Paper accepted for presentation at 2009 IEEE Bucharest Power Tech Conference, June 28th - July 2nd, Bucharest, Romania

Missing Data Treatment of the Load Profiles in Distribution Networks

G. Grigoras, G. Cartina, *Member*, *IEEE*, E.C. Bobric, *Member, IEEE*, and C. Barbulescu

***Abstract*--In distribution systems, the determination of the load is relatively simple when measurements are available. Frequently, due to various causes such as metering and transmission equipment failures, data are missing for part or all of a day. In these cases, estimation a "corrected" value must be made. The paper presents two methods (k-Nearest Neighbors, (kNNs) and Clustering methods) for treatment of missing data problems in electric distribution networks. The numerical results indicate that the methods are efficient in the estimation of the missing load values for electric distribution stations.**

***Index Terms*--clustering method, distribution networks, fuzzy kNNs method, load profile, missing data**

I. Introduction

LOAD monitoring in distribution systems is a very imported problem. The readings of the measurement points arrive at the data retrieval centre by telemetry, at regular intervals. This system of data retrieval is subject to the occurrence of irregularities that can be caused by erroneous information transmission, errors in data consolidation or unexpected load behavior (blackout). These irregularities can mainly be described as, [5]:

- Occurrence of missing values: when there is no load-value reading during an interval time.
- Occurrence of outliers: when the registered value is nonsense, outside the range of expected behavior. This can be due, for example, to measurement failure at some points or inherent problems in the information sending process. This category also includes real values corresponding to unexpected events, such as load failures or blackouts.

These irregularities can impair the performance of State Estimation. Thus, before carrying through the load modeling, a treatment module must be developed to correct for missing values and outliers. The treatment module can have two main functions: detection of outliers, declaring them to be missing values, and the estimation of all missing values. Once the values have been declared to be missing, these values must be substituted by values in conformity to the dynamics of the series, [5].

Finally, this paper presents implementation of two methods (k-Nearest Neighbors (kNNs) and Clustering methods) for missing values estimation of the electric load.

II. Classical Analysis Methods for Missing Data Treatment

The missing data from electric distribution systems can have various causes: recording error, variable out of range, metering/transmission system failures etc. Different types of missingness demand different treatment.

The missing data requires consideration of additional issues. In many instances, identifying variables that explain the cause of missing data can help to mitigate the bias. Such explanatory variables are known as "mechanism" variables, and by including them in our models, we can eliminate the bias caused by some types of missing data. If all the mechanism variables associated with a particular piece of missing data can be identified and included in a model as controls, the impact of the missing data can be statistically adjusted to the point where it is negligible. In practice, however, it is extremely unlikely that mechanism variables can be identified for all cases of missing data, [6].

If we want to analyze a data set with incomplete records (missing data), we may do it in two ways. In the first case, we build a complete data set, and then we apply traditional analysis methods. Second, we use o procedure, which can internally manage missing data.

The easiest way to obtain a complete data set from an incomplete one is to erase missing items. This conventional method is used for its simplicity. We do not want to lose data; we may try to guess missing items. This process is generally called imputation. In fact, usually missing values depend on other values, and if we find a correlation between two variables, we may use it to impute missing items. We may use different kinds of information to impute missing data. We may analyze the behavior of all the other records on the missing variable (global imputation based on missing data), or we may try to find a relationship between other variables and the missing one in all the other records, and use this relationship to impute the missing value (global imputation based on non-missing data), or we may look for similar records to see how they behave on the non-missing data (local imputation), [4], [6], [7].

III. k-Nearest Neighbors Method

In the last years, the techniques based on machine learning such as Artificial Neural Networks (ANN) have been applied

G. Grigoras is with the Department of Power Systems, "Gh. Asachi" Technical University, Iasi, Romania (e-mail:ggrigor@ee.tuaiasi.ro).

G. Cartina is with the Department of Power Systems, "Gh. Asachi" Technical University, Iasi, Romania (e-mail:gcartina@ee.tuaiasi.ro).

E.C. Bobric is with the Department of Electrical Engineering, "Stefan cel Mare" University, Suceava, Romania (e-mail:crengutab@ee.tuaiasi.ro).

D. Barbulescu is with the Transelectrica S.A., Bucuresti, Romania (e-mail: cecilia.barbulescu@ee.tuaiasi.ro).

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

to load estimation. The ANNs are trained to learn the relationships between the input variables (mainly preceding loads) and historical load patterns. The main disadvantage of ANNs is the required learning procedure.

Several papers have been published on the application of nearest neighbors' techniques to the electricity market price forecasting but there are no applications to missing load values, [3].

In this method, the missing values of a record are imputed considering a given number of records that are most similar to the record of interest.

The k-Nearest Neighbors (kNNs) method separates the dataset into an incomplete and a complete set that has or has no missing values respectively. The records in the incomplete set are imputed by the order of missing rate. The algorithm is as follows, [7]:

1. Divide the data set D into two parts. Let D_m be the set containing the records in which at least one of the features is missing. The remaining records will complete feature information form a set called D_c.
2. For each vector X in D_m:
   - Divide the vector X into observed and missing parts as $X = [X_o; X_m]$.
   - Calculate the distance between the X_o and all the vectors from the set D_c. Use only those features in the vectors from the complete set D_c, which are observed in the vector X.

Use the K closest vectors (K-nearest neighbors) and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes, replace the missing value using the mean value of the attribute in the k nearest neighborhood. The median could be used instead of the mean value.

IV. Using the Clustering Techniques in Fuzzy Modeling

Clustering can be considered the most important unsupervised learning problem. The scope of clustering is to determine the intrinsic grouping in a set of unlabeled data. A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters [8]-[10].

Clustering algorithms may be classified in: exclusive clustering, overlapping clustering, hierarchical clustering and probabilistic clustering, [8]-[10].

In the first case, data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. The overlapping clustering uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. The hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations, it reaches the final clusters wanted. The last kind of clustering uses a probabilistic approach.

Membership functions defining with the clustering techniques, in the case of the set of two dimensional objects is presented in Fig. 1.

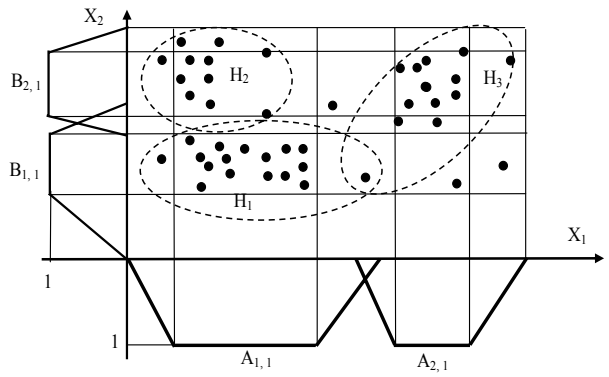

Fig. 1. Defining of the membership function using clustering techniques

Clustering is useful in several exploratory pattern-analysis, grouping, decision-making, and machine-learning situations, including data mining, document retrieval, image segmentation, and pattern classification.

V. Study Case

For the evaluation of the performance of the presented methods, we used the records of 58 load profiles of an electric station (20 February – 19 April period). The time interval of sampling load profile is one hour. Thus, the load profile is represented by 24 load values. We present some hypotheses concerning the time period of consecutive missing values:

- Case I – 3 consecutive missing values;
- Case II – 5 consecutive missing values;
- Case III - 9 consecutive missing values.

There are several ways to express the accuracy of the estimates, depending on the data available. If the actual value of the estimated quantity is available (such as during method development and testing), the following quantity can be useful to verify the method:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (1)$$

where y_i and $\hat{y}_i$ are the actual and, respectively, estimated values, and n represents the number of estimated values.

The mean absolute percentage error (MAPE) from (1) is dimensionless and thus it can be used to compare the accuracy of the model on different data sets.

A. Missing Data Treatment Based on kNNs Method

For the application of the kNNs method, the period February 20– March 31 was used as the training set for the determination of the optimal number of neighbors. The period of April 1 – 19, 2005 has been chosen as a test set for the validation of the proposed method.

For the considered training set, to evaluate the similarity between a certain day and the historical data, Euclidean distance is used. The obtained results for the training set show that the optimal number of neighbors is equal to four using the Euclidean distance.

Figs. 2 - 4 show the real and estimated load for the April 18 2005, time period of consecutives missing values has 3 hours (19-21), 5 hours (18-22), respectively 9 hours (16–24). Note

that the estimated errors for the period that has 9 missing values are larger during low consumption hours. However, it is more important to obtain an accurate estimation during peak hours (Figs. 2 - 4) because the electric energy is more expensive during these hours.

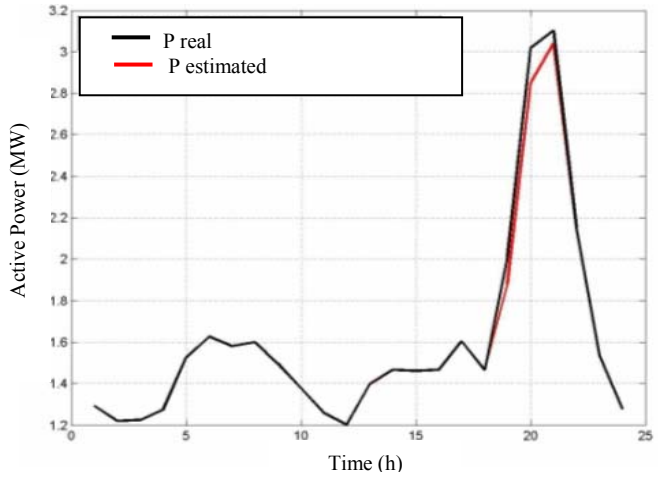

Fig. 2. The real and estimated load for 18 April 2005 (3 consecutives missing values 19-21, MAPE=3.97 %)

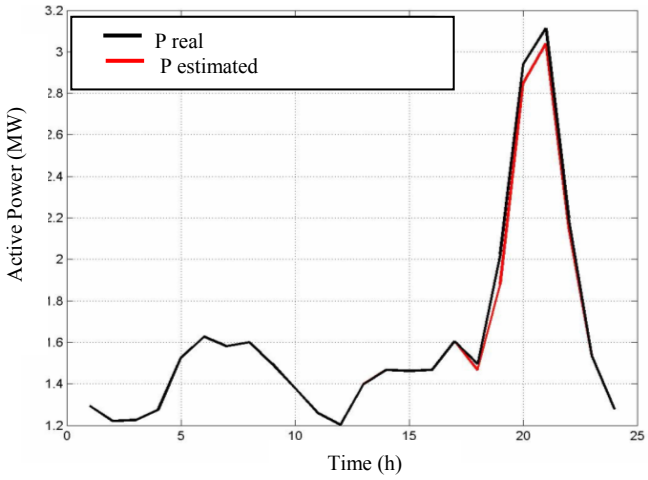

Fig. 3. The real and estimated load for 18 April 2005 (5 consecutives missing values 18-22, MAPE=4.21 %)

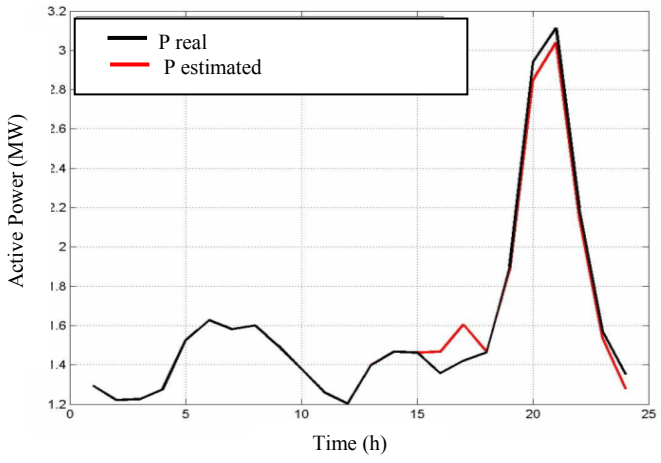

Fig. 4. The real and estimated load for 18 April (9 consecutives missing values 16-24, APE=5.99%)

B. Missing Data Treatment with Fuzzy Typical Load Profiles

In this method, the imputation comes from other records in the same database. For this purpose, the use of the hierarchic clustering method conjunctively with fuzzy techniques is applied to classify load profiles of an electric station into coherent groups – typical load profiles (TLPs), [10].

The steps to obtain the TLPs are:

- all gathered measurements are preprocessed by sorting them and normalized using a suitable normalizing factor (average power, peak power or energy over the surveyed period).
- a clustering method is used to group the normalized load profiles in the clusters. After aggregation of loads of each cluster, the TLPs are determined. The TLP for each cluster is obtained by averaging the values for each hour.
- assignation of the typical load profile to each node according to its type is performed.

The shape of TLPs is influenced by the type of day or season of the year. By knowing typical load profile, one can estimate the missing load values from an electric distribution station.

A clustering method (average method) was applied to the database corresponding to a set of 58 load profiles of an electric station (20 February – 19 April period). The time interval of sampling the load profile is one hour. Thus, the load profile is represented by 24 load values. The load profiles were normalized relative to the energy consumption of each day during the considered period. The hierarchical tree was plotted to determine the threshold where to cut the tree into groups, Fig. 5, where the units along the horizontal axis represent the objects in the original data set. The links between objects are represented on the vertical axis.

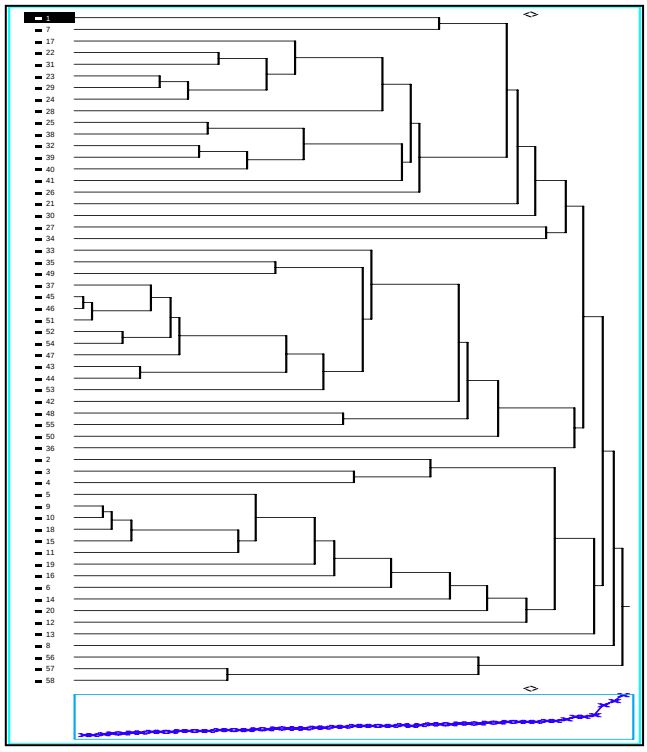

Fig. 5. Grouping of the load profiles

Thus, three groups were determined for the load profiles. The first group (C_{P1}) corresponds to the week days (Monday – Thursday), the second group (C_{p2}) corresponds to Friday and Saturday and the third group (C_{P3}) corresponds to Sunday.

For each group, the average and variance values (m and σ) were calculated. The signification of the coefficients m is: these coefficients transform the energy consumed by the medium member of the group in average active power demanded by it. These coefficients lead us to the typical load profiles (TLP) corresponding to the active power, Fig. 6.

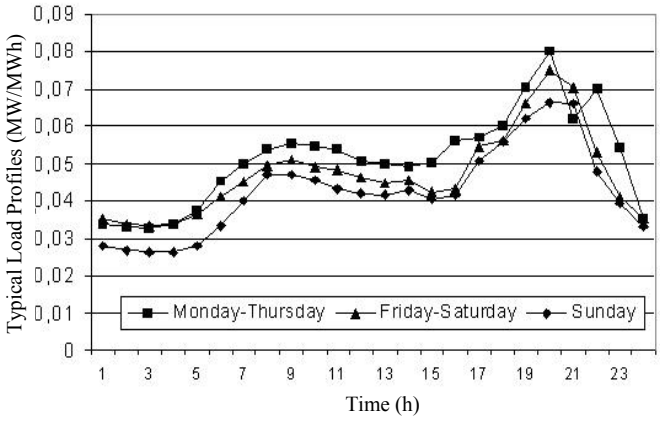

Fig. 6. The typical load profiles for obtained groups

Using the statistical values (m and σ), a fuzzy hourly model can be obtained, Fig. 7 and Table 1, [10].

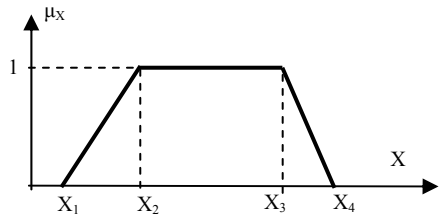

Fig. 7. Fuzzy trapezoidal model

TABLE I
BREAKING POINTS FOR FUZZY HOURLY LOAD, TLPS, C_{P1}, C_{P2} AND C_{P3}, [10]

Breaking points	Groups		
	C_{P1}	C_{P2}	C_{P3}
X_1	m^h_{P1} - 1.1 d^h_{P1}	m^h_{P2} - 1.12 d^h_{P2}	m^h_{P3} - 1.28 d^h_{P3}
X_2	m^h_{P1} - d^h_{P1}	m^h_{P2} - d^h_{P2}	m^h_{P3} - d^h_{P3}
X_3	m^h_{P1} + d^h_{P1}	m^h_{P2} + d^h_{P2}	m^h_{P3} + d^h_{P3}
X_4	m^h_{P1} + 1.1 d^h_{P1}	m^h_{P2} + 1.12 d^h_{P2}	m^h_{P3} + 1.28 d^h_{P3}

Considering the results of the grouping process presented above, we assumed that the data from April 18 are lost. For this day, the missing values of the load were imputed using the crisp values corresponding to the fuzzy typical load profiles. The obtained results were plotted in Figs. 8 - 10.

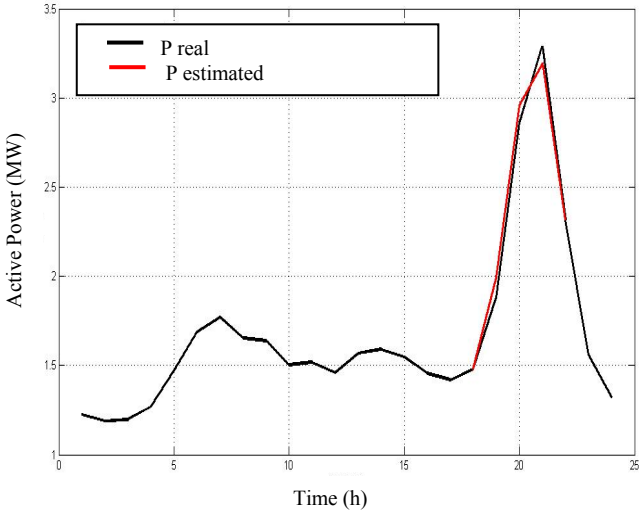

Fig. 8. The real and estimated load for 18 April 2005 (3 consecutives missing values 19-21, MAPE=3.92 %)

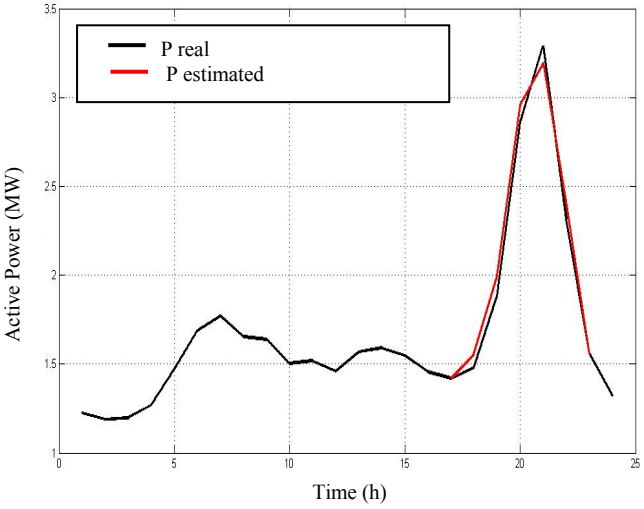

Fig. 9. The real and estimated load for 18 April 2005 (5 consecutives missing values 18-22, MAPE=4.14 %)

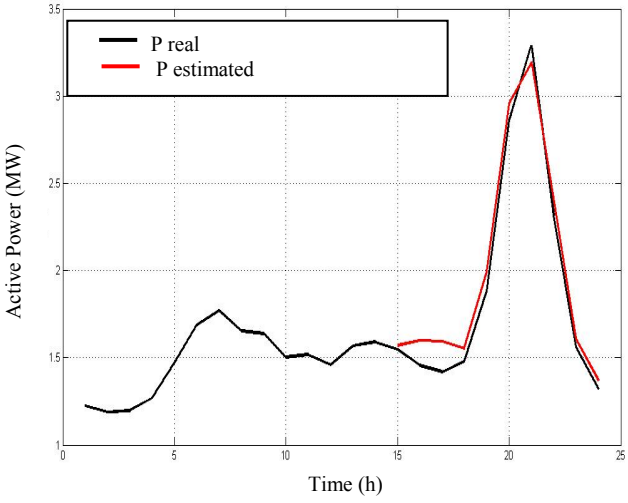

Fig. 10. The real and estimated load for 18 April (9 consecutives missing values 16-24, APE=5.28%)

VI. Conclusions

From the analysis of the obtained results, it can be observed that the largest errors correspond to the period with the 9 consecutives missing values in both methods. Note that the estimated errors are larger during valley hours and smaller during the peak hours. However, it is more important to obtain an accurate estimation during peak hours because the electric energy is more expensive during these hours. In the Fig. 11 the MAPE values obtained in the both method are represented.

Even if the errors are comparable, the method based on the fuzzy typical load profiles is more robust than KNNs method, and requires less work. In the KNNs method one must to find again the optimum nearest neighbors from the data set for every missing data case.

Smaller errors can be obtained when using it works with the larger databases.

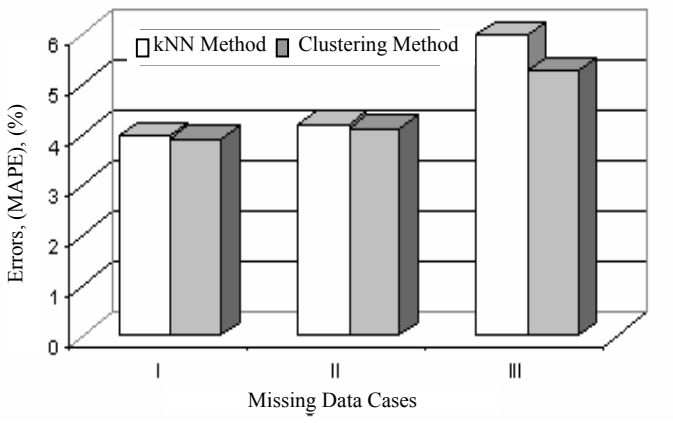

Fig. 11. Comparison between the obtained MAPE

VII. References

[1] M.E. Baran, L.A. Freeman, and F. Hanson, ″Load Estimation for Load Monitoring at Distribution Substations″, *IEEE Trans. on Power Systems*, vol. 20, pp. 164 – 170, 2005.

[2] F. Barrila, R. Jacobsen, and S. Schwarz, ″Estimation of Missing Flowmeter Data″, [Online]. Available: http://www.flowcontrolnetwork. com/PastIssues/novdec2000/1.asp.

[3] M. Suarez-Farinas, R. Lage de Sousa, and Castro Souza R, ″A Methodology to Filter Time Series: Application to Minute-by-Minute Electric Load Series″, [Online]. Available: http://www.scielo.br/ pdf/pope/v24n3/ a02v24n3.pdf.

[4] Y. Kim, *The Course of the Missing Data*, [Online]. Available: http://www.secondmoment.org/articles/missingdata.php.

[5] L. Brockmeier, J. Kromrey, and K. Hogart, ″No randomly Missing Data in Multiple Regression Analysis: An Empirical Comparison of Ten Missing Data Treatments″, *Multiple Linear Regression Viewpoints*, vol. 29, pp. 8 – 29, 2003.

[6] E. Acunã, and C. Rodriguez, ″The treatment of missing values and its effect in the classifier accuracy″, [Online]. Available: www.academic. uprm.edu/ ~eacuna/IFCSOr.pdf.

[7] J. J. Hox, ″A Review of Current Software for Handling Missing Data″. *Kwantitatieve Methoden*, vol. 62, pp. 123 – 138, 1999.

[8] ″Clustering: An Introduction″, [Online]. Available: www.elet.polimi.it/ upload/matteucc/Clustering/tutorial_html/ index.html.

[9] A.K. Jain, M.N. Murty, and P.J. Flynn, ″Data Clustering: A Review″, [Online]. Available: http://www.cermics.enpc.fr/~keriven/vision/ articles.

[10] G. Cartina, G. Grigoras, and E.C. Bobric, *Tehnici de clustering in modelarea fuzzy. Aplicatii în electroenergetica*, Iasi: Casa de Editură Venus, 2005.

Gheorghe Grigoras received the M. SC. and Ph. D. degrees in Electrical Engineering from Technical University “Gh. Asachi” of Iasi, Romania, in 2000 and 2005 respectively. He is currently lecturer in the Department of Power Systems of Electrical Engineering Faculty at the same university. His main areas of interest are analysis, planning, and optimization of power systems.

Gheorghe Cartina (Member IEEE) received the M. SC. and Ph. D. degrees in Electrical Engineering from Technical University “Gh. Asachi” of Iasi, Romania, in 1964 and 1972 respectively. He is currently professor in the Department of Power Systems of Electrical Engineering Faculty at the same university. His research interests include especially to monitoring and optimal control of power systems.

Elena-Crenguța Bobric (Member IEEE) received the M. SC. and Ph. D. degrees in Electrical Engineering from Technical University “Gh. Asachi” of Iasi, Romania, in 1995 and 2006 respectively. He is currently lecturer in the Electrical Engineering Faculty at the University “Stefan cel Mare” of Suceava. His research interests are the fuzzy technique applications in power systems.

Cecilia Barbulescu received the M. SC. In Electrical Engineering from Technical University “Gh. Asachi” of Iasi, Romania, in 1981. She is a diplomat engineer working with the Department of Facility Security and Crisis Management within the Romanian Power Grid Company Transelectrica SA, involved in the security and disaster recovery as well in the Business Continuity Planning Company wide.