SHORT-TERM LOAD FORECAST USING TREND INFORMATION AND PROCESS RECONSTRUCTION

Santos, P.J.[a], Martins, A.G.[b], Pires, A.J.[a], Martins, J. F.[a], Mendes, R. V.[c]

[a] LabSei-Department of Electrical Engineering, ESTSetúbal/InstitutoPolitécnico de Setúbal , Rua do Vale de Chaves Estefanilha 2914-508 Setúbal Portugal: e-mail: psantos@est.ips.pt; apires@est.ips.pt jmartins@est.ips.pt : tel:+351 265 790 000 fax:+351 265 721 869

[b] Department of Electrical Engineering FCTUC/INESC, Polo 2 University of Coimbra, Pinhal de Marrocos, 3030 Coimbra Portugal; e-mail: amartins@deec.uc.pt tel: +351 239 796 279 fax: +351 239 796 247

[c] Laboratório de Mecatrónica, Instituto Superior Técnico, Av. Rovisco Pais, 1096 Lisboa Codex, Lisboa Portugal e-mail: vilela@cii.fc.ul.pt: tel +3510218417435 fax: +351 218417167

ABSTRACT

The algorithms for short-term load forecast (STLF), especially within the next-hour horizon, belong to a group of methodologies that aim to render more effective the actions of planning, operating and controlling electric energy systems (EES). In the context of the progressive liberalisation of the electricity sector, unbundling of the previous monopolistic structure emphasizes the need for load forecast, particularly at the network level. Methodologies such as artificial neural networks (ANN) have been widely used in next-hour load forecast. Designing an ANN requires, amongst other things, the proper choice of input variables, avoiding overfitting and an unnecessarily complex input vector (IV). This may be achieved by trying to reduce the arbitrariness in the choice of endogenous variables. At a first stage, we have applied the mathematical techniques of process-reconstruction to the underlying stochastic process, using coding and block entropies to characterize the measure and memory range. At a second stage, the concept of consumption trend in homologous days of previous weeks has been used. The possibility to include weather-related variables in the IV has also been analyzed, the option finally being to establish a model of the non-weather sensitive type. The paper uses a real-life case study.

Keywords Distribution Systems, Load forecasting, Measure, Memory range, Consumption trend, Artificial neural networks.

1. INTRODUCTION

Distribution companies (DISCO) operating on a scenario of complete or partial unbundling of the electricity sector are confronted with increasing demands on planning, management and operation of the networks. Relations with generation, transmission and retail companies (GENCO, TRANSCO, RESCO) are now becoming increasingly complex (Gross, 1987). Therefore, DISCOs play a major role in the managing and planning of distribution, with an emphasis on the quality of the supply.

The supply quality rules that are being imposed by the regulatory authorities are becoming more and more demanding. Thus, forecast plays a key role in this sector (Philipson, 1988). Several short-term load forecast (STLF) models have been developed in the last few decades (Drezga, 1998 and Hippert, 2001). However, few amongst them have done a specific analysis of this sector (Chen, 1996, Fidalgo, 1999 and Sargunaraj, 1997). Next-hour load forecast allows DISCOs to address issues such as: network reconfiguration, voltage control, and maintenance planning and power factor correction.

Methodologies for STLF forecast are divided in three major groups (Al-Hamadi, 2004): Models that are independent of weather changes (non-weather sensitive models), models depending on weather changes (weather-sensitive models), and hybrid models. Methodologies based on ANNs have been widely used with, to some extent, satisfactory results. However, design options are not always fully justified and frequently the models have a high complexity level (Hippert, 2001).

The most important type of variable included in the input vector (IV) is the past time-series of the variable being forecast (Hippert, 2001, Senjyu, 2002, Papalexopoulos, 1994 and Khotanzad, 1994). Other variables, of an auxiliary nature, are used and, not being directly related to electricity consumption, they are usually represented by functions of the sinusoidal or binary type with the goal of helping the ANN to detect periodic features of the load behaviour (Drezga, 1998 and Fidalgo, 1999).

The model that was developed, taking into consideration the pre-established time horizon and the low correlation between active load and weather variables may be considered a non-weather sensitive model (Al-Hamadi, 2004). In fact, the active power time-series $p(t)$ itself contains the most important IV data.
In order to diminish arbitrariness in the definition of the input vector and the prediction algorithm, we have attempted a mathematical characterisation of the stochastic process underlying the data in the experimental time-series. Reconstruction of a process involves two different, but related, steps. One is the identification of the *grammar* of the process, that is, the allowed transitions in the state space. The second step is the identification of the *measure*, which concerns the occurrence frequency of each orbit in typical samples. Identification of grammars and measures (in particular Gibbs measures) has been dealt with recently, in particular in the context of hydrodynamic turbulence and market analysis (Chazottes, 1998 and Mendes, 2002). Some of these techniques will be applied in Section 3 to our experimental time-series.
The correlation between active load in homologous days of the week has also been considered.

The paper has the following structure: the case study is presented in Section 2, where the different types of substations are described, along with the collected data, the time length of the data series and the results of various correlations between consumption and weather variables. The application of the process reconstruction techniques is carried out in Section 3. In Section 4, the concept of trend is used and the input vector established. Section 5 presents results from the simulations. Finally, some conclusions on the methodology are included in the last Section.

2. DATA ANALYSIS AND CASE STUDY CHARACTERISTICS

The case study is located around the city of Coimbra, in the centre of Portugal, comprising three substations. Installed capacity and voltage level of these substations is of average dimension (Alegria (ALG), Relvinha (RLV) e Alto de São João (ASJ)) (fig.1). These three substations are responsible for the electrical power supply to the city of Coimbra. The data series obtained from each of these substations have a time span of approximately three years (from Dec. 21, 1998 to Dec. 20, 2001).

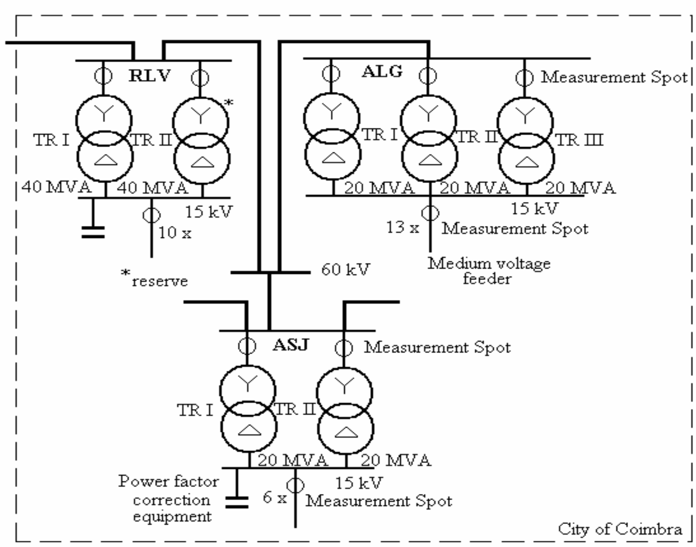

fig. 1 - Simplified one line diagram of the medium voltage network supplying the city of Coimbra.

Data on the following variables was collected: active power (MW), inductive and capacitive reactive power (Mvar), with a maximum time resolution of one hour. Several types of weather variables were also collected, with the purpose of carrying out correlation analyses, in order to assess the advantages of including these variables in the IV. Correlation of electricity consumption with climatic data may be strong in certain climates, particularly when high humidity and temperature are current in summer or very low temperatures occur in winter. In the case study, moderate temperature swings are accompanied by moderate humidity conditions as well. Hence, a strong correlation was not to be expected (Santos, 2003).

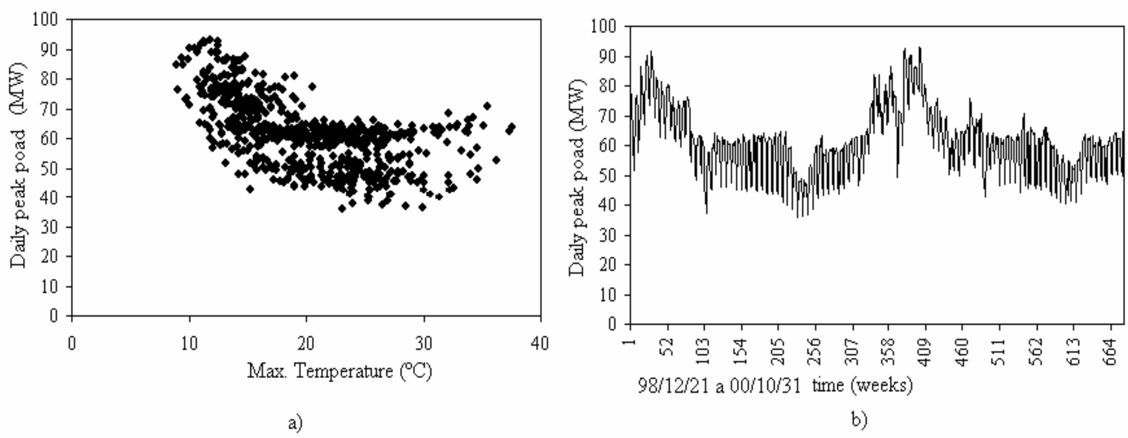

fig.2. a) - scatter plot between the peak load day and the maximum temperature; b) - representation of variation of peak demand (both between Dec. 1998 and Oct. 2000, city of Coimbra).

Analysing the diagram in figure 2, one observes that the daily peak load drops with the coming of warmer seasons, which indicates a low impact of ventilation and air conditioning loads. According to this, one would expect stronger correlations only in the wintertime. The forecast time spans, as well as the weather conditions, do not favour strong correlations between temperature and electrical energy consumption (fig.3). Based on this analysis, the composition of the IV relies essentially on the endogenous variables.

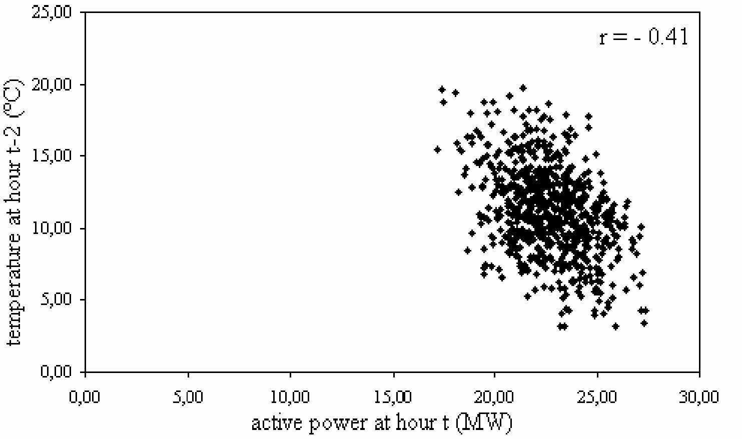

fig.3 . Scatter plot of hourly values of active power load at hour t and temperature at the hour t-2, 11:00 am to 23:00 pm, from Dec. 21, 1998 to March 15, 1999.

3. PROCESS RECONSTRUCTION AND MEMORY RANGE

Usually, the number of consumption instances, prior to the value to be estimated, that one must take into account, is established in an arbitrary manner, based on experience obtained by using correlation analysis (Drezga, 1998 and Hippert, 2001) (fig.4). What one must find out is whether the amount of contiguous information that is chosen is appropriate or whether it merely contributes to over-parameterize the model.

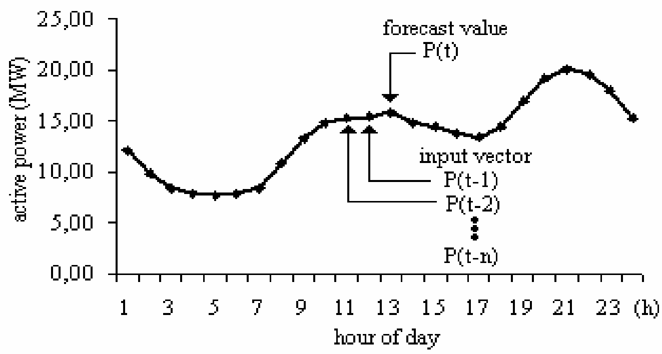

fig.4. Daily load pattern. Representation of contiguous instances of consumption.

Coding and computing block entropies as used in (Mendes, 2002) allows a rigorous estimation of the effective memory range of process. This study was carried out for the data from the three substations yielding similar results. For that reason, it has been decided to present only the results referring to the substation RLV (fig.1).

The data was collected by the existing supervisory control and data acquisition system (SCADA), with the time-scale resolution of one hour. The data collection started in Dec. 21, 1998. The period chosen for the analysis of the process was from Dec. 21, 1998 to March 15, 2001, producing a set of N=19584 values. The signal was discretized according to the alphabet: ? = {-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5}, containing eleven levels (k=11). This discretization allows a satisfactory representation of the active power load pattern (fig.5). The whole signal length is, thereby, translated by means of this alphabet and the active power series described by symbol sequences ? $=p_1, p_2, \ldots p_i, \ldots,\ p_k \in \Sigma$.

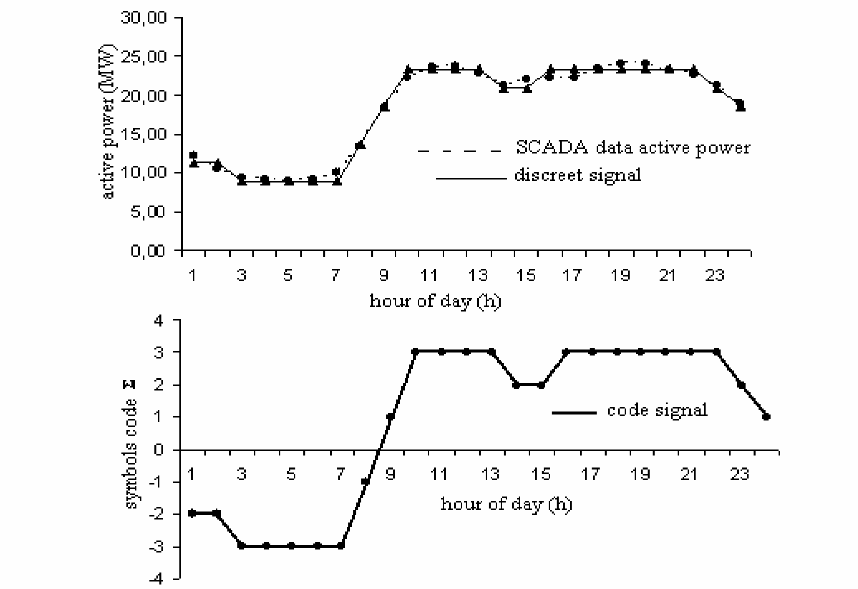

fig.5. Example of the discretizing the original SCADA data by means of the proposed alphabet (RLV Substation).

In the alphabet ? the maximal number of distinct blocks of length N is 11^N. The graph of figure 6 a) compares, for each size N, the actual number of distinct blocks that are present in the signal with the maximum 11^N . The deviation of the values from the maximum shows that the signal rather than being completely random, has a non-trivial grammar.

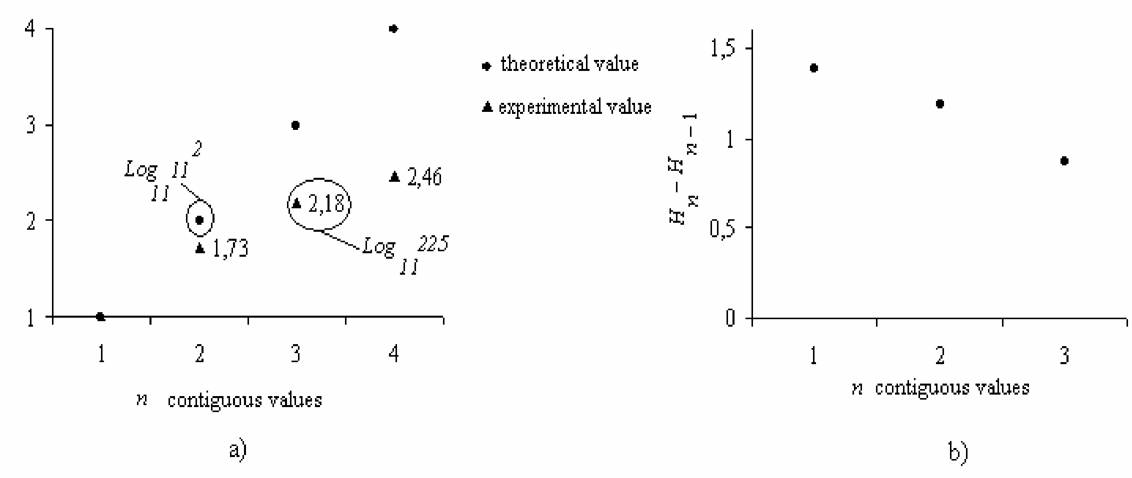

fig.6. a) possible and actual number of combinations of blocks of the alphabet; b) values of entropy.

A very general class of measures for stochastic processes is the class of Gibbs measures. In this context a very simple characterisation of the memory range of the process is obtained from the growth of the block entropies. Let

$$H_k = -\sum_{p_1 \ldots p_k} m[p_1 \ldots p_k] \log(m[p_1 \ldots p_k]) \quad (1)$$

be the entropy associated to blocks of size k, $\mu[p_1 \ldots p_k]$ representing the probability of finding within the series a sequence of contiguous values of the $p_1 \ldots p_k$. type (Chazottes, 1998 and Mendes, 2002). Using the empirical block probabilities $\mu[p_1 \ldots p_k]$ one computes H_k for successively larger k. Then, the memory range of the process is found when $H_k - H_{k-1}$ tends to a constant value. In practice, for a long memory process, this difference after converging to its constant value, starts to decrease. This is an effect of lack of statistics, because for a finite sample there is a small probability that all grammatically allowed blocks will appear in the signal.
The graph of figure 6 b) shows the $H_k - H_{k-1}$ values computed for our time series. It clearly shows the short-term memory of the signal, in the sense that the next-hour value depends essentially on the information related to the previous state. Therefore, it may be assumed that the information based on many contiguous values is of small importance and that the main focus should be on incorporating into the IV the information regarding the previous hour together with other (non time-contiguous) information as explained below.

4. THE TREND CONCEPT

The analysis of block entropies has revealed that the use of long chains of contiguous values does not result in any sort of advantage in the design of the IV of the ANN. Moreover it possibly contributes to an overparameterization of the model.
The composition of the input vector will have to rely essentially on a careful analysis of both the auto-correlations of the active power and its possible interdependences with the exogenous variables. One notices that the values of auto-correlation are more important in the cases of the two previous hours, deteriorating quite rapidly for deeper excursions into the past. This behavior is to be expected due to the fact that one is comparing different periods of the day (fig.7). The values become higher when one considers the correlation of the most recently available active power values with those of homologous instants of homologous days of previous weeks, showing two relative maxima in the two previous weeks (fig.7).

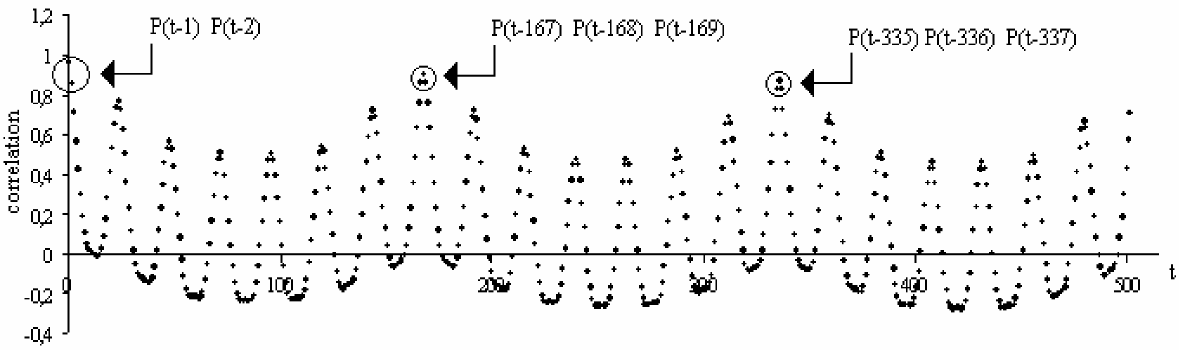

fig.7. Auto-correlation values, showing local maxima at homologous instant in the past.

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 evolution of the auto-correlation coefficients is always downward as one moves deeper into the past, which can be explained by the seasonal variation of consumption, which entails different load patterns.
We have, therefore, divided the information into periods (fig.8), according to the evolution of the average daily temperature.

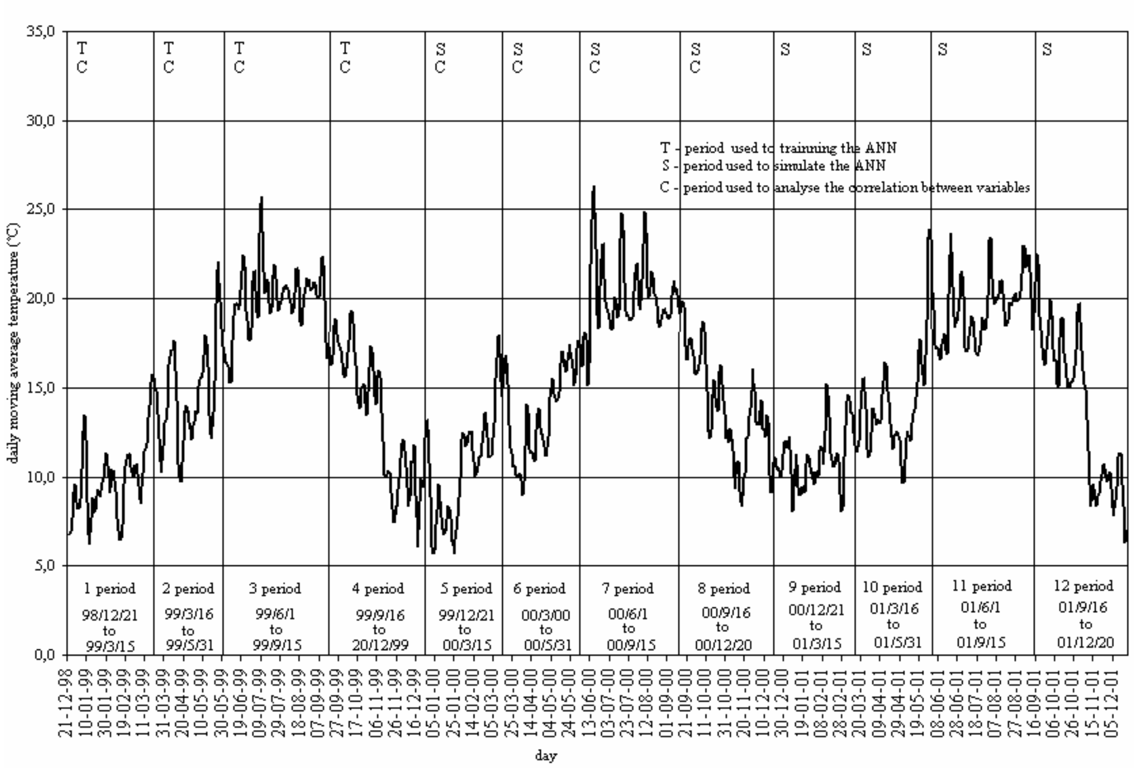

fig.8 . Time-scale divisions adopted in the active power series analysis.

In spite of the low correlation of electricity consumption with weather variables in this forecast horizon, this division into periods allows for a neural network (ANN) that has been trained in a given period to be used in a simulation in similar weather conditions, helping the network to better deal with weather-related effects. The input vector is the one defined in fig.9. It also includes reactive power instances in the two hours prior to the forecast *Q(t-1)* e *Q(t-2)*, due to the fact that they show important correlations with the target variable, and generate improvements in the performance of the model (Santos, 2003). This is shown through the beneficial influence on the MSE regarding the ANN training set of values.

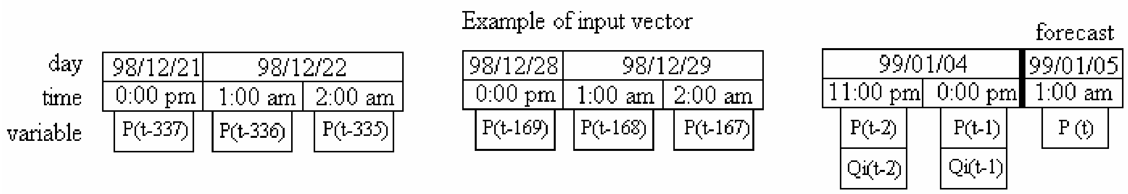

fig.9 . Composition of the input vector.

5. SIMULATION RESULTS

A standard feedforward backpropagation ANN have been used for the forecast models, with a fully connected architecture and a single hidden layer, the hyperbolic tangent being used as the common activation function. The output is activated with linear functions. 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.

Simulations have been carried out with data not used in training, testing or validating the ANN. In Table I a set of calculated parameters is presented for the same periods that help to assess the model performance.

Table I .Next hour forecast active power – data analysis.

RLV SE	ME (MW)	MAD (MW)	MSE (MW^2)	RSE (MW)	MPE (%)	MAPE (%)
Period 5 from 99/12/21 to 99/03/15	-0.03	0.36	0.25	0.50	-0.27	2.27
Period 6 from 99/03/16 to 99/05/31	-0.03	0.41	0.41	0.64	-0.43	2.76
Period 7 from 99/0/1 to 99/09/15	0.05	0.28	0.15	0.38	0.42	1.97
Period 8 from 00/09/16 to 00/12/20	-0.02	0.33	0.22	0.47	-0.33	2.16
Period 9 from 00/12/21 to 01/03/15	-0.11	0.37	0.31	0.56	-0.75	2.31
Period 10 from 01/03/16 to 01/05/31	-0.02	0.27	0.16	0.40	-0.31	2.04
Period 11 from 01/06/01 to 01/09/15	0.05	0.26	0.13	0.36	0.42	1.97
Period 12 from 01/09/16 to 01/12/20	-0.11	0.37	0.27	0.52	-1.09	2.75

Figure 10 shows some examples of the results obtained with the models of the next-hour load forecasts, which should be self-explanatory.

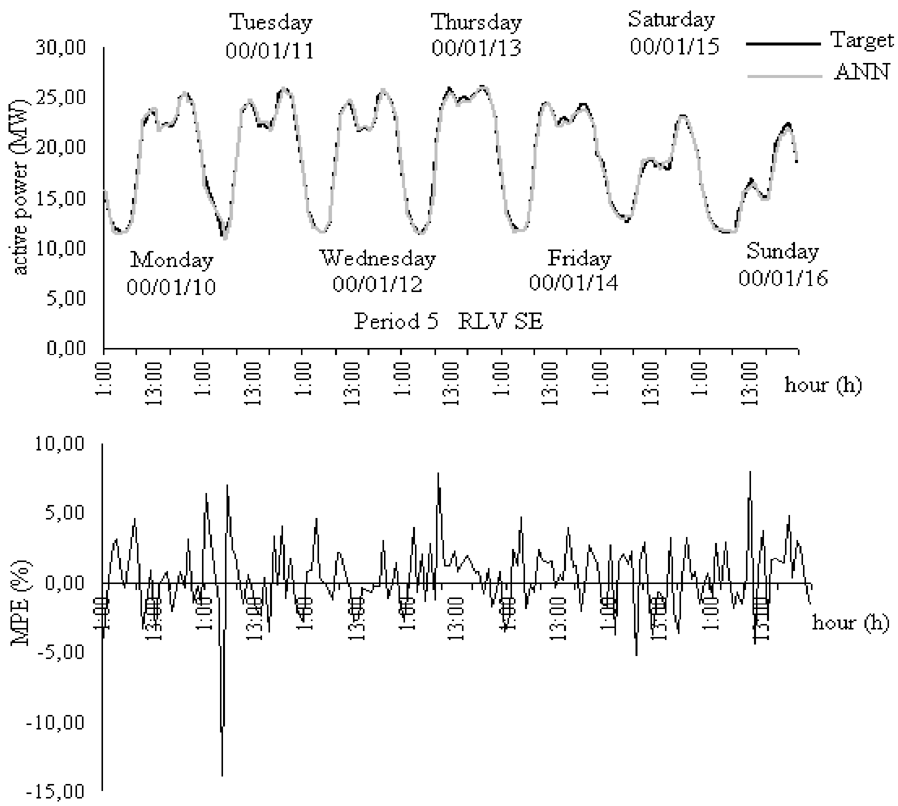

fig.10 . Simulations results, showing one week actual and predicted active power and, below, the corresponding mean percentage error (MPE) values.

6. CONCLUSIONS

STLF has an important role in the electricity distribution sector, as it is subsidiary to the control and management of networks, aiding in decision-making. The growing tendency towards electric systems unbundling makes the implementation of forecast methodologies in all levels of the EES all the more necessary.

The ANN, working as a methodology for short-term forecast, has been widely used with satisfactory results. However, there are always some arbitrary traits in the choice of the variables that constitute the input vector. To reduce this arbitrariness, the concepts of memory range (through block entropies estimation) and consumption trend have been used, with the aim of defining input vectors of small dimensions, avoiding model overparameterization.

This kind of vector has been compared to other proposals in the literature, showing in general satisfactorily improved results. The models were trained with consumption values of the year 1999 and simulation has been performed up until 2002, maintaining a good performance level throughout. They were also tested in different types of substations with different load configurations. The reactive power was also included in the composition of the IV, producing a slight improvement in the model behavior.

7. ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of EDP Distribuição - Direcção do Centro for all the provided data and João T. Martins for his careful translation.

8. REFERENCES

Al-Hamadi H, Soliman S. 2004. Short-term load forecasting based on Kalman filtering algorithm with moving window weather and load model. *EPSR- Electric Power System Research*. 68: 47-59.

Chazottes J, Floriani E, Lima R. 1998. Relative entropy and Identification of Gibbs measures in dynamical systems. *Journal of Statistical Physics*. **90** (3/4): 697-725.

Chen C, Tzeng Y, Hawang J. 1996. The application of artificial neural networks to substation load forecasting. *Electric Power Systems Research*. **38**: 153-160.

Drezga I, Rhaman S. 1998. Input variable selection for ANN- Based Short-Term load forecasting. *IEEE Trans. on Power Systems*. **13** (4): 1238-1244.

Fidalgo J. 1999. Previsão de carga em saídas de Subestações- resultados preliminares *ELAB'99 4º encontro Luso-Afro-Brasileiro de Planejamento e Exploração de Redes de Energia* :339-348.

Gross G, Galiana F. 1987. Short term load forecasting. *IEEE Proceedings*. **75** (12): 1558-1573.

Hippert H, Pereira C, Souza R. 2001. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Trans. on Power Systems*. **16** (1): 44-55.

Khotanzad A, Afkhami-Rohani R, Lu T-L, Abaye A, Davis M, Maratukulam D. 1997. ANNSTLF- A neural network based electric load forecasting system. *IEEE Transactions on Neural Networks*. 8 (4): 835-845.

Mendes R, Lima R, Araújo T. 2002. A process-reconstruction analysis of market fluctuations. *International Journal of Theoretical and Applied Finance*. 5 (8): 797-821.

Papalexopoulos A, Hao S, Peng T-M. 1994. An implementation of a neural network based load forecasting model for the EMS. *IEEE Trans. on Power Systems.* 9 (4): 1956-1962.

Philipson L, Willis H. 1998. Understanding Electric Utilities and Deregulation, Marcel Dekker, Inc: New York; 1-24.

Santos P, Martins A, Pires A. 2003. On the use of reactive power as an endogenous variable in short-term load forecasting. *IJER International Journal of Energy Research*. 27 (5): 513-529.

Sargunaraj S, Gupta D, Devi S. 1997. Short-term load forecasting for demand side management. *IEE Proc.-Gener. Tramission Distribution.* 144 (1): 68-74.

Senjyu T, Takara H, Uezato K, Funabashi T. 2002. One-hour-ahead load forecasting using neural network. *IEEE Transactions on Power Systems.* 17 (1): 113-118.