# Data mining software process used to identify power quality event traces

Dan Apetrei[1], Ralf Neurohr[2], Mihaela Albu[2], Petru Postolache[2], Valentin Rascanu[2], Nicolae Golovanov[2] and Ioan Silvas[1]

[1] SC Electrica SA, str. Grigore Alexandrescu nr. 9, Bucharest, 011621 Romania; e-mail: dan.apetrei@electrica.ro;ioan.silvas@electrica.ro

[2] "Politehnica" University of Bucharest, Splaiul Independenţei nr.313, Bucharest 060042 Romania e-mail: ralf.neurohr@gmail.com, albu@ieee.org, vrascanu@yahoo.com, postolachepetru@yahoo.com, nicolae_golovanov@yahoo.com ,

**Abstract. Power quality data gathered on-site are referred to specific standards to check if the power parameters are within the prescribed limits. From voltage or frequency time series, events are usually extracted based on fixed thresholds. The paper analyses an alternative based on data mining to the classical approach. Positive effects of the new approach are named and investigated.**

**Key words**

Power quality analysis, data mining, clustering, event trace.

## 1. Introduction

In this paper, data resulting from several voltage survey long-term monitoring campaigns are studied to highlight their relevance to the steady-state conditions of the distribution grids. Based on half-cycle voltage measurement, we present ways to improve data versatility in power quality investigation. Thanks to the latest technology developments, now it is possible to deal with significant amounts of data that can be used in improving power quality management. The analysis of the measurement results is done in two steps: multistage processing and direct information extraction. Based on this approach, the unified measurement architecture concept is presented. Advantages of the new architecture are discussed.

## 2. Data processing techniques with respect to PQ Standards - alternatives to aggregation

Standards define the methods for determination of the power quality (PQ) parameters. Relevant parameters are described, in terms that allow obtaining reliable, repeatable and comparable results. The power quality parameters considered in most application are: power frequency, magnitude of the supply voltage, flicker, supply voltage dips and swells, voltage interruptions, transient voltages, supply voltage unbalance, voltage and current harmonics and interharmonics, mains signaling on the supply voltage, and rapid voltage changes [1][2].
In order to get build "primary" data from the field measurements, one of the processes that are standardized in PQ measurement is the aggregation. As presented in figure 1, the aggregation process allows us determining a unique value that describes a whole data set. For instance, in the case of voltage measurement, 10 minutes values are standardized. Building these values is a complex process. First stage in the data acquisition provides cycle voltage root mean square (presented by small white circles). This is the primary data presented in figure 1. From a number of data grouped in a package the first aggregate value is built. This is a 3 seconds time interval value (presented by yellow circles in the figure). The second stage of the aggregation process gathers together all the 3 seconds values in a 10 minutes aggregated value (presented by red value in the figure).

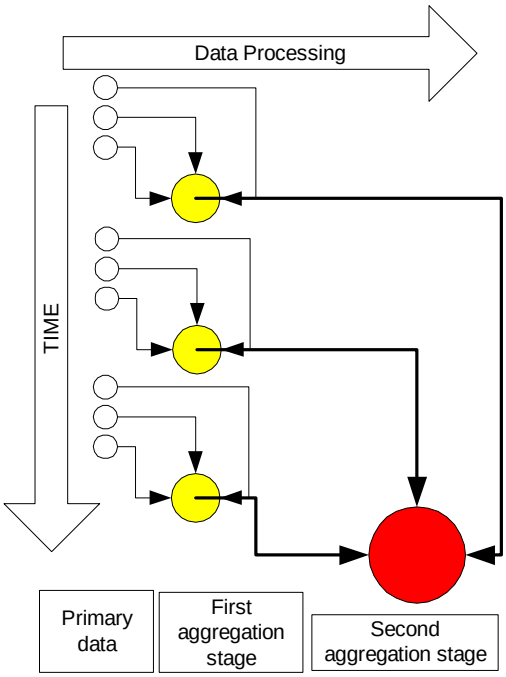


**Figure 1 aggregation process**

Each measurement campaign has some purpose. More or less collecting data from a process is a good thing but it's not enough. If from the data is not possible to extract information, or the information does not reflect the relevant aspects of the process, the measurement was not a success.
The most obvious difference between data and information is that the information needs to sustain a decision while data is sort of information raw material. The aggregation process gives a pragmatic approach to information extraction from the PQ measured data, but as presented in figure 2, this approach has a moderate risk to

lose same secondary information that could be also extracted from the data.
As could be seen in figure 2, the aggregation processes produces secondary and tertiary data. From tertiary data, extraction of the information regarding PQ state in the measurement point is possible.

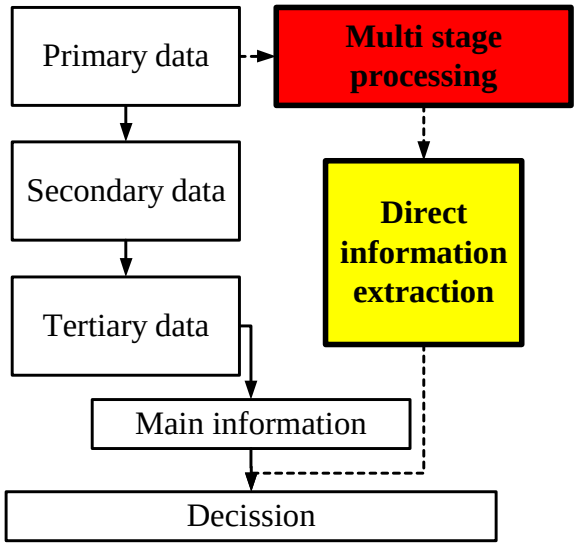


**Figure 2 data vs. information in PQ measurement**

The alternative for data processing that we are investigating in this paper is related to direct information extraction by multistage processing. As could be seen in figure 2, the red block of multi stage processing has at the input the primary data and at the output the information extraction process. That allows direct decision connection and is supposed to improve the process by gathering more accurate the information from the process.

## 3. The consequences of Multistage processing & Direct information extraction for measurement system design - Unified Measurement Architecture

In figure 3, typical node architecture in DSO applications is presented. More or less, most of the important nodes in the network contain common functions like: metering, SCADA, PQ and protection. Each of these functions is done by dedicated equipment. These equipments developed different architectures in time. Now we are close to a stable common data measurement solution. This solution is based on standardised blocks/functions like:

- Analogue to digital conversion (A/D convertor1-4);
- Processing unit (PU1-4);
- Local display (1-4);
- Upper level communication (ULC 1-4);
- Decision and control;

Just a quick look at the multiple blocks/functions presented in figure 3 shows multiple similarities between the four classes of applications. For instance, there are four analogue to digital convertors. All that four convertor that could be replaced by a single one if we get to a unified architecture. Even more, when it comes to analogue to digital conversion, it is possible to include in the conversion stage, dedicated correction curves. If that would be done, the need for different measurement transformers in protection or measurement application is no longer actual. Existing protection transformers after recalibration could be software compensated for the whole measurement domain.
Similar approaches are possible for the other blocks in the figure 3.

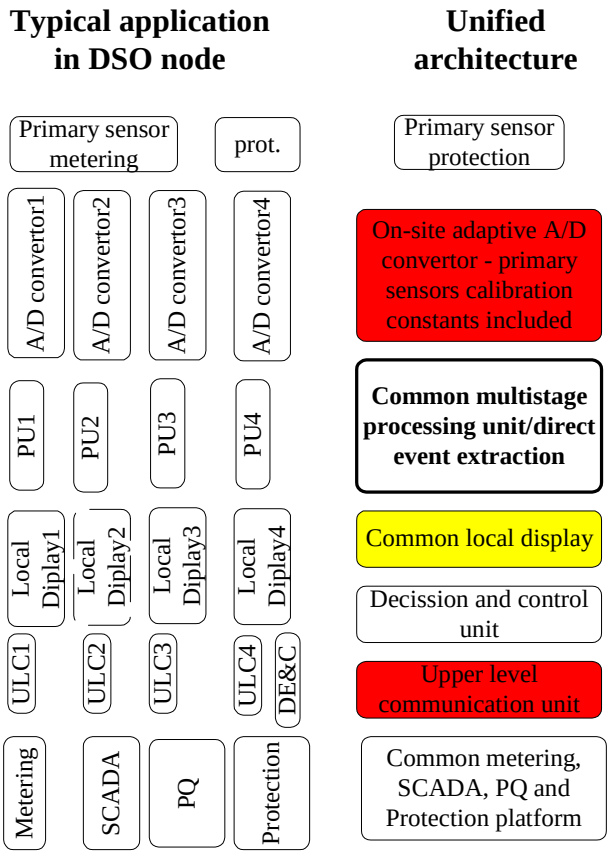


**Figure 3 compare common approach and unified architecture**

All this is possible by defining the multistage processing ain direct event extraction in detail. With that in mind, direct information extraction is possible as it would be further described.
The functions presented in red (figure 3), are the most probable to drive the change to a unified architecture. The common display function presented in yellow is less probable to get to a standardised fast approach and the remaining function are more o less of local relevance.

## 4. Multistage processing

Figure 4 presents the main actions that are involved with multistage processing of primary data. All the process is done over a dynamically adjusted rolling window. The dynamics of the window size is controlled inside the

process and depends on the changes in the measured signals.

First stage in processing is the descriptive statistics of the signal. One good option in order to do that is to use the principles established for the box-plots representation of data. Among the data that describe a time series there are mean, median (or value at the 50$^{th}$ percentile), maximum, minimum, quartiles, and so forth. In 1977, John Tukey published [3] an efficient method to display robust statistics called box plot or box whiskers.

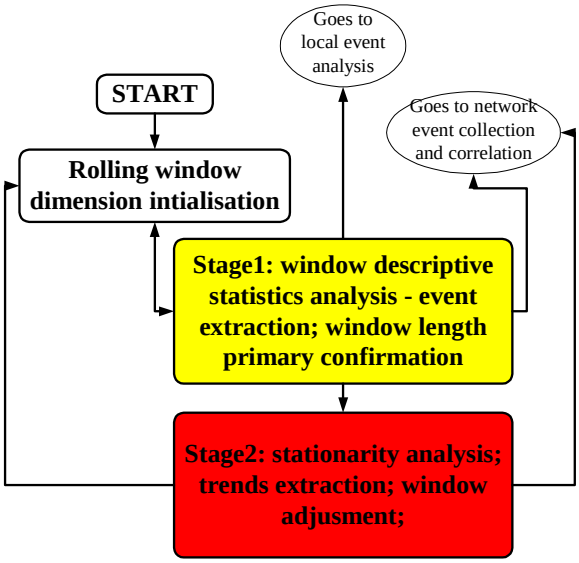


**Figure 4 multistage processing steps**

Basic features of such a graphical image are: central box includes the middle 50% of the data; whiskers show range of data; symmetry is indicated by box and whiskers and by location of the mean; it is easy to compare groups by constructing side-by-side box plots, as shown below.
Figure 5 shows a real life example of such an analysis. There are 20 intervals of 10 minutes measurements presented in the chart.

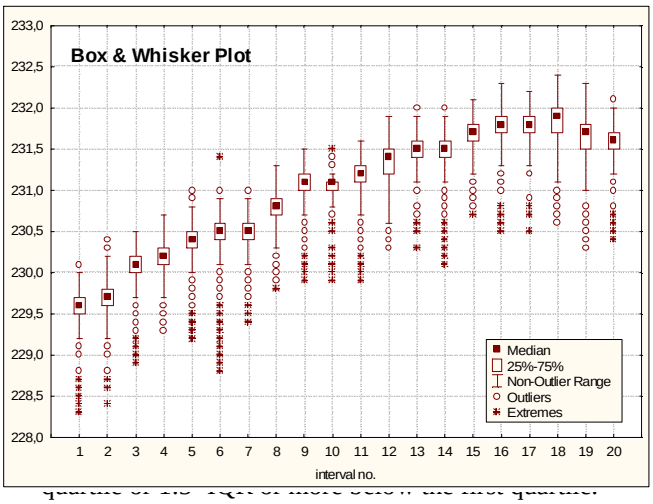


**Figure 5 voltage box plot sample for 20 inervals of 10 minutes each interval containing 60000 primary data for the half cycle voltage determination**

There are two ways to present data as a box plot: "median-based" or "mean-based". The center point in a median-based display is the group median, or the middle value. If the number of items in a group is odd, the median is the exact value of the middle number. If the number of items in a group is even, the median is the average of the two middle values. Figure 5shows the box plot for all the data selected to be analyzed.
If either type of outlier is present the whisker on the appropriate side is taken to 1.5×IQR from the quartile (the "inner fence") rather than the *max* or *min*, and individual outlying data points are displayed as unfilled circles (for suspected outliers) or filled circles (for outliers). The "outer fence" is 3×IQR from the quartile.
For median-based plots, extremes are typically defined as those individual points which are 3 times the IQR, beyond the end of the box.
The analysis done in stage one feeds four processes with data:

- rolling time window definition process;
- local event analysis process
- network event collection and analysis
- stage two processing

In the stage two the main analysis on the data is meant to check for stationarity/non-stationarity feature.

One of the hypotheses that sustain aggregation is the stationarity of the time series under analysis. In order to carry out an appropriate statistical test we choose Student's t test. IN order to do that, the actual voltage ($x_i$) a cluster of $N$ consecutive voltages in the past ($x_{i-N}$, …, $x_i$) is compared to an equally sized cluster of voltages in the future ($x_i$, …, $x_{i+N}$) by the two sample t test according to:

$$\hat{t} = \sqrt{\frac{N_1 N_2 [N_1 + N_2 - 2]}{N_1 + N_2}} \cdot \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{s_1^2 (N_1 - 1) + s_2^2 (N_2 - 1}} \quad (1)$$

where

$$\bar{x}_1 = \frac{\sum_{j=i-N_1-1}^{j=i} x_j}{N_1} \quad (2)$$

$$\bar{x}_2 = \frac{\sum_{j=i}^{j=i+N_2-1} x_j}{N_2} \quad (3)$$

$$s_1 = \sqrt{\frac{1}{N_1 - 1} \cdot \sum_{j=i-N_1-1}^{j=i} (x_j - \bar{x}_1)^2}; \quad (4)$$

$$s_2 = \sqrt{\frac{1}{N_2 - 1} \cdot \sum_{j=i}^{j=i+N_2-1} (x_j - \bar{x}_2)^2}$$

The stationarity test could by represent a filter as showed in figure 6. This filter produces *M-N* estimates for Student's t.

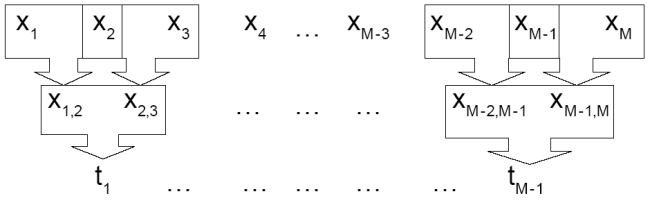


**Figure 6 stationarity filter**

The resulting *M-N* estimates for t are tested at the common error level of 5%. The results are presented in table 1 for the filter widths 2, 4, 10, 20, 100 and 200.

TABLE 1 STATIONARITY TEST RESULTS

**Two sample t test**

| $N_1=N_2$ | 2 | 4 | 10 | 20 | 100 | 200 |
|---|---|---|---|---|---|---|
| error | 1117 | 1 | 0 | 0 | 0 | 0 |
| not significant | 58880 | 59634 | 57937 | 52916 | 21376 | 15417 |
| significant | 1 | 361 | 2053 | 7064 | 38524 | 44383 |

The errors (1st row) for the filter widths 2 and 4 are the result of "division by zero" errors, caused by zero standard deviations (5) due to identical voltages for 3 respectively 5 consecutive measurements.

## 5. Direct information extraction - data mining

Figure 7 shows a simplified form of the process of direct information extraction. Of course this is only one of the multiple approaches possible. Our option for unsupervised learning techniques was based on the stage research in this domain is at the moment.

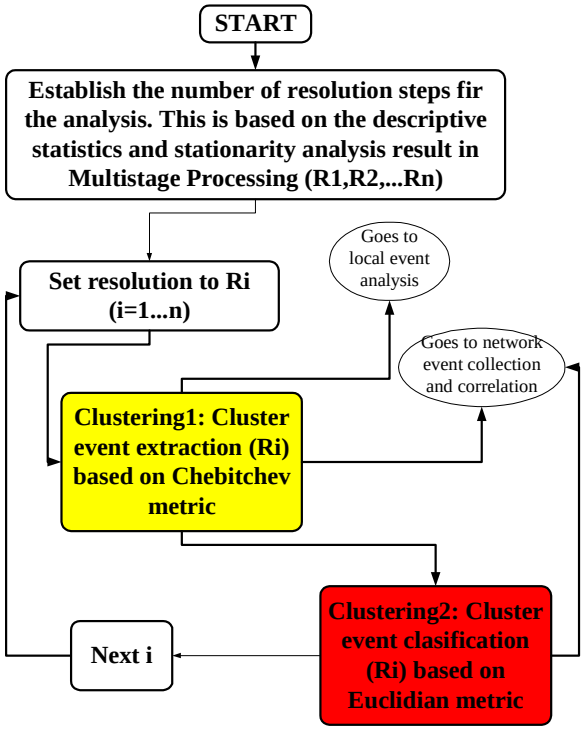


**Figure 7 Direct information extraction process**

As could be seen in the figure, the analysis is a multi-resolution process. The resolution package is established depending on the results of multistage processing. For instance, it is expected to have a reduced number of resolutions for the analysis during the low consumption periods and an increased number of resolution steps in the peak hours. That is because of the more complex energy exchanges during the peak hours.

For the practical results of the clustering 1 process we selected from the data collected the intervals six and twelve to test the clustering techniques. This decision was developed based on the information that we got in graphical box plot analysis. For the practical results we used only the 10ms resolution scale.

**Clustering 1**

The main idea is to isolate the data containing sag traces in order to go further into details with this event. The 60000 values measured during 10 minutes survey were split into 60 vectors of 1000 values. Each vector describes the measurement for 10 seconds.

The method used to form clusters was the "nearest neighbor" [4]. This means that the distance between two clusters is the distance between their closest neighboring points. Results of the analysis are presented in the form of a dendrogram. On the dendrograms the vectors marked with a blue line are the one most probable containing event traces.

Figure 8 shows the dendrograms build for intervals six and twelve. As it can be seen in figure 8, the dendrograms show the relative size of the proximity coefficients at which cases were combined. Cases with low distance/high similarity are close together. Cases showing low distance are close, with a line linking them a short distance from the upper part of the dendrogram, indicating that they are agglomerated into a cluster at a low distance coefficient, indicating alikeness [7][8]. When, on the other hand, the linking line is at the bottom of the dendrogram the linkage occurs a high distance coefficient, indicating the cases/clusters were agglomerated even though much less alike.

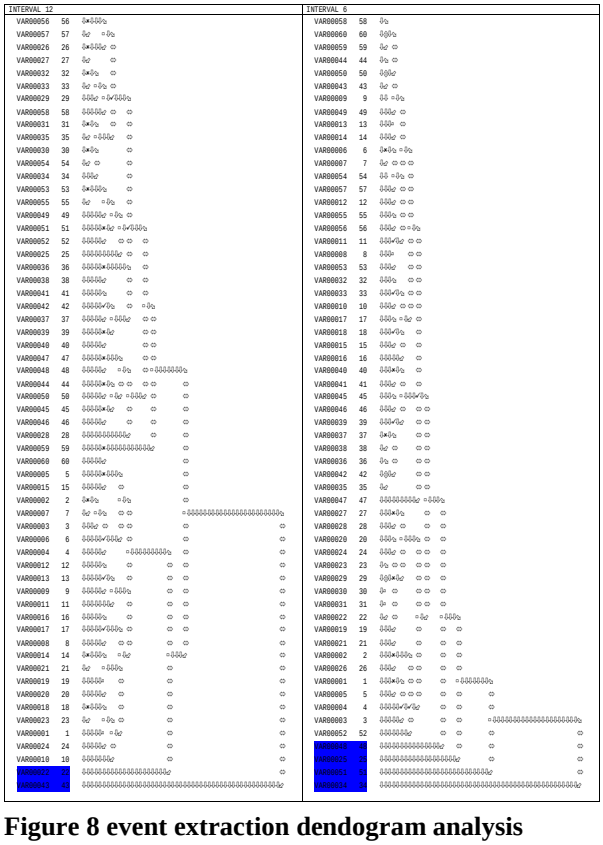


**Figure 8 event extraction dendogram analysis**

Previous papers [5][6] investigated the differences between clustering scenarios and determined the best distance used in order to select event traces. The following methods were used [7]:

- *Euclidean distance* – the most common distance measure. A given pair of cases is plotted on two variables, which form the horizontal and vertical axes.
- *Chebychev distance* – the maximum absolute difference between a pair of cases on any one of the two or more dimensions (variables) which are being used to define the distance.
- *Minkowski distance* – the generalized distance function, defined as the *p*-th root of the sum of the absolute differences to the *p*-th power between the values for the items.

Since the conclusion in [5][6] was that the Chebychev distance is the best choice (at least for the specific application), we used this distance to continue the analysis.

**Clustering 2**

Second stage of the clustering process puts all 56 detected events together and does regular clustering on the crowd. Before processing data, the selected vectors are normalised in order to conserve the shape and ignore the absolute values. Euclidean distance is the option for clustering since feature extraction is the main objective. Figure 9, presents the dendogram of the clustering process.

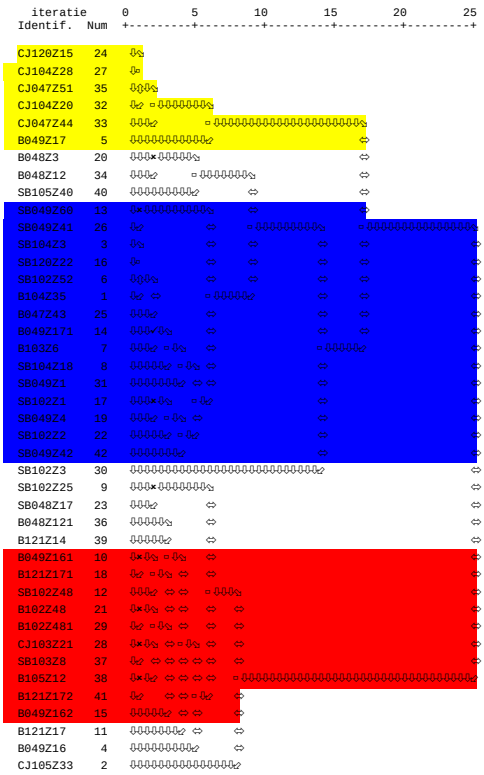


**Figure 9 frequency events clustering dendogram**

As could be seen, there are three distinct areas in the dendogram resulting in three specific signatures of the process under investigation. Yellow area gives the signature that appears only in one of the determination points since the other areas are grouping event signatures that appear in two or three measurement locations.

As an example, figure 10 present the signature identified by the yellow area of the dendogram.

Since we are discussing the frequency parameter is important to give a Remarque on the process identified by this signature.

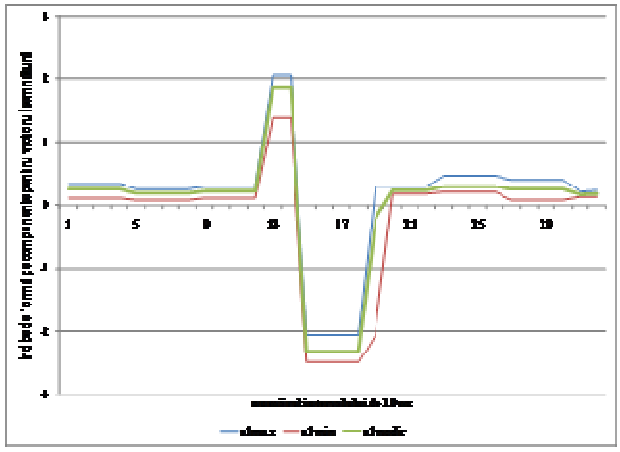

**Figure 10 - single measurement location signature**

As could be seen in the figure, we are dealing with a short reduction of the frequency followed by a short increase and then a coming back to the initial state. Since this could be also sort of noise behaviour, further investigation of the value of this signature are necessary. Best way to indentify the family of the events leading to this signature is to correlate with the real life events registered by the other equipments in the area.

## 6. Conclusion

The paper presents data mining applied in event traces extraction from data quality surveys. This is a process supplemental to typical aggregation recommended by the standard. Main consequences of the new approach are:

- By combining multistage processing and direct information extraction, unified measurement architecture is possible;
- Dynamically defining the rolling window for multistage processing could solve the aggregation/stationarity conflict;
- Cluster filtering based on Chebychev distance is an effective option for event and event traces extraction;
- Events taxonomy based on clustering could separate network from local events.

Further developments of the investigation will include the development of automated procedures for multistage processing and direct information extraction.

## References


1. IEC 61000-4-30 Ed. 1: Electromagnetic compatibility (EMC) - Part 4-30: Testing and measurement techniques - Power quality measurement methods, 2003.

2. Project IEC 61000-4-30 Ed. 2.0, Electromagnetic compatibility (EMC) - Part 4-30: Testing and measurement techniques - Power quality measurement methods, 2008.
3. J.W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, MA. 1977.
4. M.R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973.
5. D. Apetrei, G. Chicco, P. Postolache, N. Golovanov and M. M. Albu, "Cluster Analysis of Half-Cycle Duration Measurements to Classify Local and Network Events", *Proc. IEEE Powertech 2009*, Bucharest, Romania, 28 June - 2 July 2009, paper 405
6. D. Apetrei, P. Postolache, N. Golovanov , M. Albu and G. Chicco, "Hierarchical Cluster Classification of Half Cycle Measurements in Low Voltage Distribution Networks for Events Discrimination", *Proc. ICREPQ 2009*, Valencia, Spain, 15-17 April 2009.
7. G. Gan, C. Ma and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, 2007.
8. J.H. Ward, "Hierarchical grouping to optimise an objective function", *Journal Amer. Stat. Assoc.*, Vol. 58, 1963, pp. 236–244.