One step ahead forecasting using Multilayered perceptron

Stéphane Canu (1,2), Yves Grandvalet(1) and Xia Ding (2)

(1) Heudiasyc, U.R.A. C.N.R.S. 817
Université de Technologie de Compiègne
Centre de Recherches de Royallieu
B.P. 529, 60205 Compiègne Cedex, France

(2) Lyonnaise des Eaux / CITI
Technopolis - Rue du Fonds Pernant
F-60471 Compiègne cedex- France

scanu@hds.univ-compiegne.fr
http://www.univ-compiegne.fr/~scanu/NLprev.html

Abstract

When dealing with time series, the one step ahead forecasting problem based on experimental data is the problem of estimating the autoregression function of the underlying process. When minimizing the expected forecasting error is the main goal the flexible approach has to be used to be able to adjust the complexity of the model to the complexity of the data. Multilayered perceptrons are a popular example of such a flexible approach but not the only one. Other methods such as kernel approximator (e.g. Naradaya Watson regressor), regression spline or wavelet regressor can also be used. But whatever flexible approach is, the main issue remains the control of the complexity of the flexible approximator. Noise injection in the inputs is an efficient technique to do so. The complexity of the regessor is then adjusted thanks to the quantity (variance) of injected noise. This quantity is tuned using a bootstrap estimation of the forecasting error. One unexpected effect of this approach is the possibility to prove the consistency of the estimator under some assumptions about the underlying process who generates the time series. The two main assumptions are: the process is varying "significantly" and the process is bounded. Furthermore the boundary conditions imply to eliminate the tendency for the remaining process to be bounded. This theoretical result permits to design a methodology for using multilayered perceptrons to forecast. The whole approach was applied successfully to forecast the daily water consumption in the south of Paris.

1 Introduction

From a practical point of view, the problem we are interested in, is relatively simple: how to estimate today the future value of a quantity, with the utmost possible precision and reliability. This is the one-step-ahead prediction. To this end, we have at our disposal the past values of this quantity, the data of one or several time series and prior knowledge of the phenomenon which produces this time series.

For the "previsionnist" this same problem of prediction can be defined as a problem of *estimation of dependencies based on empirical data*, with two important specificities: the regularity of the sampling and the dynamic aspect due to time which creates a fundamental relationship between the data. The hypothesis of the independence of the sample which is admitted for the regression, must be rejected here. A time series is a dependent sample. What we have to do, therefore, is to find "regularities" in the available sample - supposing the phenomenon is stable enough in time, for us to retrieve and predict these regularities. To this end, a "classical" approach consists in first, setting down a model and analyzing the phenomenon which has caused the time series, then in valuing the sample-parameters, by starting from the available data. We are going to deal with a different approach giving greater importance to experimental data. Starting from these data and from a criterion to be minimized, we are going to see how to build a forecasting model of the "black box" type, a flexible model whose complexity will have to be adjusted to the complexity of the available data.

Let us, first, specify the problem, making clear what data are available, what aims are pursued and lastly what prior hypothesis are set on the nature of the solution, which will allow us to build up a model. This problem can be thus formulated:

$$\text{for a given time series: } \{x(t)\}_{t=1,T} \qquad \text{forecast} \quad x(T+1) \tag{1}$$

One can also have additional data $\{u(t)\}_{t=1,T}$ also called exogeneous variables to estimate $x(T+1)$.

Our aim is, therefore, to obtain a one-step-ahead prediction that would be as precise as possible on still unknown data. In the following study we shall choose the "quadratic" cost to quantify the quality of the prediction. In this case, at time T, the best prevision $\hat{x}(T+1)$ will be the one minimizing the quadratic criterion.

$$\min_{\hat{x}(T+1)} \|\hat{x}(T+1) - x(T+1)\|^2 \tag{2}$$

The observed value will often be considered as the realization of a random variable. We shall then try to minimize the error in average.

Before setting down a model, one must interpret the variations recorded on this time-series. These variations can be explained by various factors (figure 1):

- a regular deterministic component, depending on time (tendency, periodical or other component)

- an eventually chaotic deterministic component constituting the texture. This is the "regressive" or "autoregressive" part.
- some exceptional events or perturbations
- a stochastic component shaping the random and unobservable aspects of the phenomenon.

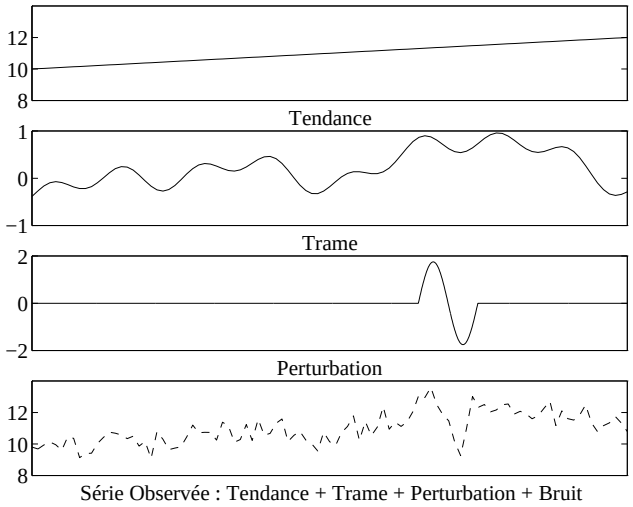

Figure 1: Exemple of the different components of a time series

These different components are not necessarily additive. It is clear that "all" is not foreseeable. In particular, the random series and the "very" chaotic series cannot be foreseen if the prediction-horizon is too remote. The phenomena which are difficult to identify from a sample, are also nearly impossible to foresee. We shall therefore have to specify the conditions under which we shall be able to give a reliable forecasting. But even among the predictable phenomena, different practicians may have different problems combined with the same time-series. According to the time constant the following cases can eventually occur:

- the long-term forecasting. The time-constant is of the order of the year, in the field of economy predictions. It is the tendency prediction.
- the mid-term forecasting. The time-constant is of the order of the month or of the day, in the field of economy predictions. It is the regular part prediction. It is the case we are interested in
- the short-time forecasting. The time-constant is of the order of the hour or of the minute. The essential aim is to detect perturbations.

The classification can be crossed with the various aims of analysis as seen by Weigend /citeWeigend95 who makes a distinction between the aims of characterization of the phenomenon and the aims of forecasting. This allows us to set the field we consider in this study: the detection of regularities for a one-step-ahead forecasting. The performance of the series will be supposed to be statistically the same in the future as it has been in the past. Moreover we shall only be interested in the phenomena which can be modeled by a short term memory, as opposed to the long-term memory which is for instance necessary for speech or language recognition. We do not wish to make other hypotheses as to the nature of the phenomenon producing the observed data. It is the available data and the criterion to be minimized that will be put forward.

We are going to develop an approach of the "black box" type, qualified as non-parametric, able to adjust its complexity to that of the data, a flexible approach of the model-capacity. The text is thus organized: we are going first recall what is a non-linear modeling of the "black box" type, and we shall explain under what respects it can prove to be useful. We shall then discuss about the "black box" approach we are interested in: forecasting using connectionist models ; we shall insist on the principal short-time memory dynamic phenomenon: the multi-layer perceptrons. Lastly, we shall present an example of use of this type of model on a problem of water day-consumption forecasting.

2 Non-linear models for forecasting

2.1 Non-linear approaches classification

In the field of time-series analysis, the main trend of research developed itself under the hypotheses on the "ergodicity", stationarity" and "linearity" of the models, considering residuals of Gaussian type [3]. In this field, *linear models of autoregressive moving average type Box and Jenkins' approach seemed able to answer most of the "previsionnists" needs* [14]. When the available data didn't seem to verify the hypotheses of the model, they were altered. A number of techniques has thus been developed so as to make the data in keeping with this type of modeling. But in some cases no linear model is fit. Tong [72] gives five limits to the ARMA type modeling which he sums up thus: the class of the ARMA models constitutes a judicious choice *if and only if the autocovariances are considered as an important characteristic of the phenomenon.* To justify the passage to non-linearity, Weigend [79] gives the example of simple models allowing to generate chaotic or pseudo uncertain series, unpredictable with linear models

$$\begin{aligned} x_t &= 4\,x_{t-1}(1 - x_{t-1}) \\ x_t &= 2\,x_{t-1}\ mod\,1 \end{aligned}$$

Moreover, even if the model is linear, the best prediction knowing the past, may not be linear. Coming back to the studies of Shepp [69] on the non gaussian A models, Tong gives a simple example enlightening the limits of linear modeling. He considers the following MA model of order one:

$$X_t = \varepsilon_t - 2\varepsilon_{t-1} \tag{3}$$

where ε_t is a i.i.d random variable uniformly distributed on $\{-1, 1\}$. In this case the conditional expectation $\mathbb{E}(X_t|X_{t-1} = x)$ is a non-linear function of the past observed value x of the time series. Lastly many authors, as much in economists [30, 62] as in automatics [47] agree on an experimental observation: the form of many time series suggests the use of non linear forecasting-models. It is nevertheless a fact that in the economic field particularly, the use of non-linear models always faces a certain number of practical and theoretical difficulties [62].

All this has therefore accounted for the development of non-linear techniques. The different non-linear approaches can be classified according to several criteria: whether they are deterministic or stochastic, of the inputs/outputs type or with an internal parametric or non-parametric representation.

- deterministic models, non linear physical models and chaotic approach [7, 11, 32].
- state-models (internal representation)
  - Hidden Markov models [27, 67]
  - generalized Kalman's filter[60].
  - recurrent neural networks [16].
- non-linear parametrics
  - bi-linear models [31],
  - exponential type models (EAR [35], EXPAR [40] or NEAR [12])
  - models with a non constant variance, autoregressive, conditionally heteroscedastic ARCH [5]
  - models with change of speed or threshold autoregressive TAR [72]
  - regression models with smooth transition SETAR ([32] page 39).

  To have a more detailed review of this model, one can read [19, 32, 33, 53] and Tong [72] who devotes the third chapter of his book to the study of about fifteen models of this type.
- non parametric of the black box type (see [70] for a review in the scope of identification)

- with a kernel basis or the use of kernel-based methods in ARCH models [59]
- splines [8], multivariable Regression Splines (MARS) [46]
- the nearest neighbours [45], eventually in an *ad hoc* feature space [42]
- the connectionist models (semi-parametric): the multi-layer perceptrons [43], the radial basis functions [10], the time delay neural networks (TDNN) [75] and other hybrid approaches
- the models based on fuzzy logic
- wavelets [63]

One notices that an important effort has been realized in the field of non-linear models: the contribution of the neural-networks has still to be precised. To this end, we are going to discuss the non-parametric point of view on prevision in order to evoke afterwards the advantages and drawbacks of connectionist models. Before, we shall first examine some "experimental" results allowing a first comparison between these various approaches.

2.2 Forecasting competitions

After a summer-course in Santa Fe in 1990 where several specialists of non-linear prediction models had met [11] it was decided to undertake a prediction "competition" to try and sort out better the various models. A competition does not solve all problems but it may provide some elements of information as to the practical interest of the various approaches. The confrontation of the results of this so-called "Santa Fe" competition led to another meeting [79]. The results may seem contradictory since the methods, based on the use of networks of the multi-layer perceptron type, gave both the best and also the worst results. This apparent contradiction has the merit of bringing to the fore, the dangers linked with the use of neural networks and particularly that of overtraining. The flexibility and great capacity of connectionist models allow them to adjust themselves to a sample as precisely as one wishes. But, if it possible to do everything, it is also possible to do anything and it is not because the model is well adjusted on the sample that it will necessarily give good prediction results on new data. The use of a connectionist model must therefore necessarily be associated with a mechanism allowing the control of this model capacity. Another competition on power demand time series has also been won by neural networks - seen under a bayesian point of view [49].

These competitions have also allowed to precise in which cases the connectionist models may prove useful i.e. when:

- the aim is the minimization of a prediction-criterion as opposed to the description of the phenomenon.

- no prior knowledge on the process having generated the data is available, particularly no model is given.
- many data are available
- the data suggest the use of a non linear model (this can be tested)
- time series is sufficiently regular in an *ad hoc* representation space. Recall that multilayer perceptron type when properly used, put into practice a smoothing technique.

Other competitions are clearly less favourable to connectionist models. On the data of the M and M2 competitions [50] the neural networks haven't shown any superiority [68]. These data are probably too short (between 50 and 100 points) and also perhaps "too specifically economic". We shall probably get some more information soon, as a competition with financial data is taking place now.

Since we have now roused interest on connectionist models, and since we have presented them as a class of non parametric models, we are going to be more precise on what we mean by non parametric model.

2.3 Models to minimize the prediction error

On important point when using neural networks is that no model of the system is available. We are now going to show how to derive the best predictive a model based on the minimization of some given cost but not on physical considerations. This analysis has been inspired by the one performed to design input/output black box models and the point of view developed in the "kernel estimator community" [36].

2.3.1 The auto regressive function

Let us consider the observed times series $\{x_t\}_{t=1,T}$ as the realization of a discrete stochastic process that is to say a series of random variable $(X_t)_{t=1,T}$ ordered in time. In a "black box" modeling approach the problem is not to estimate some parameters but to find the best forecasting value $\widehat{x}_{t+1}$ of X_{t+1} knowing the past Φ_t that is to say we are looking for the function $\widehat{x}_{t+1} = f(\Phi_t, t)$ minimizing the following criterion:

$$\min_{f \in \mathcal{F}} J_g = \mathbb{E}\Big(\|X_{t+1} - \widehat{x}_{t+1}\|^2 \mid \Phi_t = x, t \Big) \tag{4}$$

where $\Phi_t = (X_t, X_{t-1}, ..., X_1, U_t)$ and U_t denotes the exogenous variables explaining the next random variable X_{t+1}. A first important hypothesis we made now is that the distribution of the variable to be forecasted only depends on a finite set of past values. That is to say that $\Phi_t = (X_t, X_{t-1}, ..., X_{t-\tau}, U_t)$ where τ denotes the order of the model. In this case the process $(X_t)_{t=1,T}$ is said to be τ–Markovian. Note that the Markovian hypothesis is necessary to derive the following consistency results.

The solution of the minimization problem (4) is given when it exists by the autoregression function m:

$$m(t,x) \stackrel{\Delta}{=} \mathbb{E}(X_{t+1} \mid \Phi_t = x, t) \tag{5}$$

It can be shown that for any Markovian process such function exists. This autoregression function is our target and we are going to see multilayered perceptron as an estimator of this function. But this is not always possible and the double dependency of the autoregressive function on time t and other informations x may be the root of some problems. If this interaction between time and other informations x is too complex the autoregressive function won't be identifiable. Temporal component are of two kinds. Either time determine a tendency or it's effect is periodic as it is for instance for a seasonal effect. The tendency model the fact that the observed system producing the time series is not isolated according to Ramsey's definition [62]. To deal with such a problem we propose to eliminate the tendency to concentrate ourselves on the isolated system. To do so we make the hypothesis that the autoregresion function admit an additive tendency $g(t)$, so it can be written under the form:

$$m(t,x) = r(x) + g(t) \tag{6}$$

$r(x)$ being a bounded function, the autoregression function of an isolated system. In the following the autoregression function will be supposed to be independent from the time and denoted $r(x)$. The underlying process will be supposed to be Markovian (at order τ).

Since the target function is an expectation and the underlying density function is unknown the autoregressive function is impossible to compute directly. In our approach the autoregressive function $r(x)$ is estimated by $\widehat{r}(x)$ the function minimizing the empirical cost:

$$\min_{\widehat{r} \in \mathcal{F}} J_{emp} = \frac{1}{T-1} \sum_{t=1}^{T-1} \| x_{t+1} - \widehat{r}(\Phi_t) \|^2 \tag{7}$$

For this empirical mean to converge to an expectation it implies certain ergodicity properties from the process. From this modelization some problems arise:

- choice of set $\mathcal{F}$ which reflects the *a priori* and determines the type of non parametric model chosen (kernel, spline, neural network, wavelet ...).
- nature and number of explication variables. In other terms how to chose the delay τ and how to chose the other explicative variables
- consistency of the principle, i.e under which conditions and at what speed - in function of the sample size T, the minimum of empirical cost converges on the minimum of theoretical cost

- validity domain of the model, particularly when should a non-parametric and a parametric model be used.

Before trying to answer some of the above questions, we are going to precise the nature of the model we are going to put on the process in order to identify it by minimization of the empirical risk.

After trend elimination, the most general model we would like to consider would be expressed thus:

$$X_{t+1} = f(t, \Phi_t, \varepsilon_t) \tag{8}$$

where ε_t denotes some unobservable process. If function f (anticipation diagram) is non-linear and if time t doesn't appear explicitly in the equations, the series of random variables X_t then defines a Non-linear, Auto-Regressive process with exogenous entries (NARX) [15].

But this truth model is unknown. Since we are interested only by finding the best possible forecasting approach. we are dealing with the following black box model:

$$X_{t+1} = r(\Phi_t) + \varepsilon_t \tag{9}$$

where ε_t is the equivalent innovation with zero mean and finite variance. Note that we consider the process with no time component. That is to say if Y_t is the observed process, First by eliminating the time component $X_t = Y_t - g(t)$. This kind of model is called NLAR in [72] (page 98). But since no hypothesis is made about function r we prefer the to speak about Functional Auto Regressive model (FAR). From this point of view, the problem is to define a functional estimator. The function $\widehat{r}(x)$ computed by a multilayered neural networks can be seen as an estimation of function r.

2.3.2 Confidence interval on the forecasted value

In many applications a confidence interval is required to take a decision. This interval may be designed in the following way. Assume $\widehat{r}(x)$ is the function computed by the neural network trained with x_t as input and x_{t+1} as output. Assume for simplicity w.l.o.g. that the forecasted value only depends on the previous value (only one input x_t to forecast x_{t+1}). Then for a given observation x at time t the forecasting error will be

$$\mathrm{err}(x) = \mathbb{E}\left((\widehat{r}(x) - X_{t+1})^2 | X_t = x\right)$$

Assume $r(x)$ is the best forecasting function. Then we can derive for a given x

$$\mathrm{err}(x) = (\widehat{r}(x) - r(x))^2 + \mathbb{E}\left((r(x) - X_{t+1})^2 | X_t = x\right)$$

The error has two component B^2 and $\sigma^2(x)$ defined in the following way

$$B(x) = (\widehat{r}(x) - r(x))$$

$$\sigma^2(x) = \mathbb{E}\left((r(x) - X_{t+1})^2 | X_t = x\right)$$

First term is the estimation error while the second term is the incompressible error, so called the conditional variance. Both of these term may be estimated. $B(x)$ is estimated through a bootstrap procedure presented latter in this paper.

To build an estimate of $\sigma^2(x)$ it has to be noticed that

$$\sigma^2(x) = \mathbb{E}\left(X_{t+1}^2 | X_t = x\right) - r^2(x)$$

since we already have an estimate of function $r(x)$ thank to our first neural network the conditional expectation of the square of the output remains to be estimated. This can be performed with a second neural network using the same inputs than the first neural networks but using y_t^2 as target instead of y_t. Assume $\widehat{r}_2(x)$ is the function computed by the second MLP trained with y_t^2 as output, for a given x the estimator of the conditional variance is given by:

$$\widehat{\sigma}^2(x) = \widehat{r}_2(x) - \widehat{r}^2(x)$$

Note that in this case the underlying model for process $(X_t)_{t \in \mathbb{N}}$ is the following FARCH model (FARCH stands for Functional Auto Regressive with Conditional Heteroscedacity that is to say non constant variance) [36]:

$$X_{t+1} = r(X_t) + \sigma(X_t)\, \varepsilon_t \tag{10}$$

where ε_t is the equivalent innovation with zero mean and variance one. Note that higher conditional moment may theoretically be estimated in the same way. For instance, the third conditional moment give information about the disymetry of the error.

$$\mathbb{E}\left(X_{t+1}^3 | X_t = x\right)$$

But we verify empirically that the convergence rate of a functional estimator to such a third order conditional moment is so low that this estimation can not be use for real applications.

2.3.3 Functional estimators

All time-series will not possibly be expressed according to the model described in equation (9). Moreover, some supplementary hypotheses have to be set to be to insure the consistency of the principle of minimization of empirical risk in the absence of a supplementary hypothesis as to the noise law for instance. The process that has generated the data will have to verify the following properties:

- the "short term memory" therefore Markovian aspect
- ergodicity,

- a certain probability structure, invariant when translated in time (simple or strict stationarity ?) or more feebly the asymptotic independence of the Present with its Future and Past, given by the mixture character of the process.

Let us underline that a process which verifies a mixing condition is not necessarily stationary (34). We think the conditions expressed by Doukhan and Ghindès [24] are sensible for the series we deal with. These conditions are the following ones:

> *The autoregression function r of the additive model described by the equation (9) is bounded. The noise law ε is absolutely continuous in regard to Lebesgue measure.*

Under the conditions (X_t) is a φ-mixing process on which consistency results have been demonstrated for estimations with a kernel estimator.

These results where extended to non linear models in general [22] and neural networks in particular [82] under the same hypothesis. In this paper White give a clear comment about mixing process: *Mixing process are a class of time-series processes that can exhibit considerable short-run dependence, but display a form of asymptotic independence, in that events involving elements of (X_t) separated by increasingly greater time intervals are increasingly closer to independence.* Here is the main point about Doukhan and Ghindès conditions. They allow to consider after trend elimination the remaining time series as the realisation of a mixing process and therefore insure the consistency of the functional estimators.

2.4 Parametric or non-parametric model: the priors about complexity

Even if the distinction between parametric and non-parametric model is widely used, it is difficult to give a precise and satisfactory definition. The various suggested definitions are based on the number of parameters - a parametric model being according to the authors, a model for which the number of parameters is fixed or finite. Another definition can be based on the consistency of the estimator. From this point of view multilayered perceptrons are both parametric and non parametric. multilayered perceptrons with a given fixed architecture are not consistent and therefore multilayered perceptron is a parametric model. But using the sieves technic, that is to say an increasing architecture depending on the sample size, Multilayered perceptron is a consistent estimator [82]. The fact that the same neural network may be seen at the same time as a parametric and non parametric model reveals the weakness of such classification.

A different approach consists in establishing a distinction based on the nature of the prior hypothesis set on the searched solution. Seen under this point of view, the parametric and non-parametric models are bound with two types of " a priori" which are different whether we look for the most explicative (parametric) model,

or the most precise (non-parametric) model. Therefore, all depends on the chosen criterion. If it is the precision of the prediction that prevails, it is necessary to adjust the model complexity to the data complexity. Up to a point, it little matters whether the model is parametric or not, the important thing being for it to be flexible i.e adjustable in its complexity to that of the data. This complexity which is still to be defined is bound to the regularities of the data from an informational point of view [20].

There are two different ways to describe the time-series regularity [48], therefore two ways of controlling the complexity of the model. Either one is interested by the information - quantity necessary to describe the predictable part of the phenomenon, one tries to specify the information contained in its stochastic component. In the first case, the complexity is the one meant by Kolmogorov, i.e the size of the smallest program implementing the function which generates the data. In the second case the complexity is described by the quantity of Shannon information. In the first case, one looks for a compromise between the size of the program and the searched precision. This size is to be related with the number of the parameters of the model. It is the parametric point of view on the control of the complexity. On the other hand, in the second case, if we favour the informational content of the solution the non-parametric approach is imperative. In this case, even if the structure of the model and therefore its number of parameters remains fix it is the effective number of the parameters of the model which will have to vary, allowing the adjustment of the model complexity to the complexity of the data. Following Friedman classification [28], we prefer to use the distinction between rigid or flexible model.

Therefore, the choice of a rigid or flexible model depends on the criteria and prior knowledge introduced about the nature of the solution. If the criterion is the precision of the prediction, in the absence of physical model of the phenomenon, a flexible model must be used. But the most important distinction concerns the mode of control of the complexity which will depend on the aim to attain. This distinction will give the model its rigid or flexible character. The consequences regarding the use of neural networks of multi-layered perceptron type are important. We shall see that this type of model can be used in two different ways according to the type of complexity control chosen, whether we try to adjust the number of real parameters of the number of effective parameters of the network.

3 The connectionist approach of prediction

Owing to the results of the competitions we have already been able to underline some specificities bound with the connectionist approach: an aim of precision in the sense of a certain criterion (often the least squares). If these conditions are fulfilled, then a neural network type model will be organized. It is for us now to choose the model.

3.1 Various connectionist models - various architectures

3.1.1 Historical approach

It is to Lapedes and Farber [43] that we owe the first work showing the possibility to identify and predict the future of chaotic determinist time series, owing to multi-layered perceptrons. This study has opened the way to applications in real-world such as, for instance, the forecasting of stock-return for IBM by White [80]. The latter has gone further, proposing on the following year, a test to decide on the eventual presence of non-linearities impossible to detect with a linear model [81]. This test, based on the use of a multi-layerd perceptron, has been called the neural network test. Yet in this articles - as in the others available in that time [9, 73, 84] the description of the employed method is brief. Some good results in prediction are also reported, that have been obtained by using other more complex connectionist models such as the recurrent models [16]. Concerning the one-step-ahead prediction we have to wait for Andreas Weigend and Eric Wan and their related papers [76, 78] to find detailed descriptions of significative applications including systematic comparisons with other approaches. Since then, the applications have become more and more numerous and about sixty articles on this question have been registered in 1993 and about forty in 1994. At the same time some "automaticians" have taken an interest in the same type of models, applying them to identification with similar results [15, 56, 71].

3.1.2 Connectionnists models of time

Parallel to this main trend of application, another question has mobilized the connectionist community: what is the best architecture for modeling a dynamic system [54] ? Obviously all depends on the problem. In an article dealing with this subject [13] the authors suggest distinguishing three different mechanisms allowing the treatment of time informations corresponding to three different levels of complexity of the time effect. Time can be treated either from the outside or internally by the use of architecture, or, lastly, directly at the level of the formal neuron itself. At the first level, we rediscover time modelized by a time-window of fixer size, the model already suggested in [64] with its time-delay variant TDNN [75]. The neural network of multi-layer perceptron type can then be seen either as a non-linear autoregressive model (NARX) [15], or as a filter with infinite impulsional response (IIR) [76]. other equivalent networks have also been used such as networks with radial basis functions [10]. To take time into account in the very architecture of the network an internal memory. This is made possible owing to various closed loop systems making of the network a recurrent network. These closed loops have been introduced in different ways. Either to take into account the state - notion for controlling a dynamic system looping the outputs towards the inputs [41].

Or for the "context" modelization in the scope of the analysis and understanding of the language (looping of the hidden units towards the entries [26]). Then "unifying"

studies have allowed to define totally connected networks and to precise different algorithms of identification for this type of looped architecture [52, 57][1]. Other models of still more complex dynamic recurrent networks have been put forward as the "dynamic recurrent neural network" DRNN [2].

In a recent review paper [70] the authors present these various architectures by referring to the existing models. The non-linear models of the black box type are classified according to the type of "regressor" chosen.

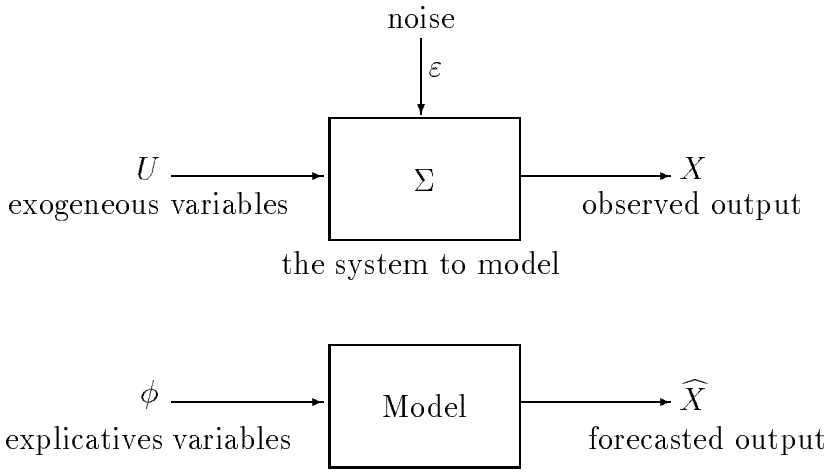

Figure 1 : Black box models

- NFIR non-linear with finite impulsional response : $\phi = U$
- NARX or parallel series : $\phi = (X, U)$
- NOE non-linear with output error or parallel model : $\phi = (U, \widehat{X})$
- NARMAX : $\phi = (U, X, \varepsilon)$
- NBJ Box and Jenkins non-linear model : $\phi = (U, \widehat{X}, \varepsilon)$
- Non-linear models with state representation.

These various models can be carried out in a neural network or in another non-parametric model. The models of the types NOE, NARMAX, NBJ with state-representation correspond to recurrent networks.

The third way consisting in introducing a dynamic aspect at the level of the neuron itself, leads to still more complex systems which are therefore all the more difficult to control [61]. Their application to prediction is therefore delicate [74]. Let us underline lastly, that many applications use hybrid models in which only one component is of the connectionist type.

[1] Les réseaux récurrents ont fait notamment l'objet d'un numéro spécial de la revue *IEEE Transaction on Neural Networks* en Mars 1994.

3.1.3 Choice of a connectionist model

The choice of an architecture is ruled by the type of problem to be resolved. We are going to consider the simplest architecture: the multi-layer perceptron without any loop. This model corresponds to short-term memory phenomena and to a rather to a weak informational content i.e to a weak relationship between signal and noise. On the other hand, for some other phenomena, such as the language or speech analysis, the presence of a "context" seems indispensable. In this case, we shall choose a recurent network.

From now, we shall consider only the multi-layer perceptrons because their architecture implements the models we are interested in. The mathematic formulation of this class of estimators is the following one $\mathcal{F}$ denotes the set of perceptrons with one hidden layer, defined by $\mathcal{F} = \bigcup_{k=0}^{+\infty} \mathcal{F}_k$ with:

$$\mathcal{F}_k = \Big\{ f \;\Big|\; \mathbb{R}^d \to \mathbb{R}\ ,\ f(x) = \sum_{i=1}^{k} c_i \alpha(W_i x + w_{i0}) + w_0, W_i \in C \ \text{ et } \ c_i, w_0, w_{i0} \in \mathbb{R} \Big\} \tag{11}$$

where α denotes the activation function (typically the hyperbolic tangent). In some cases architectures with two hidden layers are reset.

3.2 Identification Algorithm and learning

The basic algorithm for the identification of the multi-layer perceptron parameters remains the retropropagation of the gradient [64]. In a study on this matter [15] it is suggested to apply this same principle of prediction-error-minimization, in using Gauss-Newton algorithm, so as to minimize the risk, without explicitly precising which mechanisms are used to control the complexity. In that work, the complexity-control was implicitly achieved by using the smallest number of units in the hidden layer, therefore a reducer number of parameters. A certain lack of precautions naturally led to overlearning denounced by some authors [62]. Still in this parametric setting, several algorithms for the elimination of some parameters have been proposed, starting from heuristics, such as the so-called "optimal Brain Damage" [44] or based on a statistical test [51, 17]. It is always the same principle: the research of the smallest real parameters by building up or cutting down method. The number of real parameters of the model can also be reduced by selecting the explicative variables presented to the network [18].

Another method for controling the complexity is inspired by the regularization theory and is known as penalization [77]. This method consists in favouring the "simple" solutions, adding a term of penalization to the function to be minimized which may, for instance become :

$$\sum_{t=1}^{T-1} \|x_{t+1} - \widehat{f}(\Phi_t, W)\|^2 + \lambda \sum_{j=1}^{|W|} \frac{w_j^2/w_0^2}{1 + w_j^2/w_0^2} \tag{12}$$

in which $|W|$ represents the number of parameters to be adjusted. At the same time, the authors propose a heuristic to decrease λ during the optimization. The aim is to penalize not only the important connections but also the too weak connections which are supposed to correspond to more complex models, i.e having a more important number of efficient parameters. The "favoured" areas are around the values 1 and -1. This approach also consists, to a certain extent, in eliminating the non significant parameters. We have chosen a different approach which, to our eyes, seems easier to put into practice: that is noise-injection [29]. Noise-injection is a particularly attractive heuristic because of its null algorithmic cost. The method consists in adding noise in the examples presented at the entrance of the multi-layer perceptron. The model complexity is then controlled by the variance of the additional noise. Other techniques do exist but it is difficult to prove the advantage of one approach over another one. Even if it is possible to justify theoretically the noise-injection, it is our experience which also led us to choose this technique for complexity-adjustment. Up to a point, the method matters little as long as it allows the control of the number of the effective parameters of the flexible model.

As we have just seen the same neural network can be used in two different ways, corresponding to two different objectives: the parametric point of view, allowing the determination of the smallest architecture, and the non-parametric approach which looks for the optimal number of effective parameters for an over dimensioned network. This second approach will give better results. And it is no wonder, as we have already seen that its aim is to minimize the predictional error.

3.3 Consistency of the principle

We are going to show, in the particular case of the algorithm of noise-injection in the multi-layer perceptrons, and for some processes, the consistency of the empirical risk minimization principle, that is to say the complete, uniform convergence on a compact $C \subset \mathbb{R}^d$. of the function $\widehat{f}_T(x)$ calculated by the network identified at time T towards the regression function. In other terms, a real positive ε exists such that :

$$\sum_{T=1}^{\infty} \mathbb{P}(\sup_{x \in C} |\widehat{f}_T(x) - r(x)| > \varepsilon) < \infty \tag{13}$$

Whatever the marginal distribution denoted $\mu(x)$ supposed to be common to all the random variables , as long as they verify certain conditions of regularity (condition A1 in [34]). To obtain this kind of result, we shall have to lean on the fact that a neural network is a universal approximator and therefore let his architecture grow just as the regressor of Naradaya Watson grows according to the size of the sample. We shall thus start giving a result in consistency. The evaluation of the speed of convergence should come later.

Noise injection consists in modifying the empirical cost and therefore in minimizing

the noised cost, thus:

$$\widetilde{J}_{\text{NI}} = \frac{1}{N}\frac{1}{T-1}\sum_{t=1}^{T-1}\sum_{n=1}^{N_t}\|x_{t+1} - \widehat{f}(\Phi_t + \xi_{t,n})\|^2 \tag{14}$$

in which ξ designates a realization of a random variable drawn according to the distribution K_σ with variance σ. To a certain point when often injecting noise (N large) one minimizes:

$$J_{\text{NI}} = \mathbb{E}\Big(\frac{1}{T-1}\sum_{t=1}^{T-1}\|x_{t+1} - \widehat{f}(\Phi_t + \xi)\|^2\Big) \tag{15}$$

that is to say defining $\zeta_t = \Phi_t + \xi$

$$J_{\text{NI}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\int \|x_{t+1} - \widehat{f}(\zeta_t)\|^2 K_\sigma(\zeta_t - \Phi_t)d\zeta_t \tag{16}$$

The minimum of J_{NI} is then given by [29] :

$$\widehat{f}_{\text{NW}}(x) = \frac{\sum_{t=1}^{T-1} x_{t+1}K_\sigma(x - \Phi_t)}{\sum_{t=1}^{T-1} K_\sigma(x - \Phi_t)} \tag{17}$$

A multi-layer perceptron being a universal approximator if only the number of units in the hidden layer is sufficient, the network will converge to the Naradaya Watson regressor $\widehat{f}_{\text{NW}}(x)$. The variance σ plays the part of the bandwidth of the Naradaya Watson regressor. Now we know that under certain conditions this estimator is a consistent estimator of the autoregression - function [34]. This result can thus be directly adapted for the noise-injected multi-layer perceptrons.

Theoreme

If the random processus $(X_t)_{t\in\mathbb{N}}$ is a φ–mixing process, if all random variable X_t all admit the same marginal density $\mu(x)$, if this density μ and the autoregression function r are "regular enough", if the kernel $K\sigma$ is "well-chosen", if the capacity of the multi-layer perceptron is sufficient, if the variance of the injected noise verifies : $\sigma = aT^{-b}$ with $0 < a$ and $0 < b < \tau/2$,

Then the principle of the minimization of the empirical risk is consistent i.e the multilayer perceptron obtained in minimizing the empirical risk uniformly converges uniformly in probability towards the autoregression function when the size of the sample tends towards infinity. The (13) property is verified.

The same result can be obtained with other types of less restrictive mixture conditions such as α-mixture conditions. Compared to the kernels, the use of the multi-layer perceptron offers two advantages. On one hand, they allow a sort of "compilation" of the Naradaya-Watson regressor. On another hand, when the regression-function verifies certain properties, the neural approximation allows us to defeat the "curse of dimensionality" [4] thanks to the non linearities of multilayered perceptrons. On this class of functions it should be possible to calculate more advantageous rates of convergence than those obtained by the kernel-method. To sum it up neural networks are preferable when the number of explicative variables is important and in a lesser measure, when the size of the sample is important.

3.4 Validation and confidence intervals

The problem of the validation of a model and that of the prediction-precision are bound. To select a model for a given application it is necessary to be able to compare it with others. The comparison of the models can be achieved but with the help of a "good" estimator of the error in generalization. This estimator will also give us the precision of the prediction.

3.4.1 Test set

The use of a test ensemble to validate a model is both the simplest and the most frequently used approach. It consists in separating the sample into two parts, and keeping part of the data for the validation. The empirical error on the test ensemble being independent from the model identification process is an estimator without bias of the error in generalization. But there is sometimes a lack of data to build up a representative test ensemble. In the case of solar eruptions for instance, the data given for the test had to be split into two parts. It seems that the behaviour of the series from 1955, is not comparable to the rest of the series [77] ! These test datas are not significant and should rather be used to identify the model.

3.4.2 Forecasting resampling methods

When the data are too few to build up a representative test-ensemble it is possible to resample them so as to build another estimator of this error in generalization. Among the three most used resampling techniques: Jackknife, crossed validation and bootstrap, we shall study only the last technique which costs less in calculation on our big samples. In prediction, the dependent character of the sample imposes some adjustments before using these techniques.

The bootstrap principle consists in duplicating the sample [25]. To do so, we suppose the data to have been generated according to the following model

$$X_{t+1} = f(\Phi_t) + \varepsilon_t \qquad t = \tau, T-1 \tag{18}$$

in which the ε_t are supposed residuals i.i.d. with an unknown distribution d. To resample, we must first identify a model $\widehat{f}$ from the time series x_t, $t = 1, T$. The residuals $\widehat{\varepsilon}_t$ are then estimated by the difference :

$$\widehat{\varepsilon}_t = x_{t+1} - \widehat{f}(\Phi_t) \qquad t = \tau, T-1 \tag{19}$$

An empirical estimator $\widehat{d}$ of d can then be built, starting from the estimation of the residuals $\widehat{\varepsilon}_t$.

$$\widehat{d}(x) = \begin{cases} \frac{1}{T-\tau} & \text{if } \exists\, t \in \{\tau, \cdots, T\} \text{ such that } \quad x = x_t \\ 0 & \text{else} \end{cases} \tag{20}$$

It is now possible to resample, starting from the first observed values in which $\Phi_1^* = \Phi_1$ is a sample drawn :

$$x_{t+1}^* = \widehat{f}(\Phi_t^*) + \varepsilon_t^* \qquad t = \tau, T-1 \tag{21}$$

where ε_t^* is a sample sorted from distribution $\widehat{d}$. It remains to make the algorithm work again to find $\widehat{f}^*$ the estimator of $\widehat{f}$. When repeating the operation B times, we obtain B different models whose statistic properties are the same as those of the initial sample. For instance, a prediction-precision will be obtainable by observing the randomizing of the predictions provided by the B "replicas" obtained by the resampling. There exist several estimators of error in generalization all built from a replicated sample ([25] chapter 17). The naïve estimator consists in estimating the error owing to the mean of the empirical errors calculated with the initial sample and the "bootstrap" functions $\widehat{f}^*$. This method is relatively costy in calculation time for it multiplies by B the times of learning but it gives good results.

If one is only interested by the determination of a confidence interval about the prediction, it is then possible to identify the error-variance, i.e the g^2 component of the FARCH model presented by the equation (10). This technique has already been successfully used [58, 18].

4 Forecasting on real data: the water consumption case

4.1 Supply prediction: the problem

The knowledge of the morrow water-demand is a concrete prediction problem bound with the operating of the systems of distribution of drinking water. To build up a prediction model we have a time series taking again the daily consumptions from 1976 to 1992 as well as the average temperatures and the pluviometry of those days. The datas have been divided into two subsets, the years 1976-1982 being used for the training and the years 1983-1992 constituting the test set.

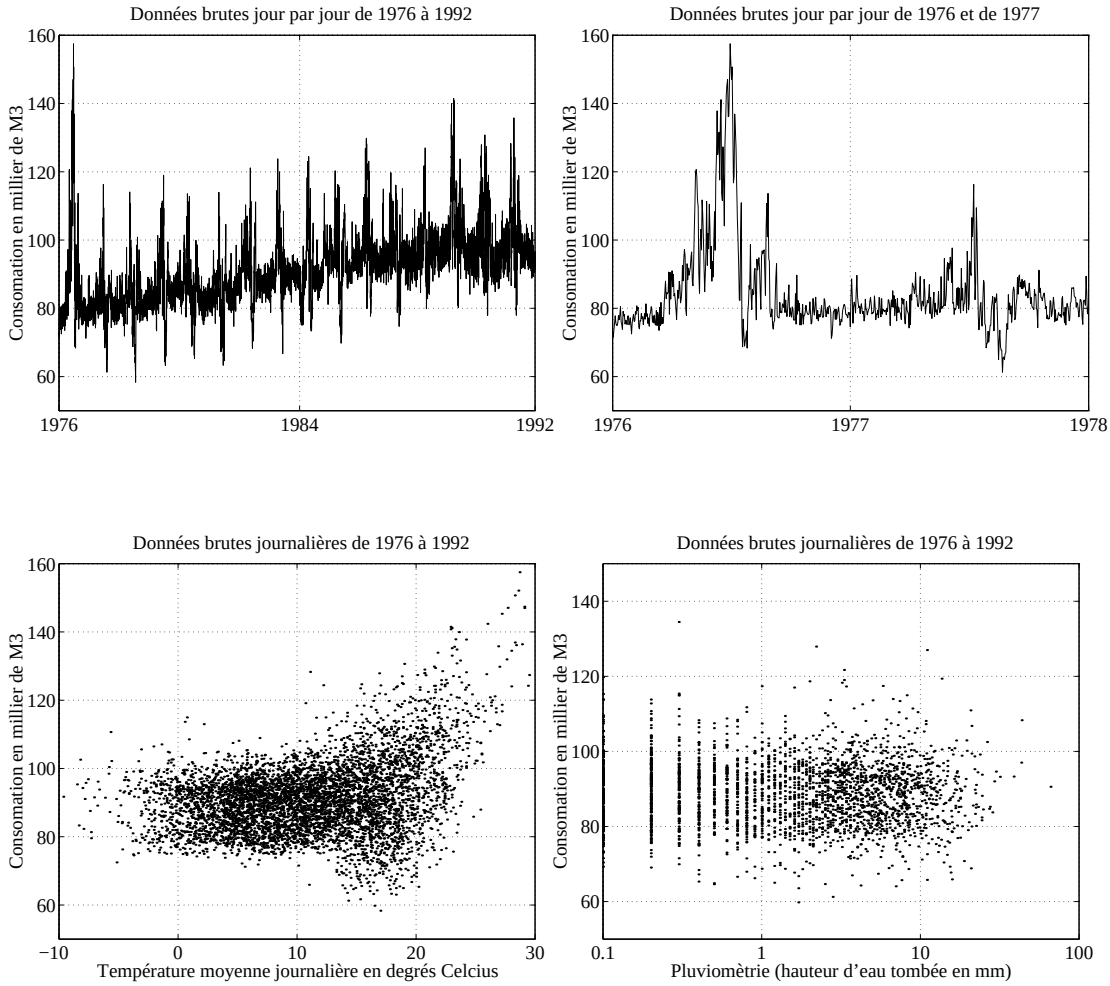

Figure 2: Raw water demand data

Figure 2 presents the raw data of water consumption in an under network serving an area in the south suburbs of Paris. The series admits a tendency, i.e a regular increase of the water consumption through out time. This tendency is due to two factors: the increase of the number of connection points linked with the seasons and the other is weekly, visible on the data of 1976 and 1977. These data show an increase of the consumption at the beginning of summer, then a sudden fall linked with the holiday departures.

The relationship between the consumption and the temperature is non-linear. In other studies, this relationship has been modelized in a threshold system. Under about 15^o Celsus temperature has no correlation with consumption. Above 15^o the consumption increases in a linear way, according to the temperature. But how can we define this threshold precisely knowing that it varies according to the season and

the number of persons present in the area ? The non parametric approach may improve the analysis of this non-linear aspect of the dependency.

Regarding the pluviometry, the relationship is not as clearly established and it seems according to figure (2) that rain and the consumption are independent. It is not exactly so. By experience, the operating services know that, in summer, when it rains the consumption tends to decrease whereas, in winter rain has no influence on the water-demand. On another hand, the available information is not quite sufficient for the 10 mm of water fallen in five minutes during a storm will not at all have the same effect on the consumption, as the 10 mm fallen continuously during a day. The factors influencing the water-demand are thus:

- Periodical factors
  - period of the year, season effect
  - day of the week
- Sociological factors
  - Bank holidays
  - School holidays
- Meteorological factors
  - temperature
  - pluviometry
  - eventually other non available variables such as hours of sunshine (or cloud covering) or the successive number of rainless days.

The raw data present a tendency which has to be eliminated. To achieve this aim, the tendency is estimated by a linear regression, eventually by bit. The year 1976 is not taken into account because being too exceptional it would perturb the results. The estimation of tendency is then expressed in the following way:

$$T = 3,7j + 80000 \tag{22}$$

in which j represents Time in days counter from the origin of the calculation: January 1st 1977. If the residual series represented on diagram 3 is not necessarily stationary it is at least bound and verifies the hypotheses bound with the flexible estimation of the autoregression $r(x)$.

There is still to discuss about the "normal" character of the data. Indeed, if we can accept the gaussian modelization (if we admit the joint law of any complete family of observation is Gaussian). Then the linear prediction techniques are optimal [3]. To study this hypothesis we can resort to a normal using Henry's line [66]. The purpose of a normal probability plot is to graphically assess whether the data

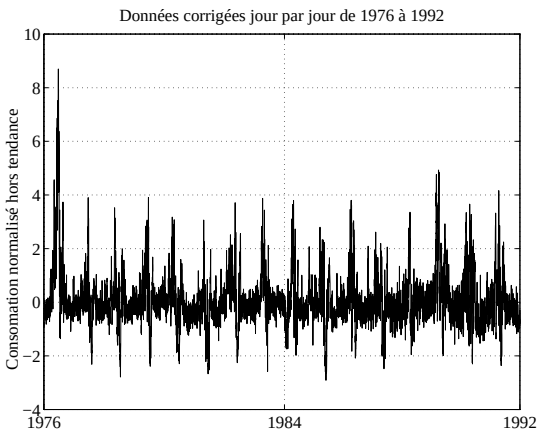

Figure 3: Corrected data after trend elimination

could come from a normal distribution. If the data are normal the plot will be linear. The study of the plot (figure 3) allows us to reject the normality hypothesis. Indeed the empirical distribution admits a significant deviation in comparison with the theoretical curve for high consumptions. This deviation was predictable when examining the data historiogram. This can be explained by the fact that there are peaks of consumption but that, on the other hand, the demand very rarely lowers under a certain threshold. The water-demand is therefore a dissymetrical phenomenon. The nature of the problem and the operators' experience suggest that we should, in fact, use two models: one for winter and one for summer/ According to them the hypothesis of normality would be acceptable in winter, whereas, in summer the effects of temperature and pluviometry should be eliminated first. The study of Henry's straight lines only from the sole data of the months of November, December, January and February invalidates this "a priori": in winter too, the consumptions do not follow a normal law.

As to the examination of the summer data, it throws us into the heart of the problem: how to model the relationships between the consumption and the temperature and rain data, how to take into account the interactions between the variables and how to decide that summer has come ? There are two solutions to this problem. The first one consists in eliminating these phenomena as we have eliminated the tendency to make the series stationary. This preliminary work allows us after to use a linear model. The second solution chooses to abandon the linear model to the profit of a non parametric modeling of the phenomenon. But let us remember, not with standing - and this is probably the most important argument - that the aim pursued in this study is the minimization of the prediction error.

The physical phenomenon ruling over the drinking water consumption being unknown, it is the non parametric modeling which will give us the best precision

since precision is its sole and unique objective for the flexible models the criterion comes before the model ! To finish with the study of the data a revealing and important aspect of their nature is provided by the analysis of the autocorrelations and periodograms presented in figure 4. The "week-day" and season effects are quite

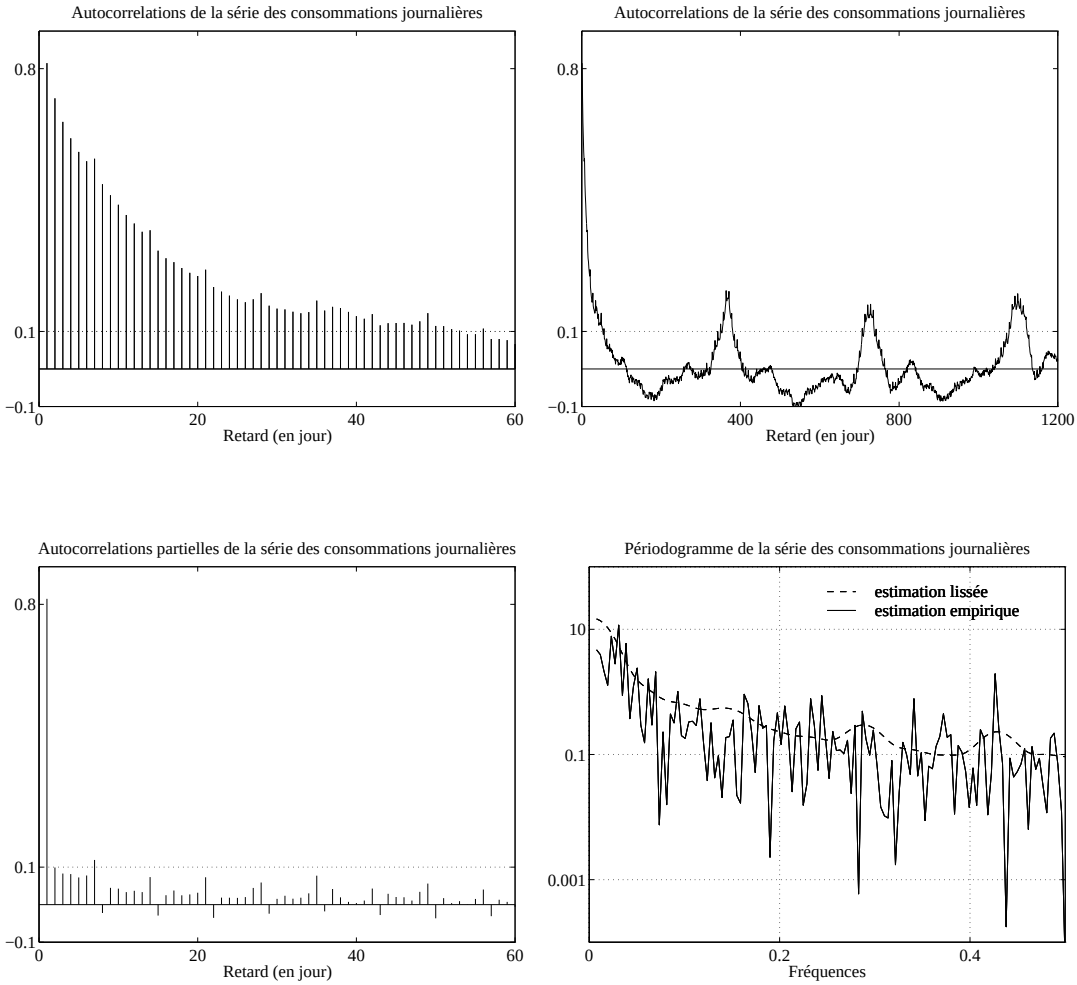

Figure 4: Autocorelation, partial autocorelation and power spectrum

visible. The data appear cyclo-stationary with a lot of energy in the low frequencies. No simple structure emerges from these representations.

4.2 Used method

Input Selection : The partial autocorrelations suggest to use an at least 7 days delay to identify the function of autoregression. In a similar study on energy consumption filtered data have been used as explicative variables [49]. This seems to

open on theoretical problems [49]. The three last days' temperatures and pluviometry have also been taken into account. Lastly the week-day has been coded by seven Booleans. The explicative variables thus come to the number of 20.

4.2.1 Pretreatments

: The tendency of the series has been eliminated. The variables have been normalized and the rain logarithmically coded according to our results, the other considered coding - particularly for the day of the week - don't bring anything more on this series. Still to be considered: the public holidays. The initially chosen approach consists in considering the "holiday" effect as an additive effect to be corrected. This effect was therefore linearly estimated then eliminated. In the example presented here, this pre-treatment has not been performed.

4.2.2 Parameters identification

The general principle is that of the minimization of the empirical risk by a gradient process called retropropagation. But, as we have seen, the principal problem linked with the use of multi-layer perceptrons is that of complexity-control. In this paper, we have compared two approaches of complexity-control which reflect two different points of view on neural networks: the parametric and non parametric views.

If we consider a multi-layer perceptrons as a parametric model, we shall control its complexity if we find the smallest model giving the best result. We can then use an empirical incremental approach which consists in starting with one unit in the hidden layer, then in increasing this number so as to attain the minimum on an estimator of error in generalization. In our example, the minimum was obtained for 4 units in the hidden layer.

The alternative approach considers the intrinsic flexibility of the model and aims at adjusting no more the real parameters of the model but its number of effective parameters. We have used the noise-injection technique trying to find the optimal variance. In this application, noise has been injected but on the random components. Let us underline that this algorithm is of course slower to converge - would it be only because of the size of the network (we took 20 units) - but it is more reliable than the "classical" retropropagation in the sense that it almost always converges on the same minimum in the sense of generalization.

4.2.3 Prediction precision

The confidence intervals have been estimated through a resampling of "bootstraping" type. The prediction-validity is also ensured by the definition of the field of validity of the application implemented through a simple mechanism of distance-reject.

Year	Persistant	ARMAX(8,11,2)	"parametric" MLP (4 units)	Noise Injection MLP
1983	4.09 %	3.38 %	3.00 %	2.95 %
1984	3.89 %	3.45 %	3.37 %	3.29 %
1985	3.73 %	3.42 %	3.14 %	3.12 %
1986	3.94 %	3.39 %	3.28 %	3.17 %
1987	4.64 %	4.19 %	3.66 %	3.55 %
1988	4.67 %	4.01 %	3.76 %	3.57 %
1989	4.35 %	4.13 %	3.78 %	3.73 %
1990	4.98 %	4.50 %	4.01 %	4.13 %
1991	4.44 %	4.07 %	3.74 %	3.60 %
Mean	4.28 %	3.79 %	3.49 %	**3.42** %

Table 1: MERA error comparison

Year	Persistant	ARMAX(8,11,2)	"parametric" MLP (4 units)	Noise Injection MLP
Mean	0.424	0.341	0.326	**0.279**

Table 2: ARV error comparison

4.3 Results

To compare results, we have looked for the best ARMAX model on the stationarized series with temperature as another explicative series. The best model for the test-error has been the ARMAX model (8, 11, 2). For comparison, we have used the minimized criterion and a criterion of the absolute value of the relative error on a test-set. This criterion is the one recommended by the prediction users multiplied by a hundred to get percentages. It has the merit to possess, besides, good statistic properties. The results as seen in table 1 show a significative advantage for the non-parametric approach of neuron-networks. The study of the residuals shows that there is still some information to be drawn.

4.4 Other applications

This type of technique has already been used for other applications of prediction for the daily water consumption in France [6], in England [39], in Germany [37] and in the United States [38]. It has also been used for electricity consumption [18, 23, 55] and for the stock-exchange [65]. In all these applications, three fundamental points bound with a prediction-application are found. Besides the prediction itself we must have a confidence-interval available and we must be able to define the validity field of the model. Other prediction-applications have also been developed owing to connectionist models, particularly for the prediction of a daily consumption profile

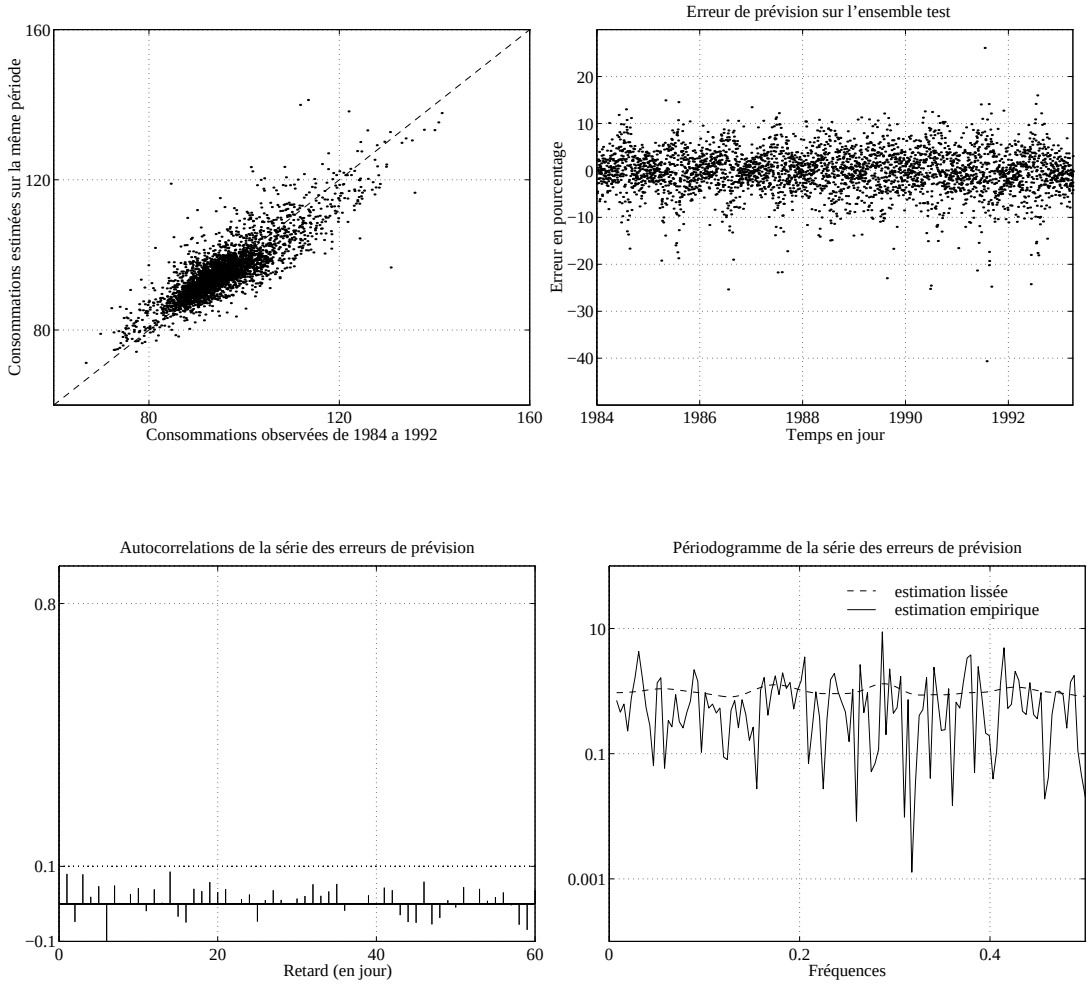

Figure 5: Forecasting error analysis

[21].

5 Conclusion

To make a one-step-ahead prediction, when the precision of the prediction is the factor and no model is available, a flexible estimator of the autoregression function must be used. If a great number of experimental data is available and, above all, if the number of explicative variables is important, one can use a multi-layer perceptron as an estimator if implementing the principle of minimization of the empirical risk. Under certain conditions this estimator is consistent. This type of so called connectionist model will therefore be generally preferred to other model which are

more complex and more difficult to use as networks. To implement this approach efficiently it is preferable to eliminate an eventual tendency. The principal difficulty is the adjustment of the model complexity. To this end, we have used noise-injection, and a method of resampling, the "bootstrap" to compare different adjustments. This last technique moreover allows us to give a confidence interval on prediction. Yet, because of its empirical character, it is still only case after case that this technique can be valued. Moreover, our results but confirm those underlined in [62]: if, when looking at the data, there are many evidences of the presence of non linearities the use of a non-linear prediction model gives significant but weakly precise savings and practically no improvement as to the reliability of the prediction. Where from the interest of a coupling with a parametric modeling to the fields of application of the non-parametric model. When using a flexible model such as a multi-layer perceptron, the essential point is the mechanism of the complexity control. But this control, being implicit, is difficult to be mastered with neural networks. In some areas, the solution may be oversmoothed, and in others it may be under smoothed. Therefore, we think important to have at our disposal a complexity adjustment mechanism which is at one local and global. It is this type of mechanism which is explicit in the wavelet representation. This representation, already used to solve the one-step-ahead prediction problems [63] defines a new class of models which might prove useful to solve the complexity-control problems raised by the flexible regression.

ACKNOWLEDGEMENTS

The authors would like to thank H.N. Pham (Lyonnaise des Eaux) to have given them access to the consumption data. They also address their thanks to all those who, under a respect on another, have helped them to achieve this study Sergio Alvares, Bernard Butchy, Thomas Czernichow, Paul Doukhan, Larry S. Liebovitch, Nicolas Limnios, James B. Ramsey, Patrice Simard, Philippe Vieu, et Stéphane Boucheron as well as all the members of the Group "Formal Approaches to learning in neural networks" supported by the french PRC-IA.

References

[1] J. Anderson and E. Rosenfeld, eds., *Neurocomputing: Foundations of Research*, MIT Press, Cambridge, 1988.

[2] A. Aussem, *Théorie et Application des Réseaux de Neurones Récurrents et Dynamiques à la prédiction, à la Modélisation et au Contrôle Adaptatif des Processus Dynamiques*, PhD thesis, Université René Descartes, Paris V, 1995.

[3] R. Azencott and D. Dacunha-castelle, *Séries d'observations irrégulières : Modèlisation et prévision*, Techniques Stochastiques, Masson, 1994.

[4] A. R. Barron, *Universal approximation bounds for superposition of sigmoidal functions*, IEEE transactions on Information Theory, (1991).

[5] T. Bollerslev, R. Y. Chou, and K. F. Kroner, *ARCH modeling in finance*, Journal of Econometrics, 52 (1992), pp. 5–59.

[6] L. Bottou. ICANN'95 proceedings, 1995.

[7] W. A. Brock, D. Hseih, and B. LeBaron, *Nonlinear Dynamics,Chaos and Instability: Statistical Theory and Economic Evidence*, M.I.T. Press, 1991.

[8] D. M. Caffrey and S. Ellner, *Estimating the lyapunov exponent of a chaotic system with nonparametric regression*, Journal of the American Statistical Society, 87 (1992), pp. 682–695.

[9] S. Canu, R. Sobral, and R. Lengellé, *Formal neural network as an adaptative model for water demand*, in Proceedings INNC'90, Kluwer Academic Publishers, 1990, pp. I–131–136.

[10] M. Casdagli, *Nonlinear prediction of chaotic time series*, Physica D, 35 (1989), pp. 335–356.

[11] M. Casdagli and S. Eubank, eds., *Nonlinear Modeling and Forecasting*, vol. XII of SFI Studies in the Sciences of Complexity, Addison-Wesley, 1992.

[12] K. S. Chan, *On the existence of the stationnary and ergodic near(p) model*, J. Time Series Analysis, 9 (1988), pp. 319–328.

[13] J.-C. Chappelier and A. Grumbach, *Time in neural networks*, SIGART buletin, 5 (1995), pp. 3–11.

[14] C. Chatfield, *Neural networks: Forecasting breakthrough or passing fad ?*, Int. J. of Forecasting, 9 (1993), pp. 1–3.

[15] S. Chen, S. A. Billings, and P. M. Grant, *Non-linear system identification using neural networks*, Interational Journal of Control, 51 (1990), pp. 1191–1214.

[16] J. Connor and L. Atlas, *Recurrent neural networks and time series prediction*, in Proceeding of IJCNN, vol. 1, 1991, pp. 301–306.

[17] M. Cottrell, Y. Girard, and M. Mangeas, *Time series and neural networks: a statistical method for weigth elimination*, in Proceeding of ESANN, 1993.

[18] T. Czernichow and A. Muñoz, *Variable selection through statistical sensivity analysis: Application to feedforward and recurrent networks*, Tech. Rep. 95-07-01, INT-SIM, 1995.

[19] J. G. DeGooijer and K. Kumar, *Some recent developments in non-linear time series modeling, testing and forecasting*, Int. J. of Forecasting, 8 (1992), pp. 135–156.

[20] J.-P. Delahaye, *Informations, Complexité et Hasard*, Langue, Raisonnement, Calcul, Hermès, 1994.

[21] X. Ding, S. Canu, and T. Denœux, *Neural network based models for forecasting*, in Proceedings of ADT'95, 1995.

[22] I. Domowitz and H. White, *Misspecified models with dependent observations*, Journal of Econometric, 20 (1982), pp. 35–58.

[23] B. Dorizzi, J.-M. Duval, and H. Debar, *Use of recurrent networks for electricity consumption forecasting*, in Neuro Nimes'92, EC2, 1992, pp. 141–150.

[24] P. Doukhan, *Mixing: Properties and examples*, Springer Verlag, 1994.

[25] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*, no. 57 in Monographs on Statistics and Applied Probability, Chapman and Hall, New York, 1993.

[26] J. Elman and D. Zipser, *Learning the hidden structure of speech*, Journal of the Acoustical Society of America, 83 (1988), pp. 1615–1626.

[27] A. M. Fraser and A. Dimitriadis, *Forecasting probability densities by using hidden markov models with mixed states*, in Time series prediction: Forecasting the future and understanding the past, A. S. Weigend and N. A. Gershenfeld, eds., Addison Wesley, Reading, MA, 1993, pp. 265–282.

[28] J. H. Friedman, *An overview of predictive learning and function approximation*, in From Statistics to Neural Networks, NATO ASI, series F, Springer Verlag, 1994, pp. 1–61.

[29] Y. Grandvalet, *Injection de Bruit dans les Perceptrons Multicouches*, PhD thesis, UTC, Laboratoire Heudiasyc, 1995.

[30] C. W. J. Granger, *Forecasting stock market prices: Lessons for forecasters*, Int. J. of Forecasting, 8 (1992), pp. 3–13.

[31] C. W. J. Granger and A. P. Andersen, *An Introduction to Bilinear Time Series Models*, Vandenhoeck and Ruprecht, 1978.

[32] C. W. J. Granger and T. Teräsvirta, *Modeling Nonlinear Economic Relationships*, Advanced texts in econometrics, Oxford Univeritary Press, 1993.

[33] D. Guégan, *Séries Chronologiques Non Linéaires à Temps Discret*, Statistique Mathématique et Probabilité, Economica, 1994.

[34] L. Györfi, W. Härdle, P. Sarda, and P. Vieu, *Nonparametric curve estimation from time series*, springer-verlag, 1989.

[35] V. Haggan and T. Okazi, *Modeling non linear vibration using amplitude dependant autoregressive time series model*, Biometrika, 768 (1981), pp. 189–196.

[36] W. Härdle, *Forecasting using kernels estimators*, in Rencontres Franco-Belges de Statistiques, 1995.

[37] C. Hausen, *Communication personnelle*, 1995.

[38] M. Heller and H. S. Thind, *Forecasting with cascade correlation: an application to potable water demand*, in Intelligent Engineering systems through artificiel neural networks, vol. 4, AMSE Press, 1994, pp. 1155–1160.

[39] C. Holmes. ICANN'95 proceedings, 1995.

[40] R. H. Jones, *Nonlinear autoregressive processes*, Proc. R. Soc A, 360 (1976), pp. 71–95.

[41] M. Jordan, *Attractor dynamics and parallelism in a connectionist sequential machine*, in Proceedings of the Eighth Annual Conference of the Cognitive Science Society, Amherst 1986, 1986, Lawrence Erlbaum, Hillsdale, pp. 531–546.

[42] E. J. Kostelich and D. P. Lathorp, *Time series prediction by using the method of analogues*, in Time series prediction: Forecasting the future and understanding the past, A. S. Weigend and N. A. Gershenfeld, eds., Addison Wesley, Reading, MA, 1993, pp. 283–295.

[43] A. Lapedes and R. Farber, *Nonlinear signal processing using neural networks: Prediction and system modelling*, Tech. Rep. LA–UR–87–2662, Los Alamos National Laboratory, Los Alamos, NM, 1987.

[44] Y. Le Cun, J. Denker, and S. Solla, *Optimal brain damage*, in Advances in Neural Information Processing Systems, D. Touretzky, ed., vol. 2, Denver 1989, 1990, Morgan Kaufmann, San Mateo, pp. 598–605.

[45] B. LeBaron, *Nonlinear forecast for the S & P stock market*, in Nonlinear Modeling and Forecasting, M. Casdagli and S. Eubank, eds., vol. XII of SFI Studies in the Sciences of Complexity, Addison-Wesley, 1992, pp. 381–394.

[46] P. A. W. Lewis, B. K. Ray, and J. G. Stevens, *Modeling time series by using multivariable regression splines (mars)*, in Time series prediction: forecasting the future and understanding the past, A. S. Weigend and N. A. Gershenfeld, eds., Addison Wesley, Reading, MA, 1993, pp. 297–322.

[47] L. Ljung, *Perspectives on the process of identification*, Tech. Rep. 1993-09-10, Linköping University, 1993.

[48] s. Lloyd and J.-J. Slotine, *Information theoretic tools for stable adaptation and learning*, Tech. Rep. MIT NSL-950201, MIT, 1995.

[49] D. J. C. MacKay, *Bayesian non-linear modeling for the energy prediction competition*, ASHRAE Transactions, (1993), pp. 1053–1062.

[50] S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler, *The forecasting accuracy of major time series methods*, Wiley, 1984.

[51] M. Mangeas, M. Cotrell, B. Girard, Y. Girard, and C. Muller, *Advantages of multilayered perceptron for modeling and forecasting time series: Application to the daily electrical consumption in france*, in Proceeding of Neuro Nîmes'93, EC2, 1993.

[52] S. Marcos, O. Macchi, C. Viognat, G. Dreyfus, L. Personaz, and P. Roussel-Ragot, *A unified framework for gradient algorithms used for filter adaptation and neural network training*, International Journal of Circuit Theory and Applications, (1991).

[53] R. R. Mohler, *Nonlinear time series and signal processing*, no. 106 in lecture notes in control and information science, springer-verlag, 1988.

[54] M. C. Mozer, *Neural net architectures for temporal sequence processings*, in Time series prediction: Forecasting the future and understanding the past, A. S. Weigend and N. A. Gershenfeld, eds., Addison Wesley, Reading, MA, 1993, pp. 243–264.

[55] C. Muller and M. Mangeas, *Neural networks and time series forecasting: a theoretical approach*, in International Conference on System, Man and Cybernetics. System Engineering in the service of humains, vol. 2, IEEE, 1993, pp. 590–594.

[56] K. S. Narendra and K. Parthasarathy, *Identification and control of dynamical systems using neyral networks*, IEEE Transaction in Neural Networks, 1 (1990), pp. 4–27.

[57] O. Nerrand, P. Roussel-Ragot, D. Urbani, L. Personnaz, and G. Dreyfus, *Training recurrent neural networks and nonlinear adaptive filtering: why and how ? an illustation in dynamical process modeling*, IEEE Transactions on Neural Networks, 5 (1994), pp. 178–182.

[58] D. A. Nix and A. S. Weigend, *Learning local error bars for nonlinear regression*, in NIPS-7, 1995, pp. 489–496.

[59] A. R. Pagan and Y. S. Hong, *Non-parametric estimation in the risk premium*, in Semiparametric and Nonparametric methods in Econometrics and Statistics, W. Barnett, J. Powell, and G. Tauchen, eds., Cambridge University Press, 1990.

[60] M. B. Priestley, *Nonlinear and non-stationnary time series*, accademic press, 1988.

[61] U. Ramacher, *Hamiltonian dynamics of neural networks*, Neural Networks, 6 (1993), pp. 547–557.

[62] J. B. Ramsey, *If nonlinear models cannot predict, what use are they*, preprint, Dept. of Economics, New York Univ., New York, 1995.

[63] J. B. Ramsey and Z. Zhang, *The application of wave form dictionaries to stock market data*, report, C. V. Starr Center for Applied Economics, New York University, New York, 1994.

[64] D. Rumelhart, G. Hinton, and R. Williams, *Learning internal representations by error propagation*, in Parallel Distributed Processing, D. Rumelhart and J. McClelland, eds., vol. 1, MIT Press, Cambridge, 1986, ch. 8, pp. 318–362. Reprinted in [1].

[65] I. Salun and P. Sequier, *Stock selection using neural networks: the french stockmarket market*, in Avignon'92, vol. 4, EC2, May 1993, pp. 105–113.

[66] G. Saporta, *Probabilités, Analyse des Données et Statistiques*, Technip, 1990.

[67] T. Sauer, *Time series prediction by using delay coordinate embedding*, in Time series prediction: Forecasting the future and understanding the past, A. S. Weigend and N. A. Gershenfeld, eds., Addison Wesley, Reading, MA, 1993, pp. 175–193.

[68] R. Sharda and R. Patil, *Neural networks as forecasting experts: an empirical test*, in IJCNN, vol. 2, IEEE press, 1990, pp. 491–494.

[69] L. A. Shepp, D. Slepian, and A. D. Wyner, *On prediction of moving average prosecces*, Bell System Technical Journal, 59 (1980), pp. 367–415.

[70] J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Delion, P.-Y. Glorenec, H. Hjalmarsson, and A. Juditski, *Nonlinear black box modeling in system identification: a unified overview*, Tech. Rep. LiTH-ISY-R-1742, Linköping University, 1995.

[71] E. Sontag, *Neural networks for control*, in Essays on Control: Perspectives in the Theory and its Applications, H. L. Trentelman and J. C. Willems, eds., Birkhauser, 1993, pp. 339–380.

[72] H. Tong, *Non-linear Time Series. A dynamical System Approach*, vol. 6 of Oxford Statistical sciences series, Oxford Science Publications, Oxford, UK, 1990.

[73] A. Varfis and C. Versino, *Univariate economic time series forecasting by connectionist methods*, in Proceedings INNC'90, Kluwer Academic Publishers, 1990, pp. 342–345.

[74] G. Vaucher, *Study of a self-learning artificial neuron model*, in Proceedings of the ICANN'93, S. Gielan and B. Kappen, eds., Springer-Varlag, 1993, p. 204.

[75] A. Waibel, *Modular construction of time-delay neural networks for speech recognition*, Neural Computation, 1 (1989), pp. 39–46.

[76] E. A. Wan, *Finite Impulse Response Neural Networks with Applications in Time Series*, PhD thesis, Stanford University, 1993.

[77] A. Weigend, B. A. Huberman, and D. E. Rummelhart, *Predicting sunspots and exchange rates with connectionist networks*, in Nonlinear Modeling and Forecasting, M. Casdagli and S. Eubank, eds., vol. XII of SFI Studies in the Sciences of Complexity, Addison-Wesley, 1992, pp. 397–434.

[78] A. Weigend, D. Rumelhart, and B. Huberman, *Generalization by weight-elimination with application to forecasting*, in Neural Information Processing 3, R. Lippman, J. Moody, and D. Touretzky, eds., Morgan Kaufmann, San Mateo, CA, 1991, pp. 875–882.

[79] A. S. Weigend and N. A. Gershenfeld, eds., *Time Series Prediction: Forecasting the Future and Understanding the Past*, vol. XV of SFI Studies in the Sciences of Complexity, Addison-Wesley, 1993.

[80] H. White, *Economic prediction using neural networks: the case of IBM daily stock returns*, in International conference on neural networks, vol. II, IEEE, IEEE, 1988, pp. 451–458.

[81] ——, *An additional hidden unit test for neglected nonlinearity in multilayered feedforward networks*, in International Joint Conference on Neural Networks, vol. II, SOS printing, 1989, pp. 451–455.

[82] ——, *Connectionist nonparametric regression: Multilayered feedforward networks can learn arbitrary mapping*, Neural Networks, 3 (1990), pp. 535–550. Reprint as chapter 10 in [83].

[83] ——, *Artificial Neural Networks*, Blackwell, 1992.

[84] C. G. Winsor and A. H. Harker, *Multi-variate financial index prediction - a neural network study*, in Proceedings INNC'90, Kluwer Academic Publishers, 1990, pp. I–357–360.