A data mining approach to support the development of long-term load forecasting

M. R. Maia[1], K. de Oliveira Gonçalves Veloso[1], M. T. Okamoto[1], A. dos Santos Rigueira[2], G. M. Tavares[2], Â. M. Cister[3], M. A. F. Zarur[3], F. T. de Souza[3], G. S. Terra[3,4], A. G. Evsukoff[3] & N. F. F. Ebecken[3]
[1]*FURNAS Centrais Elétricas S.A., Brazil*
[2]*UFF, Fluminense Federal University, Brazil*
[3]*NTT, COPPE/Federal University of Rio de Janeiro, Brazil*
[4]*CEFET-Campos/UNED-Macaé, Brazil*

Abstract

Load forecasting is an important subject for power distribution systems and has been studied comparing different points of view. In general, load forecasts should be performed over a broad spectrum of time intervals, which could be classified into short-term, medium-term and long-term forecasts. Several research groups have proposed various techniques for either short-term load forecasting or medium-term load forecasting or long-term load forecasting.

This paper presents two approaches for modelling the long-term load forecasting: a neural network (NN) and a non-linear (cause/effect) model. The data used by the models are gross domestic product (GDP), the national minimum salary, the electrical energy price, the estimated national population and the total number of electrical connections.

The suitability of the proposed approach is illustrated through a long-term load forecasting application (electricity consumption in Brazil ten years ahead).
Keywords: neural networks; long-term; load forecasting; power distribution systems.

1 Introduction

Brazil's electric power system (EPS) consists of two major interconnected systems (SIN) and many small isolated systems. The energy is mainly generated

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060341

by hydraulic power plants (90%). The National Power System Operator (ONS) is responsible for the control and transmission of this generation planning and programming the entire operation. The government controls the company EPE which is responsible for implementing Brazil's electrical power policy.

In the beginning of 1995, the Brazilian electrical sector was in a deep structural crisis. As a solution, the restructuring of the electrical power sector lead to a liberalization process where the State assumes a strategic position. Lots of changes have been made such as privatizations (focused on electrical distribution company), the creation of new companies such as CCEE/MAE (created in order to oversee competition in the future wholesale market) and EPE (created to be in charge of the planning of power sector supply), building the Brazil-Bolivia gas pipeline (increasing the thermal power generation using natural gas) and conceiving a new act, the Public Private Partnership (PPP) which establishes the partnership in the infrastructure area between the public and private sectors.

Considering those comments load forecasting is an important subject for power distribution systems and has been studied from different perspectives. In general, load forecasting should be performed over a broad spectrum of time intervals, which could be classified into short-term (hours, days ahead), medium-term (month, year ahead) and long-term (five, ten or more years ahead) forecasts.

Long-term load forecasting has and important commercial use, once power supply companies make their generation/distribution contracts on long-term bases and by the end of the contract the balance is negotiated, at present time rates, with CCEE.

It also has a strategically application, in view of the fact that the study, project and operation of a new power generation unit take many years to be concluded.

An also important factor is that the national economic growth is strictly related to the increase of the electrical energy consumption which can be measured by the gross domestic product (GDP) indicator (ex.: the economical impacts caused by energy crisis and the rationing program in 2001).

In the last years, many articles were published describing load forecast systems [1–4]. Short, medium and long-term forecasting have been predicted by traditional statistical models and especial data mining techniques like neural networks [5], genetic algorithms, neuro-fuzzy model [6], and others.

Particularly, the neural model presents some characteristics that make it attractive to the forecast areas [7], such as: (a) it is a self-adaptive method directed by the data itself. The knowledge is captured by the model through examples, in other words, learning by experience; (b) after the learning, it presents generalization capacity; (c) it approaches any continuous function in the desired precision; (d) it is a non-linear model, thus much more generic.

Considering the existing relation between the different data related to the phenomenon, the model based on this technique offers great potential to make an accurate prediction of the energy consumption as it will be shown in this article.

This paper presents two approaches, based on computational intelligence techniques – neural nets and genetic algorithm – and statistical methods (cause-effect/econometric model), for modelling the long-term load forecasting (ten

year ahead). The data used by the models are gross domestic product (GDP), the national minimum salary, the electrical energy price, the national estimated population and the total number of electrical connections.

The development of this work was made in some stages. The first step was the selection of the input data. Then, the selected data were checked in order to identify errors and inconsistencies. The following phase corresponds to the analysis of the data when some studies were carried through for a better understanding of its behaviour, economic effects, tendencies, etc. In the next stage, the models were generated and tested.

After having generated the models, the best models, according to the chosen evaluation convention, were selected and turned into a procedure to forecast the future electric power consumption. The results were compared to those obtained with the top-down and bottom-up forecasting approaches by EPE (Energy Research Company).

The load forecasting models has been used by FURNAS; a company controlled by ELETROBRAS with a regional activity range and responsible for developing large generation and transmission projects in the Southeast and Middle-West regions. FURNAS is responsible for 20% of electricity generation in Brazil providing services in the southeast region responsible for 65% of Gross Domestic Product (GDP) where 48% of consumers live (Electricity Connections).

2 Data selection and preparation

The main principles to select the historical data sets are: relevance with the problem, availability, confidence, continuity, and detailed information (national, regional and state values). These data are summarized in table 1. The long-term load forecast (ten year ahead) is shown at the first line (the grey one). Some providers' information is shown below.

Table 1: Historical data sets.

Historical Data	Period	Covering BR	Reg	UF	Unit	Data provider
Electrical Power Consumption (EPC)	62-04	x	x	x	MWh	ELETROBRAS (SIESE)
Electrical Power Generation (EPG)	70-04	x	x	x	MWh	
Number of Electrical Connections (NEC)	70-03	x	x	x	-	
Population	60-50	x	x	x	-	IBGE, MME (BEN), IPEADATA
National Gross Domestic Product (GDP)	70-05	x	x	x	R$	
Average monthly income (AMI)	82-05	x		x	R$/month	IBGE (PME), IPEADATA
National Consumer Price Index (INPC)	63-05	x			%	IBGE, IPEADATA
National Minimum Salary (NMS)	63-05	x			R$/month	IPEADATA
Tariff - Price for electrical power service	70-05	x			R$/MWh	ANEEL

ELETROBRÁS(SIESE)	Brazilian Business Information System for the Energy Sector	www.eletrobras.gov.br
MME (report BEN)	Brazil's Minister of Mines and Energy / National Energy Balance	www.mme.gov.br
EPE	Energy Research Company	www.epe.gov.br
ONS	National Power System Operator	www.ons.org.br
IBGE (PME)	Brazilian Institute of Geography and Statistics / Monthly Employment Survey	www.ibge.gov.br
IPEA/IPEADATA	Institute of Applied Economic Research	www.ipeadata.gov.br
MAE	Power market administrator	www.mae.org.br
CCEE	Power Chamber of Commerce	www.ccee.org.br
ANEEL	Government Power Agency	www.aneel.gov.br

2.1 Population x Number of Electrical Connections (NEC)

The national electrical power consumption grows when the number of electrical connections (NEC, or number of electrical customer) or the electrical power consumption per connection (ρ) increase as formalized in equation (1).

$$EPC_{region}^{sectors}(t) = \rho_{region}^{sectors}(t) \times NEC_{region}^{sectors}(t) \quad (1)$$

where

$\rho_{region}^{sectors}$	the electrical power consumption per connection, (MWh/year)
Sectors	residential or non-residential (industrial+commercial+others)
region	Brazil or State

The electrical power service reaches 95% of households in the country. The governmental program 'Light for Everyone' aims to provide until 2015 energy to all the households in the country, especially to the north and northeast regions.

There are many factors, population and Gross Domestic Product (GDP) for example, that contribute to increase the number of electrical customers such as those above mentioned. The asymptotic relation between population and number of electrical connections presented during the past 34 years (see figure 1), inspired the choice for the auto-regressive models as the one formalized in equation (2).

$$R(t) = Pop(t)/NCE(t) = a_0 + a_1 \times R(t-1) + a_2 \times R(t-2) \quad (2)$$

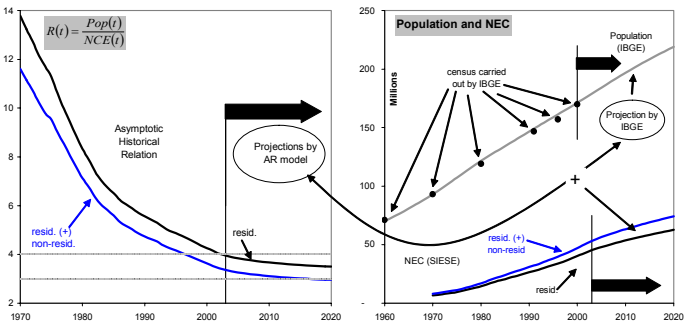

Figure 1: Population and NEC (projections).

Evaluating the projections, estimated population by IBGE and $R(t)$ by AR-models, the number of electrical connections is projected in equation (3).

$$NCE(t) = \frac{Pop(t)}{R(t)} = \frac{Pop(t)}{a_0 + a_1 \times R(t-1) + a_2 \times R(t-2)} \quad (3)$$

2.2 Gross domestic product x electrical power consumption

The historical data about GDP and EPC is available at different governmental sources. The detailed published reports by SIESE describe generation and,

mainly, consumed electrical power – difference between generation and losses – by places and by sectors (residential, commercial, industrial and others).

However, important differences exist among them regarding data sources and concepts. The National Energy Balance (BEN), printed by MME, for example, displays EPC considering the classical auto-production (the production of electrical plants set up by, generally, industrial establishments to supply their own needs). The SIESE considers auto-production, called transported auto-production, but only the part that is transported throw SIN and, consequently, controlled by ONS.

The market share of auto-production electrical generation increased from 5% to 10% in the last ten years [8]. The new scenery brings "uncertainty" about load forecasting models because the historical data sets was related to different politic, economic and market reality.

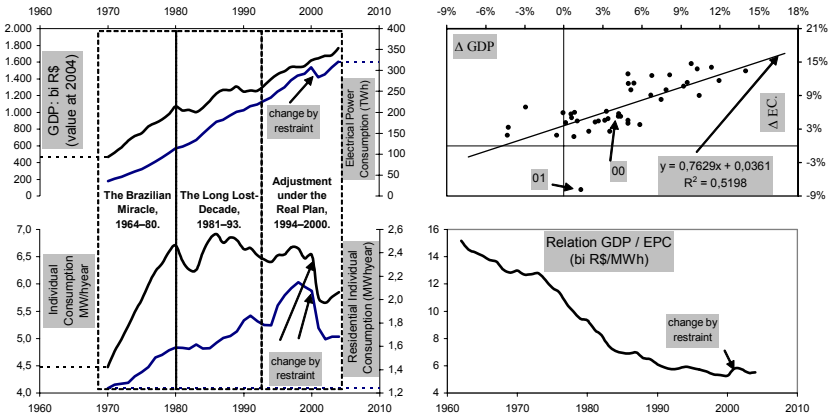

Figure 2: GDP x EPC.

The GDP is a financial information (sources: IBGE, MME (BEN), etc) and contemplates the meaning time value of money – present, past and future value. Those concepts must be considered when series like that (Tariff, Prices and National Minimum Salaries) are used by forecasting models.

The GDP published by IBGE and MME/BEN is detailed by sector (Industrial, Agricultural, Services and Energetic) and by place (Brazil and states). Brazil's growth pattern during the 20th century [9] has direct influenced over power electrical consumption. The historical individual power consumption, for example, shows patterns changes including exogenous effects, like power restriction in 2001. The relationship between these dataset was explored in figure 2.

2.3 Financial data: MAI, NMS, TARIFF, INPC

The National Consumer Price Index (INPC) was select to build the deflator index transforming the historical value into present value of money for these

three financial data – monthly average income (MAI), national minimum salary (NMS) and the price of electrical power (tariff).

The MAI and NMS are related to the income distribution and population purchase power. These data and tariff affect the pattern of EPC, represented by individual EPC, and are affected by GDP. There is a dynamic relation between all these data. To map them is beyond the scope of this work. Restricting the focus on residential sector it is possible to map, qualitatively, some relation between the dataset as expressed in equation (4), which represents the relation between tariffs, pattern of EPC and monthly average income.

$$Resid\ Electricity\ Cost_{\%}(t) = \rho_{year}^{resid.}(t-1) \times Tariff_{year}^{resid.}(t) \Big/ MAI(t) \qquad (4)$$

These indexes represent how much a consumer would pay for the same electrical consumption of the year before considering the new up dated tariff. The MAI increased (1993–1998) when Real Plan started (a government economic plan). At the same time, the price of electrical power grew but less then the consumer income until 1998 when the MAI reaches its maximum value followed by a down movement. These economic facts could explain, partially, the small decrease in the residential consumption pattern in 1999 and 2000. The power restriction (2001) forces the consumption to decrease. When the power restriction stops, 2002, the power consumption did not increase. This fact could be explained, partially, by the decrease of the income and the augment of tariff as show in figure 3.

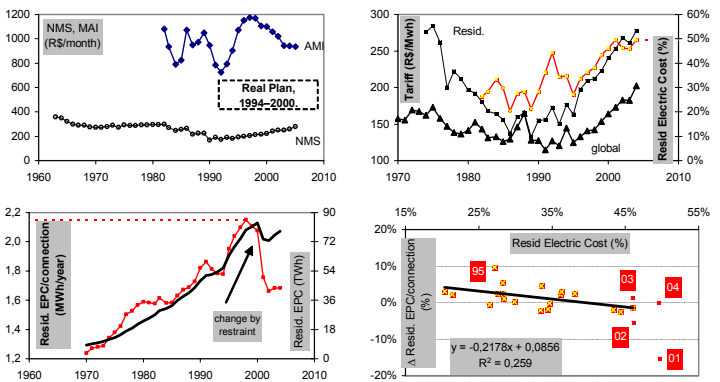

Figure 3: Financial data.

The 1980s and 1990s was characterized by sharp fluctuations in economic activity. The combination of hyperinflation and economic stagnation in the 1980s forced the government to try many heterodox inflation stabilization programs controlling electrical power prices.

3 Cause-effect model

The data mining process demands a large dedication to prepare the data base which may involve cleaning data, data transformations, checking inconsistencies,

etc. The procedure of looking for historical, national or regional facts is very important because some apparent inconsistencies could be explained by the facts such as: restrictions of power, electric power consumption and the national power plant capacity, macroeconomics politicize, etc. Another way to find inconsistencies is to analyze graphically the simple historical data with the derived one which is a new series created by relation with another. Those transformations motivated the study presented before and the cause-effect model. The cause-effect model was developed in terms of the linear relationship between GDP (bi R$) and the EPC/connections (MWh/year) expressed by equation (2):

$$I(t) = \frac{GDP(t)}{\rho_{sector}(t)} = \frac{GDP(t)}{EPC_{sector}(t)/NEC_{sector}(t)} = A*t + B + error \quad (2)$$

The linear relationship present by $I(t)$ from 1970 to 2000 changes after the power restriction period as shown in Figure 4. The NEC was predicted as described at previous section. The new value for coefficients A (slope) and B (intercept) should be calculated including values from 2002. The inclusion of 2001 values incorporates inconsistencies to the model because of the imposed power restriction.

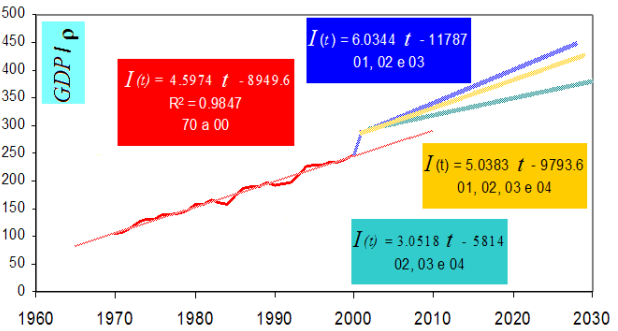

Figure 4: Cause-effect model.

The future value for EPC (forecasting value) depends on how GDP is projected and can be expressed as equation (3)

$$EPC_{sector}(t) = NEC_{sector}(t) \times GDP_{projected}(t) / (A*t + B) \quad (3)$$

4 Data mining technique (ANN and GA)

Artificial Neural Networks (ANN) (as compared to real ones) are mathematical systems comprised of a number of "processing units" that are linked via weighted interconnections. A processing unit is essentially an equation which is often referred to as a "transfer function". A processing unit takes weighted signals from other neurons, possibly combines and transforms them and then outputs a numeric result. Processing units are often considered crudely analogous

to real neurons and since they are linked together in a mesh or network, the name Neural Networks was coined.

Many neural networks have their neurons structured in "layers" that have similar characteristics and execute their transfer functions in synchronization (at the same time, virtually speaking). Nearly all neural networks have neurons that accept data and neurons that produce outputs.

The behavior of neural networks, how they map input data to output data, is influenced primarily by the transfer functions of neurons, how they are interconnected and the weights of those interconnections.

Typically, an architecture or structure of a neural network is established and one of a variety of mathematical algorithms is used to determine what the weights of the interconnections should be to maximize the accuracy of the outputs produced. Neural networks are "trained", meaning they use previous examples to establish (learn) the relationship between the input variables and the predicted variables by setting these weights. Once the relationship is established (the neural network is trained), the neural network can be presented with new input variables and it will generate predictions.

This application uses Genetic Algorithms (GA) to evolve neural network structures while simultaneously searches for significant input variables to maximize the predictive accuracy of the resulting neural network models.

The effectiveness of the genetic algorithms capabilities cannot be overstated. For example, a problem consisting of finding the best combination (subset) of 20 inputs and up to 15 hidden nodes in a back propagation neural network is a combinatorial problem with over 16 million permutations. To train a network in a full search of all permutation would be a good project for a super computer. But with genetic algorithms, an excellent solution often appears in less than 1500 evaluations which is 0.009% of the total possible configurations.

Using some statistical data analysis to assist, highly fit networks are often found in the first 30 to 50 neural networks evaluated. This is clearly an efficient means for discovering effective network structure/input combinations.
Several tests were performed and some results are discussed in section 5.

5 Network neural model

The electric power price (tariff) and the national minimum salary (NMS) have direct impact over residential EPC and other sectors. The commercial and industrial economic activity is responsible for a great share of the GDP, and is also related to the population purchasing power. The neural models were developed to capture the non-trivial relationship between the variables: GDP (all and the energetic share), national minimum salary (NMS), electric power prices and population.

The samples presented to neural network model during the training phase were composed by information obtained at a certain period. The input and output values are at time *t*. The best neural model was select to forecast electrical power consumption using project values for the input data sets. This approach allows the study of the influence of a particular input over output value (sensibility

analysis). The 34 years of data (1970–2004) was segmented in training, testing and validation of data sets.

6 Results

The cause-effect and neural model results were compared with the EPE forecasting values. The EPE prints technical reports about long-term load forecasting (ten years ahead). The historical series has been analysed. The forecasting value was project by scenery methodology and three results were obtained: optimist (A), realistic (TRA) and pessimist (B). Each scenery has its own projection to the correlated data (GDP, NMS, and Tariff). The forecasting value (ten years ahead) of EPC and the related index are shown in figure 5. The cause-effect and neural forecast are very close to the pessimist forecast by EPE.

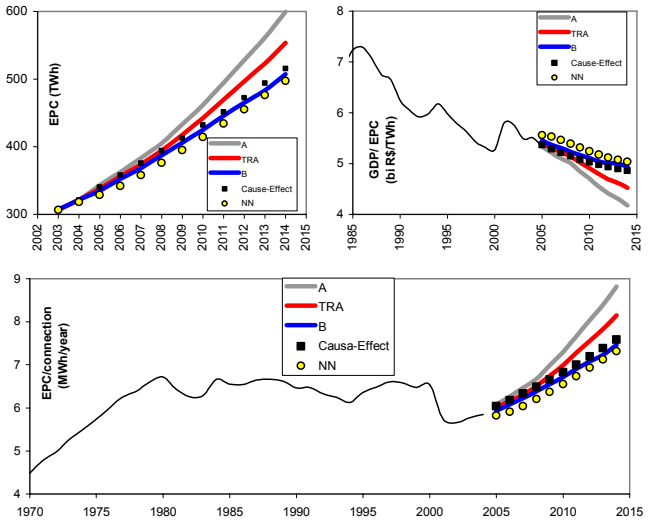

Figure 5: Consumption forecast by EPE.

7 Conclusions

Traditional forecasting models requires training sample of input and output values delayed by Δt lag interval. This was the first idea when the project started. When the preparation of data base was concluded and some relationship between variables was studied the scope of the models changed. Then, the models were built with samples of input and output values at the same period of time. The non-trivial relationship between series like the ones evolved in that study was captured by the cause-effect and neural model. The forecasting with those models was more realistic because a scenery study could be made with the projection of the relevant variables.

The forecasting values by cause-effect and neural model are very close to the pessimist projection by EPE. The very high projection for the realistic and optimist results may be explained once the information has been obtained from

the Government Company. The government is an opinion-maker so it needs to be very cautious when making such projections.

By the other hand, FURNAS forecasting models should be more realistic and conservative. Investments and contracts will be made based on the forecasting values.

Another important factor to be considered when analysing the forecasting is the study of derived index.

The procedure developed by that project which consists in acquiring and cleaning data, pre-processing values, index study and forecasting EPC has been largely used by FURNAS.

Future improvements on the forecasting models could be obtained by the use of fuzzy logic and neuro-fuzzy techniques. Regional models could be developed using the same techniques.

Acknowledgement

This research and development project was supported by FURNAS S.A. (R&D program 2002-2003 cycle).

References

[1] Chen, G. J., Li, K. K., Chung, T. S., Sun, H. B. & Tang, G. Q., Application of an innovative combined forecasting method in power system load forecasting, *Electric Power Systems Research*, **59**; pp. 131-137, 2001.

[2] Hippert, H. S., Pedreira, C. E. & Souza, R. C., Neural Net-works for short-term load forecasting: A Review and Evaluation, *In IEEE Transactions of Power Systems*, **16(1)**, pp. 44-55, 2001.

[3] Gavrilas, M., Ciutea, I. & Tanasa, C., Medium-term load forecasting with artificial neural network models , *Proc. of the 16th International Conference and Exhibition on Electricity Distribution*, Part 1: Contributions. CIRED. (IEE Conf. Publ No. 482), **6**, pp. 167–171, 2001.

[4] Chen, G. L., Li, K. K., Chung, T. S., Sun, H. B. & Tang, G. Q., Application of an innovative combined forecasting method in power system load forecasting, *In Electric Power Systems Research*, **59**, pp. 131-137, 2001.

[5] Al-Saba, T., El-Amin, I., Artificial Neural Networks as applied to long-term demand forecasting, *In Artificial Intelligence in Engineering*, **13**, pp. 189-197, 1999.

[6] Padmakumari, K., Mohandas, K. P. & Thiruvengadam, S., Long Term distribution demand forecasting using neuro fuzzy computations, *In Electrical Power and Energy Systems*, **21**, pp. 315-322, 1999.

[7] Zhang, G., Patuwo, B. E. & Hu, M. Y., Forecasting with artificial neural networks: The state of the art, *International Journal of Forecasting*, **14**, pp. 35–62, 1998.

[8] CTEM/CCPE, Demand forecasts for Expansion Plan 2004-2014, 2004.

[9] Pinheiro, A. C., Gill I. S., Servén L. and Thomas M. R., Brazilian Economic Growth, 1900-2000: Lessons and Policy Implications, Inter-American Development Bank, 2004.