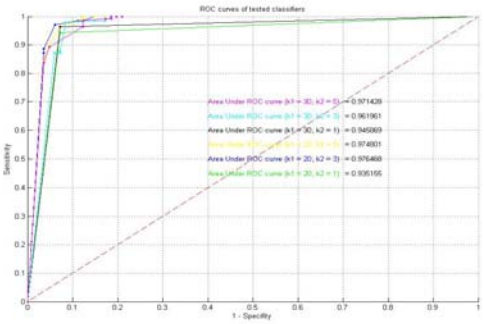

Fig. 5. ROC curves and AUC of several classifiers

Table 2: AUC values of the classifiers

(k1, k2)	AUC
(20,5)	0.975
(20,3)	0.976
(20,1)	0.935
(30,5)	0.971
(30,3)	0.962
(30,1)	0.946

Once the best parameter for the classifier is known, the next step is to choose the best decision threshold. As exposed before, the easiest way is to find the nearest point to the point (0,1) because there are no errors in the classification. For this classifier the Sensitivity and specifity of Table 3 have been obtained.

Table 3: Table of Sensitivities and 1 - Specificities of the best tested classifier

Threshold	1 - Specificity	Sensitivity
1	0	0
0,8	0,036	0,872
0,9	0,036	0,872
0,7	0,036	0,886
0,6	0,061	0,971
0,5	0,111	0,986
0,4	0,124	0,986
0	0,148	1
0,1	0,148	1
0,2	0,148	1
0,3	0,148	1

As shown in Table 3, the classifier has a better classification ratio with low threshold values (threshold with a maximum value of 0.3). So, any of these values can be chosen as the best operating point to classify sags according to HV and MV.

7. Conclusions

The methodology presented in this paper for sag classification has been demonstrated to be very good with the data set used in the experiments. The usage of the ROC curves and the AUC has been used to compare different configurations of the classifier with excellent performance and have been used to decide the best configuration for this location of origin of sags.

Although these good results encourages to proceed improving the methodology, the data used only represents a part of all the available cases, and future experiments will be done in order to evaluate the correlation among the subset used in this paper and the whole set of cases. Another point to take into account will be to use more distances in the k-NN selection step.

Acknowledgement

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 projects DPI2005-08922-C02-02 and DPI2006-09370 funded by the Spanish Government.

References

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

[2] Mora, J.; Llanos, D.; Melendez, J.; Colomer, J.; Sanchez, J. and Corbella, X., “Classification of Sags Measured in a Distribution Substation based on Qualitative and Temporal Descriptors”, in 17th Int. Conf. On Electricity Distribution, May 12-15, Barcelona, Spain, 2003.

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

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

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

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

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

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

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

[10] Fukunaga K., Introduction to Statistical Pattern Recongition, in Sand Diego (California) Academic Press, 1990.

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

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