



Hierarchical Cluster Classification of Half Cycle Measurements in Low Voltage Distribution Networks for Events Discrimination

Dan Apetrei¹, Petru Postolache², Nicolae Golovanov², Mihaela Albu² and Gianfranco Chicco³

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

² "Politehnica" University of Bucharest, Splaiul Independenței nr.313, Bucharest 060042 Romania e-mail: postolachepetru@yahoo.com, nicolae_golovanov@yahoo.com, albu@ieee.org

³ Politecnico di Torino, Dipartimento di Ingegneria Elettrica, corso Duca degli Abruzzi 24, 10129 Torino, Italy, e-mail gianfranco.chicco@polito.it

Abstract. The paper presents a case study on half cycle RMS voltage classification. The study is based on clustering-based elaboration of data gathered during more than one year at two different locations in a low voltage distribution network. In order to perform the measurements, a specific facility (OSC function) of the custom designed equipment called MOT has been used. After recalling the characteristics of hierarchical clustering techniques, the paper illustrates the main requirements of European regulations for voltage characteristics measurement and presents the MOT equipment. The analysis technique used to process the gathered data and a method to classify the relevant events are then illustrated and discussed, showing a set of significant results obtained in the study.

Keywords

Voltage half cycle RMS, Hierarchical cluster classification, nodal event.

1. Introduction

Voltage measurement is one of the most important sources of information in power systems [1]. Nowadays, the volume of available data is much higher than the amount of data possible to be analysed with classical processing tools. Therefore, new algorithms are necessary. For this purpose, the analyses can be assisted by the use of clustering methods [2][3], which include many alternatives of classification. Among them, *hierarchical clustering* has emerged as a powerful method in many applications [3].

This paper presents a comparison among the clustering results obtained on half-cycle voltage data, in order to identify and classify abnormal events due to disturbances in the electrical network. Different definitions of distance have been used within a hierarchical clustering framework, in order to find out the most effective clustering method. The effectiveness is judged on the basis of the capability of the clustering methods tested of isolating abnormal events from the groups of recorded events.

The details of the clustering methods and of the data used are illustrated in the remainder of the paper. Section 2 recalls the basic concepts of hierarchical clustering. Section 3 summarizes the characteristics and basic requirements of the regulations concerning voltage measurements. Section 4 describes the measurement system. Section 5 presents and discusses the results obtained from processing the real data gathered at different locations.

2. Hierarchical clustering

In the typical structure of hierarchical clustering applied to a set of N data points, initially the N data points are associated to N separate clusters. Then, the clustering procedure proceeds by successive steps, grouping together one pair of clusters at each step (reducing the number of clusters by one), until the predefined number of clusters has been reached, or until all data point have been grouped into a single cluster. The criterion used to group the clusters is based on a similarity measure. In order to define this measure, at each step a square similarity matrix of dimensions equal to the number of clusters existing at the current step of the procedure is built by using a notion of *distance* defined under a specified metric. Then, the similarity measure is obtained by applying a selected *linkage criterion* to the contents of the similarity matrix. A common characteristic of these criteria is that, after forming the similarity matrix, the two clusters exhibiting the maximum similarity (or the minimum distance) are merged.

Various linkage criteria have been defined. The simplest one is the *single* linkage (or nearest neighbour) criterion, that calculates the minimum distance between the elements of a pair of clusters. The calculation is repeated for each pairs of clusters to fill the similarity matrix. An alternative is the *complete* linkage (or farthest neighbour) criterion, that forms the similarity matrix by calculating the maximum distance between the elements of each pair of clusters. Moreover, other linkage criteria have been defined to compute the similarity information by taking

into account all the elements of the clusters and not only the extreme cases. In the *average* linkage criterion, each location of the similarity matrix, corresponding to a pair of clusters, is filled with the average distances between all pairs of elements of the two clusters considered. In the *Ward* linkage criterion, the rationale for forming the clusters is the minimization of the increase of the within-cluster sums of squares. For this purpose, the distance considered to fill a position of the similarity matrix corresponding to a pair of clusters is measured as the increase of these sums of squares obtained if the two clusters were merged [4]. Other rationales based on the calculation of a central or median value could be used to calculate the entries of the similarity matrix.

In order to apply one of the above hierarchical clustering variants, besides choosing the linkage criterion, one has to rely on a definition for distance. A common choice for the available software tools is computing one of the classical distances defined in the literature, such as interval data, Euclidean distance, squared Euclidean distance, cosine, Pearson correlation, Chebyshev, block, or Minkowski. Further customized definitions of distance could be added.

3. Voltage Measurement Requirements

Voltage measurement is regulated in Romania by SR EN 50160 [5] and by IEC 61000-4-30 [6]. This couple of standards makes a good combination of threshold requirements and measurement methods.

A. SR EN 50160

This standard gives the main characteristics of the voltage at the customer's supply terminals connected to the public LV and MV power distribution systems. The standard is used for quantities describing the system under normal operating conditions. The objective of the standard is to define and describe the characteristics of the supply voltage concerning [1][5] frequency, magnitude, waveform shape, and symmetry. These characteristics are subject to variations during the normal operation of a supply system due to changes of load, of various disturbances generated by some equipment and the occurrence of faults which are mainly caused by external events.

The Annex A of the standard gives also a comment concerning the "Special nature of electricity": "Electricity as delivered to the customers has several characteristics which are variable and which affect its usefulness to the customer. [...] In practice, there are many factors which cause departures from this. In contrast to normal products, application is one of the main factors which influence the variation of "characteristics".

For the purpose of this paper, the voltage under normal operating conditions is analysed according to SR EN50160 as follows:

- during each period of one week, 95% of the 10 min mean RMS values of the supply voltage shall be within the range of $U_n \pm 10\%$, where U_n is the nominal RMS voltage of the network at the PCC¹;

- all 10 minutes mean RMS values of the supply voltage shall be within the range $[U_n + 10\%, U_n - 15\%]$.

The rest of the parameters described in the standard could not be prescribed, since they have a local behaviour. Dips for instance are generally caused by faults occurring in the customers' installations or in the public distribution system. They are unpredictable, largely random events. The dip annual frequency varies greatly depending on the type of supply system and on the point of observation. Moreover, the distribution over the year can be very irregular. Under normal operating conditions, the expected number of voltage dips in a year may be from up to a few tens to up to one thousand. The majority of voltage dips have a duration less than 1s and a depth less than 60% of U_n . However, voltage dips with greater depth and duration can seldom occur. In some areas, voltage dips with depths between 10 % and 15% of U_n can occur very frequently as a result of the switching of loads in customers' installations [7].

B. CEI 61000-4-30

This standard defines the methods for measurement and interpretation of results for power quality parameters. Measurement methods are described for each relevant type of parameter, in terms that will make it possible to obtain reliable, repeatable and comparable results. The power quality parameters considered in this standard 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 signalling on the supply voltage, and rapid voltage changes [6].

Since this paper is dealing with voltage measurement, we will briefly present the requirements for this type of measurement.

There are three classes of performance for voltage measurement: A, B, and C. The following rule must be applied for class A measurement:

- the measurement shall be the RMS value of the voltage magnitude over a 10-cycle time interval for 50 Hz power system;
- every 10-cycle interval shall be contiguous with, and not overlap, adjacent 10-cycle intervals;
- the measurement uncertainty shall not exceed $\pm 0.1\%$ over the range of influence quantity conditions;
- aggregation intervals shall be used.

Measurement time intervals are aggregated over 3 different time intervals. The aggregation time intervals are: 3s interval (150 cycles for 50 Hz nominal or 180 cycles for 60 Hz nominal), 10-min interval, and 2-hour interval.

Aggregations are performed using the square root of the arithmetic mean of the squared input values.

Three categories of aggregation are necessary:

- *Package aggregation*: 10 cycle time interval aggregation; this time interval is power system frequency-based;
- *Cycle aggregation*: the data for the 150 cycle time interval shall be aggregated from 15, 10-cycle time intervals; this time interval is not a "time clock"

¹ PCC stands for Point of Common Coupling.

interval; it is based on the frequency characteristic. Because the time interval is not a "time clock" interval, a cycle to time-clock aggregation is needed. According to the standard, the 10-min value shall be tagged with the absolute time (for example, 01H10.00). The time tag is the time at the end of the 10-min aggregation. If the last 10-cycle value in a 10-min aggregation period overlaps in time with the absolute 10-min clock boundary, that 10-cycle value is included in the aggregation for this 10-min interval. At the beginning of measurement, the 10/12-cycle measurement shall be started at the boundary of the absolute 10-min clock, and shall be re-synchronized at every subsequent 10-min boundary. This implies that a very small amount of data may overlap and appear in two adjacent 10-min aggregations.

- *Time-clock aggregation*: data for the "2-h interval" shall be aggregated from twelve 10-min intervals.

Two notes in the standard are important for the purposes of this paper:

- NOTE 1: the measurement method is used for quasi-stationary signals, and is not used for the detection and measurement of disturbances such as dips, swells, voltage interruptions and transients.
- NOTE 2: the RMS value includes, by definition [8], effects of non-sinusoidal shape of the voltage waveform: harmonics, interharmonics, mains signalling etc.

As a consequence, according to the IEC6100-4-30, the stationarity of the voltage is only affected by dips, swells, interruptions and transients. The rest of the intervals could be considered stationary despite the fact that rigorous mathematical definition of stationarity implies statistical parameters testing.

For the class B measurements, the requirements are more relaxed and the responsibilities of data processing are passed to the producer.

4. Measurement System

For the measurements performed, the MOT-103B/BG equipment was used. This is a custom designed, made in Romania measurement system [9].

The system was designed to meet SREN 50160 requirements even for long term surveys. There are many versions of the system depending on the front panel and number of inputs. In order to be connected to a higher processing level, MOT has factory configurable serial interfaces (RS-232 or RS-422/485) implementing a MODBUS protocol. MOT is treated as a MODBUS slave. As it can be seen in Fig. 1, MOT has dedicated internal blocks for:

- power supply;
- input voltage monitoring;
- RMS value recording;
- frequency measurement;
- operator console;
- processing and storage;
- clock and synchronization.

The equipment has dedicated inputs for voltage to be monitored, time synchronization and dedicated bidirectional interfaces for communication.

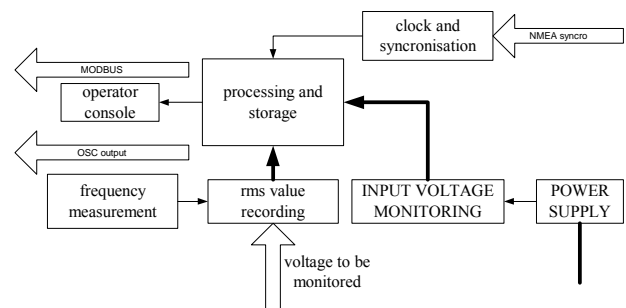


Fig. 2. MOT internal blocks [9]

Besides voltage monitoring according to SR EN50160, the equipment can give on OSC output the RMS value of the voltage and the duration of the half cycle, every half cycle. This secondary function is used for the study presented in this paper. The configuration of the system dedicated is presented in Fig. 2. As it can be seen, there is a Windows XP - PC running a dedicated software module.

The software is custom made and is called VCTest. The link between computer and MOT is made through serial interface. Since the equipment has no dedicated memory

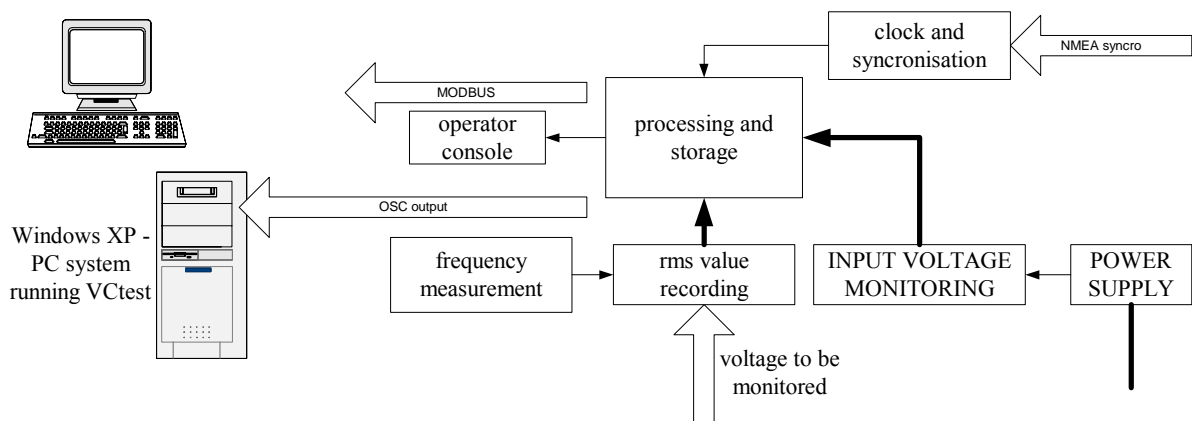


Fig. 1. Running VCTest configuration [9]

for half cycle RMS storage, the computer must be continuously connected during the data acquisition. Fig. 3 presents the location for the measurement points used in the one-year campaign. The three locations are Bucharest, Sibiu and Cluj, relatively far from each other. The data gathered have been stored and then processed. In the present study, results obtained from processing the data collected in Bucharest and Cluj are illustrated and discussed.

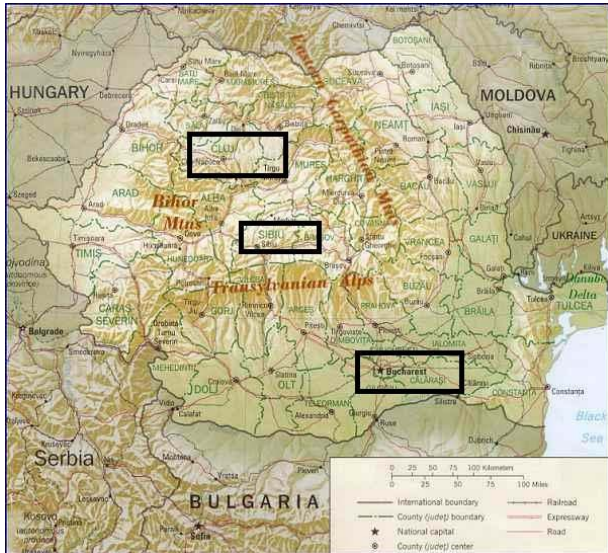


Fig. 3. Measurement points location on the Romania map.

5. Measurement Results

Voltage measurement is an indirect survey of the energy transfer process from producer to customer. As it can be seen in Fig. 4, depending on the balance between production and consumption, during stationary regime, the number of events is not very high when compared to other networks. The stationarity of this process is judged according to the standard [1][5] by the +/-10% limits.

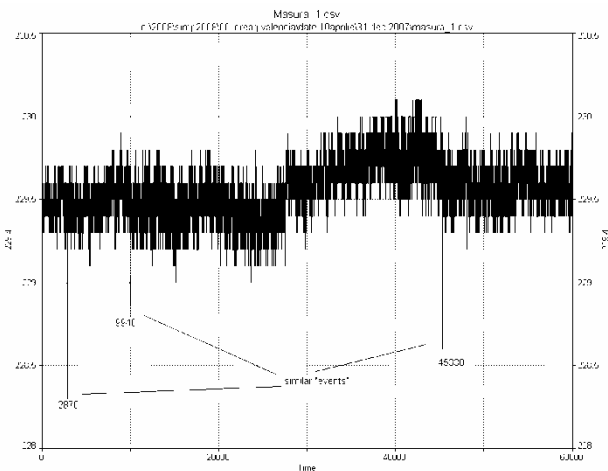


Fig. 4. Example of time variation of the RMS voltages during the measurement campaign [Bucharest, 31st of December 2007 00:00 - 00:10]

In Fig. 4, the 60,000 half cycle measurements performed during 10-minute intervals are presented. Using a subjective method of investigation, it can be seen that around intervals 2870, 9940 and 45330 something happened. In order to get a more detailed image of what happened around these intervals, Table 1 presents the expanded view of the events.

As it can be seen in Table 1, the form of the chart representing voltage over time is similar in all events-located waveforms.

TABLE I - Selected events shape from Fig. 4.

Nb	First "sag"	Second "sag"	Third "sag"
chart			
Explain	first event details	second event details	third event details

This similarity leads to the idea of cluster analysis over the set of data. The purpose is to check the rationale behind the standard limits and to make a distinction between local and regional events. Dedicated software was used in order to carry out clustering analysis.

6. Data processing

Selected data from the huge amount available data as result of the measurement campaign was further processed with statistical routines. The purpose was to build and compare clusters based on different methods. The main idea is to isolate the data containing a sag in order to go further into details with this event. The 60,000 values measured during 10-minute survey were split into 60 vectors of 1000 values each. In this way, each vector describes the measurements for 10 seconds. The criterion used to form the clusters was the single linkage, or "nearest neighbour". The differences between the scenarios are determined by the way the distance between the vectors is calculated. After initial tests, the following distances were used:

- Euclidean distance – this is the most common distance measure. A given pair of cases is plotted on two variables, which form the horizontal and vertical axes. The Euclidean distance is the square root of the sum of the square of the difference on the horizontal axis plus the square of the distance on the vertical axis.
- Chebyshev distance – this is 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 distance.
- Minkowski distance – this is the generalized distance function (1), given by the p -th root of the sum of the absolute differences to the p -th power between the values for the items. Considering two vectors (\mathbf{x}_i and \mathbf{x}_j) of K components each, the Minkowski distance is:

$$d_{ij}^{(p)} = \left(\sum_{k=1}^K (x_{ik} - x_{jk})^p \right)^{1/p} \quad (1)$$

When $p = 2$, the Minkowski distance becomes the Euclidean distance.

In order to determine the best distance definition used for clustering with the objective of separating the sags, a set of dendrograms have been built. Each dendrogram represents in graphical form the evolution of the clustering process for a specific clustering variant. These dendrograms are shown in Fig. 5.

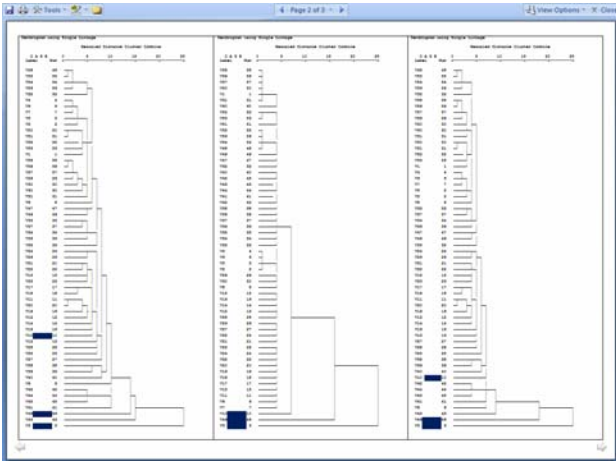


Fig. 5. Example of dendrograms built using different metrics. From the left to the right, the distances used are Euclidean, Chebyshev and Minkowski (with $p = 3$).

Dendrograms show the relative size of the proximity coefficients at which cases were combined. Cases with low distance/high similarity are close to each other. Similar cases are connected with a line linking them at a short distance from the upper part of the dendrogram. This indicates that they are agglomerated into a cluster at a low distance coefficient, indicating alikeness.

When, on the other hand, the linking line is at the bottom of the dendrogram, the linkage occurs with a high distance coefficient, indicating the cases/clusters were agglomerated even though much less alike.

On the dendrograms in Fig. 5, vectors 3, 10 and 46 are marked with a dark line. The same vectors are presented in Fig. 6. These are the vectors containing sags.

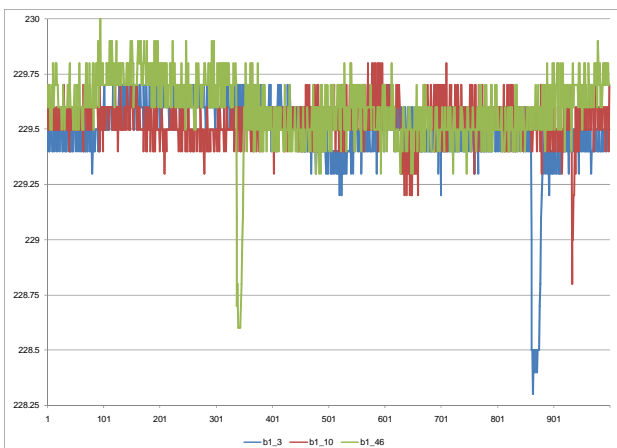


Fig. 6. Clustering target intervals

It appears that the most effective of the three methods to identify a sag for hierarchical clustering with single linkage criterion on the data analysed is the Chebyshev definition of the metric. In this case, the three events containing sags are left as the last vectors to be grouped by the clustering process. This reflects the more effective capability of isolating events with uncommon shape, shown by the use of the Chebyshev distance with reference to the set of data tested.

7. Identification of local and network events

The analysis described in Section 5 was extended to a 30-minute interval for two locations (Bucharest and Cluj). The result of first stage clustering was 18 vectors related to sags. There are 10 events in Bucharest and 8 in Cluj. In order to select local events from network events, a second clustering analysis was conducted. The analysis was lead on the 10-second vectors extracted from the measurement. The analysis pool was completed for every vector with the similar time interval vector in the opposite measuring point. Vector naming rule was:

- the vector name starts with a letter or group of letters identifying the measurement point;
- the number following the letters is the number of the 10-minute measurement interval, followed by an underscore;
- the number after the underscore is the number of the 10-second interval inside the 10-minute interval.

Table 2 presents some selected results of the second clustering process. The distances used were Chebyshev, Euclidean, Minkowski ($p = 3$), Minkowski ($p = 4$). The last vectors remaining in the dendrogram are taken for further analysis. As it can be seen in Table 2, there are chances that B1_3 and CJ1_4 describe an event that can be sensed at the network level.

TABLE 2 - Last vectors in the dendrogram

	clustering distance	v31	v32	v33	v34	v35
1	Chebyshev	B2_29	B1_46	B1_3	CJ1_4	CJ1_59
2	Euclidean	B3_43	B3_48	B3_9	B3_17	CJ1_4
3	Minkowski p=3	B1_4	B2_29	B1_46	B1_3	CJ1_4
4	Minkowski p=4	B3_9	B2_29	B1_46	CJ1_4	B1_3

The problem raised by this result is that the events recorded are not in the same time interval. Considering the adjacent time intervals, it means that we have to accept a time-localization uncertainty in the order of one second.

Fig. 7 presents the B1_3 and CJ1_4 curves registered as it resulted from the second clustering analysis. Despite the fact similarity of the events seems obvious, without a correct estimate of time uncertainty we cannot jump to the conclusion that the events are not local. Time uncertainty could give a fair estimate of the truthfulness of the assumption that the events are related to each other. If we suppose the two sags are related to the same event, as it could be seen in Fig. 7, the sag duration is not

the same. Based on sag recovery time and on further information on the structure of the electrical system, it could be possible to make some assumptions regarding the location of the event between the two measured points. Availability of more measurement locations could provide further information for performing this task.

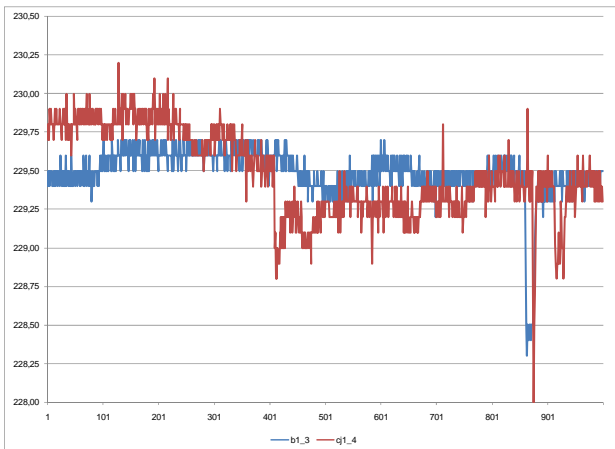


Fig. 7. B1_3 and CJ1_4 voltage measurement

8. Conclusion

After investigating the results of measurements taken on half-cycle voltage data in two locations, the authors have established a method to extract uncommon events using hierarchical clustering. Considering the single linkage variant, the Chebyshev distance has been the best suited distance for the purpose of isolating the events studied. Starting from the results obtained, it has been shown that a second stage clustering analysis conducted on a selected subset of events could lead to the identification of local versus network events. For this purpose, possible information concerning the details of the time instant of the events and the electrical system structure could allow for estimating the location of the events classified as system events and refining the conclusions drawn from the first stage of data analysis.

References

- [1] CENELEC (European Committee for Electrotechnical Standardisation), *Voltage characteristics of electricity supplied by public distribution systems*, European Norm EN 50160, 2003.
- [2] G. Gan, C. Ma and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, 2007.
- [3] M.R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973.
- [4] J.H. Ward, "Hierarchical grouping to optimise an objective function", *Journal Amer. Stat. Assoc.*, Vol. 58, 1963, pp. 236–244.
- [5] SR EN 50160:1998 *Caracteristicile tensiunii furnizate de rețelele publice de distribuție*, 1998.
- [6] IEC 61000-4-30 Ed. 1: *Electromagnetic compatibility (EMC) - Part 4-30: Testing and measurement techniques - Power quality measurement methods*, 2003.
- [7] M. Albu, "Aspects concerning the siting of distorting and non-symmetrical consumers" (in Romanian: Aspecte privind localizarea consumatorilor deformați și nesimetri), *Energetica*, Vol. 50, No. 5, 2002, pp. 216-220.
- [8] M. Albu and G.T. Heydt, "On the Use of RMS Values in Power Quality Assessment", *IEEE Transactions on Power Delivery*, Vol. 18, No. 4, Oct. 2003, pp. 1586-1588.
- [9] MOT Operating Console, <http://www.felix.ro/cd-ice.en/pdf/consola.pdf>.