Cluster Analysis of Half-Cycle Duration Measurements to Classify Local and Network Events

Dan Apetrei, *Member, IEEE,* Gianfranco Chicco, *Senior Member, IEEE*, Petru Postolache, *Member, IEEE*, Nicolae Golovanov, and Mihaela M. Albu, *Senior Member, IEEE*

Abstract **– This paper presents a real case study of the half-cycle duration measurement and classification. The study is based on data gathered at three different low voltage locations. In order to carry out the measurements, a special facility (OSC function) of a custom designed equipment called MOT was used. After the introductory aspects concerning the regulation for frequency measurement, the paper presents the structure of the MOT equipment and illustrates the characteristics of the data gathered. These data are then analysed through a hierarchical clustering method to identify and classify the events. The results obtained and the main ideas for future developments are illustrated and discussed in the paper.**

Index Terms **— half cycle time duration, frequency, hierarchical clustering, classification, local event, network event.**

I. Introduction

Frequency is a measure of the number of occurrences of a repeating event per unit time. According to IEV (International Electrotechnical Vocabulary) [1], the definition (101-14-08) states that frequency is the reciprocal of the period. The same source defines period (101-14-07) as the smallest difference between two values of the independent variable at which the values of a periodic quantity are identically repeated. Notwithstanding these simple definitions, the assessment of frequency from measured samples is generally not trivial, since even small differences in the system frequency can be relevant for the operation of some components, e.g., for synchronization purposes.

The system frequency varies as load and generation change. Increasing the mechanical input power to a synchronous generator will not greatly affect the system frequency but will produce more electric power from that unit. Short-time frequency variations are more pronounced during severe disturbances in the power system, while frequency dynamics depending on thermal processes can be experienced in the long-term.

Frequency protection relays may sense the decline of frequency in the power system and automatically initiate load shedding or tripping of interconnection lines, to preserve the operation of at least part of the network. Specific protections can be equipped with volt/hertz relays. Small frequency deviations (i.e., 0.5 Hz on a 50 Hz network) could result in automatic load shedding or other control actions to restore the system frequency.

Smaller power systems, not extensively interconnected with many generators and loads, are unable to maintain frequency with the same degree of accuracy. Where system frequency is not tightly regulated during heavy load periods, the system operators may allow system frequency to rise during periods of light load, to maintain a daily average frequency of acceptable accuracy.

Frequency measurement is one of the most important sources of information in electrical systems. Nowadays, the volume of data that can be gathered by using specific analyzers is higher and higher. For this reason, new algorithms for data processing are necessary. Among these algorithms, different types of clustering methods can be applied. In particular, hierarchical clustering has proven to be effective in several applications to time series classification [2,3]. Hierarchical clustering is applied in this paper to classify the events occurred in the network on the basis of the half-cycle duration measured at different locations.

II. Frequency Measurement

Frequency as characteristic of the voltage waveform is regulated by various standards, like EN 50160 [4] and IEC 61000-4-30 [5,6]. These standards combine threshold prescription and measurement methods.

The standard EN 50160 provides requirements for the main characteristics of the voltage at the customer's supply terminals in public Low Voltage and Medium Voltage electricity distribution systems under normal operating conditions. The objective of the standard is to define and describe the characteristics of the supply voltage concerning frequency, magnitude, waveform and symmetry of the three-phase voltages. Regarding frequency, the values measured under normal conditions during a 10 s interval have to remain within:

- 50 Hz ± 1% (i.e., 49.5 Hz-50.5 Hz) during 99.5% of a year;
- 50 Hz +4%̃/6% (i.e., 47 Hz-52 Hz) during 100% of time.

D. Apetrei is with SC Electrica SA, str. Grigore Alexandrescu, nr. 9, sector 1, Bucureşti, Romania, tel. 2085250 (e-mail: dan.apetrei@electrica.ro).

G. Chicco is with Politecnico di Torino, Dipartimento di Ingegneria Elettrica, corso Duca degli Abruzzi 24, 10129 Torino, Italy (e-mail gianfranco.chicco@polito.it).

P. Postolache and N. Golovanov are with Universitatea Politehnica din Bucureşti, Facultatea Energetica, Splaiul Independenţei nr.313, sector 6, Bucureşti, Romania, tel. 4029482 (e-mail petrupostolache@yahoo.com, nicolae_golovanov@yahoo.com).

M.M. Albu is with Universitatea Politehnica din Bucureşti, Department of Electrical Engineering, Splaiul Independenţei 313, Bucureşti, Romania (e-mail albu@ieee.org).

978-1-4244-2235-7/09/$25.00 ©2009 IEEE

These characteristics are subject to variations during normal operation of a supply system with changing load and generation, as well as for disturbances generated by certain equipment and the occurrence of faults mainly caused by external events. The Annex A of the standard gives a comment about 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 contrast to normal products, application is one of the main factors which influence the variation of the "characteristics".

The standard IEC 61000-4-30 defines the methods for measurement and interpretation of the results of power quality assessment for different 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 requirements for this type of measurement are set up by defining two classes of performance for voltage measurement, namely, *class A* and *class B*.

For *class A* measurements, the frequency reading shall be obtained every 10 s. As power frequency may not be exactly 50 Hz within the 10 s time clock interval, the number of cycles may not be an integer number. The fundamental frequency output is the ratio of the number of integral cycles counted during the 10 s time clock interval, divided by the cumulative duration of the integer cycles. Before each assessment, harmonics and interharmonics shall be attenuated to minimize the effects of multiple zero crossings. The measurement time intervals shall be non-overlapping. Individual cycles that overlap the 10 s time clock are discarded. Each 10 s interval shall begin on an absolute 10 s time clock, ±20 ms for 50 Hz. Over the range of influence quantities, and under the conditions described in [5], the measurement uncertainty Δf shall not exceed ±10 mHz.

For *class B* performance, the manufacturer shall indicate the process used for frequency measurement. The manufacturer shall specify the uncertainty Δf over the range of influence quantities. Measurement time intervals are aggregated over 3 different time intervals. The aggregation time intervals are 3 s interval (150 cycles for 50 Hz nominal), 10 min interval, 2 h interval. The aggregations are performed by using the square root of the arithmetic mean of the squared input values. Three different 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" 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 [5], 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. To begin the 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*: the data for the “2-h interval” shall be aggregated from twelve 10 min intervals.

Specific comments and results will be provided in accordance to the second edition of the IEC 61000-4-30 [6], under publication, that will replace the first edition published in 2003. The new edition introduces significant technical changes concerning the adjustments, clarifications, and corrections to class A and class B measurement methods, and adds a new category (class S) for survey instruments and a new Annex C providing guidance on instruments.

III. Measurement System Description

The MOT-103B/BG equipment [7] was used for the measurements. This is a custom designed measurement system, made in Romania. It was designed to meet the EN 50160 requirements even for long-term surveys. There are four versions of the equipment 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) treated with the MODBUS protocol. MOT is treated as a MODBUS slave. As could be seen in Fig. 1, the MOT has dedicated internal blocks for power supply, input voltage monitoring, RMS value recording, frequency measurement, operator console, processing and storage, clock and synchronization.

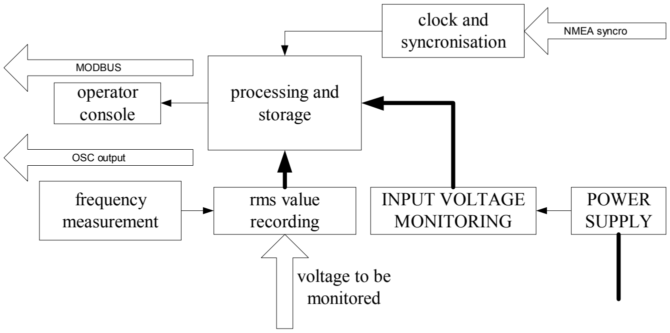


Fig. 1. MOT internal blocks.

The equipment has dedicated inputs for voltage to be monitored, time synchronization and dedicated bidirectional interfaces for communication. Besides voltage monitoring according to the standard EN50160, the equipment can give on the OSC output the RMS value of the voltage and the duration of the half-cycle, for every half cycle. This secondary function is used for the study presented in this paper. The system configuration is presented in Fig. 2. As it can be seen, there is a Windows XP - PC running a dedicated piece of software, called VCtest. This software links the computer and the MOT through a serial interface. Since the equipment has no dedicated memory for half-cycle measurement storage, the computer must be continuously connected during the data acquisition process.

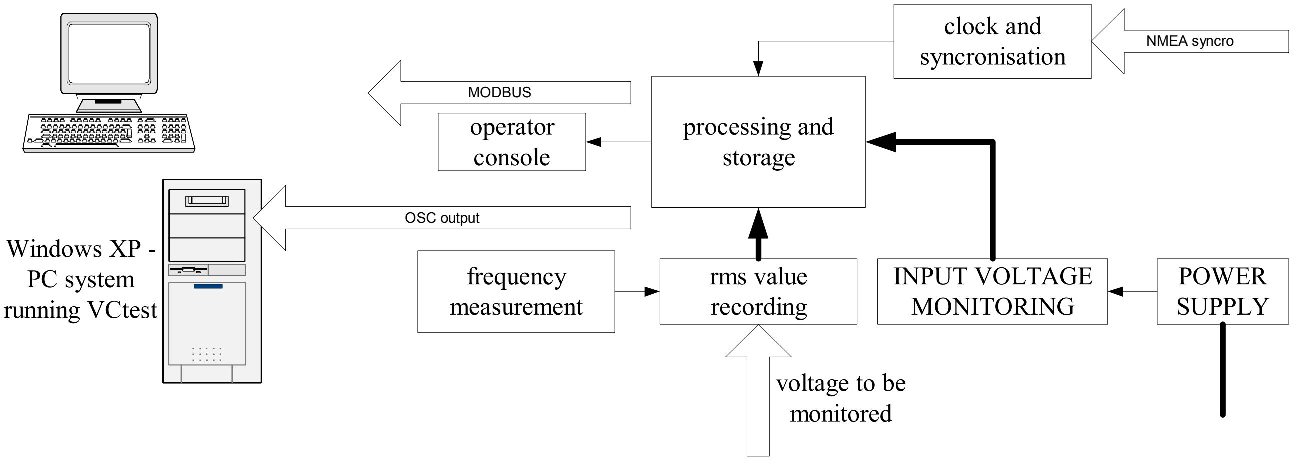


Fig. 2. VCtest configuration.

IV. Hierarchical Clustering

Let us consider a set of M patterns, each one containing H data points. Let us represent each pattern by using a vector $\mathbf{x}^{(m)}$ = $\{x_h^{(m)}$, h = 1,…,$H\}$, for m = 1,..., M. The hierarchical clustering procedure starts from considering each pattern as a separate cluster. Then, the procedure groups together the pair of patterns exhibiting the highest similarity according to a predefined metric, thus reducing the number of clusters by one. The procedure continues until a predefined number of clusters is reached, or until all the patterns are grouped into a single cluster. A pictorial view of the evolution of the hierarchical clustering process can be obtained by drawing a graph representing the hierarchical tree showing the successive grouping of the patterns at each step of the clustering procedure, called *dendrogram* [8].

The similarity measure that can be adopted depends on the notion of distance defined by the user. The Euclidean distance is a typical choice in most applications. However, in this paper, we consider the possibility of applying some non-Euclidean metrics. In particular, the set of distances addressed is those of the so-called Minkowski distances of order $p \geq 1$ (or p-norm distances), defined as follows, taking two vectors $\mathbf{x}$ and $\mathbf{y}$, each of which having H components:

$$d^{(p)}(\mathbf{y},\mathbf{x})=\sqrt[p]{\frac{1}{H}\sum_{h=1}^{H}(y_h-x_h)^p} \qquad (1)$$

The cases considered include the Manhattan distance (p = 1), the Euclidean distance (p = 2), as well as two further cases with p = 3 and p = 4. In addition, the set of distances is completed with the *normalized Euclidean distance*, defined as:

$$\tilde{d}(\mathbf{y},\mathbf{x})=\sqrt{\frac{1}{H}\sum_{h=1}^{H}\left(\frac{y_h-x_h}{\sigma_k}\right)^2} \qquad (2)$$

where σ_k is the standard deviation of the k-th component of the data evaluated over the sample set.

In hierarchical clustering, the similarity between the patterns is addressed by building at each step of the clustering procedure a square matrix with dimensions equal to the number of clusters existing at that step of the procedure. The entries of the similarity matrix are calculated by using the chosen notion of distance. The entries of the similarity matrix are then used to identify the most suitable pair of clusters to be grouped, according to a user-defined *linkage criterion* [8]. The criteria tested in this paper are the *single* linkage (or nearest neighbour), *complete* linkage (or farthest neighbour), *average* linkage, and *Ward* linkage.

During the construction of the dendrogram, the distance between the two clusters $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be grouped together, called *cophenetic distance*, is used to represent the height of the corresponding link in the hierarchical tree. This distance can be expressed as the average distance between all pairs of patterns of the two sets [9], namely, $\mathbf{x}_w$ of the set $\boldsymbol{X}$ (assumed to contain W patterns) and $\mathbf{y}_j$ of the set $\boldsymbol{Y}$ (containing J patterns):

$$d(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{W\,J}\sum_{w=1}^{W}\sum_{j=1}^{J}d\left(\mathbf{x}_w,\mathbf{y}_j\right) \qquad (3)$$

In order to assist the comparison among the various types of metrics, a widely used indicator specifically applied to hierarchical clustering is the *cophenetic correlation factor* [8]. Given the dendrogram resulting from the hierarchical clustering process, the cophenetic correlation factor χ is the linear correlation coefficient between the cophenetic distances calculated in the tree and the distances among the initial patterns used in the clustering process. Values of χ close to unity represent high correlation between the original distances and the cophenetic distances, and as such indicate effective performance of the hierarchical clustering process.

V. Measurement Results

Fig. 3 indicates the measurement points. Recordings were made at three locations (Bucharest, Cluj and Sibiu) during one year. The results were stored and then processed. Data analysis to get the half-cycle time duration values is not easy, since generally the waveforms at low voltage points are relatively more distorted than the ones at high voltage points.

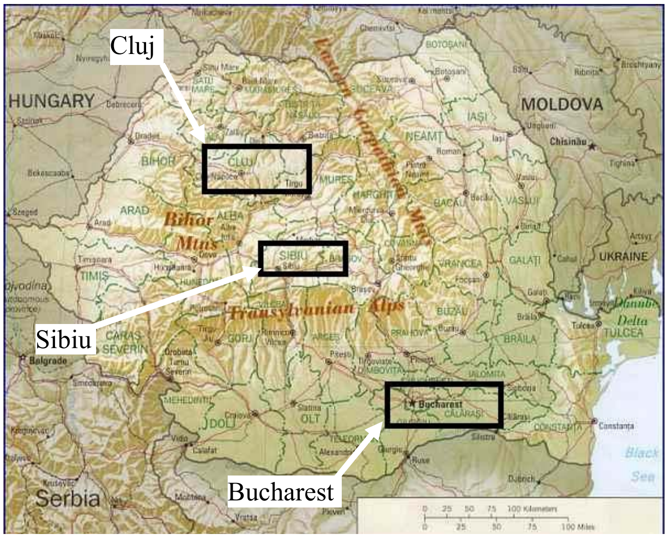


Fig. 3. Measurement points location.

The half-cycle data used in this application have been gathered from the three locations on the same time period (starting at hour 0:00) on Monday, January 14, 2008. More specifically, the whole amount of data includes 180,000 successive points for each location. For the purpose of data processing, this amount of data has been partitioned into data sets of 60,000 points each, labelled as data sets 1, 2 and 3, respectively. The synthetic labelling of these data is constructed by using the initial letter of the city and the number of the data set, so that the data gathered in Bucharest are labelled as B1, B2 and B3 (Fig. 4), whereas the data gathered in Cluj are labelled as C1, C2 and C3 (Fig. 5), and finally the data gathered in Sibiu are labelled as S1, S2 and S3 (Fig. 6).

However, because of lack of synchronization among the measurements it is not possible to associate the points of the sets of data at the different locations exactly, nor to proceed by difference of the measured groups of data in the various sites in order to endeavour the occurrence of local events. Thus, the clustering procedure is run in order to recognize the anomalous events and to enable the assessment of their local-based or system-based nature. This is done by constructing specific patterns representing a relatively significant duration in time (3 seconds), thus proceeding to the classification of the events by isolating the uncommon situations through clustering in each of the data sets. Then, the possible similarity among particular events occurred in the same 3 s time period at the various locations, singled out or separated by clustering at the various locations, could be further investigated to check whether they can represent the effects of events at the network-scale level.

The patterns to be classified have been formed in such a way that each pattern corresponds to a duration of 3 s, and as such is composed of 300 half-cycle time duration values. Thus, the 60,000 points of any data set are partitioned into 200 resulting patterns. The nature of these patterns is partially affected by the discretization error in the measurements due to the accuracy of the digital recording (for which the minimum step of variation is 1 μs, so that the information is coded with relatively large steps of variation from one discrete level to another). In order to mitigate this source of inaccuracy, the measured values could be grouped together by averaging them on a predefined time interval, for instance at each 6 successive half cycles (60 ms). This would also make the operation of the clustering procedures easier. However, in order to capture possible short-duration effects more effectively, in this application the data have been maintained in their original format.

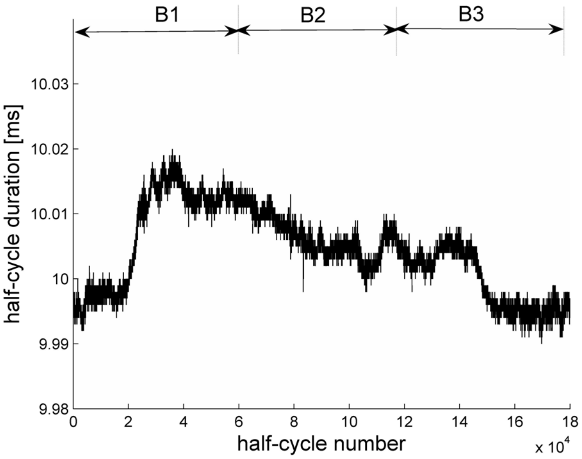


Fig. 4. Data gathered in Bucharest.

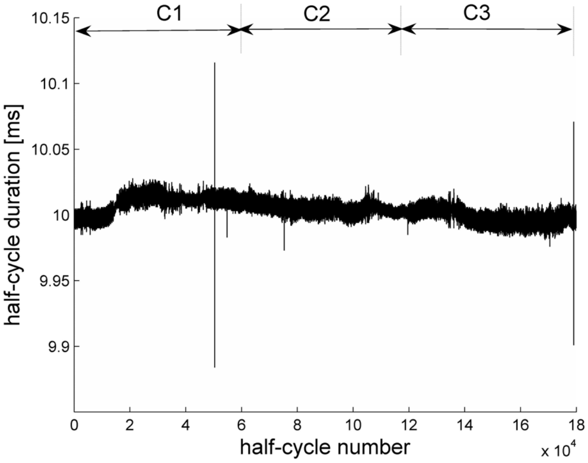


Fig. 5. Data gathered in Cluj.

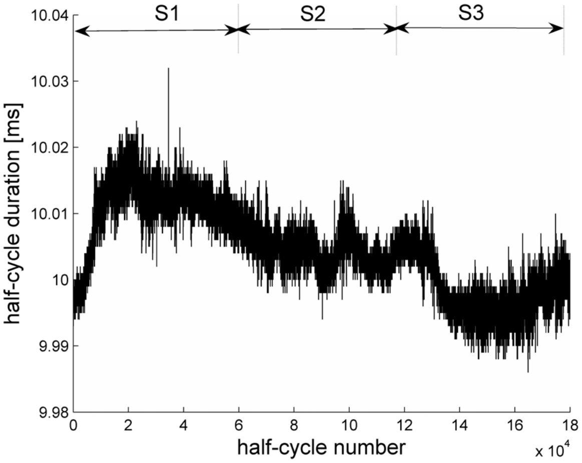


Fig. 6. Data gathered in Sibiu.

VI. Data Processing and Clustering Outcomes

Cluster partitioning has been obtained by applying the hierarchical clustering algorithm with different distances and different linkage criteria [8] in a comparative analysis. The cophenetic correlation factor (shortly *cophenetic index*) has been used to rank the clustering procedures. The comparison is consistent when the clustering procedures operate on the same data set. Table I reports a synthetic view of the results obtained on the nine data sets. The indication emerging from all these plots is that the average linkage criterion is the most suitable one, while the single linkage criterion proves to be in general the least efficient one. In addition, the Minkowski distance with $p = 4$ presents a very interesting performance in most cases with the average linkage criterion. These results enable us to consider investigating non-Euclidean distances for setting up the hierarchical clustering procedures with the set of data under analysis, going beyond the classical view of Euclidean distances. The average linkage criterion with Minkowski distance will be used in the sequel for presenting more detailed aspects of the clustering process and results.

As an example, Fig. 7 and Fig. 8 show the dendrograms obtained for the data sets C1 and C2, respectively. The higher cophenetic index and the convey the information that cluster partitioning in the data set C1 is simpler to be found than in the data set C2. In other words, it is expected to be easier to single out uncommon patterns in the data set C1 with respect to C2. Indeed, dealing with a huge amount of data divided into segments of 3 seconds each for clustering purposes, this suggestion can be useful to understand in which segments one or more uncommon events are more likely to be found.

Another reason for getting relatively high values of the cophenetic index is the presence of data that can be easily merged into separated groups, as it happens in the Bucharest data, especially in the data sets B1 and B3, in which there is a remarkable variation of the half-cycle duration level during the three seconds, even though this fact does not indicate the presence of uncommon events (Fig. 9).

It can be noticed that the above considerations can be indeed drawn from the structures of the clustering procedures, that can be represented by the corresponding cophenetic indices and, for visual purposes, by the dendrograms, even before or without setting up the final number of clusters of possible interest. In fact, the clustering procedure should be able to single out the uncommon events of interest very easily, even by setting up a relatively low number of final clusters. In other terms, following the dendrogram from the top to the bottom, one can guess how many uncommon events are likely to be, and how much the patterns representing these events can be different with respect to the other patterns: the more the cophenetic distance representing a vertical link is high, the more the pattern or patterns representing a specific event clustered with others at the corresponding link should exhibit an individual and distinct shape with respect to the others.

In order to verify the practical applicability of the above concepts, Fig. 10 and Fig. 11 show the clustering results corresponding to the data sets C1 and C2 and to the dendrograms of Fig. 7 and Fig. 8, respectively. The same number of final clusters $C = 16$ is considered.

Indeed, in the data set C1 it is possible to find at least two events affecting the pattern shape at the macroscopic level, one of which is clearly evident and corresponds to the first event singled out in the dendrogram of Fig. 7.

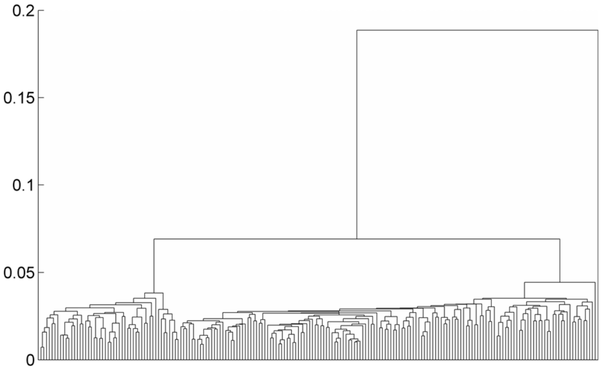


Fig. 7. Dendrogram for the C1 data set (average linkage criterion, Minkowski distance with $p = 4$). Horizontal axis: initial patterns; vertical axis: the heights of the link in the hierarchical tree are the cophenetic distances.

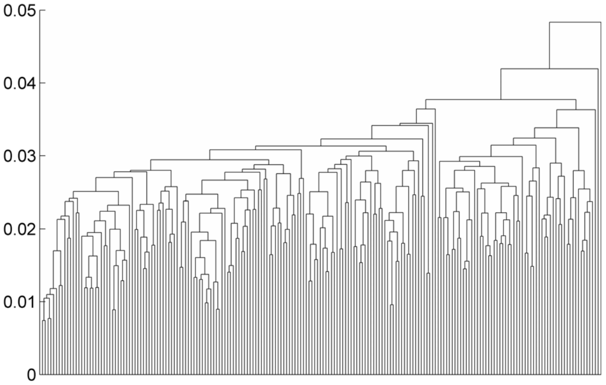


Fig. 8. Dendrogram for the C2 data set (average linkage criterion, Minkowski distance with $p = 4$). Horizontal axis: initial patterns; vertical axis: the heights of the link in the hierarchical tree are the cophenetic distances.

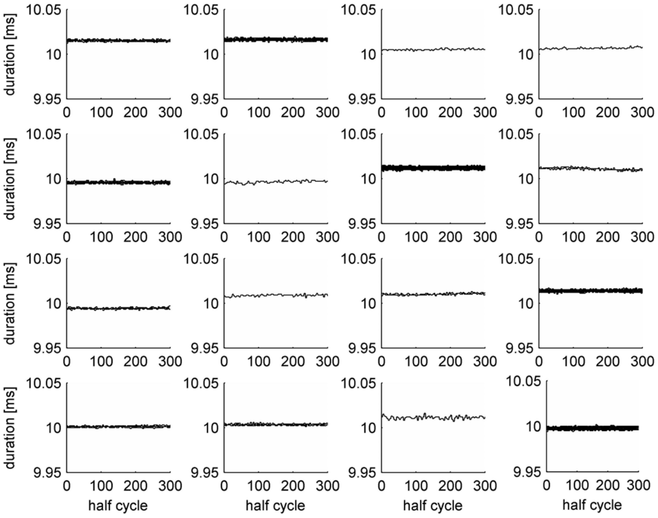
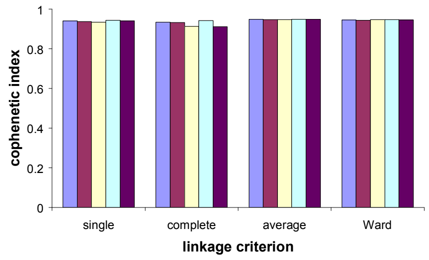
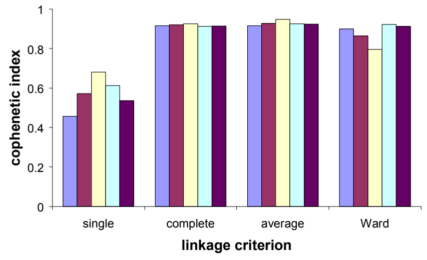
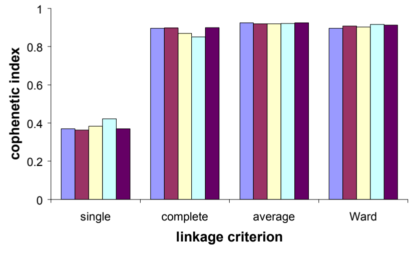
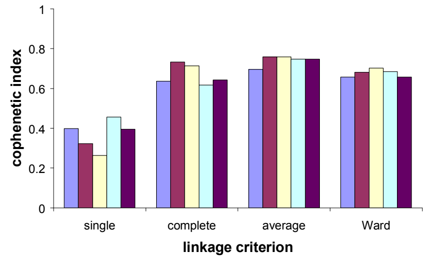
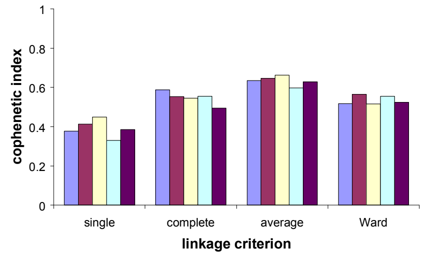
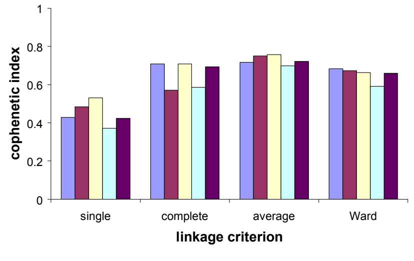
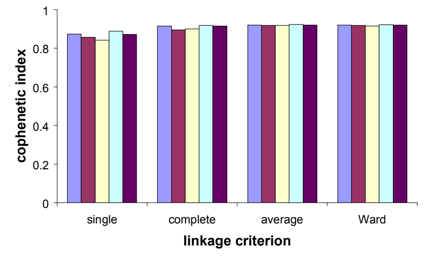
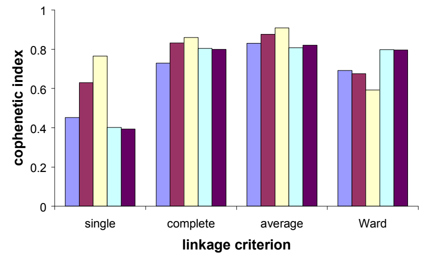
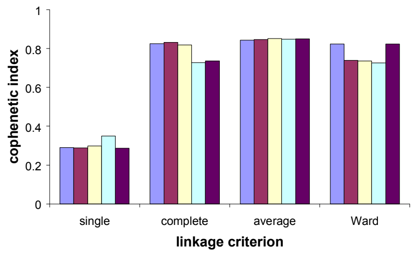


Fig. 9. Clustering results for the B1 data set (average linkage criterion, Minkowski distance with $p = 4$), $C = 16$ clusters.

From the clustering results and pattern labelling, it is immediate to recognize the exact period in which the uncommon events occurred. In particular, the event indicated in the last pattern of Fig. 10 (truncated on the vertical side to the common vertical limits) corresponds to the with respect to

TABLE I
VALUES OF THE COPHENETIC INDEX FOR DIFFERENT DATA SETS AND LOCATIONS

data set	Bucharest	Cluj	Sibiu
1			
2			
3			

▨ Euclidean ▨ Minkowski (3) □ Minkowski (4) □ Manhattan ■ Std Euclidean

large spikes indicated in Fig. 5, and could be classified a bad data for an overall analysis, while the other event is of local type (no corresponding similar pattern exists in the data sets B1 and S1 for the same time period). This kind of analysis has been carried out for the various available data sets.

About global events, from Fig. 5 it can be seen that at the macroscopic level the evolution of the half-cycle duration over time follows a consistent behaviour. This behaviour could be recognised averaging the original data for longer time periods, and repeating the cluster analysis.

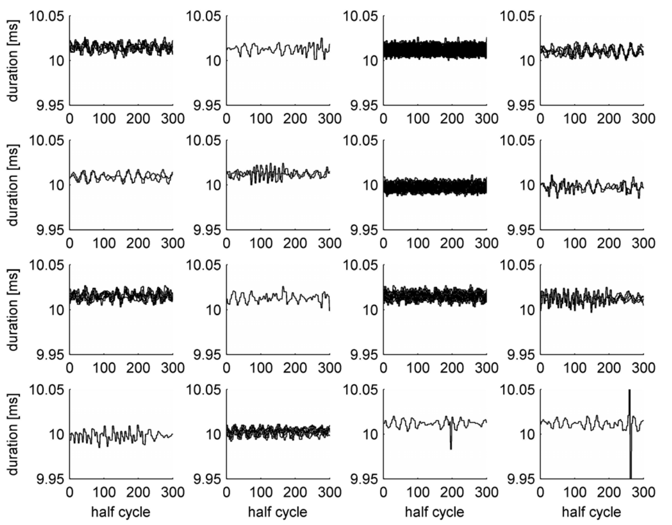


Fig. 10. Clustering results for the C1 data set (average linkage criterion, Minkowski distance with $p = 4$), $C = 16$ clusters.

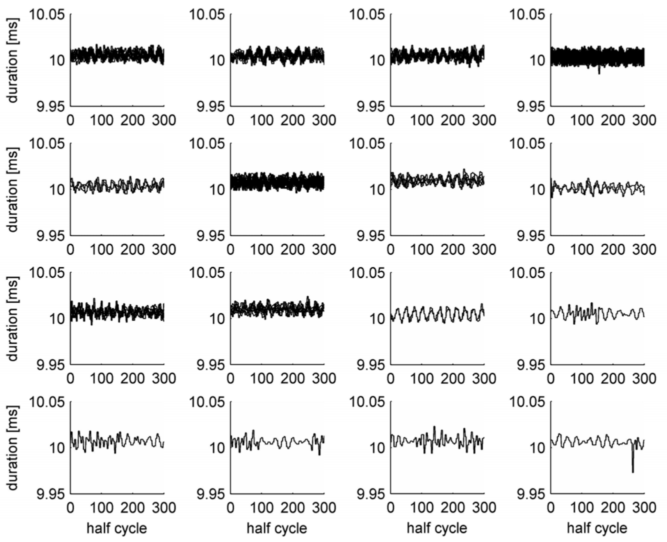


Fig. 11. Clustering results for the C2 data set (average linkage criterion, Minkowski distance with $p = 4$), $C = 16$ clusters.

VII. Conclusions

This paper has presented an application of the hierarchical clustering procedure with different types of metrics to classify half-cycle duration data. The half-cycle values have been used to detect short-term events in a large amount of data gathered from the field. Clustering procedures have been used to analyse the data, because lack of automatic synchronization among the measurement makes it possible that related events are located in instants with slightly different time labelling gathered at the different locations.

The clustering results confirm the good properties of hierarchical clustering in grouping together consistent patterns and isolating the uncommon patterns, even within a large group of initial patterns. Moreover, the use of non-Euclidean metrics based on the Minkowski distance ($p = 4$) has proven to be effective. The results shown are a small sample of the total recordings available. Further tests are in progress to extend the analysis to all the gathered data.

By using the approach outlined in this paper, it could be possible to detect events at relatively longer time scales. For this purpose, the original data could be pre-processed by averaging them on given time intervals, then repeating the same clustering procedure for identifying and classifying the events. Indeed, the proposed clustering procedure and the metrics adopted are a viable tool for performing studies on the measured data on any time scale, provided that the original data are suitably averaged to reduce the size of the data set. More detailed results in this respect will be presented in the near future.

VIII. References

[1] International Electrotechnical Commission (IEC), Electropedia: The World's Online Electrotechnical Vocabulary, IEC 60050, web site http://www.electropedia.org/. (available March 2009).

[2] G. Chicco, R. Napoli, F. Piglione, M. Scutariu, P. Postolache and C. Toader, Emergent Electricity Customer Classification, *IEE Proc. Gener. Transm. and Distrib.* 152 (2), March 2005, 164-172.

[3] G. Chicco, R. Napoli, F. Piglione, P. Postolache, M. Scutariu and C. Toader, Load pattern-based classification of electricity customers, *IEEE Trans. on Power Systems* 19 (2), May 2004, 1232-1239.

[4] CENELEC (European Committee for Electrotechnical Standardisation), *Voltage characteristics of electricity supplied by public distribution systems*, European Norm EN 50160, 2003.

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

[6] Project IEC 61000-4-30 Ed. 2.0, Electromagnetic compatibility (EMC) - Part 4-30: Testing and measurement techniques - Power quality measurement methods, 2008.

[7] MOT Operating Console, http://www.felix.ro/cd-ice.en/pdf/consola.pdf.

[8] R.R. Sokal and F.J. Rohlf, The comparison of dendrograms by objective methods, *Taxon.* 11, 1962, 33-40.

[9] M.R. Anderberg, *Cluster Analysis for Applications*, Academic Press, New York, 1973.

[10] G. Chicco, R. Napoli and F. Piglione, Comparison among Clustering Techniques for Electricity Customer Classification, *IEEE Trans. on Power Systems* 21 (2), May 2006, 933-940.