Wind Power Prediction Using ARX Models and Neural Networks

Henrik Aalborg Nielsen, Ph.D. student
Henrik Madsen, Associate Professor

Institute of Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Abstract

Prediction of the wind power production at a wind farm placed near the west coast of Denmark is considered. The wind farm consists of 27 wind mills each with a power capacity of $225kW$.

Based on previous work regarding autoregressive models with external signals (ARX models), models based on feed-forward neural networks with one hidden layer are formulated. The size of the network is determined by the Bayes Information Criterion. The prediction performance of the selected networks are compared with the performance of the ARX models. Furthermore the naive predictor has been used as a reference of prediction performance. The criterion used for evaluating the prediction performance is the Root Mean Square of the prediction errors.

For most horizons three to four hidden units are found optimal with respect to the Bayes Information Criterion. Comparing the optimal neural network predictors with the ARX-based and naive predictors it is concluded that the neural network type investigated is inferior in prediction performance to the other prediction procedures investigated. Finally neural networks with only one hidden unit has been compared with the other prediction procedures. Also these networks prove to be inferior.

Key Words
Wind power, prediction, neural networks.

1 Introduction

In Denmark wind energy is becoming of increasing importance and hence it is important to be able to perform short term predictions of the wind power production. Adaptive prediction procedures based on autoregressive models with external signals (ARX models) have been developed and implemented for on-line wind power prediction in the western part of Denmark, see [1]. In this paper predictors based on neural networks are compared with ARX-based predictors. Furthermore the naive predictor, which corresponds to predicting the future value as the most resent observed value, has been used as a reference of prediction performance. The data used is half-hourly averages of wind speed and wind power production. Prediction horizons from 30 minutes to 3 hours are considered.

The paper contains a brief description of the ARX models. These models use the wind speed and a diurnal profile (representing a time-varying mean) as inputs. The parameters of the models are estimated using the adaptive least squares method with exponential forgetting [2]. The estimation method is modified in [1] to handle multi-step predictions.

A feed-forward neural network with one hidden layer has been used. This kind of network is described in [3] and a software package for S-Plus is available, see [4]. The networks considered all use the same variables as the ARX models. In this context the potential advantage of neural networks over ARX models is a more adequate description of any non-linear relationships between the variables. The disadvantages are a larger number of parameters and the non-adaptive estimation.

2 Predictors based on ARX models

In [1] a careful investigation of the problem of wind power prediction for the ELSAM (power distributor for the western part Denmark) area are described. Based on this investigation the following models has been implemented and used for k-step wind power predictions:

$$\begin{aligned}\sqrt{p_{t+k}} &= \mu^k + a_1^k\sqrt{p_t} + b_1^k\sqrt{w_t} + b_2^k w_t + \\ &\quad c_1^k \sin\tfrac{2\pi h_{t+k}}{24} + c_2^k \cos\tfrac{2\pi h_{t+k}}{24} + e_{t+k}^k,\end{aligned} \tag{1}$$

where μ^k, a_1^k, $b_\cdot^k$, and $c_\cdot^k$ are constants. The symbols p_t and w_t represent average wind power production and wind speed in the interval $[t-1, t[$. h_t is the 24-hour clock at time t. $\{e_t^k\}$ is a sequence of independent identically distributed

random variables with zero mean and variance σ_k^2. One time step corresponds to 30 minutes.

The parameters of the models (1) are estimated adaptively using recursive least squares with exponential forgetting, see [2]. The algorithm has, however, been modified in order to handle multi-step predictions. This modification consists of updating only the most resent parameter estimate, say $\hat{\theta}_{t-1}$. In order to make this feasible a pseudo prediction of $\sqrt{p_t}$ is used in the update of parameters; this prediction is constructed from p_{t-k}, w_{t-k}, h_t, and $\hat{\theta}_{t-1}$. See [1] for further details. A forgetting factor of 0.999, as suggested in [1], is used in this paper.

3 Neural Networks

3.1 Type of Neural Network

A feed-forward neural network with one hidden layer and without connections directly from input to output is used, see e.g. the documentation on the software [4]. Suppose that observations (indexed by i) of the independent variables (indexed by j) x_{ij} and the dependent variable y_i are present. The dependence of y on x can then be modelled by a neural network of the above type as:

$$y_i = \phi_o\left(\alpha_o + \sum_{h=1}^{n_h} w_{ho}\,\phi_h\left(\alpha_h + \sum_{j=1}^{n_j} w_{jh}x_{ij}\right)\right) + e_i, \tag{2}$$

where $w_{\cdot o}$ are the weights on the connections from the hidden layer to the output layer, $w_{\cdot h}$ are the weights on the connections from the input layer to unit h in the hidden layer, α_o is the bias on the output unit, and α_h is the bias on the hidden units. n_h and n_j are the No. of hidden units and inputs, respectively. It is seen that the weights and the biases are just parameters of the model. Considering (2) as a statistical model one would assume that the residuals (e_i) are independent identical distributed.

The functions $\phi_h(\cdot)$ and $\phi_o(\cdot)$ are predefined functions associated with the units in the hidden and output layer, respectively. Most frequently these functions are sigmoid (also called logistic), i.e. the output of the network is restricted to the interval $]0, 1[$. This is, however, not desirable in this application since the future output of the network is then restricted to the range of observations in the data set used for estimating the parameters. For this reason the output unit has been chosen to be linear, i.e. $\phi_o(z) = z$.

3.2 Estimation of Parameters

The parameters (weights and biases) of the model (2) are estimated by non-linear least-squares. The initial values of the estimates are rather important since the minimization problem may contain local minima due to the fact that the model is non-linear in the parameters. For this reason each model should be estimated several times using different initial parameter estimates.

Since it is rather difficult to suggest appropriate values it seems reasonable to select these values at random. In this case the data has been scaled to the interval $[0, 1]$ (see Section 4) and according to the documentation on the software used (see Section 3.4 and [4]) it should be sufficient to sample from the $U(-1, 1)$ distribution. In spite of this it was decided to sample from the $U(-5, 5)$ distribution in order to cover a wider interval of initial parameter estimates. The parameters of each network were estimated 20 times with initial parameter estimates chosen at random.

3.3 Selection of Network Size

To use a neural network it remains to decide upon the independent variables to include in the model and on the No. of hidden units. This may be done by using some kind of information criteria. Here the Bayes Information Criterion (BIC) has been used, see [5]. With L^*, n_p, and N being the value of the likelihood in the optimum, the No. of parameters, and the No. of observations used in the estimation, respectively, the criteria corresponds to chose the model so that

$$\log L^* - \frac{n_p}{2}\log N, \tag{3}$$

is maximized. For a large class of linear time series models and other linear models with the residuals being normally distributed the criteria is equivalent to minimizing

$$BIC = N\log\left(\frac{1}{N}\sum_{i=1}^{N}\hat{e}_i^2\right) + n_p\log N, \tag{4}$$

where $\hat{e}_i$ is the prediction errors from the model (2), with the unknown parameters replaced by the estimates. Note that $\hat{e}_i$ must be based on maximum likelihood estimates.

In this case the procedure used for estimation of the parameters (see Section 3.2) is not a maximum likelihood procedure. Hence the above criteria must be considered as an approximation.

It was decided to use the criteria (4) to select the appropriate No. of hidden units only. The independent variables have been considered fixed, see Section 4.

3.4 Software

The neural network software used is written by Professor of Applied Statistics, B.D. Ripley, University of Oxford. The software can be obtained from StatLib by anonymous ftp from `lib.stat.cmu.edu`. The software is written for S-Plus and briefly described in [4].

The adaptive predictions have been calculated using the software Off-line Wind Power Prediction Tool, version 1.0, see [6].

4 Variables in the Models

Based on the investigation described in [1] (see also Section 2) it was decided to use the following indepen-

dent variables: The present wind power production (p_t) scaled to approximately $[0, 1]$, the present wind speed (w_t) scaled to approximately $[0, 1]$, $\frac{1}{2}\sin(2\pi h_{t+k}/24) + \frac{1}{2}$, and $\frac{1}{2}\cos(2\pi h_{t+k}/24) + \frac{1}{2}$. The wind power production k step ahead (p_{t+k}) scaled to approximately $[0, 1]$ was used as the dependent variable.

5 Validation

The models have been validated using a different data set than the one on which the selection of the No. of hidden units and the estimation of parameters is based. This data set is called the validation set.

The neural network model selected for each prediction horizon k has been compared with the naive k-step predictor and with the adaptive predictor described in Section 2.

The estimation and validation set are just two parts of one time series. Therefore it was possible to allow the adaptive predictions to settle before the validation was initiated. This method was chosen since this corresponds to the real application.

Based on the validation set the k-step residuals (or prediction errors) were calculated on the original scale and based on these the Root Mean Square (RMS) were calculated. For the residuals $(r_1, r_2, \ldots, r_N)$ the RMS of the residuals is defined as $\sqrt{\frac{1}{N}\sum_i r_i^2}$.

6 Data

The data used in this investigation has been collected in the Vedersø Kær wind farm in the ELSAM area during the period July 2, 1993, 5.30 p.m. until October 11, 1993, 7 a.m. The original sampling time was 5 minutes. Based on these values half-hourly averages were calculated. The data until September 6 at 7 a.m. (3148 averages) is used for estimation purposes whereas the remaining data (1680 averages) has been used for validation.

The maximum wind speed observed is 15.9 m/s and 75% of the time it did not exceed 8.5 m/s. The corresponding values for the wind power production are 5789 and 1783 kW.

7 Results

7.1 Estimation

In Figure 1 plots of the resulting values of BIC are shown. It is seen that for prediction horizons $k = 1, 2, 3$ the lowest value of BIC is observed for a network with three hidden units. For $k = 4, 6$ a network with four hidden units results in the lowest observed BIC, and for $k = 5$ a network with five hidden units results in a marginally lower BIC, than for $k = 4$.

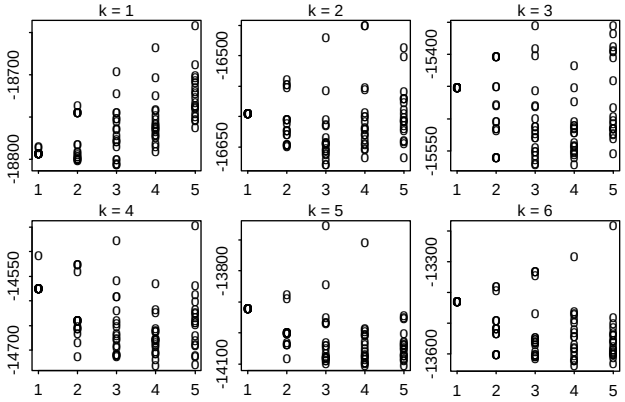

Figure 1: Bayes Information Criterion versus No. of hidden units. A few extreme (high) values are not shown.

7.2 Validation

For all prediction horizons the neural network with the lowest BIC was validated as described in Section 5. The results are shown in Table 1.

k	RMS_{nn} (kW)	RMS_{naive} (kW)	RMS_{adap} (kW)
1	297.6	261.7	262.8
2	448.6	375.7	377.0
3	534.9	440.9	442.0
4	621.2	507.6	507.5
5	678.4	569.2	565.1
6	705.8	623.1	614.1

Table 1: Validation set; RMS of prediction errors of the best neural network, naive, and adaptive predictor.

It is seen that the neural network models investigated are all inferior to both the naive and the adaptive predictors. It is noted that the naive predictor is slightly better than the adaptive for $k = 1, 2, 3$. For $k = 5, 4, 6$ the adaptive predictor is better than the naive. However working out the ratios between RMS_{adap} and RMS_{naive} will reveal that the difference is minor.

7.3 Networks with One Hidden Unit

From the validation of the network of optimal size it is seen that the naive predictor performs well compared to the other methods investigated. It is therefore peculiar that the selection procedure does not lead to a selection of the most simple network; a network with one hidden unit only. It was therefore decided to compare this kind of network with the network selected according to BIC. Results are indexed by opt and 1 for the optimal and the simple network, respectively. The validation results are displayed in Table 2.

From the table it is seen that for $k = 1, 2, 3$ the neural network with one hidden unit actually performs better than the network selected according to BIC. However comparing Table 2 with Table 1 it is seen that the neural network with one hidden unit is inferior to the naive and the adaptive predictor.

k	RMS_{opt} (kW)	RMS_1 (kW)
1	297.6	268.8
2	448.6	411.6
3	534.9	523.3
4	621.2	622.8
5	678.4	699.9
6	705.8	758.3

Table 2: Validation set; comparison of optimal network with a network with one hidden unit.

8 Conclusion

The type of neural networks investigated is inferior in prediction performance to both the adaptive predictor and the simple naive predictor for the prediction horizons investigated (1/2 to 3 hours).

For the prediction horizons investigated the naive predictor performs better than the adaptive predictor for the short prediction horizons (up to $1\frac{1}{2}$ hour). However, the difference between the two predictors is minor. For horizons larger than 2 hours the adaptive predictor is better than the naive.

9 Discussion

Performance of predictors: Apart from the non-linear response of the hidden units, a neural network predictor includes the naive predictor. The reason why the neural network predictor performs considerably worse than the naive and the adaptive predictors is probably that: (i) The estimation of the parameters in the neural network is not adaptive, (ii) the No. of parameters which must be estimated in the neural network is large (seven or larger), and/or (iii) the non-linear response of the hidden units is inappropriate for wind power predictions.

For the low horizons investigated the naive predictor performs slightly better than the adaptive predictor based on the autoregressive model. For the larger horizons the adaptive predictor is slightly superior. For very large horizons a simple profile (probably containing harmonics corresponding to daily and yearly periods) will probably be the best predictor. The adaptive predictor processes the characteristic of being able to interpolate between these extremes. For this reason the adaptive predictor is attractive.

Maximum size of neural network investigated: The largest No. of hidden units investigated is five. In most cases the optimal network size is found to be less than five. Since the sum of the squared prediction errors for the estimation data set is a non-increasing function of the No. of hidden units BIC will have one minima only. Therefore the maximum size of the networks investigated is sufficient.

Estimation of parameters in neural networks: For one particular neural network model the estimation with random initial values of the parameters results in different values of the mean square of the residuals. This is seen from the random scatter of the values of Bayes Information Criterion (BIC). This clearly reveals that the surface on which the minimization is performed in order to obtain the parameter estimates contains local minima. If this was not true the final estimates and hence BIC should be independent of the initial values of the estimates.

References

[1] ELSAM. *Final Report, Wind Power Prediction Tool in Central Dispatch Centres, JOU2-CT92-0083*, (ELSAM, Skærbæk, Denmark, 1995).

[2] L. Ljung. *System Identification, Theory for the User*, (Prentice-Hall, Englewood Cliffs, New Jersey, 1987).

[3] B.D. Ripley. "Statistical aspects of neural networks". In *Chaos and Networks - Statistical and Probabilistic Aspects*, (Chapman and Hall, London, 1993, pp. 40-123).

[4] B.D. Ripley. Software for neural networks. Statlib Index (`lib.stat.cmu.edu`), S library, `nnet` package, 1994. Updated 24 January 1993, 19 September 1993, 13 February 1994.

[5] G. Schwarz. "Estimating the dimension of a model", *The Annals of Statistics*, 6(2), 1978, 461–464.

[6] H.Aa. Nielsen and H. Madsen. *Off-line Wind Power Prediction Tool - Users Manual*. (Institute of Mathematical Modelling, Technical University of Denmark, Lyngby, Denmark, 1995).