ELSEVIER

International Journal of Forecasting 16 (2000) 71–83

international journal of forecasting

www.elsevier.com/locate/ijforecast

Forecasting the short-term demand for electricity
Do neural networks stand a better chance?

Georges A. Darbellay*, Marek Slama

Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic

Abstract

We address a problem faced by every supplier of electricity, i.e. forecasting the short-term electricity consumption. The introduction of new techniques has often been justified by invoking the nonlinearity of the problem. Our focus is directed to the question of deciding whether the problem is indeed nonlinear. First, we introduce a *nonlinear* measure of statistical dependence. Second, we analyse the linear and the nonlinear autocorrelation functions of the Czech electric consumption. Third, we compare the predictions of nonlinear models (artificial neural networks) with linear models (of the ARMA type). The correlational analysis suggests that forecasting the short-term evolution of the Czech electric load is primarily a linear problem. This is confirmed by the comparison of the predictions. In the light of this case study, the conditions under which neural networks could be superior to linear models are discussed.

Keywords: Energy forecasting; Time series; Nonlinearity; Artificial neural networks; ARIMA models

1. Introduction

The estimation of the future demand for electric energy is central to the planning of a regional or a national system for generating electricity. Shorter-term forecasting is equally central to operating a power generation system. In the context of running such a system, engineers refer to the electricity demand as the electric load. Managing it is a complex task. It has security and economic aspects. In order to guarantee a regular supply, it is necessary to keep a reserve. Depending on how fast it is available, this reserve bears the names of spinning reserve or of cold reserve. Overestimating the future load results in unused spinning reserve, and it being 'burnt' for nothing. An unexpected supply to the international energy network is usually not welcome. Underestimating the future load is equally detrimental, because high starting costs are incurred if the cold reserve has to be drawn upon. Buying at the last minute from other suppliers is obviously expensive. Good cooperation on the international electricity grid requires that every member is able to foresee his own needs.

*Corresponding author. Tel.: +420-2-6605-2431; fax: +420-2-688-4702.

E-mail address: dbe@utia.cas.cz (G.A. Darbellay)

0169-2070/00/$ – see front matter
PII: S0169-2070(99)00045-X

In this work, we consider the electric load time series of the Czech Republic. It covers the years 1994 and 1995, and it is measured at hourly intervals. We are interested in its short-term evolution, where short-term means up to 36 h ahead. The Czech Republic has about 10 million inhabitants and the level of the electric load, averaged over the entire year, lies at about 7000 MW. Electric load time series have periodic components. They display an annual cycle, a weekly cycle and a daily cycle. These are visible upon visual inspection of the data. The last two cycles can be observed in Fig. 1. The periodic part of the signal is, of course, amenable to linear modelling. However, the great variety of users' needs produces random fluctuations. It is not clear whether the randomness in the signal is also of a linear character or whether it is of a nonlinear nature. This is an important question because the choice of the model is determined by the nature of the data. Most of the body of the statistical methods available for analysing time series rests on the assumption of linearity, be it in the original data or after having subjected them to some (invertible) transformation. The same applies to the stationarity assumption. This methodology is referred to as classical time series analysis. The 'classical' linearity assumption considerably simplified modelling. It also enabled the development of an elegant mathematical theory of time series, and gave birth to well-established methods, the most prominent of which probably being the Box and Jenkins approach, widely known as the autoregressive integrated moving average (ARIMA) procedure.

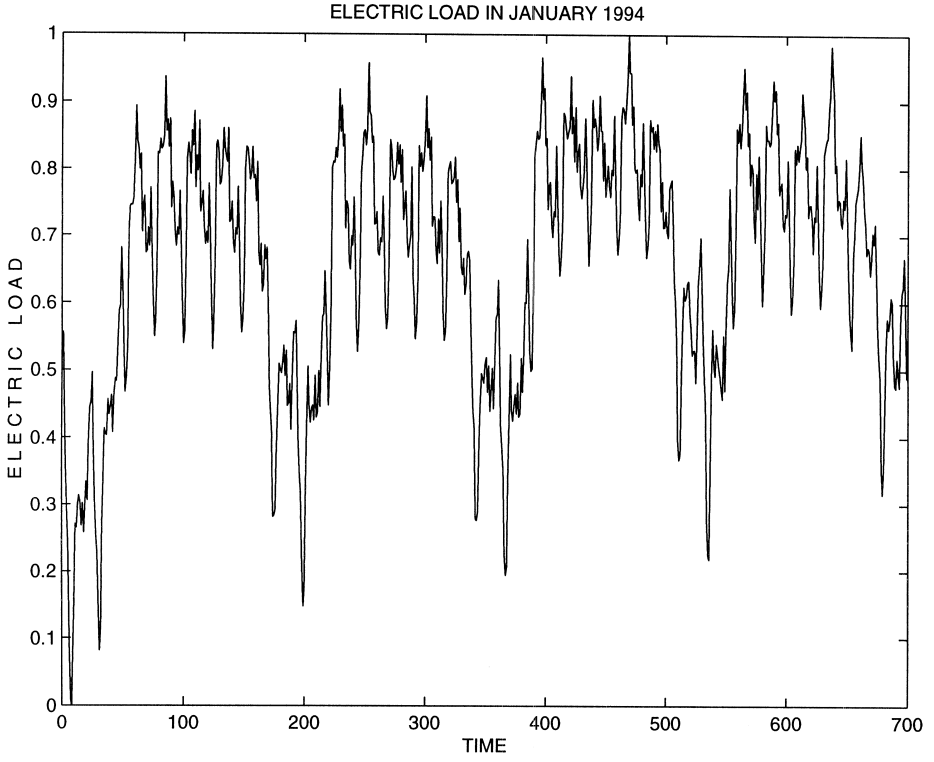

Fig. 1. National electricity consumption for the first 700 h of January 1994 in the Czech Republic. It has been scaled between 0 and 1.

We shall first be concerned with univariate modelling. In a second stage, we will turn our attention to multivariate modelling with the addition of an external input variable, namely the temperature. We start, in Section 2, by analysing both the linear and the nonlinear dependence structure of the data. In Section 3, we build and compare linear and nonlinear models. For univariate modelling, we consider an ARIMA model, and, for multivariate modelling, we consider an ARMAX model, where the letter X stands for any external variable added to the ARIMA formulation. In the class of nonlinear models, we consider artificial neural networks. Our conclusions are summarised in Section 4.

Artificial neural networks are one of the nonlinear statistical methods. Their name is inspired by a (highly simplified) formal analogy to the working of the human brain. Neural networks have become very popular over the last 10 years, not only in time series analysis. This surge of interest in neural networks is linked to the publication of the backpropagation algorithm (Rummelhart, McClelland & the PDP Research Group, 1986), which has become the most popular method for building neural networks. This algorithm makes it possible to estimate the parameters of the network. This is what 'training' or 'learning' means. In forecasting, neural networks have been heralded as a breakthrough because of their ability to model nonlinear relationships without much theoretical effort on the part of the modeller, since they should be able to extract, so to speak, statistical relationships from data. In this paper, we will investigate the area of validity of such claims for electric load forecasting (starting with, e.g., Park, El-Sharkawi, Marks, Atlas & Damborg, 1991; for a review of neural networks in the power industry, not only in forecasting, see Niebur, 1995). Current research in load forecasting is still somewhat divided on the issue, whether sophisticated nonlinear models represent progress (e.g., Piras, Germond, Buchenel, Imhof & Jaccard, 1996), or whether more effort should be spent in improving linear models (e.g., Yang, Yuang & Huang, 1996; Ramanathan, Engle, Granger, Vahid-Araghi & Brace, 1997). The same applies to forecasting in general. The finding as to whether and when artificial neural networks are better than classical methods remains inconclusive (Swanson & White, 1997; Zhang, Patuwo & Hu, 1998). Artificial neural networks are highly flexible. They can model an extremely wide class of functional relationships, in particular nonlinear ones. However, flexibility has a price. There is no established method for identifying the network structure that can best approximate the function mapping the inputs to the outputs. As a result, tedious experiments and time-consuming trial-and-error procedures are unavoidable. One key contribution of this paper is to present a nonparametric statistical method that helps in deciding whether the autocorrelation function of a time series is linear or not. It is clear that if it is linear there is little point in going through the trouble of building nonlinear neural networks.

2. Model identification

In classical time series analysis, model identification centers on the study of the correlational structure. For univariate analysis, this means the study of the autocorrelation function, and possibly of its three associates, the partial, the inverse and the extended autocorrelation functions (e.g., Wei, 1990). The electric load can be considered as a random process $\{X(t)\}$ in continuous time, from which a time series $\{x_t\}$ is obtained by sampling at discrete times. Obviously, the symbol t denotes the time. The linear autocorrelation function of a time series $\{x_t\}$

$$r(X(t),X(t-k)) = \frac{\mathrm{Cov}(X(t),X(t-k))}{\sqrt{\mathrm{Var}(X(t))}\sqrt{\mathrm{Var}(X(t-k))}} \quad (1)$$

is a function of the lag k. The lag may be positive or negative. For a given k, the linear autocorrelation coefficient $r_k = r(X(t),X(t-k))$, which lies between -1 and 1, describes how much on average two values x_t that are k steps apart co-vary with each other, i.e. it measures their linear dependence. For simplicity, we will also speak of the linear autocorrelation r_k.

In order to investigate whether the electric load time series contain nonlinear dependences, we borrow a concept from the mathematical theory of information (for a very readable book on the subject we refer to Cover & Thomas, 1991). Let X and Y be two continuous random variables with joint probability density function $p(x,y)$ and marginal density functions $u(x)$ and $v(y)$. The mutual information between X and Y is defined as

$$I(X,Y) = \int\int p(x,y) \ln \frac{p(x,y)}{u(x)v(y)} \mathrm{d}x\, \mathrm{d}y \qquad (2)$$

where the integrals are taken over the whole domain of the joint density function. The mutual information takes values between 0 and ∞, but it can be normalised to a number between 0 and 1 by the following invertible transformation:

$$\rho(X,Y) = \sqrt{1 - e^{-2I(X,Y)}}. \qquad (3)$$

The coefficient ρ can be shown to satisfy the following properties (Rényi, 1959; see also Granger & Teräsvirta, 1993):

(a) $\rho(X,Y) = \rho(Y,X)$,
(b) $0 \le \rho(X,Y) \le 1$,
(c) $\rho(X,Y) = 0$ if and only if X and Y are independent,
(d) $\rho(X,Y) = 1$ if $Y = f(X)$, where f is a one-to-one function,
(e) $\rho(f(X),g(Y)) = \rho(X,Y)$ if f and g are one-to-one functions,
(f) $\rho(X,Y) = |r(X,Y)|$ if the joint density of X and Y is Gaussian.

To avoid any confusion, we note that the functions f and g are deterministic transformations. Properties (c), (d) and (e) must be contrasted with those of the linear correlation coefficient $r(X,Y) = \mathrm{Cov}(X,Y)/\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$

(c′) $r(X,Y) = 0$ if and only if X and Y are linearly independent,
(d′) $r(X,Y) = \pm 1$ if and only if $\exists$ constants a and b such that $Y = aX + b$,
(e′) $r(f(X),g(Y)) = r(X,Y)$ if f and g are linear transformations.

The coefficient $\rho(X,Y)$ has the ability of capturing both the linear and the nonlinear dependences between X and Y. As one would expect, there is a price to pay for the additional abilities of ρ: estimating ρ from a data sample is far more difficult than estimating r, because the estimation of the underlying probability densities is now required. Yet, an algorithm for estimating ρ from a data sample has been developed recently (Darbellay, 1999; Darbellay & Vajda, 1999). It is based on a data-dependent partitioning of the two-dimensional observation space, $\mathbb{R}^2$ in our case. The accuracy (bias) and precision (variance) of this estimator for ρ were found to be comparable to the standard estimator for r. It is thus possible to compare the estimated ρ, which we shall denote as $\hat{\rho}$, and the estimated r, to be denoted by $\hat{r}$. Below, we will not give error bars on the values of $\hat{r}$ and $\hat{\rho}$: these errors were very small because we used long time series (with about 10 000 points).

As an illustration, we consider a well-known nonlinear time series, namely the Mackey Glass time series. It is generated by the time-delay differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{ax(t)}{1 + x^c(t-\tau)} - bx(t) \qquad (4)$$

where the constants are often (as here) taken to be $a = 0.2$, $b = 0.1$ and $c = 10$. The delay parameter τ determines the behaviour of the time series, and for $\tau > 16.8$ the Mackey Glass

equation has a chaotic attractor. Here, we chose $\tau = 30$ and generated a time series from (4) with the time interval $\Delta t = 1$. For the estimation of the autocorrelation, 10 000 data points were used. In Fig. 2 we display the estimated *linear* autocorrelation function, $\hat{r}_k$, and the estimated *general* autocorrelation function, $\hat{\rho}_k$, which will also be referred to as the *nonlinear* autocorrelation function. The difference between the two speaks for itself. For example, around lag 70 the linear autocorrelation is very weak, with a value of about -0.1, yet the nonlinear autocorrelation has a peak nearing 0.8. This is so because the distribution of the pairs (x_t, x_{t-70}) has a strongly nonlinear shape (a circular shape, to be explicit). For predicting the evolution of this time series, it is necessary to use a nonlinear modelling approach.

We start our analysis of the electric load data by comparing the sample general autocorrelation $\hat{\rho}_k$ with the sample linear autocorrelation $\hat{r}_k$. $\{x_t\}$ denotes from now on the electric load time series under study. It covers the years 1994 and 1995. The load was recorded at hourly intervals. The lag k is thus measured in hours. The sample used for the calculation of $\hat{\rho}$ and $\hat{r}$ included about 10 000 points. In Fig. 3 we show the sample autocorrelations $\hat{\rho}_k$ and $\hat{r}_k$ as functions of the lag k for the first 200 lags. The series is quite strongly autocorrelated as the autocorrelation level is above 0.5. The strongest dependence is at lag $k = 1$. The second highest peak is at 1 week (168 h). The other peaks of the autocorrelation are located, in decreasing order of importance, at 1 day, 6, 8, 2, 5, 3, 4 days. Both $\hat{\rho}$ and $\hat{r}$ give the same relative importance

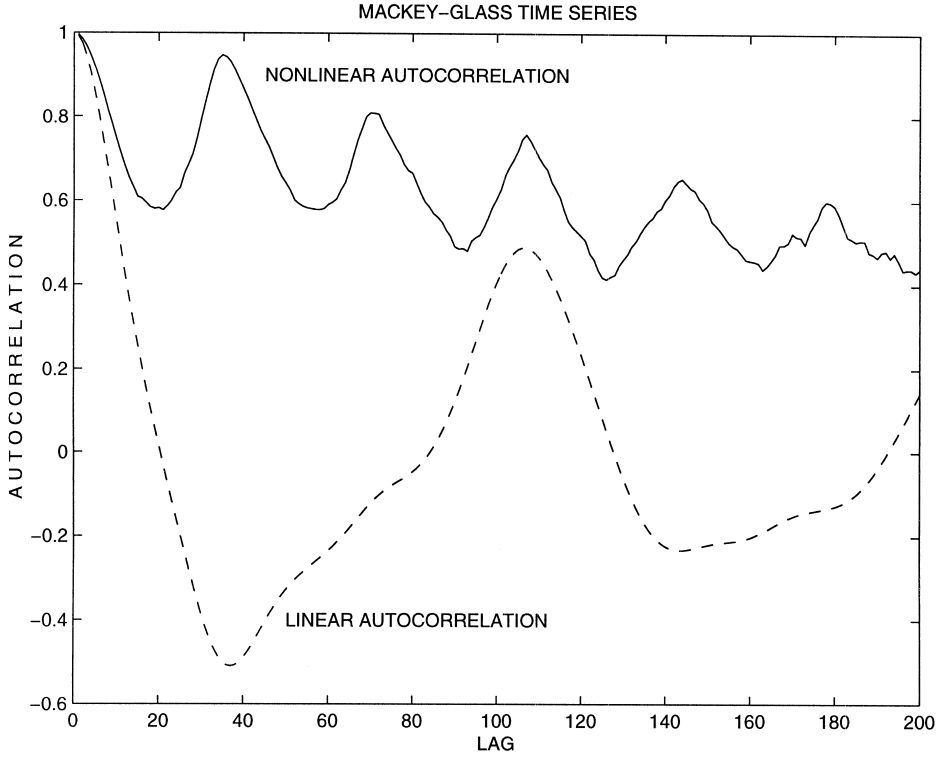

Fig. 2. Autocorrelation functions of the Mackey Glass time series for the first 200 lags. The dashed curve is the *linear* autocorrelation function, $\hat{r}_k$. The upper curve is the *nonlinear* autocorrelation function, $\hat{\rho}_k$. Note that $\hat{\rho} > |\hat{r}|$.

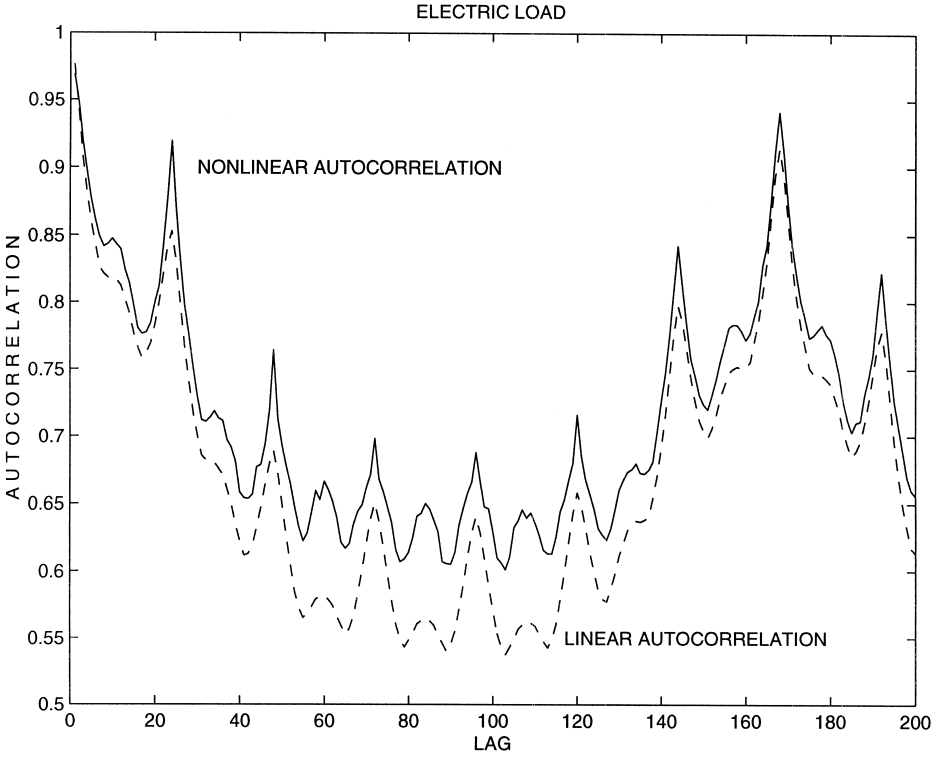

Fig. 3. Comparison of the nonlinear and the linear autocorrelation functions, $\hat{\rho}_k$ and $\hat{r}_k$ respectively, of the original electric load data, $\{x_t\}$, for the first 200 lags. The lags k are measured in hours. The upper line is $\hat{\rho}$, the lower (dashed) line $\hat{r}$.

to the peaks. For making predictions, it is clear that the two peaks at 1 day and 1 week are of the highest importance. The strong 1 week peak will doubtlessly play a crucial role in forecasting for *any* horizon between now and 1 week ahead. Plotting the pairs $(x_t, x_{t-\tau})$ confirms that their distribution has a linear shape. Overall, it can be said that the electric load time series is predominantly linear. Series, such as, for instance, those of nonlinear dynamical systems, would produce an entirely different picture with $\hat{\rho}$ and $\hat{r}$ lying widely apart, as can be witnessed in Fig. 2, for instance.

The electric load time evolution is nonstationary, as indicated by the strong and barely decaying autocorrelations in Fig. 3. Within the ARIMA procedure, nonstationarity is removed by differencing. Given the seasonality of the original time series $\{x_t\}$, we considered the following differenced time series:

$$y_t = x_t - x_{t-1} \tag{5}$$

$$z_t = y_t - y_{t-24} \tag{6}$$

$$w_t = z_t - z_{t-168}. \tag{7}$$

First differencing makes the spike at $k = 1$ in the linear and nonlinear autocorrelations disappear. Obviously, the daily spikes remain. Nonlinearities are essentially confined to the intraday intervals. The dependence pattern of the differenced time series $\{y_t\}$ is, as was the case for $\{x_t\}$, dominated by the linear relationships at the daily peaks. Since there is again no tailing off, further differencing is needed. Differencing a second time does not yield a station-

ary time series. Differencing once more produces the correlation structure displayed in Fig. 4. There are five spikes, at 1 week ($k = 168$ h), 1 day, 1 h, 6 and 8 days. There seems to be some remaining nonlinearity at lag $k = 1$ h. Otherwise the series $\{w_t\}$ can be regarded as linear. Beyond $k = 200$, the autocorrelation functions, both linear and nonlinear, are not significantly different from zero, as may be expected for a stationary time series. The linear autocorrelation function displays two large (negative) spikes at 1 day and at 7 days, with smaller flanking (positive) spikes at 6 and 8 days, and another negative spike at 1 h. This indicates that a multiplicative ARIMA model could be appropriate.

3. Models and results

In the previous section, we investigated the linear and nonlinear correlation structure of the electric load time series $\{x_t\}$, and of its differenced versions. By comparing $\hat{\rho}$ and $\hat{r}$, we found the nonlinearities to be weak. It is, however, not clear whether these weak nonlinearities could be entirely neglected in building a predictive model. The idea of comparing $\hat{\rho}$ and $\hat{r}$ is new, and there is at present no experience in drawing conclusions from such a comparison. Therefore, we need to check if indeed nonlinear models would not achieve a better performance than linear models. It is beyond our scope to verify this for the entire

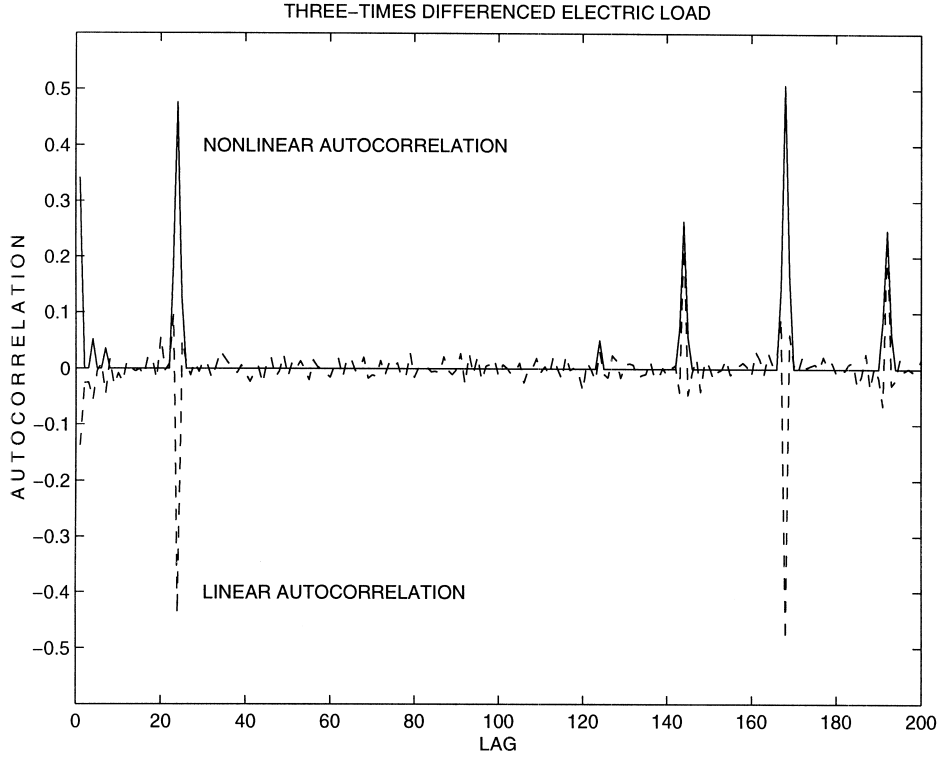

Fig. 4. Same as Fig. 3, except that it is now with the three times-differenced data $\{w_t\}$ defined in Eq. (7). The dashed line is the linear autocorrelation function. It has three negative spikes, at 1 h, 1 day and 1 week, and two positive spikes at 6 and 8 days. It is otherwise not significantly different from zero. The nonlinear autocorrelation function has five positive spikes, at 1 h, 1 day, 6, 7 and 8 days. Otherwise, it is not significantly different from zero.

class of linear models and the even larger class of nonlinear models. We chose two well-established representatives of each class, namely an ARIMA model and an artificial neural network model. The former is linear by construction, while the latter belongs to a class which has been designed to incorporate nonlinear relationships. The models were estimated during the year 1994. They were tested in the year 1995, during which they were not updated.

The analysis conducted in Section 2 suggests the following multiplicative seasonal ARIMA model:

$$(1-\theta_1 B^1)(1-\theta_2 B^{24})(1-\theta_3 B^{168})w_t = (1-\phi_1 B^1)(1-\phi_2 B^{24})(1-\phi_3 B^{168})e_t. \quad (8)$$

B denotes the usual backshift operator, i.e. $B^i x_t = x_{t-i}$. The process e_t is Gaussian white noise. The nature of the electric load time series further suggests splitting the data into two subsets, working days and holidays. Splitting does not affect the results of the correlational analysis in Section 2, nor does it affect the overall conclusions of this paper. For the sake of conciseness, we will only report results for working days, in which case the model becomes

$$(1-\theta_1 B^1)(1-\theta_2 B^{24})(1-\theta_3 B^{120})w_t = (1-\phi_1 B^1)(1-\phi_2 B^{24})(1-\phi_3 B^{120})e_t. \quad (9)$$

We estimated the parameters θ_i and ϕ_i with the generalised least squares method. For the working days, we obtained

$$\theta_1 = 0.752, \quad \theta_2 = 0.026, \quad \theta_3 = 0.027$$

$$\phi_1 = 0.938, \quad \phi_2 = 0.611, \quad \phi_3 = 0.826.$$

The model (9) was then integrated to obtain a model for the original time series $\{x_t\}$. The results are displayed in Table 1.

The performance is measured in several complementary ways. The first measure is the widely used normalised mean square error (NMSE)

$$\mathrm{NMSE} = \frac{\sum_{t=n+1}^{N}(x_t-\hat{x}_t)^2}{\sum_{t=n+1}^{N}(x_t-\bar{x})^2} = \frac{1}{\hat{\sigma}^2}\frac{1}{N-n}\sum_{t=n+1}^{N}(x_t-\hat{x}_t)^2 \quad (10)$$

where N is the number of points in the estimation set, or the test set, and n the largest lag used in the model. This means, for instance, that the effective number of points $N-n$ is about 6000 for the working day model. $\hat{\sigma}^2$ is the variance of the load data on the estimation set, or the test set. In the year 1994, its value was $\hat{\sigma}^2 = 1.24\times 10^6$. In 1995, the variance rose to 1.63×10^6. We chose to normalise the MSE by dividing it by the variance of the time series because this facilitates comparisons across time series. However, the MSE is not the last word in the world of error measures (Armstrong & Collopy, 1992; Fildes, 1992). For each problem, it is necessary to consider error measures which

Table 1
Forecasting results for the working day ARIMA model

Horizon (h)	NMSE (1994)	NMSE (1995)	MAPE (1994)	MAPE (1995)	maxAPE (1994)	maxAPE (1995)
$k=1$	0.8%	0.6%	1.1%	1.0%	15%	12%
$k=12$	2.8%	2.7%	2.0%	2.2%	15%	14%
$k=24$	3.5%	3.5%	2.3%	2.5%	16%	14%
$k=36$	5.4%	5.8%	3.0%	3.3%	15%	16%

are directly relevant to the users. In our case, practitioners are particularly interested in the mean absolute percentage error

$$\text{MAPE} = \frac{1}{N-n}\sum_{t=n+1}^{N}\frac{|x_t - \hat{x}_t|}{x_t} \tag{11}$$

and the maximum absolute percentage error

$$\text{maxAPE} = \max_{t=n+1,\ldots,N}\frac{|x_t - \hat{x}_t|}{x_t}. \tag{12}$$

If there were outliers, which was not the case in our time series, it would make sense to replace the MAPE by the median absolute percentage error. The MAPE, which has become a standard error measure in load forecasting studies, is sometimes simply referred to as the average percentage error.

Four forecasting horizons were considered, 1, 12, 24 and 36 h. The forecasts above the 1 h horizon are obtained by iteration: the current output is used as an input for predicting the next output. The second, fourth and sixth columns of Table 1 give the performance on the estimation set, from January 1994 to December 1994. The third, fifth and seventh columns show the performance achieved on the test set, from January 1995 to December 1995. As would be expected, the performance deteriorates a little on the test set relative to the estimation set. The 36 h horizon is given because of its importance in practice. Engineers in the dispatching room make a preliminary scheduling plan for the entire next day at 12 o'clock the day before.

We now turn our attention to artificial neural networks. For an overview of this field, we refer to, for example, Bishop (1995). A recent review of their use in forecasting is that of Zhang et al. (1998). In time series analysis, neural networks are usually used as nonlinear function approximators. They map an input space (present and past values of the time series) onto an output space (future value(s)). Our network will be of the following feedforward type:

$$\begin{aligned}\hat{x}_t &= f(x_{t-1},\ldots,x_{t-n})\\ &= b_0 + \sum_{i=1}^{m} b_i \tanh\left(a_{i0} + \sum_{j\geq 1} a_{ij}x_{t-j}\right)\end{aligned} \tag{13}$$

where tanh is the tangent hyperbolicus function. It is a nonlinear transformation, with a sigmoid shape. The inputs are x_{t-j}, with j running over an index set of not necessarily sequential positive integers (n denotes the largest lag in the model). These inputs form the so-called input layer. In the second layer, which is referred to as the hidden layer, there are m nonlinear processing units. These units transform the inputs, by means of the multiplicative weights a_{ij}, the additive weights a_{i0} and the sigmoid functions. A weighted sum, with weights b_i, over the outputs of these hidden units plus a shift, b_0, produces the final output. The network parameters are estimated by minimising the error function

$$\sum_{t=n+1}^{N}[f(x_{t-1},\ldots,x_{t-n}) - x_t]^2 \tag{14}$$

where N is the number of elements in the estimation set. This error function takes all input vectors of the estimation set into account. It takes, of course, several passages over the estimation set to obtain reasonable values for the parameters a_{ij} and b_i.

Building neural networks has a far more exploratory nature than building linear models. As a result, it takes much more time and computational effort to obtain a neural network model. As yet, there is no universally accepted methodology for an optimal selection of the input space. Linear approaches are not quite convincing, precisely because they are linear. Another difficult choice concerns the number (m in Eq. (13)) of nonlinear processing units. There is unfortunately no general method for choosing the number m. It is clear that adding more and more nodes leads to a more complex network. This will make the learning more difficult and

less reliable, in the sense that the performance of the network is liable to drop quite seriously outside the estimation set. Our experiments indicated that it was unnecessary to use more than 10 hidden units. We used networks with m between 6 and 10. Inspired by our study in Section 2, we selected, for the working day model, the following 15 lagged inputs: 1 to 6, 23, 24, 25, 48, 72, 96, 120, 144, 240 h (we also tried a neural network with exactly the same inputs as the ARIMA model above). We used a fully connected network. This meant a network with up to 171 parameters. To estimate them, we used the Levenberg Marquardt method. This is a second-order method that uses both the gradient vector and the Hessian matrix. We found it to be superior to simple gradient descent. The programs in the MATLAB neural network toolbox were used.

Ten neural networks were trained, starting with different initial conditions and number of hidden units. The network with the lowest error on the estimation set was selected and used on the test set. It was not updated in the course of the test. The results are shown in Table 2. The neural network made only one-step forecasts. These were iteratively used as inputs for predicting the next value up to 36 h. The same iterative procedure was used for the linear model as well. A comparison of Tables 1 and 2 shows that the neural network performs slightly less well for the 1 h and 12 h horizons. At $k = 24$ and 36 h the results are about the same. Overall, it can be said that both models achieve fairly comparable results with respect to the performance measures NMSE, MAPE and maxAPE.

It could be argued that an autoregressive moving average model would be more properly compared to a recurrent neural network, the recurrent links in the network corresponding to the moving average terms in the linear model. This is only partly true because the standard feedforward network, as used above, is not a straightforward nonlinear generalisation of the autoregressive part of the ARMA model. In any case, for completeness, we built an AR model and a recurrent neural network. We found that the AR model was worse than the neural network whose results are given in Table 2. We also found that the recurrent neural network was slightly better than the feedforward neural network but no better than the ARIMA model whose results are given in Table 1. It is worth noting that recurrent networks are more difficult to design and train than feedforward networks.

Electricity consumption is affected by many factors. Only a minority of them can be harnessed in a forecasting model. Temperature is one such factor. We did not have hourly measurements for the temperature. Unfortunately, it is impossible to use a daily input in an hourly ARMA model with (in our case three times) differenced data. Neural networks are more

Table 2
Forecasting results for the working day feedforward neural network model

Horizon (h)	NMSE (1994)	NMSE (1995)	MAPE (1994)	MAPE (1995)	maxAPE (1994)	maxAPE (1995)
$k = 1$	0.9%	0.8%	1.1%	1.1%	13%	10%
$k = 12$	3.3%	3.4%	2.2%	2.5%	13%	14%
$k = 24$	3.5%	3.7%	2.3%	2.6%	14%	14%
$k = 36$	5.1%	5.5%	2.8%	3.2%	14%	14%

Table 3
Forecasting results for the working day feedforward neural network model with temperature input

Horizon (h)	NMSE (1994)	NMSE (1995)	MAPE (1994)	MAPE (1995)	maxAPE (1994)	maxAPE (1995)
$k = 1$	0.8%	0.7%	1.1%	1.1%	11%	11%
$k = 12$	2.3%	2.5%	1.9%	2.1%	11%	13%
$k = 24$	2.4%	2.6%	1.9%	2.2%	11%	13%
$k = 36$	2.9%	3.4%	2.2%	2.6%	16%	15%

flexible, as they do not require any differencing. They can therefore make use of daily temperatures for hourly forecasts. In order to extend our study towards multivariate modelling, we decided to proceed along two paths.

First, we constructed a neural network with daily temperature inputs for hourly load forecasts. The neural network had the same architecture as before, but with three new inputs, the temperature today, yesterday and 1 week ago. The results are shown in Table 3. With respect to Tables 1 and 2, the inclusion of the temperature leads to a better performance for the horizons $k = 24$ and $k = 36$ h.

Second, in view of this improvement, we built two models, an ARMAX model and a neural network, for *daily* load forecasts, using, of course, daily inputs. Two-times differencing was needed and an adequate ARMAX model was found to be

$$z_t = az_{t-1} + be_{t-5} + c_0 v_t + c_1 v_{t-1} + c_2 v_{t-2} \tag{15}$$

where $z_t = y_t - y_{t-5}$ with $y_t = x_t - x_{t-1}$. x_t is now the value of the average daily load on day t. Similarly, $v_t = u_t - u_{t-5}$ and $u_t = T_t - T_{t-1}$, where T_t is the average daily temperature on day t. Model (15) is for working days only, but T_{t-1} and T_{t-2} are always the temperatures of, respectively, 1 and 2 days ago, even if these days are Saturdays or Sundays. The parameters were $a = -0.0788$, $b = -0.9366$, $c_0 = -7.9421$, $c_1 = -13.2675$ and $c_2 = 4.5376$. The results are shown in the upper part of Table 4. The results of the neural network for daily forecasts with daily load and temperature inputs are shown in the lower part of Table 4. The neural network was again of the type defined by Eq. (13). It had six inputs (z_{t-1}, z_{t-2}, z_{t-3}, v_{t-1}, v_{t-2}, v_{t-3}), one hidden layer with six nodes and

Table 4
Daily forecasting results for the ARMAX model and the feedforward neural network model, both with temperature input, working days only

Horizon (days)	NMSE (1994)	NMSE (1995)	MAPE (1994)	MAPE (1995)	maxAPE (1994)	maxAPE (1995)
$k = 1$ (ARMAX)	1.1%	1.2%	1.2%	1.4%	4%	7%
$k = 2$ (ARMAX)	2.0%	1.8%	1.7%	1.9%	6%	6%
$k = 1$ (neur. net.)	1.4%	1.6%	1.3%	1.5%	4%	6%
$k = 2$ (neur. net.)	2.4%	2.5%	1.9%	1.9%	5%	8%

one output (z_t). We started several networks with different initial conditions and chose the best network on the estimation set. As can be seen in Table 4, the ARMAX model appears to be slightly superior to the neural network. One may expect that experimenting further with neural networks could yield results which are as good as those of the ARMAX model. Yet, this would take more time than that spent building the ARMAX model.

4. Conclusions

Our analysis of the dependence structure of the electric load time series of the Czech Republic indicated that the autocorrelations in this time series are predominantly linear. For univariate modelling, we found that, indeed, the forecasting abilities of a linear model and a nonlinear model were not very different. These models were, respectively, an ARIMA model and a neural network. Artificial neural networks are not the only nonlinear paradigm (for others, see, e.g., Priestley, 1988, or Grassberger, Schreiber & Schaffrath, 1991). However, our analysis of the autocorrelation structure suggests that no univariate nonlinear model will be much better than a simpler linear cousin.

Multivariate modelling was considered by adding the temperature as an external input. The relationship between electricity consumption and temperature is linear in the Czech Republic. One would thus expect that the temperature would not enable a neural network to beat a linear model. Actual modelling confirmed it. Of course, external variables could be related to the load in a nonlinear fashion, in which case a nonlinear model should outperform a linear model. There are also cases where the data cannot be preprocessed in a form suitable for linear models because of the need for differencing, for example with hourly load data and daily temperatures. In such a case, artificial neural networks are able to integrate more information and thus produce better forecasts.

Although, in principle, any slight nonlinearity could be used by a neural network to improve upon a linear model, it appears that, in practice, weak nonlinearities will not produce perceptible improvements. In this respect, one has to bear in mind that finding the optimal parameters of a neural network is a far more time-consuming enterprise than estimating those of a linear model.

Our main point is the following: before embarking on some complex nonlinear model, one would be better advised to check that the problem is indeed nonlinear. In our study, we found that forecasting the electric load in the Czech Republic is primarily a linear problem. Consequently, for all practical purposes, linear models will be adequate and there is only limited scope for improvement in using nonlinear models. The situation in other countries warrants an inquiry. In further work, it would also be interesting to use other nonlinear tests in parallel with the nonlinear autocorrelation ρ. Furthermore, still with the mutual information (2), it is possible to calculate multivariate nonlinear correlations, as well as conditional nonlinear correlations (Darbellay, 1998). These could prove invaluable in selecting the smallest set of inputs that contains most of the information. For nonlinear models, such as artificial neural networks, this is particularly important.

Acknowledgements

The authors would like to thank Jaroslav Franěk for his programming assistance and Emil Pelikan for a critical reading of the manuscript. G.A.D. is grateful to the *Fonds National Suisse de la Recherche Scientifique*, and M.S. to the *Grantova Agenturá České Republiky* under grant No. 102/95/1311, for financial support. Thanks are due to the Czech power company, ČEZ, for providing the data.

References

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *International Journal of Forecasting 8*, 69–80.

Bishop, C. M. (1995). Neural networks for pattern recognition, Oxford University Press, Oxford.

Cover, T. M., & Thomas, J. A. (1991). Elements of information theory, Wiley, New York.

Darbellay, G. A. (1998). Predictability: an information-theoretic perspective. In: Procházka, A., Uhlíř, J., Rayner, P. J. W., & Kingsbury, N. G. (Eds.), Signal analysis and prediction, Birkhäuser, Boston, pp. 249–262.

Darbellay, G. A. (1999). An estimator of the mutual information based on a criterion for independence. *Journal of Computational Statistics and Data Analysis* (http://siprint.utia.cas.cz/darbellay) (in press).

Darbellay, G. A., & Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory 45*(4), 1315–1320.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting 8*, 81–98.

Granger, C. W. J., & Teräsvirta, T. (1993). Modelling nonlinear economic relationships, Oxford University Press, Oxford.

Grassberger, P., Schreiber, T., & Schaffrath, C. (1991). Nonlinear time sequence analysis. *International Journal of Bifurcation and Chaos 1*, 521–547.

Yang, H. -T., Yuang, C. -M., & Huang, C. -L. (1996). Identification of ARMAX model for short term load forecasting: an evolutionary programming approach. *IEEE Transactions on Power Systems 11*(1), 403–408.

Niebur, D. (1995). Artificial neural networks in the power industry, survey and applications. *Neural Networks World 6*, 945–950.

Park, D. C., El-Sharkawi, M. A., Marks II, R. J., Atlas, L. E., & Damborg, M. J. (1991). Electric load forecasting using an artificial neural network. *IEEE Transactions on Power Systems 6*(2), 442–449.

Piras, A., Germond, A., Buchenel, B., Imhof, K., & Jaccard, Y. (1996). Heterogeneous artificial neural network for short term electric load forecasting. *IEEE Transactions on Power Systems 11*(1), 397–402.

Priestley, M. B. (1988). Nonlinear and nonstationary time series analysis, Academic Press, London.

Ramanathan, R., Engle, R., Granger, C. W. J., Vahid-Araghi, F., & Brace, C. (1997). Short-run forecasts of electricity loads and peaks. *International Journal of Forecasting 13*, 161–174.

Rényi, A. (1959). A measure of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae 10*, 441–451.

the PDP Research Group, Rummelhart, D. E., & McClelland, J. L. (1986). Foundations, Parallel distributed processing, explorations in the microstructure of cognition, vol. 1, MIT Press, Cambridge, MA.

Swanson, M. R., & White, H. (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting 13*, 439–461.

Wei, W. W. S. (1990). Time series analysis, Addison-Wesley, New York.

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

Biographies: Georges A. DARBELLAY holds an MSc. in Physics and Mathematics from the University of Fribourg, Switzerland, and a Ph.D. in Theoretical Physics from the University of Oxford, UK. He is now a researcher at the Czech Academy of Sciences in Prague. His research interests include information theory, statistical physics and time series analysis.

Marek SLAMA holds an MSc. in Chemical Physics from the Faculty of Mathematics and Physics of Charles University, Prague. He was a researcher at the Czech Academy of Sciences and now works for Reuters, in Prague. His research interests include data modelling, forecasting and time series analysis.