A densification trick using mean shift to allow demand forecasting in special days with scarce data

Jean Sumaili, *Member, IEEE*, Vladimiro Miranda, *Fellow, IEEE*,
Liviane Rego, Ádamo Santana and Carlos Francês.

***Abstract* – In this paper, the Information Theoretical Learning Mean Shift algorithm is used to populate, with virtual data, a scarce set related to daily energy consumption in special days, such as holidays, in a process denoted *densification trick*. This allows the proper training of neural networks with the virtual data, reserving all the scarce real data for validation purposes. The networks are then used to predict consumption in special days. An example with real data from a Brazilian distribution utility illustrates the technique.**

***Index Terms* – Mean shift, Information Theoretic Learning, Neural Networks, demand forecasting.**

I. Introduction

IN any approach to electricity demand forecasting, it is universally recognized that the prediction for special days (mostly holidays and the adjacent days) is the most difficult exercise. The reason for this is that such days are scarce even in long historic records – a particular holiday (with fixed date) not only occurs once a year but also it does not occur in the same day of the week every year. Therefore, methods based on similar patterns or on system training suffer form an acute lack of data to be tuned and validated.

This deficiency in data affects the proper use of neural networks to perform demand forecasting. Having available only 10 samples from historic records for a particular special day is not statistically acceptable. To give an example: considering that 7 samples are used as a training set and 3 as the validation set to train a simple neural network with some 30 weights, there simply is not enough redundancy to train and validate the network's capacity for generalization.

For this reason, a 24 hour ahead prediction for special days is usually accomplished with some historical analysis together with the application of heuristic rules.

This paper offers an alternative, which is to generate virtual data sharing some statistical properties with the known real data. The scarce data set can thus be populated with a high number of virtual samples, in what is denoted as the *densification trick*. This allows the proper training of neural networks with the virtual data, reserving all the scarce real data for validation purposes.

This work is partially funded by the ERDF from the EU through the Programme COMPETE and by the Portuguese Government through FCT - Foundation for Science and Technology, project ref. LASCA PTDC/EEA-EEL/104278/2008.

Jean Sumaili (jean.sumaili@inescporto.pt) and Vladimiro Miranda (vmiranda@inescporto.pt) are with INESC TEC (INESC Technology and Science, coordinated by INESC Porto), Portugal. V. Miranda is also with FEUP, Faculty of Engineering of the University of Porto.

L. Rego (liviane@ufpa.br), A. Santana (adamo@ufpa.br) and C. Francês (rfrances@ufpa.br) are with UFPA, Federal University of Pará, Brazil.

The generation of virtual data requires the new samples to be distributed in space in such a way as to form a cluster exhibiting the identifying properties that should represent the special day in analysis. The known real data are considered as samples extracted from such hypothetical dense cluster. Their knowledge allows one to estimate a probability density function (pdf) that one assumes similar to the real hidden pdf of the dense cluster.

The Information Theoretic Learning (ITL) Mean Shift (MS) algorithm allows to uncover properties from such an estimate of the pdf of a cluster. One simple property is the mode, but other properties may be useful to recognize, such as the principal lines of a cluster. During the process of revealing such properties, new sets of (virtual) points are generate that share those properties among each other and with the original set. These virtual samples can then be used to train the neural networks, preserving all scarce real data to a later use in the testing phase. This allows one to compose an enough dense set together with some more virtual data and verify the generalization ability of the trained system.

The paper describes how this method is applied in a demand forecasting problem in Brazil, where the prediction for special days was difficult due to the lack of data in historical records.

II. Forecasting special days

Regularly, surveys on publications dealing with load forecasting models and methods come to light. Without a concern for being exhaustive, we may mention as examples two recent surveys included in [1] and [2]. This continuous flow of works in this domain demonstrates how this is still felt as an active area of research for a problem not definitely solved and remaining with economic impact. Particularly, consumption forecasting for special days is an important issue yet not much addressed in a systematic way. Few publications are related with this specific topic – some examples are [3][4][5][6].

As a general comment, most authors have recognized that special systems should be devoted to 24 hour-ahead prediction in special days. In some cases, *ad hoc* rules are applied to correct the prediction made, as if the day were not special. Many approaches propose a prediction derived from some form of selection of similar days in the historical data.

Weekends are sometimes included in the class of special days but the really difficult days are holidays with a fixed calendar date. It is common knowledge that when a holiday occurs in the middle of a week, the previous and the following days have their consumption affected also. Furthermore, a holiday on a Wednesday induces a behavior distinct from a holiday on a Friday or Tuesday, especially in the neighboring days. Additionally, there are seasonal effects superimposed on the holiday effect: a holiday on Thursday in August may not have the same effect as a holiday on a Thursday in February. The big issue, then, is always the lack of enough data to adequately build a consistent model for each particular case.

III. Information Theoretic Mean Shift and cluster densification

The Gaussian Mean Shift and the Blurred Gaussian Mean Shift algorithms were introduced as processes to cluster data non-parametrically. The Information Theoretic Mean Shift algorithm [7][8] not only put these algorithms in perspective, but provided a theoretical explanation for their convergence (or lack of) and, based on Information Theoretic Learning concepts, provided a means to capture the dominant structures in the data set, as embedded in its estimated probability density function (pdf).

Renyi's quadratic entropy [9] for a pdf is defined as

$$H(X) = -\log \int_{-\infty}^{+\infty} p^2(x)dx \qquad (1)$$

and the pdf p(X) can be estimated by the Parzen windows technique [10]

$$\hat{p}(X) = \frac{1}{N}\sum_{i=1}^{N} G_\sigma(x - x_i) \qquad (2)$$

where G_σ is a Gaussian kernel having bandwidth $\sigma > 0$. Replacing (2) into (1) gives

$$H(X) = -\log V(X) \qquad (3)$$

with $$V(X) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma'}(x_i - x_j) \qquad (4)$$

having $\sigma' = \sqrt{2}\sigma$. V(X) is called *information potential* of the pdf p(x). Its derivative with respect to x_i gives the *information force* exerted by all data particles on x_i [11][12].

The *cross entropy* between two pdf can be defined by

$$H(X, X_0) = -\log V(X, X_0) \qquad (5)$$

with $$V(X, X_0) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sigma'}(x_i - x_{0j}) \qquad (6)$$

The Cauchy-Schwartz distance measure between two pdf p and q is given by

$$D_{CS}(X, X_0) = \log\left(\frac{\left(\int p^2(x)dx\right)\left(\int q^2(x)dx\right)}{\left(\int p(x)q(x)dx\right)^2}\right) \qquad (7)$$

It is an extension of the concept of the cosine of the angle between two vectors as a measure of distance between them.

It is easy to show that

$$D_{CS}(X, X_0) = -\left[H(X) + H(X_0) - 2H(X, X_0)\right] \qquad (8)$$

The Information Theoretic Mean Shift algorithm acts under a two-criteria optimization, minimizing the entropy of X while keeping the Cauchy-Schwartz distance at some value k. An unconstrained optimization formulation, under a parameter λ that represents the trade-off between the two objectives, is

$$J(X) = \min H(X) + \lambda\left[D_{CS}(X, X_0) - k\right] \qquad (9)$$

Differentiating J(X) with respect to each x_i gives an algorithmic rule that allows the transformation of X_0 into another set at iteration t+1, making use of the information contained in the pdf of X at iteration t, estimated by (2):

$$x_i^{t+1} = \frac{c_1\sum_{j=1}^{N} G_{\sigma'}(\| x_i^t - x_j^t \|)x_j^t + c_2\sum_{j=1}^{N} G_{\sigma'}(\| x_i^t - x_{0j} \|)x_{0j}}{c_1\sum_{j=1}^{N} G_{\sigma'}(\| x_i^t - x_j^t \|) + c_2\sum_{j=1}^{N} G_{\sigma'}(\| x_i^t - x_{0j} \|)}$$

where $$c_1 = \frac{(1-\lambda)}{V(X)} \text{ and } c_2 = \frac{2\lambda}{V(X, X_0)} \qquad (10)$$

It is if the information particles, in a field of information potential, move under the influence of the information forces like the derivatives of (4) referred to above.

With λ = 1 in (10), the particles converge to the modes of the pdf p(X). When λ is increased between 1 and 2, the algorithm tends to make the points x converge to the principal curve of the data, and a further increase in λ will concentrate them around the pdf denser regions. The iterations produce a succession of sets $X_0, X^1,\ldots, X^t,\ldots$; each set X^t is associated to a pdf – $p(X^t)$ – that retains information from $p(X_0)$.

The set $X_V = X^1 \cup X^2 \ldots \cup X^t$ is the set of virtual data generated by the ITL Mean Shift algorithm. It forms a dense cluster that shares properties with the original X_0. It becomes especially useful when data is scarce, as it may be used as the training set to train a neural network to learn such properties, preserving X_0 for the validation phase.

This use of X_V is called the *densification trick*.

IV. Case study: predicting special day consumption

The *densification trick* was successfully applied in a problem of incipient fault diagnosis in power transformers [13], where scarce data on failures existed. This suggested its application in the construction of a neural network system for the 1 day-ahead prediction of electric energy consumption in special days, for a Brazilian distribution utility. The following sections describe the new approach.

A. The problem and data available

The problem presented by a Brazilian distribution utility relates to the prediction of electric energy global consumption for the next day (1 day-ahead prediction). The historical data refer to about 10 years of consumption (from January 2002 to September 2012), and the task at hand was to predict the consumption in special days, namely holidays. Fig. 1 presents a set of special day patterns. To form the patterns, sequences of 7 days previous to the special day were organized. The window of 7 days was ideally identified from an iterative window size analysis, considering quality/representative metrics such as R^2 and conditional entropy.

The horizon 1 corresponds to the special day consumption to be predicted and the other horizons are the daily demand of the previous week. These data need to be normalized in order to eliminate the effect of the consumption growth during the 10 years of measurements. The normalization was made with respect to the total consumption of the previous week. The normalization results are shown in Fig. 2.

B. Mean Shift for classification

Although the calendar location of holidays is known, it was important to confirm that it was possible to identify distinct patterns for special days, and to cluster them in similar classes as much as possible. This way, each holiday/special day would become associated with a particular pattern (cluster). Fig. 3 presents the clusters organized with the application of the ITL MS algorithm to the patterns in Fig. 2. Setting λ=0.1 in (10), the identification of 10 modes was possible. All the patterns converging to a common mode were grouped in individual clusters. It was thus possible to form 5 proper clusters, and 5 groups with the remaining outliers.

Clusters 1 to 5 represent the classical patterns with a special day located on Tuesday, Thursday, Friday, Wednesday and Monday, respectively. Notice that the special day is always at the end of the sequence. The remaining clusters (6 to 10) correspond to very special cases which should deserve individual analysis. It must be said that some of these cases possibly correspond to blackouts (occurring in the Brazilian region of analysis), which severely reduced the daily consumption; also, holidays that do not have a fixed week day distorted the observed pattern.

The first five clusters were easily confirmed as corresponding to holidays on distinct week days. For instance, cluster 1 is associated to holidays on Tuesday.

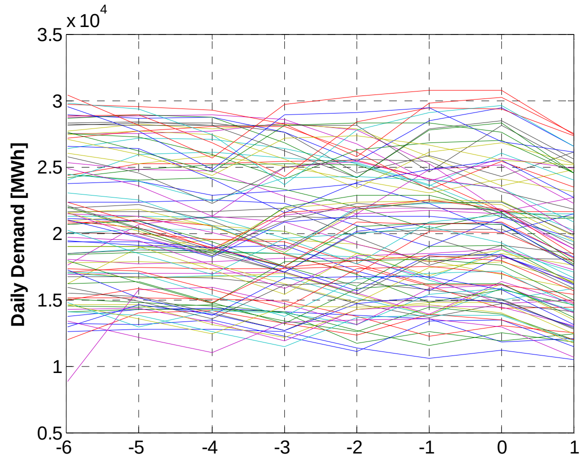

Fig. 1 – Set of consumption patterns with the special day at the end

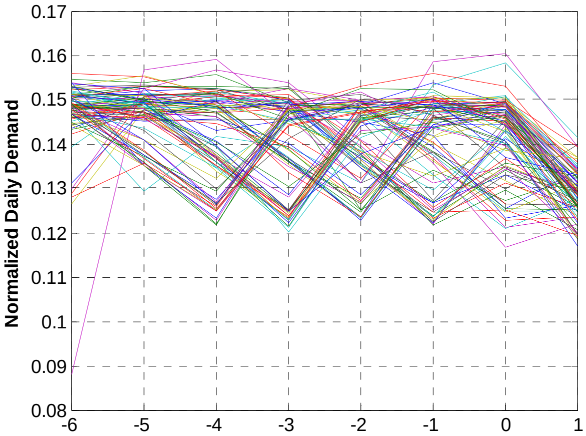

Fig. 2 – Normalized special day patterns.

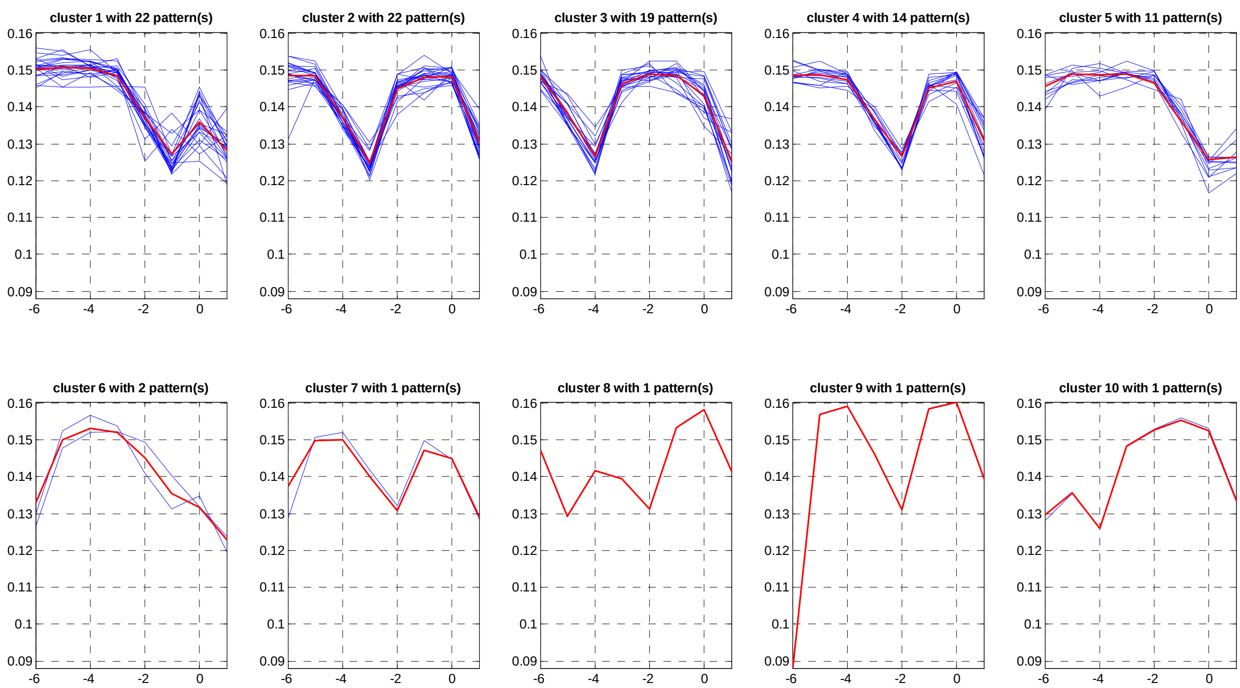

Fig. 3 – ITL Mean Shift clustering results. Five proper clusters were discovered, with 22, 22, 19, 14 and 11 members each. Five distinct outlier groups were also identified

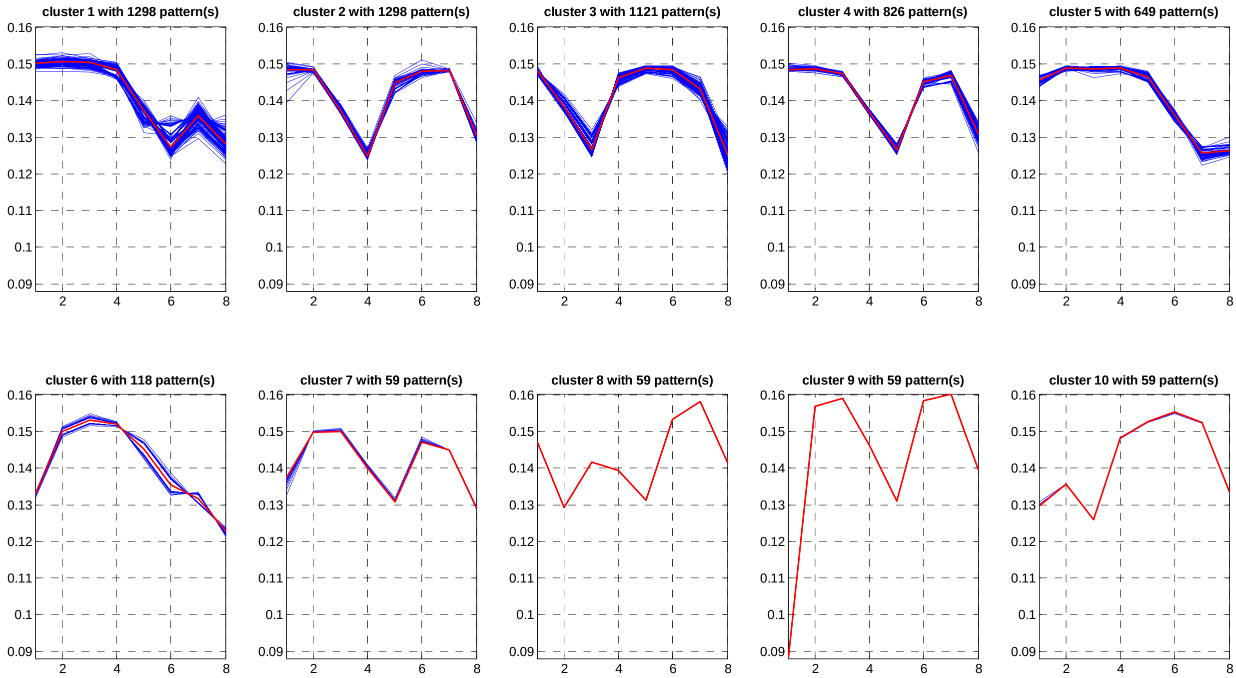

Fig. 4 – Virtual data generated by the ITL Mean Shift algorithm, after 59 iterations, with 1298, 1298, 1121, 826 and 649 virtual patterns forming the five first clusters.

It is interesting to observe that a phenomenon called "extended weekend" is easily detected in the Tuesday cluster: the consumption on Monday is on average smaller that on the other working days. Considering that the lowest value in the middle of each diagram corresponds to the consumption on a Sunday, it is straightforward to identify each diagram in association to each day of the week.

C. *Densification trick applied*

The ITL MS algorithm was run for each cluster in order to create a dense cluster of virtual data. The convergence was reached after 59 iterations, generating 59 virtual patterns for each original data point; i.e. the total number of virtual patterns in each cluster is the number of original real points times the number of iterations performed. The clusters of virtual patterns are represented in Fig. 4..

The virtual data were then used in the training of a feedforward neural network, having 7 neurons in the input layer, 1 neuron as output and one middle layer forming a 7–3–1 architecture for each cluster. The input and output layers had linear activation functions and the other neurons had *tanh* activation functions.

In the validation process, all real data were used (while in the training phase only virtual data were used). This deserves to be underlined once again. In the original data the largest cluster had only 22 members while the number of weights to be tuned in its corresponding neural network was 24. Adopting the usual rule of thumb, one should reserve one third, or 7 patterns, to validate and keep 15 patterns to train the neural network. This means a case of lack of redundancy to properly train the network and a clearly insufficient number of cases to validate the capacity of the network to perform a proper generalization. Using the virtual data for training and reserving the real data for validation, this largely overcomes the insufficiency in data discussed.

D. *Consumption prediction with neural networks*

The validation process is based on assessing the error in consumption prediction, compared to observed values. Table 1 summarizes the significant results obtained. These are divided in 5 classes, associated to the day of the week where the special day was located. The normalized mean absolute error (NMAE) varies from 1.85% (Monday) to 3.92% (Friday), while the variation range of the corresponding standard deviation is from 1.66% to 2.50%. This accuracy is quite satisfactory, even considering that the prediction model (a simple neural network for each cluster of special days) is quite simple.

TABLE 1 – PREDICTION RESULTS SUMMARY

Location of the special day	**Number of real days tested**	**NMAE**	**Standard Deviation**
Monday	11	1.85%	1.76%
Tuesday	22	2.82%	2.50%
Wednesday	14	3.41%	1.90%
Thursday	22	2.55%	1.66%
Friday	19	3.92%	2.46%

V. Conclusions

This paper demonstrates how the Information Theoretic Mean Shift algorithm could be used, in the form of a *densification trick*, to allow for the proper training of neural networks and produce a useful consumption prediction system for special days, like holidays and surrounding days.

The positive results reported can be summarized as follows:

- It is possible to train neural network systems even when faced with scarce data sets;
- The ITL MS algorithm could be used to identify distinct clusters in the consumption data, which were easily associated with holidays occurring in distinct days of the week;
- The same ITL MS algorithm allowed collecting virtual data representing distinct specific clusters.
- The systems trained with virtual data performed as if they were trained with real data;
- The prediction of electricity consumption for special days, in a real case in Brazil, is a case of scarce data and could be solved with this approach.

The remarkable accuracy achieved in forecasting for special days confirms the correctness of the new approach. It is quite possible that results with better accuracy may yet be obtainable, considering that the neural network adopted has a very simple architecture. More sophisticated arrangements are likely to allow further improvement with a narrower accuracy.

Furthermore, the results clearly demonstrate that the application of the *densification trick* allows building more robust models and training neural networks for producing consumption forecasts when in the presence of scarce data to train the models.

Acknowledgment

The authors gratefully acknowledge the utility that supplied the data for this work, whose identification is not disclosed for reasons of confidentiality.

The authors also acknowledge Joana Hora and Vera Palma for the support in training and testing the neural network for consumption prediction.

References

[1] M. Sheikhan and N. Mohammadi, "Neural-based electricity load forecasting using hybrid of GA and ACO for feature selection", Neural Computing & Applications, Vol. 21 Issue 8, p1961, Nov2012.

[2] H. Hahn, S. Meyer-Nieberg, and S.Pickl, "Electric load forecasting methods: Tools for decision making", European Journal of Operational Research. Vol. 199, Issue 3, Pages 902–907, Dec 2009.

[3] A. Laukkanen, "The use of special day information in a demand forecasting model for Nordic power market", Report Mat-2 108, Helsikin University of Technology, Feb 2004

[4] J.N. Fidalgo and J.A.P. Lopes, "Load forecasting performance enhancement when facing anomalous events", IEEE Transactions on Power Systems, Vol. 20, Issue 1, pp. 408-415, Feb 2005

[5] S. Nahi et al., "Load Forecasting on Special Days & Holidays in Power Distribution Substation Using Neural & Fuzzy Networks", Proc. International Conference on Computational Intelligence for Modeling, Control and Automation, Sidney, Australia, Nov 2006

[6] N.A. Salim, "Case study of Short Term Load Forecasting for weekends", IEEE Student Conference on Research and Development (SCOReD), 2009

[7] Sudhir Rao, Allan de Medeiros Martins, Weifeng Liu, Jose C. Principe, Information Theoretic Mean Shift Algorithm ", Intl. Work. on Neural Networks for Signal Processing", Maynooth, Ireland, pp. -, 9 2006

[8] Sudhir Rao, Allan de Medeiros Martins, Jose C. Principe, " Mean shift: An information theoretic perspective", Pattern Recognition Letters, no. 30, pp. 222–230, 2009.

[9] A. Renyi, "Some Fundamental Questions of Information Theory", *Selected Papers of Alfred Renyi*, vol 2, pp. 526-552, Akademia Kiado, Budapest, 1976.

[10] E. Parzen, "On the estimation of a probability density function and the mode", *Annals Math Statistics*, vol 33, 1962

[11] D. Erdogmus and J. C. Principe, "Generalized Information Potential Criterion for Adaptive System Training", *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1035-1044, Sep 2002.

[12] J. C. Principe, "Information Theoretic Learning - Renyi's Entropy and Kernel Perspectives", New York: Springer, 2010

[13] V. Miranda, S. L. Lima, "Diagnosing faults in power transformers with autoassociative neural networks and mean shift", IEEE Transactions on Power Delivery, vol.27, no.3, pp.1350-1357, July 2012.

Jean Sumaili (M'10) graduated in "Sciences Appliquées (option: Electricité)" from the University of Kinshasa, Kinshasa, Democratic Republic of the Congo, in 1998, and received the M.Sc. and Ph.D. degrees in electrical engineering from the Politecnico di Torino, Turin, Italy, in 2004 and 2008, respectively. He is currently a Senior Researcher at the Power Systems Unit of INESC Porto, the leading institution of INESC TEC – INESC Technology and Science, Portugal. His research activities include distribution systems analysis, distributed generation applications, electricity customer classifications, and renewable source integration (namely, photovoltaic and wind power systems).

Vladimiro Miranda (M'90, SM'04, F'06) received his Ph.D. degree from the Faculty of Engineering of the University of Porto, Portugal (FEUP) in 1982 in Electrical Engineering. In 1981 he joined FEUP and currently holds the position of Full Professor. He is also a researcher at INESC since 1985 and is currently Director at INESC Porto, the leading institution of INESC TEC – INESC Technology and Science, an advanced research network in Portugal. He is also President of INESC P&D Brasil. He has authored many papers and been responsible for many projects in areas related with the application of Computational Intelligence to Power Systems. Prof. Miranda is an IEEE Fellow and the receiver of the IEEE PES Ramakumar Family Renewable Energy Excellence Award.

Liviane Rego graduated in Computer Engineering (2006) and received the Master degree in electrical engineering (2009) from the Federal University of Para, where she is a Ph.D. candidate. She has experience in Data Mining (including load forecast, optimization, pattern discovery and decision support) and in telecommunications, working in design and deployment of wireless.

Ádamo Santana received the bachelor's degree in Computer Science from the University of the Amazon in 2002; the Master's (2005) and Ph.D. (2008) in Applied Computing for Electric Engineering from the Federal University of Pará and University College Cork. He is currently a Professor at the Faculty of Computer Engineering of the Federal University of Pará. He has experience in the area of Computational Intelligence, acting on the following subjects: game theory, load forecasting, optimization, pattern analysis and decision support.

Carlos Francês received the bachelor's degree in Computer Science from the Federal University of Pará (1995), the Master degree (1998) and the Ph.D. in Computer Science from the University of São Paulo, in 1998 and 2001, respectively. He has experience in Computer Science, acting on the following subjects: Last mile networks, performance models (Markovian and simulation) and Bayesian networks.