

# Designing the input vector to ANN-based models for short-term load forecast in electricity distribution systems

P. J. Santos, A. G. Martins, and A. J. Pires.

**Abstract**— The present trend to electricity market restructuring increases the need for reliable short-term load forecast (STLF) algorithms, in order to assist electric utilities in activities such as planning, operating and controlling electric energy systems. Methodologies such as artificial neural networks (ANN) have been widely used in the next hour load forecast horizon with satisfactory results. However, this type of approach has had some shortcomings. Usually, the input vector (IV) is defined in an arbitrary way, mainly based on experience, on engineering judgment criteria and on concern about the ANN dimension, always taking into consideration the apparent correlations within the available endogenous and exogenous data. In this paper, a proposal is made of an approach to define the IV composition, with the main focus on reducing the influence of trial-and-error and common sense judgments, which usually are not based on sufficient evidence of comparative advantages over previous alternatives. The proposal includes the assessment of the strictly necessary instances of the endogenous variable, both from the point of view of the contiguous values prior to the forecast to be made, and of the past values representing the trend of consumption at homologous time intervals of the past. It also assesses the influence of exogenous variables, again limiting its presence at the IV to the indispensable minimum. A comparison is made with two alternative IV structures previously proposed in the literature, also applied to the distribution sector. The paper is supported by a real case study at the distribution sector.

**Index Terms**— Electric energy systems, Distribution networks, Load forecast, ART neural network.

## I. INTRODUCTION

The transformations occurred during the last 10-15 years in electrical energy systems (EES), make different operators face a scenario of a completely or partially de-regulated environment. This situation increases the complexity of actions like planning, management and operation of the networks [1] [2]. In fact, liberalization tends to eliminate some of the certitudes that formerly were taken for granted in the utility business. Commercial transactions take place on the networks with a reasonable independence of the technical issues of network management. System operators have to use as much as possible reliable data, namely on load forecast results, having in mind simultaneously that uncertainty is a key issue to most decisions [3]. Forecast methodologies have registered an evolution, during the last three decades, which is influenced both by the increasing complexity of the factors that affect consumption and by the trend to the application of an increasing number of different methodologies that have been proposed and tested. Among these, ANN deal with uncertainty in a convenient way [2]. Despite the many publications and models of STLF that have been developed in the last few decades [4], [5] few amongst those have dealt with specific approaches to the sub-sector of distribution companies (DISCO) [6], [7], [8]. This paper deals with a methodological approach, based on ANN, to forecast the next-hour load, applied to three electrical substations (ES) of average dimension (60 kV/15kV), located in the center region of Portugal, more precisely in the city of Coimbra [9]. Next-hour load forecast allows DISCOs to answer in a substantiated manner to issues such as: network reconfiguration, voltage control, maintenance planning and power factor correction, amongst others. However, the design of the input vector (IV) to the ANN is usually carried out in a discretionary way, mainly based on trial-and-error procedures and on engineering judgment criteria, supported on the usual data pre-processing that allows capturing significant correlations within the available data [10] [11]. This paper is mainly concerned with the issue of IV design, presenting a methodological proposal, tested against a real-life case study, which aims at achieving a good level of forecast precision, simultaneously using only the strictly necessary input data. For this purpose, entropy analysis has been used to measure the level of complexity of the finite length time series of the active power, the signal under study. This type of methodology is also applied, for example, in physical and physiologic data series, [12], being also used to analyze econometric time series [13]. In face of the results obtained by the entropy analysis, revealing a short range memory of the signal, it was decided to define the composition of the IV based on the criterion of the best auto-correlation results, leading to a short stream of three contiguous values of active power, immediately preceding the value to be forecast. Also, auto-correlation analyses performed backwards into the time series have revealed the virtue of capturing information on the trend of consumption, by using what the authors have designated "tendency concept", leading to the inclusion, in the IV, of past values collected at

homologous periods of the two preceding weeks.

In order to evaluate the performance of the obtained IV, a benchmark analysis was carried out, involving two other IV structures developed by different authors [7], [10]. These two other approaches include some climatic variables and also variables obtained through auxiliary sinusoidal functions, in order to help the ANN capturing the periodic behavior of load.

## II. DATA ANALYSIS ORIENTED TOWARDS INPUT VECTOR DESIGN

The methodology proposed for defining the structure and composition of the input vector (IV) to the ANN-based model, demands a sequence of tasks to be followed [4]. First, data pre-processing is performed, mainly in order to fill the gaps caused by failures occurred in the acquisition system (SCADA) operation. In a second phase, a correlation analysis is made, between the endogenous and exogenous variables, to identify relevant influences of independent random factors on consumption dynamics. In a third phase, the available time-series is subject to an entropy analysis in order to find the length of the short-term memory of the active power series and, thus, to identify the maximum length of the relevant data sequence to be included in the IV. In a fourth phase, an auto-correlation analysis is carried out backwards (to the past) in the time series, seeking for the best correlation coefficients, in such a way as to identify short sequences of data representing the likely trend of consumption evolution, based on the behavior of load in homologous periods of past weeks (three values per past week for each forecast value, in the present case). Finally, after the definition of the ANN structure, simulation of the model performance under real life conditions is carried out, based on known data not used in the previous phases.

### A. Data preparation

The data on the endogenous variables was collected in three ES, from December 21, 1998, to January 31, 2004. In a data sequence of this length, there are usually some occurrences of lost data that have to be reconstructed in some way [4].

The rules used for this purpose have been, briefly, the following: if the lost data corresponds to up to one hour, the average value of the previous four intervals of 15 minutes is used; if the lost data period is longer than one hour, then the whole day is replaced with the average load profile of the two homologous days of the two previous weeks; if the lost data corresponds to a special day, a custom procedure is used – the closest similar day or the closest Sunday is used. An example of this type of data reconstruction is depicted in Fig.1. Besides, according to the current practice, some isolated abnormal values have been replaced, each with the average of the two boundary values.

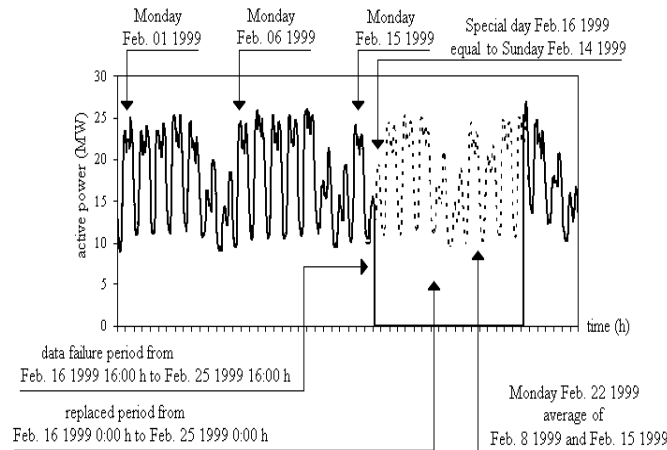


Fig. 1. Missing data replacement.

### B. Correlation between climatic data and consumption data

The definition of the IV structure involves options with a certain degree of arbitrariness. Different types of endogenous and exogenous variables are normally used [11]. In some regions with certain climatic conditions, correlation of consumption data with climatic data may be strong, particularly when high humidity and temperature are current in Summer or very low temperatures occur in Winter. Normally, this interdependence is expressed by including the temperature or humidity in the IV [14].

In the case-study, moderate temperature swings are accompanied by also moderate humidity conditions. Hence, a strong correlation was not expected, particularly because short-term forecasts were sought [9].

Daily peak loads decrease in Spring and Summer, which reveals a weak penetration of cooling loads. In Winter, peak load increases as a consequence of the strong presence of air heating loads (Fig. 2 b)). In the next hour horizon, correlations between these variables are very tenuous (Fig.2 a)).

Beyond this analysis a daily average correlation between active power and temperature, with different time shifts, was

performed, which showed very small correlation coefficients. Based on this analysis, the composition of the IV will have to rely essentially on the endogenous variables, based on a careful analysis of the auto-correlation values.

### C. Entropy analysis

Usually, in the construction of the IV, some contiguous past values of the forecast variable are used [4], [7], [10].

The number of the corresponding contiguous instants is defined based on auto-correlation analyses. The final accuracy of the results determines to what extent this number should be increased. In the particular case under consideration the auto-correlation analysis shows a rapidly decreasing trend of the auto-correlation coefficient with a deeper incursion into the past (Fig.3.).

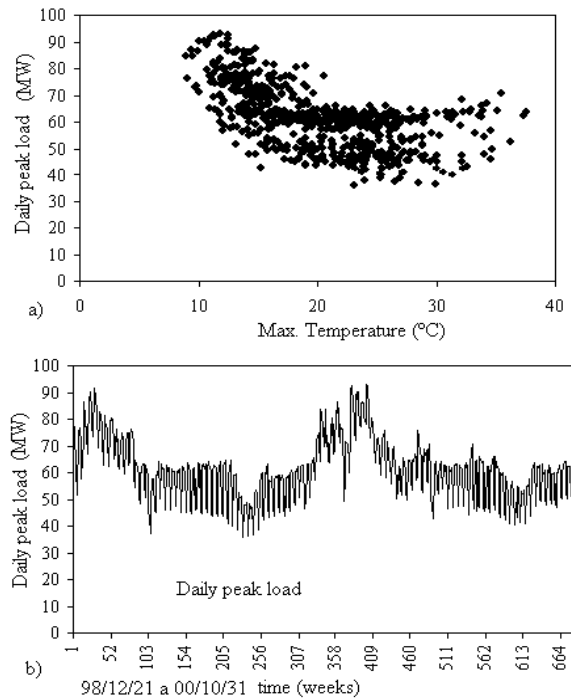


Fig. 2. a) - Scatter plot between the peak load day and the maximum temperature; b) - representation of variation of peak demand (both between Dec. 1998 and Oct. 2000, city of Coimbra).

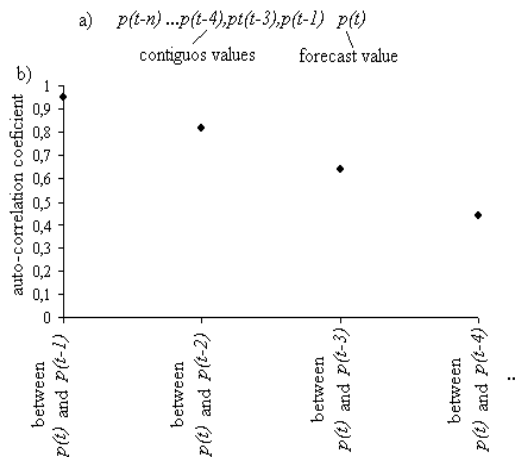


Fig. 3. auto-correlation values for contiguous information of active power time series.

In order to evaluate the influence of increasing the number of contiguous data values in the IV, an entropy analysis of the time series was performed, based on the sample entropy (SampEn) concept [15], [16]. SampEn is given by the following equation (1).

$$E = -\log(A_k / B_{k-1}) \quad (1)$$

where  $A_k$  and  $B_{k-1}$  are the probabilities associated to sequences of values of lengths  $k$  and  $k-1$  ( $k=1, \dots, m$ ). If  $E$  decreases as  $k$  increases, this denotes signal regularity. This means that increasing the length of contiguous information is redundant and certainly not the best way to explain the signal. The SampEn approach is a simplification of the Kolmogorov-Sinai entropy (KS) [17], [12].

To illustrate the application of SampEn, consider the active power diagram represented in Fig.4. The application of the SampEn algorithm implies the identification, in the time series, of sequences of contiguous values, of incrementally growing length, and their respective probabilities of occurrence. If, for instance,  $(p[1], p[2])$  are taken with their exact values, (Fig.4), there is little chance of finding exactly equal sequences in the series. Hence, a tolerance  $r$  is defined around the crisp original values in the series, leading to a relaxation in the search for equivalent sequences, based on the set  $p[1] \pm r, p[2] \pm r$  ( $r \in \mathfrak{R}$ ), (Fig.4).

Let us consider, in Fig. 4, the sequence of two elements ( $m=2$ ) as  $(p[1], p[2])$ . There are three similar sequences of two elements:  $(p[25], p[26])$ ,  $(p[73], p[74])$  and  $(p[107], p[108])$ . Analogously, there are several similar sequences of three elements ( $m=3$ ), etc.

It is necessary to identify all different sequences, calculate their probabilities, and evaluate the evolution of sample entropy values thereof.

Within the case-study, an active power series with 43752 hourly records, from Dec. 21, 1998 to Dec. 17, 2003, was used. Several tolerance values  $r$  have been tried. An example of the evolution of SampEn is represented in Fig.5, where the decreasing value denotes regularity of the time series. A strong reduction of entropy values, from scale factor 1 ( $E_1$ ) to scale factor 2 ( $E_2$ ), is evident. Several tolerance values  $r$  have been tried and, in all cases, the same behavior was verified. An extensive use of contiguous values to build an IV is therefore unnecessary because of the short-term memory of the active power time series. The choice of the values to include in the IV should be based on non-contiguous data, namely as a result of exploring the auto-correlation coefficients of the time series, as described in the next section

#### D. The tendency concept – from autocorrelation analysis to input vector refinement

The composition of the IV must be defined through a careful analysis of the auto-correlation coefficients of active power, as well as of the correlations with the exogenous variables. In accordance with the analyses made in section B, correlations with temperature and with relative humidity are weak in the case of the next hour horizon. In Fig.6, it may be seen that the auto-correlation values become higher when one considers the correlation of the most recently available consumption values with those occurred at the past instants homologous to  $p$ , showing two relative maxima at the previous weeks (Fig.6).

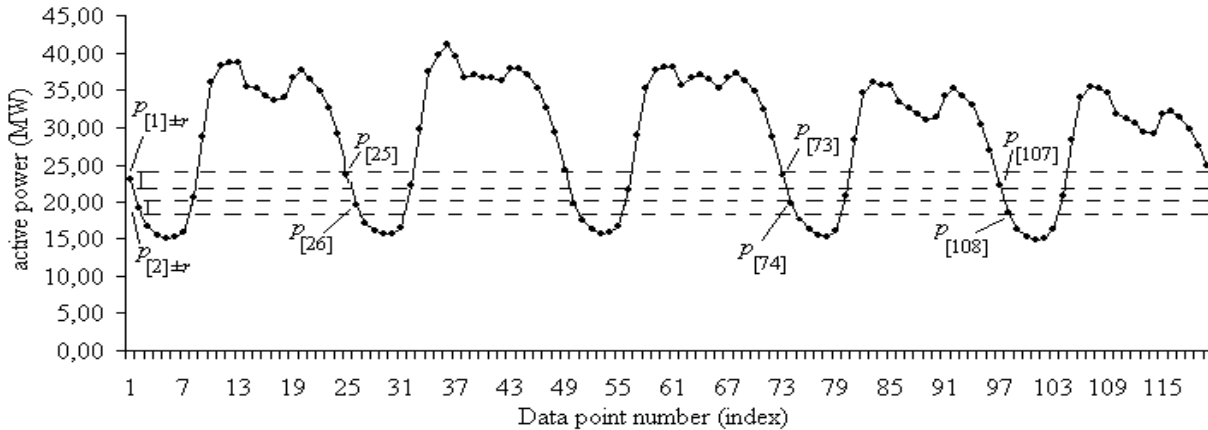


Fig. 4. Active power time-series – example of similar sequences of two elements.

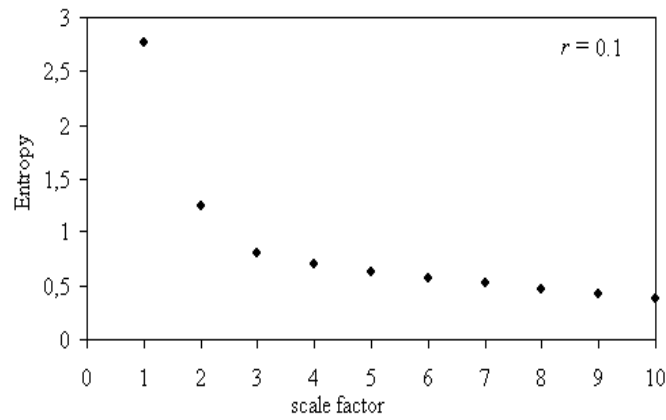


Fig. 5. Entropy values for active power time series.

One should point out the similarity between the coefficients around the homologous values  $p(t-168)$  and  $p(t-336)$ . The inclusion of these values  $p(t-167)$ ,  $p(t-169)$ ,  $p(t-335)$  and  $p(t-337)$  in the IV provides information regarding the consumption trend in past homologous periods.

The auto-correlation coefficients always decrease as one moves deeper into the past, which can be explained by the seasonal variation of consumption, entailing different load patterns. The final IV (#1) was defined as depicted in Fig. 7.

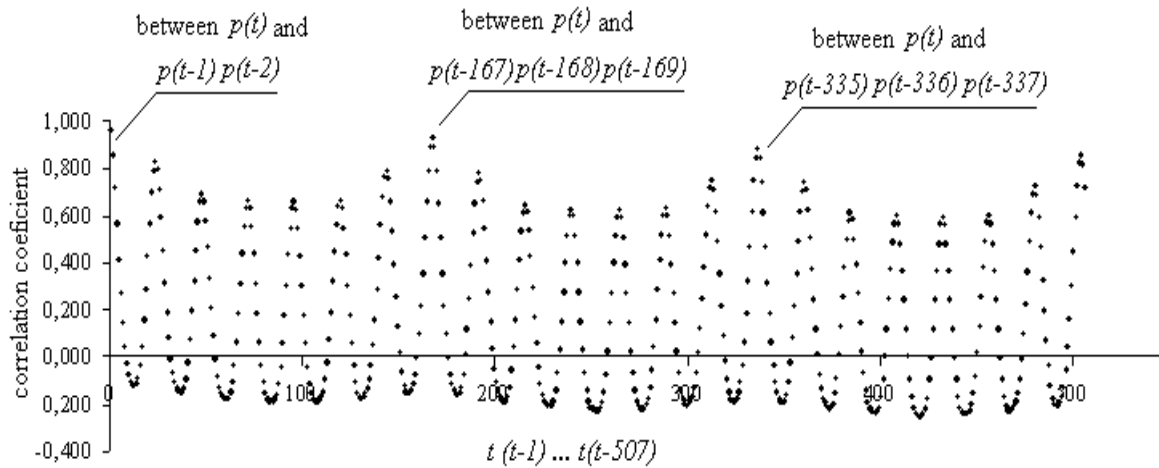


Fig. 6. Auto-correlation values showing local maxima at homologous instants in the past.

It was decided to include the reactive power values at the two previous hours, in order to take profit of the good correlations with the active power [9]. The improvement of the performance resulting of the inclusion of this last variable is detected by the result of the decreasing of MSE error of the training set.

#### E. ANN design option

A standard feedforward backpropagation (BP) ANN has been used for building the forecast models, having a fully connected architecture with a single hidden layer.

In face of the learning algorithm, the hyperbolic tangent function was chosen for the middle layer. For the output layer a linear function was used.

The number of neurons in the hidden layer was half of the one of the input layer [4]. The input vector was normalized between  $-1$  and  $1$ . This is a well-proven arrangement, adequate when, as in the present case, the relations between the variables at stake have a strong non-linear behavior. This type of structure is widely used, in similar models with this forecast horizon [18].

#### F. Data management options for model tuning

The training set was defined from December 1998 to December 2002. The forecast period occurs from January, 2003 to January, 2004.

It is very important to evaluate the relative position of the training data set as regards to the data set used in forecast simulation (Fig.8) (next page). If all values forecasted are “contained” in the region defined by the values of the training set, a better

performance of the model is to be expected than in situations when the training set of values does not "cover" the values used in forecast.

This lack of "coverage" is natural in this type of signal. Several factors may lead to it, or management actions like network reconfigurations. This lack of coverage strengthens the need of proceeding to careful evaluation of the results and, if it is the case, to make a new definition of the training set in order to include these effects. This procedure increases the capacity of generalization of the ANN.

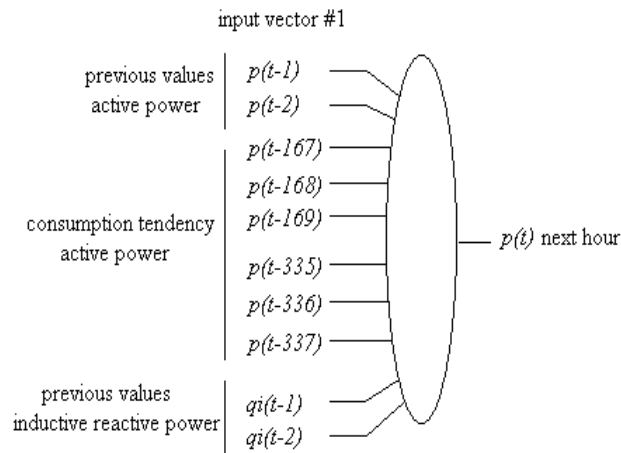


Fig. 7 . Structure of input vector #1.

### G.Forecast simulation results

In order to analyse the performance of the forecast model, the values obtained for some of the most current statistical indicators are presented in table I.

The ANN was trained and forecast results for IV #1 are presented in tables II and III. In Fig. 9 an example is represented of a load profile obtained with this type of IV.

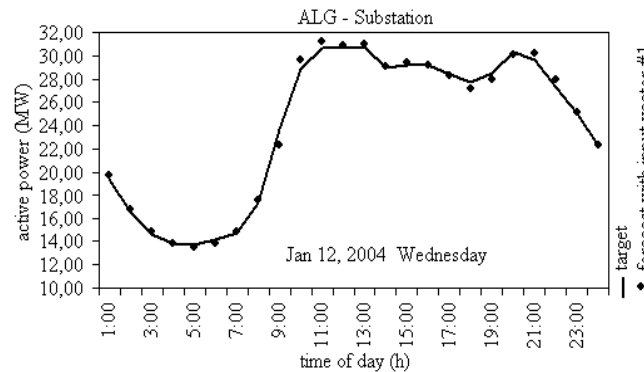


Fig. 9 . Example of Load diagram obtained using input vector #1.

The most significant indicator, as is generally accepted for comparing different forecast approaches, is the mean average percentage error (MAPE) [4]. Other statistical indicators are also necessary, however, in order to provide a more comprehensive view of the forecast results.

Parameters such as the mean error (ME) or the mean percentage error (MPE) should not deviate much from zero, as a sign of a desirable lack of bias in the forecast series of values. Other relevant indicators are the SSE and the MSE that, though not easily interpreted by themselves [19], are normally considered jointly with the other indicators for a balanced evaluation of the performance of the forecast.

The forecast results and errors obtained with IV #1 are in accordance with the expectable values for this type of methodology.

Table I - statistical parameters

error	$e_t = Y_t - \hat{Y}_t$
mean error	$ME = \sum_{t=1}^n \frac{e_t}{n}$
mean absolute deviation	$MAD = \sum_{t=1}^n \frac{ e_t }{n}$
sum of squared errors	$SSE = \sum_{t=1}^n e_t^2$
mean squared errors	$MSE = \sum_{t=1}^n \frac{e_t^2}{n}$
residual standard error	$RSE = \sqrt{\frac{\sum_{t=1}^n e_t^2}{(n-1)}}$
percentage error	$PE_t = \frac{(Y_t - \hat{Y}_t)}{Y} (100)$
mean percentage error	$MPE = \sum_{t=1}^n \frac{PE_t}{n}$
mean absolute percentage error	$MAPE = \sum_{t=1}^n \frac{ PE_t }{n}$

Table II- Next hour forecast active power - data analysis  
ALG-Substation

input vector #1		Period of simulation		1680 records		
		from Jan. 5, 2003 to March 15, 2003				
ME (MW)	MAD (MW)	SSE (MW <sup>2</sup> )	MSE (MW <sup>2</sup> )	RSE (MW)	MPE (%)	MAPE (%)
0.12	0.46	833.67	0.50	0.64	0.31	1.63

Table III - Next hour forecast active power - data analysis  
ALG-Substation

input vector #1		Period of simulation		407 records		
		from Jan. 15, 2004 to Jan. 30, 2004				
ME (MW)	MAD (MW)	SSE (MW <sup>2</sup> )	MSE (MW <sup>2</sup> )	RSE (MW)	MPE (%)	MAPE (%)
0.04	0.38	161.56	0.40	0.63	0.10	1.38

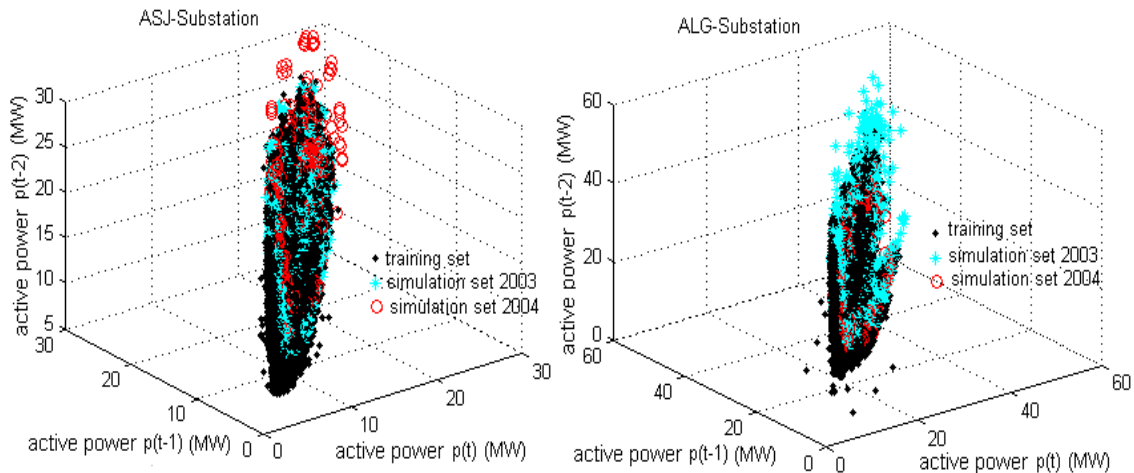


Fig. 8 Scatter plot between forecast and training sets. ASJ and ALG - Substation.

### III. A COMPARATIVE EVALUATION OF TWO ADDITIONAL PROPOSALS

It is difficult to make an exact comparative study with different documented approaches to ANN input vectors. In fact, there are many factors influencing the design of the ANN which are unknown, such as, for example, the internal structure or the number of training epochs used. Most of the times it is also impossible to have coincident data samples.

Therefore, the following comparative study is based solely on the test of different IV structures, as published by other authors. The additional IV designs are applied to the same ANN structure. Also, the training epochs and the validation process are maintained throughout. For the sake of comparability, data samples are the same for each pair of compared IVs.

Input vector #1 was compared with two other IV structures, developed by different authors. The first one (designated vector #2) (Fig.10), was developed by Fidalgo [7]. This IV was formerly applied to the distribution sector in the north region of Portugal. It uses four contiguous precedent values of the active power ( $p(t-1)$ ,  $p(t-2)$ ,  $p(t-3)$ ,  $p(t-4)$ ) and those of the two past weeks  $p(t-168)$  and  $p(t-336)$ ). It also uses information of four sinusoidal functions with the aim of informing the ANN of the consumption cycles, according to the hour of the day and the day of the week. The forecast results are presented in table IV.

In this case of vector #2, comparisons with input vector #1 are possible for all the forecast data because, in these two structures, it was possible to have the same data samples.

The error values in tables II and IV show a notorious satisfactory approximation between the target values and the results of the forecast in both cases (Fig.11). However, the MAPE value is better in the case of vector #1, as are the remaining significant indicators. An illustration of this is depicted in Fig. 12 for SSE and MAPE.

The usual discrepancies may still be found between the target values and the forecast results, as is noticeable when considering together Fig. 11 and Fig. 12. These differences are more expressive at the peak load hours. A second IV structure proposal (designated vector #3) developed by Drezga [10] was also tested.

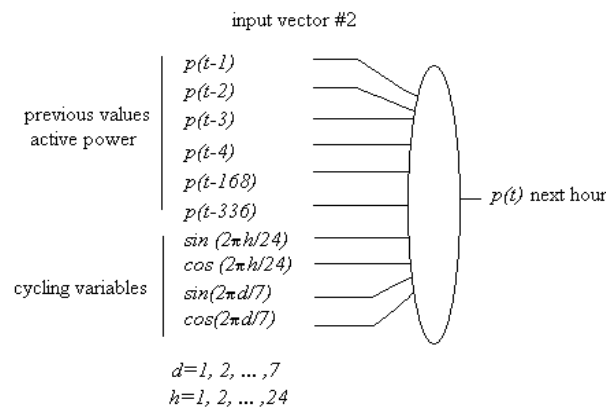


Fig.10 Structure of input vector #2.

Table IV- Next hour forecast active power - data analysis  
ALG-Substation

input vector #2		Period of simulation		1680 records		
		from Jan. 5, 2003 to March 15, 2003				
ME (MW)	MAD (MW)	SSE <sub>t</sub> (MW <sup>2</sup> )	MSE <sub>t</sub> (MW <sup>2</sup> )	RSE (MW)	MPE (%)	MAPE (%)
0.00	0.69	1474.95	0.88	0.85	0.02	2.53

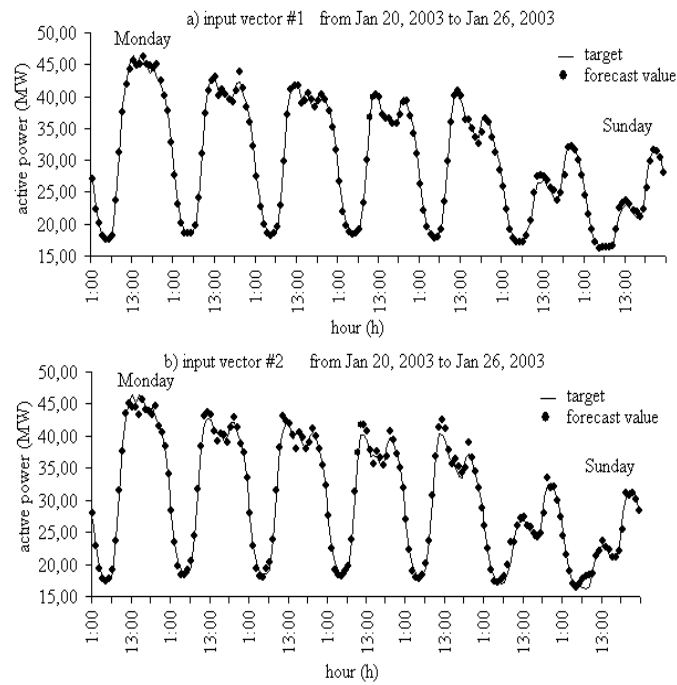


Fig. 11 Forecast results between input a) vector #1 and b) vector #2.

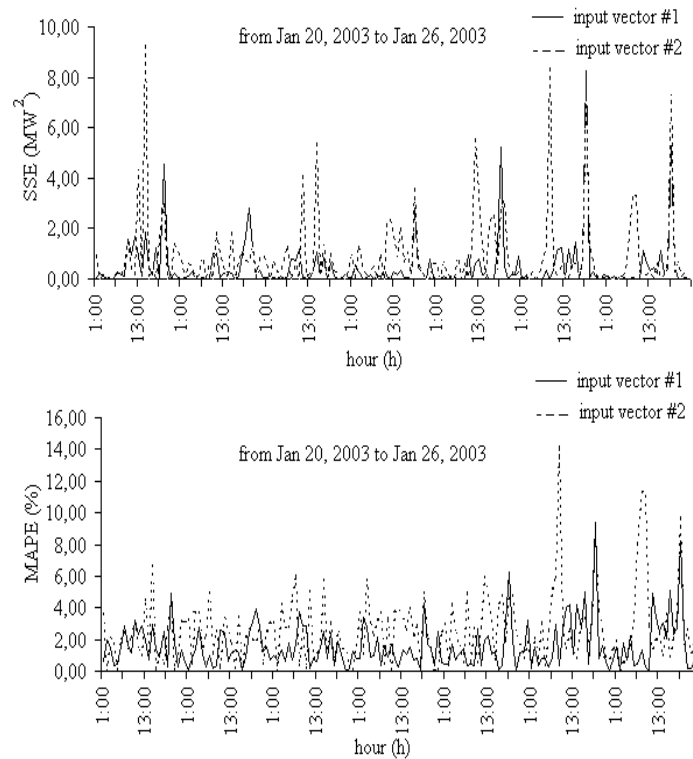


Fig. 12 . Evolution of SSE and MAPE for the forecast period between input vector #1 an #2.

This vector (Fig.13) was not originally applied by its authors to the distribution sector. The available climatic data did not allow establishing such an extensive training as in the previous one. The model was trained with data between December 1998 and May 2001, and forecast was carried out for the period from July, 1 2001 to July, 31 2001. The statistical results are presented in table V.

Once the period of forecast is not the same as in the previous pair wise comparison, a less ambitious comparative analysis was

made. Table V presents the MAPE values for both IVs for the same number of forecast records.

A single season of the year was used for the sake of comparability. The value of MAPE his higher for the structure of IV #3 in this particular case. All the other statistical parameters are within the expectable range. In Fig. 14 an example is represented of a load diagram obtained with these last vectors.

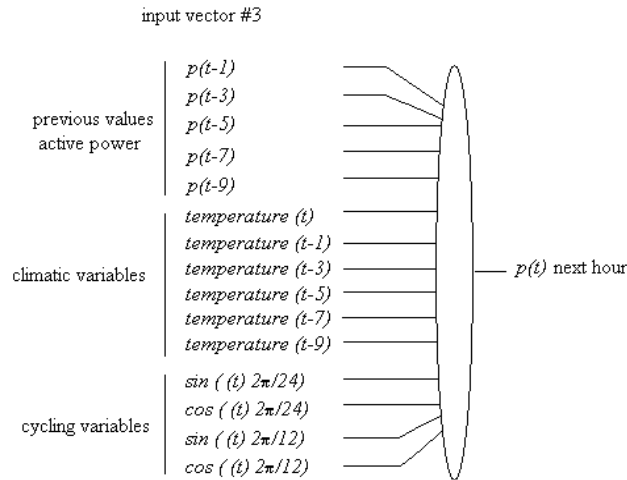


Fig. 13 Structure of input vector #3.

Table V - Next hour forecast active power - data analysis ALG-Substation	MAPE (%)
vector #3 Simulation period from July 1, 2001 to July 31, 2001 744 records	2.53
vector #1 Simulation period from July 1, 2003 to July 31, 2003 744 records	1.30

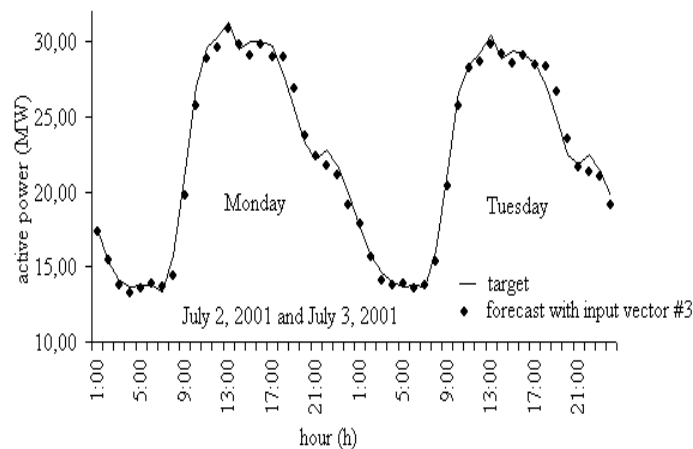


Fig. 14 Example of load diagram obtained using input vector #3.

#### IV. CONCLUSIONS

STLF has an important role in the electricity distribution sector, aiding in decision-making in actions like control and management of networks, among others. The growing trend towards electric systems unbundling makes the implementation of forecast methodologies more important.

The ANN, as a methodology for short-term forecast, has been widely used with satisfactory results. However, there is always some arbitrariness in the choice of the variables that make up the input vector. To reduce this arbitrariness, the concepts of memory range (through block entropies estimation) and consumption tendency have been used, with the aim of defining input vectors of small dimensions, avoiding model overparameterization. This kind of vector was compared to other proposals in the

literature, showing in general satisfactory results. The models were trained with three years of data values (Dec 1998 to Dec. 2002), simulated in one year (Jan. 2003 to Jan. 2004), maintaining a good performance level throughout. The models using input sinusoidal functions represent a good approach in order to account for the influence of the regularity of the time dependent variations of consumption in the model. However, the alternative use of the consumption trend concept provides a “natural” way of including this regularity, simultaneously improving the quality of the forecast results.

#### ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of EDP Distribuição - Direcção do Centro for the provided data.

#### REFERENCES

- [1] L. Philipson and H. Willis, “Understanding Electric Utilities and Deregulation,” Marcel Dekker, Inc., 1998, pp 1-24.
- [2] G. Gross and F. D. Galiana, "Short term load forecasting," *IEEE Proceedings*, vol. 75, n° 12, pp. 1558-1573, Dec. 1987.
- [3] S. Rahman, "Formulation and analysis of a rule-based short-term load forecasting algorithm," *IEEE Proceedings*, vol. 78, n° 5, pp. 805-816, May. 1990.
- [4] H. S. Hippert, C. E. Pereira and R. Castro Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. on Power Systems*, vol. 16, pp. 44-55, Feb. 2001.
- [5] IEEE load forecasting working group, “Load forecast bibliography phase I”. *IEEE Trans. on Power Apparatus and Systems*, vol. 99, pp 53-58, Jan/Feb 1980.
- [6] C. S. Chen, Y. M. Tzeng and J. C. Hwang, "The application of artificial neural networks to substation load forecasting," *Electric Power Systems Research-EPSR*, vol. 38, pp. 153-160, 1996.
- [7] J. N. Fidalgo, "Previsão de carga em saídas de subestações- resultados preliminares," presented at the 4º encontro Luso-Afro-Brasileiro de Planejamento e Exploração de Redes de Energia- ELAB'99, Rio de Janeiro, Jun 7-10, 1999, Paper ST 7-2.
- [8] Salvador Año Villalba and Carlos Álvarez Bel, “Hybrid Demand Model for Load Estimation and Short-Term Load Forecasting in Distribution Electric Systems,” *IEEE Trans. on Power Delivery*, vol. 15 n°2, pp. 764-769, April. 2000.
- [9] P. Santos, A. Martins and A.Pires “On the use of reactive power as an endogenous variable in short-term load forecasting,” *IJER-International Journal of Energy Research*, vol 27, pp 513-529, 2003.
- [10] I. Drezga, S. Rhaman S. “Input variable selection for ANN- Based Short-Term load forecasting.”. *IEEE Trans. on Power Systems*, vol. 13, pp 1238-1244, Nov .1998.
- [11] Radwan E. and Abdel-Aal, “Short-term hourly load forecasting using abductive networks,” *IEEE Trans. on Power Systems*, vol. 19, pp. 164-173, Feb. 2004.
- [12] Joshua S.Richman and J. Randall Moorman, “Physiological time-series analysis using approximate entropy and sample entropy,” *American Physiological Society*, 278: H2039-H2049, 2000.
- [13] R.Mendes, R. Lima, T. Araújo “A process-reconstruction analysis of market fluctuations,” *International Journal of Theoretical and Applied Finance*, vol 5, pp 797-821, 2002.
- [14] H.M. Al-Hamadi and S.A. Soliman, “Short-term load forecasting based on Kalman filtering algorithm with moving window weather and load model.” *Electric Power Systems Research- EPSR*, vol. 68, pp. 47-59, 2004.
- [15] Madalena Costa, ary L. Goldberger and C.-K. Peng, “Multiscale Entropy Analysis of Complex Physiologic Time Series,” *Physical Review Letters*, vol 89, n°6, pp.068102, August 2002.
- [16] Goldberger A, Amaral L, Glass L, Hausdorff J, Ivanov P, Mark R, Miteus J, Moody G, Peng C-K, Stanley I,,” *PhysioToolkit, and Physionet: components of a new research resource for complex physiologic signals*”, *Circulation* vol.101 n°3: e215-e220, June 2000
- [17] Grassberger P, Procaccia I. “Estimation of the Kolgomorov entropy from a chaotic signal”. *The American Physical Society*. vol. 28, n°4: pp. 2591-2593,1983.
- [18] C. Lu, H. Wu and S. Vemuri, “Neural network based on short term load forecasting,” *IEEE Trans. on Power Systems*, vol. 8, pp. 336-342, Feb. 1993.
- [19] De Lurgio S., “Forecasting principles and applications,” McGraw-Hill International editions, Statistics & Probability Series. Singapore, 1998, pp 36-58.