# Modified Hierarchical Clustering Algorithm for Partial Discharge Separation

Alessio Di Fatta[1], Antonino Imburgia[1], Giuseppe Rizzo[2], Ghulam Akbar[1], Vincenzo Li Vigni[2], Pietro Romano[1], and Guido Ala[1]

[1]L.E.PR.E. HV Laboratory, Department of Engineering, University of Palermo, Italy
[2]EOSS, Prysmian Group, Milan, Italy

***Abstract*—To date, one of the main tools for evaluating the reliability of an insulation system is the continuous monitoring of those phenomena which, by interacting with the elements of the system, can induce aging processes or failures. For power grids, a signal that identifies possible aging or improper use of the component is Partial Discharge (PD) activity. Generally, the evaluation of the PD phenomenon is carried out through a two-step procedure: measurement and data analysis. To optimize the PD analysis process, increasingly sophisticated PD separation/classification algorithms are needed. Especially for the measurements carried out in HVDC systems for which the absence of a phase reference makes more difficult to identify the different types of discharge. The purpose of this article is to investigate the possibility of optimizing the input data to a hierarchical clustering algorithm in order to obtain a subdivision of the dataset more faithful to the real behavior of the phenomena. Specifically, the proposed approach is based on the use of the cross-correlation matrix to carry out the clustering operation. This matrix replaces the matrix of the distances among the points distributed in the map used for the representation of the data. Results show that with this modification it is possible to separate phenomena that present partially or completely overlapping patterns. Moreover, the algorithm turns out to be automatic and does not require the choice of references or thresholds to define the similarity among pulses.**

***Index Terms*—Cross-Correlation, Hierarchical Clustering, HVDC, Partial Discharge, Pattern Recognition.**

## I. Introduction

The increased sensitivity, acquired over the years, to environmental issues has raised new challenges in the field of technological and industrial development. Indeed, today's goal is to make the globalization process compatible with the impact that certain activities, industrial and otherwise, have on the planet. In the energy sphere, this translates into a transition to a system supported by source diversification, reduced fossil fuel use, and optimization and upgrading of grids. In this scenario, High-Voltage-Direct-Current (HVDC) technology for power transmission over long distances and with cable lines (underground or submarine) is going through a period of diffusion and expansion due to its advantages over a traditional transmission system [1], [2]. An important aspect of assessing the reliability of these systems is the continuous monitoring of phenomena that, by interacting with sensitive elements or components, can lead to failure. In power grids, one of the most failure-prone elements is the electrical insulation system, and a signal that identifies possible aging or improper use of the component is Partial Discharge (PD) activity [3], [4]. Generally speaking, evaluation of PD on components can be traced back to two operations, which are summarized in Fig. 1 [3]. First operation is measurement, in which some devices and sensors are used to detect and acquire a dataset of variables or quantities related to the phenomenon to be analyzed. In the case of PD, for example, acoustic, electrical, chemical or thermal measurements can be made [5]. In this paper, focus is on the second operation, the data analysis. Since PD can occur with different features, an analysis phase of the collected data is also necessary in order to recognize the amount and type of involved phenomena. In addition to the identification among the main discharge phenomena (corona, internal, surface and treeing) it is indeed advisable to evaluate and eliminate any background noise acquired during the measurement. The typical structure of PD data analysis process is also shown in detail in Fig. 1. The first step is to choose suitable features that can provide the user with useful information about the observed phenomenon. A distinction is made between selected or extracted features. In the former case it is a subset of magnitudes selected from those measured. For example, the amplitudes of each pulse or the occurrence phase are features provided directly by the instrumentation. In the second case, the features are obtained by processing the original data. Statistical features are derived by processing the measured features. The next step is the representation of the features through tables or maps (2D or 3D) useful for identifying the presence of patterns within the observed phenomenon. Whenever it is not possible to identify the discharge phenomena directly from the patterns, these maps are used as support for the application of non-supervised (clustering) or supervised (deep neural networks) algorithms that carry out the procedure of data separation and classification [6], [7]. In AC field, the most used method for data representation is the Phase-Resolved-Partial-Discharge (PRPD) Pattern, in which the apparent charges or amplitudes of voltage signals acquired by instrumentation are plotted as

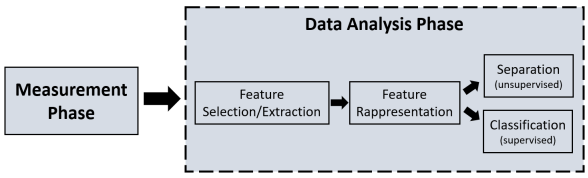


Fig. 1. Main operations for the analysis of a partial discharge phenomenon.

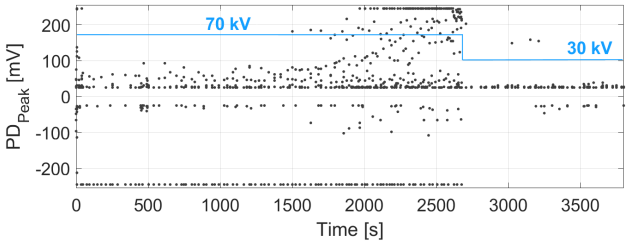


Fig. 2. TRPD pattern and applied voltage profile (in blue colour).

a function of the occurrence phase angle. In most cases, this map does not require the application of more sophisticated techniques for classification, since a correlation between the shape and position of the patterns and the type of discharge phenomenon is shown to exist. A phase reference, on the other hand, is not available in DC. Therefore, the main challenge for PD analysis in HVDC systems is to find a representation, separation, and classification procedure that can guarantee good accuracy and reliability in the obtained result. To date, there is no reference standard for the PD analysis under HVDC stress and the proposals in the literature are quite varied. Some examples for feature representation that can be used in DC are the Time-Resolved-Partial-Discharge (TRPD) pattern, the Time-Frequency Map (TF Map) or the space of principal components extracted with the Principal Component Analysis (PCA) [8]–[10]. As mentioned above, maps are used as support for the application of algorithms. Thus, these algorithms often rely on processing the data provided by the maps to conduct separation and recognition operations. The main information is the distances among points. Distance is used as a metric to evaluate the similarity among the acquired signals. In section III, the application of the hierarchical clustering algorithm for separating data from a HVDC PD measurement represented on the TF Map is given as an example. The aim of this paper is to propose a different approach based on the cross-correlation operation for evaluating the similarity among pulses and thus the data that are provided as input to the hierarchical clustering algorithm. This approach is also illustrated in section III and compared with the original approach. Section II shows the measurement from which the data have been obtained. Instead, section IV shows the results for the two different approaches.

## II. PD Measurement

The data analyzed in this paper have been obtained through a PD measurement performed on XLPE insulated model cable subjected to HVDC stress. The test lasted one hour in which the voltage profile reported in Fig. 2 has been applied at the terminations spliced in order to form a loop. An artificial defect is made on the cable to simulate the presence of a cavity between the dielectric and the outer semiconductive layer. The measurement has been carried out with a PD detection device consisting of an ultra-wide band spherical antenna sensor [12]. The use of ultra-wide band instrumentation allows to acquire the PD pulses with a good fidelity to the original signal. This is an important key aspect since the aim is to divide a dataset into subgroups based on the similarity among pulses. At the end of the measurement, the acquired dataset consists of 1035 pulses.

## III. PD Separation

### A. Original approach

In this paper, the dataset obtained as a result of the PD measurement are represented through the TF Map, which is constructed through the analysis of each individual pulse in the time and frequency domain, the information of which is summarized in two parameters called Equivalent Timelength (T) and Equivalent Bandwidth (F). The expressions for calculating these quantities are given in 1 and 2, respectively.

$$T = \sqrt{\int_{-\infty}^{\infty} (t - t_0)^2 \cdot \tilde{s}(t)^2 \cdot dt} \tag{1}$$

$$F = \sqrt{\int_{-\infty}^{\infty} (f - f_0)^2 \cdot |\tilde{S}(f)|^2 \cdot df} \tag{2}$$

where $\tilde{s}(t)$ e $|\tilde{S}(f)|$ are respectively the normalized pulse with respect to its own energy and the modulus of the related Fourier transform. $t_0$ and $f_0$ are the temporal and spectral centroids of the pulse and its transform [11]. A hierarchical clustering algorithm has been chosen for partitioning the dataset into subgroups. This algorithm performs a partitioning of the dataset by following two different procedures called agglomerative and divisive. In the first case each object is initially considered a cluster unto itself and the algorithm proceeds by clustering subgroups of objects at each iteration until a single cluster is obtained. The divisive procedure, on the other hand, follows the opposite philosophy. In this paper the agglomerative procedure has been applied. The main input data of the algorithm is the matrix of distances among the points given in the TF Map and representative of each pulse. This matrix, of size $n$x$n$, is symmetric, characterized by a diagonal of null elements and, if the map is normalized, consisting of elements enclosed in the 0 - 1 range. From this matrix, the algorithm evaluates the position (i,j) of the smallest element (pair of elements due to symmetry) and thus of the two nearest points on the map. These points are subsequently grouped into a single object (cluster). As shown in Fig. 3, in numerical terms, this operation results in the union of the i-th and j-th rows and columns, reducing the size of the matrix by 1. The distances among the remaining points and the identified object, or among objects for subsequent iterations, can be determined in different ways. Typical choices are shown in Tab. I. Step by step the algorithm gathers subgroups of objects until a single cluster is obtained. As a result, the algorithm provides the hierarchical structure by which the clusters were identified and grouped. This structure is represented with a scheme called dendogram. Fig. 4 shows the dendogram obtained for the measurement described in section II. Using this diagram, it is possible to choose how many clusters to subdivide the dataset into. The procedure described is a typical procedure used to analyze a dataset with a hierarchical clustering algorithm.

TABLE I
MAIN CHOICES FOR CALCULATING THE DISTANCE BETWEEN TWO CLUSTERS A AND B

| *Distance Dist(A, B)* | Formula |
|---|---|
| Single distance | $\min_{a\in A, b\in B} dist(a,b)$ |
| Complete distance | $\max_{a\in A, b\in B} dist(a,b)$ |
| Average distance | $\frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{a\in A}\sum_{b\in B} dist(a,b)$ |
| Centroid distance | $dist(C_A, C_B)$ |

### B. Improved approach

In the proposed approach, the similarity among pulses is evaluated through the cross-correlation operation and not in terms of distances on a map. The similarity between two signals can be evaluated by solving the following integral:

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t)\cdot y(t+\tau)\cdot dt \tag{3}$$

Where $R_{xy}$ is the cross-correlation function, $x$ and $y$ the two signals to be compared. By considering a time shift of $y$ with the $\tau$ factor, it is possible to take into account all possible lags between the two signals and thus evaluate their best overlap [13]. This information is quantified with the maximum value of the $R_{xy}$ function. If one chooses to normalize the pulses with respect to their energy, the correlation function turns out to be within the range 0 - 1. By calculating all possible correlations among all the pulses in the dataset it is possible to construct the correlation matrix. The properties of this matrix, under certain conditions, are similar to that of the distances matrix. The matrix is symmetric and with elements on the diagonal equal to 1. Due to normalization all elements of the correlation matrix have values included between 0 and 1. A value close to one indicates a high degree of similarity, while a value close to zero indicates a low similarity. Consequently, before supplying the input matrix to the algorithm, its complement to unity is performed. To evaluate the correlation between two objects, the average value has been chosen.

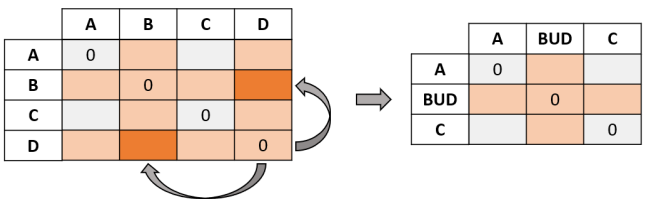


Fig. 3. In one iteration of the hierarchical clustering algorithm, the rows and columns identified by the smallest pair of elements (orange), are overlaid. The values of the elements in the new row and column are determined by one of the formulas in Tab. I.

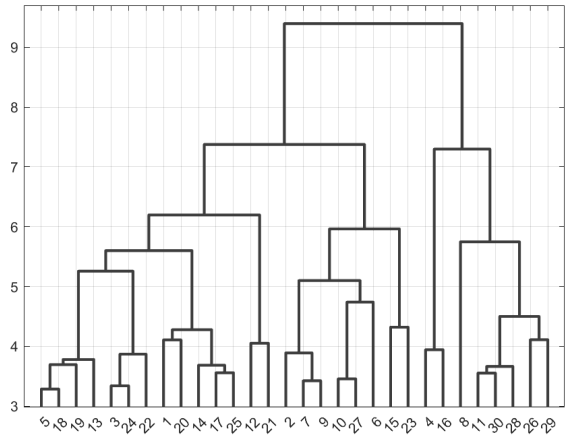


Fig. 4. Example of dendrogram.

## IV. RESULTS ANS DISCUSSION

Fig. 5 shows the comparison between the result obtained by clustering with the classical approach and the improved approach. The first difference that can be seen is that the distance matrix approach results in the formation of clusters that are adjacent to each other but never overlapping. However, this separation does not reflect the real phenomena behavior. Whereas with the correlation matrix, clusters are identified on the basis of the similarity between the waveforms of the pulses, and thus, if the phenomena generate partially overlapping patterns, the algorithm is not affected by this arrangement. Specifically, in the map there are two main patterns of linear and quasi-parallel shape, which are clearly separated using the cross correlation matrix, while they are broken into several clusters using the distance matrix. Being able to separate phenomena that generate overlapping patterns is an important feature for PD analysis, as it allows more information to be extracted from the acquired dataset. Examples of the waveforms based on which the algorithm divided the dataset are shown in Fig. 6 and Fig. 7. The first case shows signals potentially related to partial discharge phenomena, while the second case shows some waveforms typical of a background noise.

## V. CONCLUSION

In this work separation process of PD data based on the analysis of the cross-correlation among the acquired pulses is investigated. This information is provided, in the form of a matrix, as input to a hierarchical clustering algorithm, typically used for data separation based on the distances among points on the data representation map. This in-depth analysis of the phenomenon is necessary when with the map it is not possible to identify specific patterns for the discharge phenomena. To date this issue is typical for HVDC systems for which it is not possible to recognize the PD on the basis of a phase reference. The results obtained show how it is possible with this approach to separate overlapping patterns, thus guaranteeing a subdivision of the dataset into subgroups characterized by similar impulses. This separation process makes it possible to easily identify the number of main dis-

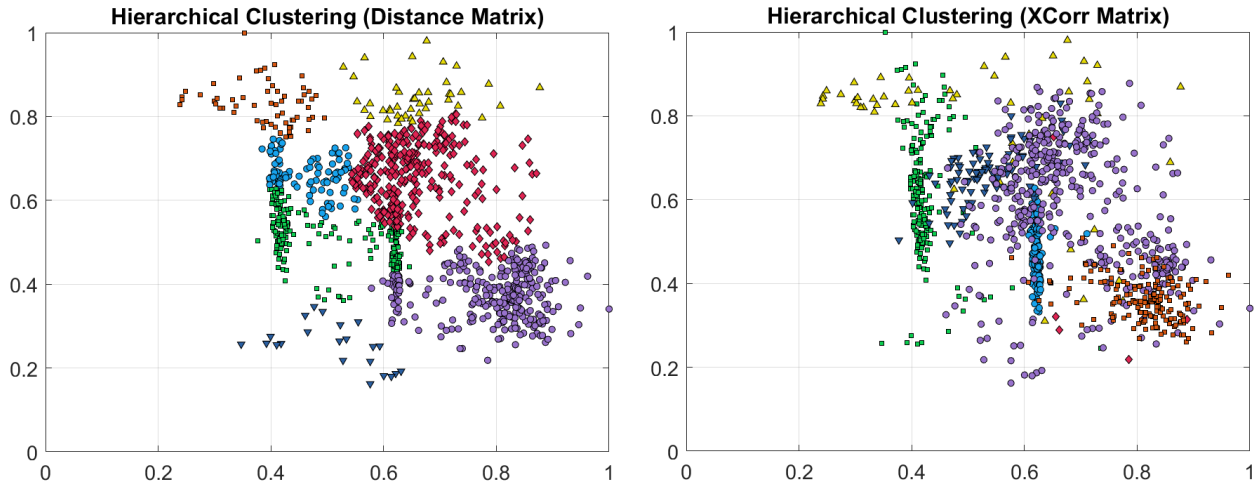


Fig. 5. Clustered and normalized TF Maps. On the left is the result of the distance matrix approach. On the right the result based on correlation matrix.

charges phenomena present and can also be used to eliminate the background noise acquired during the measurement.

## ACKNOWLEDGMENT

This work was realized with the co-financing from European Union – FSE, PON Research and Innovation 2014-2020 – DM 1062/2021.

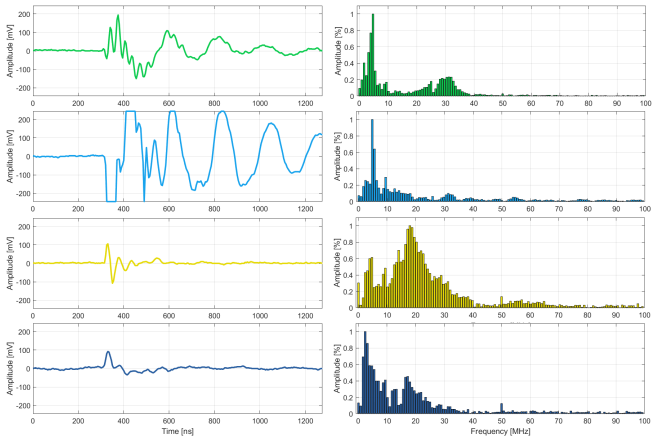

Fig. 6. Pulses and signal spectra characterizing the clusters related to PD activity.

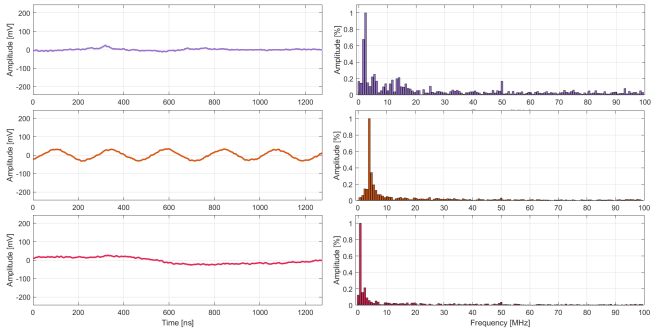

Fig. 7. Pulses and signal spectra characterizing the clusters correlated to background noise.

## REFERENCES

[1] Arrillaga, Jos, and Jos Arrillaga. High voltage direct current transmission. Vol. 29. Iet, 1998.

[2] Mazzanti, G., & Marzinotto, M. (2013). Fundamentals of HVDC cable transmission.

[3] M. Wu, H. Cao, J. Cao, H. -L. Nguyen, J. B. Gomes and S. P. Krishnaswamy, "An overview of state-of-the-art partial discharge analysis techniques for condition monitoring," in IEEE Electrical Insulation Magazine, vol. 31, no. 6, pp. 22-35, November-December 2015, doi: 10.1109/MEI.2015.7303259.

[4] W. Koltunowicz, L. -V. Badicu, U. Broniecki and A. Belkov, "Increased operation reliability of HV apparatus through PD monitoring," in IEEE Transactions on Dielectrics and Electrical Insulation, vol. 23, no. 3, pp. 1347-1354, June 2016, doi: 10.1109/TDEI.2015.005579.

[5] Yaacob, M. M., et al. "Review on partial discharge detection techniques related to high voltage power equipment using different sensors." Photonic sensors 4 (2014): 325-337.

[6] A. Nagpal, A. Jatain and D. Gaur, "Review based on data clustering algorithms," 2013 IEEE Conference on Information & Communication Technologies, Thuckalay, India, 2013, pp. 298-303, doi: 10.1109/CICT.2013.6558109.

[7] S. Lu, H. Chai, A. Sahoo and B. T. Phung, "Condition Monitoring Based on Partial Discharge Diagnostics Using Machine Learning Methods: A Comprehensive State-of-the-Art Review," in IEEE Transactions on Dielectrics and Electrical Insulation, vol. 27, no. 6, pp. 1861-1888, December 2020, doi: 10.1109/TDEI.2020.009070.

[8] G. C. Montanari, P. Seri, R. Ghosh and L. Cirioni, "Noise rejection and partial discharge source identification in insulation system under DC voltage supply," in IEEE Transactions on Dielectrics and Electrical Insulation, vol. 26, no. 6, pp. 1894-1902, Dec. 2019, doi: 10.1109/TDEI.2019.008210.

[9] A. Di Fatta et al., "Use of Time-Frequency map combined with DBSCAN algorithm for separation of partial discharge pulses under DC voltage," 2022 IEEE Conference on Electrical Insulation and Dielectric Phenomena (CEIDP), Denver, CO, USA, 2022, pp. 427-430, doi: 10.1109/CEIDP55452.2022.9985393.

[10] K. X. Lai, B. T. Phung and T. R. Blackburn, "Partial Discharge Analysis using PCA and SOM," 2007 IEEE Lausanne Power Tech, Lausanne, Switzerland, 2007, pp. 2133-2138, doi: 10.1109/PCT.2007.4538648.

[11] Franks, Lewis Embree. Signal theory. Dowden & Culver, 1981.

[12] Romano, P.; Imburgia, A.; Ala, G. Partial Discharge Detection Using a Spherical Electromagnetic Sensor. Sensors 2019, 19, 1014. https://doi.org/10.3390/s19051014

[13] Bracewell, R. "Pentagram notation for cross correlation. The Fourier transform and its applications." New York: McGraw-Hill 46 (1965): 243.