STATISTICAL METHODS FOR WIND SPEED MODELING AND FORECASTING

Petre-Cristian RĂZUŞI[1], Gianfranco CHICCO[2]

In power systems with wind power integrated, in order to take full advantage of the wind energy is necessary to understand and control this energy. Due to the stochastic character of wind, this can only be achieved by predicting its speed. One of the methods for this is the usage of Box-Jenkins models.

In this paper a parallel between the modeling and prediction capacities of three of these models (AR, ARMA and ARIMA) is drawn. Also, the influence of the sampling rate of the measured wind speed data on the wind speed predictions is presented.

Keywords: AR, wind speed forecasting, ARMA

1. Introduction

The whole world is confronting itself with an energy and climatic crisis. Between 2005 and 2030 it is expected that the world's primary energy demand to increase by 55%. Fossil fuels will account for 84% of this increase, thus remaining the major source of primary energy. The CO_2 emissions will also increase by 57%, of which the power sector will be responsible for 45% [1].

In European Union, the percentage of imported energy will rise from 50% in 2006 to around 70% in the next 30 years. A part of this energy will come from regions threatened by insecurity. Also, the prices of fossil fuels are increasing and so are the ones for electricity [2]. Moreover, a large part of Europe's electricity generating capacity has reached its projected life span and needs to be replaced.

All of these conditions lead to an increase in the usage of renewable sources of energy, which are clean and virtually inexhaustible. Among these, wind energy has a great potential that is reflected in its fast development. Only in 2008 more than 27,000 MW in wind energy capacity were installed globally, making the global cumulative installed capacity reaching the staggering value of 120,798 MW [3].

The main goal, when installing renewable energy sources in a power system, is to replace the conventional sources used to generate electricity. But, in the case of wind energy, in order to take full advantage of it, the generated power of large grid-connected wind turbines and wind farms must be known. However,

[1] Ph.D. student, Power Engineering Faculty, University "Politehnica" of Bucharest, Romania
[2] Prof., Department of Electrical Engineering, Politecnico di Torino, Torino, Italy

as we all know, this power depends directly on the wind speed, which gives it a random character and makes it almost impossible to control. Developing methods to forecast the behavior of wind can solve this problem.

This paper analyses the prediction capacity of statistical methods with focus on three models – autoregressive (AR), autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) – as well as the effect of the sampling rate of the measured data on the modeled wind speed and on predictions. The entire study is based on the models developed by statisticians George Box and Gwilym Jenkins and on time series analysis.

2. Time series analysis

A time series is a sequence of values of a variable, typically measured at successive times and at equal time intervals. The analysis of a time series includes methods to understand the series in order to identify the source of the data points, and to forecast the next values.

The basic idea of time series analysis is the fact that data points taken over time have an internal structure and a random noise, the latter making the existing pattern difficult to identify. These patterns can be divided into two basic types of components: trend and seasonality. The first represents a general systematic linear or nonlinear component that changes over time and does not repeat. The second one has a similar nature with the former but it repeats itself in systematic intervals over time. These two types of time series components coexist in real time data.

Taking into account that the data is corrupted with noise, smoothing is the first step in the process of trend identification. This is done by using some sort of local averaging of data so that the random variation components of individual observations to cancel each other out. The most common smoothing technique is moving average which replaces each element of the series by the simple or weighted average of n surrounding elements, where n is the width of the smoothing window.

Seasonality, the other type of components of the time series pattern, is formally defined as correlation dependency of order k between each i^{th} element of the series and the $(i\text{-}k)^{th}$ element, where k is called the lag. Once the noise is reduced through a smoothing process like the one previously described, seasonality can be visually identified in the series as a pattern that repeats every k elements. Seasonal patterns of time series can be detected and examined using the graphical and/or numerical displays of the autocorrelation function.

At the end of the analysis a model, which can somewhat replicate the way the time series varies and which can be used for predicting the future behavior of that variable, will be yielded. The fitting of time series models can be a difficult task. However, there are many methods for this purpose among which the Box-Jenkins ARIMA models is one of the most general.

3. Box-Jenkins models

ARIMA models, which are also called Box-Jenkins models, are one of the most powerful methods used for time series analysis and forecasting. The idea behind these models is that observations close together in time will be more closely related than observations further apart. Hence, an element of the series can be described as a linear combination of the past elements, with the closest ones having greater weight.

These models have the great advantage that they can be used for both stationary and non-stationary series. This general character comes from the fact that different models can be obtained from ARIMA, models which can be used for series with different characteristics. To be more specific, by modifying the orders of this model, ARMA, AR, and MA (moving average) models can be attained.

Let us consider a series z containing observations of a variable made at discrete and equal time intervals, μ being its mean value. By z_t, z_{t-1}, z_{t-2},… will be denoted the values of the observations for the t, t-1, t-2,… times, and by $\tilde{z}_t$, $\tilde{z}_{t-1}$, $\tilde{z}_{t-2}$,… their deviations from the mean value. For this series, the equation of the autoregressive model, AR(p), is [4]

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + ... + \phi_p \tilde{z}_{t-p} + \varepsilon_t , \qquad (1)$$

where: p is the order of the model; ϕ_1, ϕ_2,…, ϕ_p are the parameters of the model; and ε_t is an error term which is actually white noise.

In order to improve the description of the connection between the elements of the series, by joining together the autoregressive and the moving average models a more complex model was created. The new model, written as ARMA(p,q), is described by the following equation [4]:

$$\begin{aligned} \tilde{z}_t &= \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + ... + \phi_p \tilde{z}_{t-p} + \\ &+ \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q} \end{aligned} \qquad (2)$$

where: p is the order of the AR part; ϕ_1, ϕ_2,…, ϕ_p are the parameters of the AR part; q is the order of the MA part; θ_1, θ_2,…, θ_q are the parameters of the MA part; and ε_t, ε_{t-1}, ε_{t-2}…, ε_{t-q} are error terms or white noise for the t, t-1,…,t-q terms of the time series.

The two models presented above can be used only for stationary time series. In order for them to be used on non-stationary series, the methodology developed by Box and Jenkins states that the series must first be differenced for d times. Hence, the equation for the ARIMA(p,d,q) model, which is a combination between the autoregressive moving average model and a method to make series stationary, are [4]:

$$\begin{aligned} w_t &= \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} \\ &+ \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} \end{aligned}, \tag{3}$$

where:

$$\begin{aligned} w_t &= \nabla^d z_t; \\ \nabla z_t &= z_t - z_{t-1}. \end{aligned} \tag{4}$$

4. Case study applications

In this paper two issues were studied: the assessment of AR, ARMA and ARIMA models capacity of modeling and predicting wind speeds, and the sampling rate effect of the measured wind data on this capacity. The issues were studied in two hypotheses:

- the wind turbine has a fixed position – in this case, wind speed must be projected on the direction of the wind turbine;
- the wind turbine has the possibility to instantly and freely rotate around its vertical axis – in this purely ideal case, wind direction has no influence in the study, so it is neglected.

For the undertaken studies two sets of data were used. The first consist of 9,356 measured values for wind speed and direction that were taken from two by two minutes from 07:50 on 6th of April 2009 to 14:50 on 21st of April 2009 at the Chek Lap Kok station by the Hong Kong Observatory.

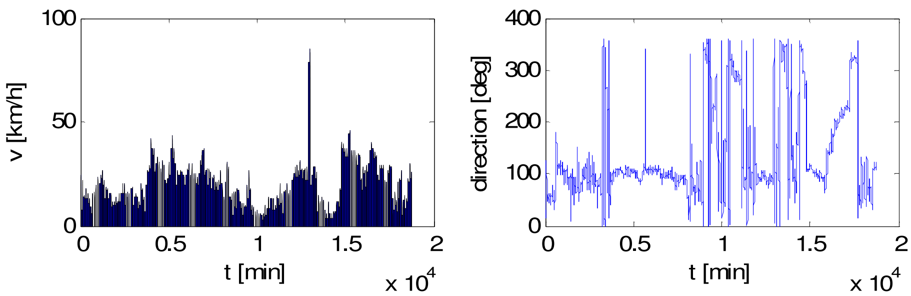

Fig. 1. Hong Kong data set

The second set of data that was used consists of 71,640 measured values for wind speed and direction taken at the Valkenburg station, Netherlands. These hourly average measurements, taken at a 10 m height above ground, cover the period from 2001 until 2009 [5].

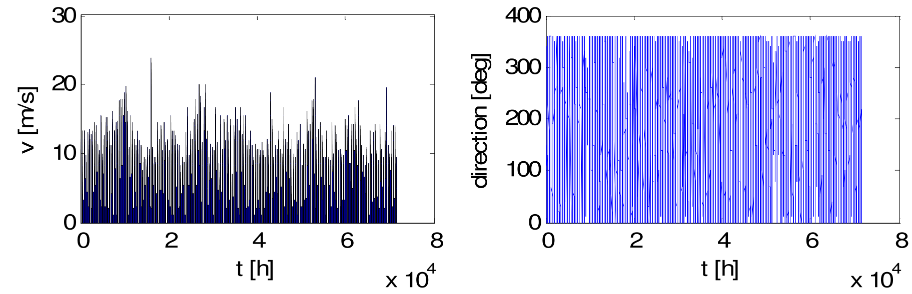

Fig. 2. The Netherlands data set

The two sets of data were divided into two halves. The first half was used to compute the orders and parameters of the models, while the second was used to validate these models.

Taking into account the Box-Jenkins methodology, the two data sets were analysed to see whether they were stationary or not. This was done using the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots [6]. For the Chek Lap Kok data set, ACF and PACF were calculated for each of the sampling rates from 2 minutes up to 60 minutes. These plots showed that both series, regardless the sampling rate, are not stationary.

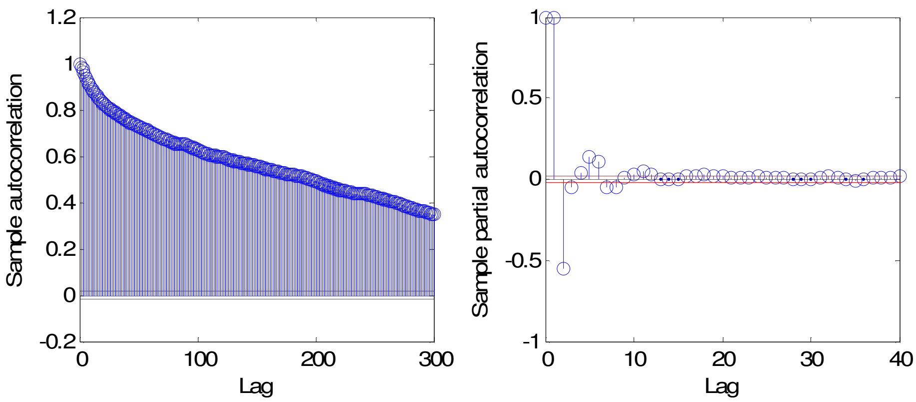

Fig. 3. ACF/PACF plots for the unprojected Hong Kong wind speed; sampling rate – 2 minutes

After differencing the series once, the ACF and PACF were plotted again. The results showed that the series became stationary, and that the most appropriate model to be used is ARIMA(p,1,q). These plots also showed that the orders p and q should not have values greater than four.

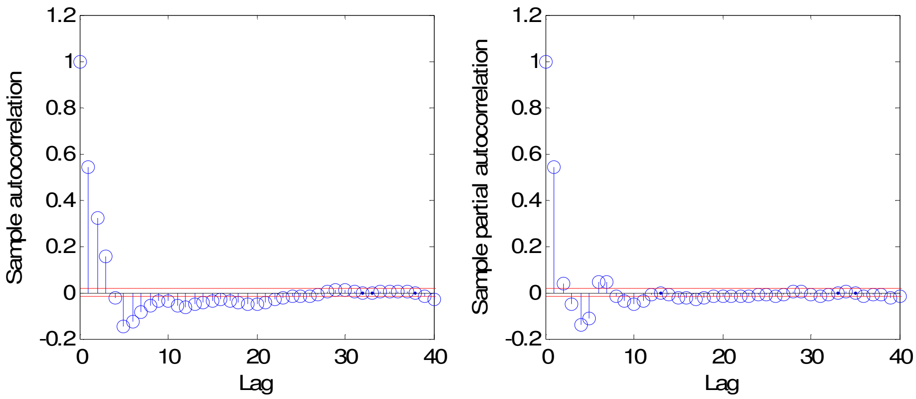

Fig. 4. ACF/PACF plots for the differenced unprojected Hong Kong data set; sampling rate – 2 minutes

The models orders were chosen to maximize the curve fitting between the modeled and real data, but also to keep the computation time at a reasonable value (an increase in the order is leading to an increase in the time necessary to estimate the parameters). Using these orders, the variation of curve fit of the modeled data with respect to the sampling rate was plotted. It can be seen that, as the sampling rate increases, the accuracy of the modeled data drops.

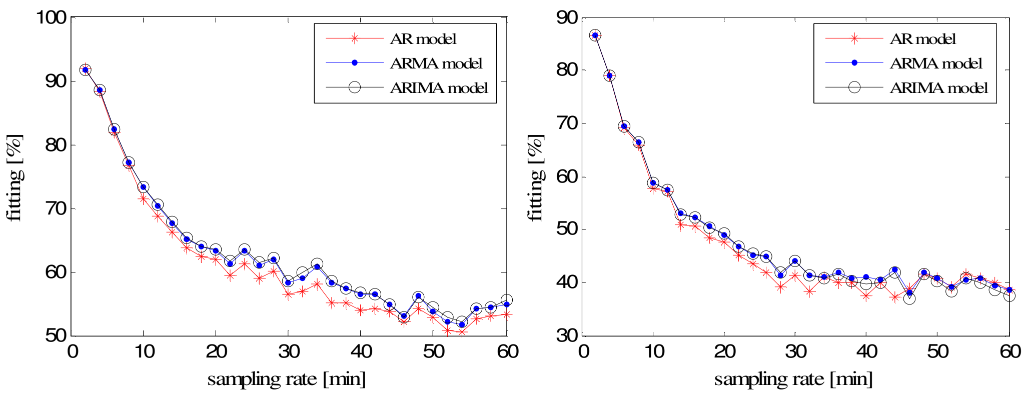

Fig. 5. Variation of curve fit with respect to the sampling rate for the unprojected (left) and projected (right) wind speed from the Hong Kong data set

From Figure 5 it can be seen that at small sampling rates all three models provide almost the same curve fit, whereas for higher values of the sampling rate ARMA and ARIMA models provide better curve fits than AR model. The same results were found at the validation of the models when the accuracy of the predicted curve was calculated. The results are presented in Table 1.

Table 1

Curve fit [%] for the predicted wind speed from the Hong Kong data set

Type of model	sampling rate [min]	2	10	20	30	40	50	60
AR	unprojected wind speed	92,158	72,259	64,662	59,928	57,865	50,582	53,347
	projected wind speed	91,678	66,927	54,507	47,123	45,666	38,994	45,767
ARMA	unprojected wind speed	91,905	73,118	65,282	60,309	58,532	50,559	54,240
	projected wind speed	91,906	69,424	56,689	49,226	47,785	38,900	47,652
ARIMA	unprojected wind speed	86,680	61,260	54,132	43,069	44,029	27,150	39,029
	projected wind speed	89,163	62,246	41,490	27,205	25,567	7,177	30,450

This drop in the accuracy of the predictions and in the fit of the modeled curve can be explained in two ways. First, when talking about forecasting and Box-Jenkins models, the sampling rate is equivalent to the prediction horizon, so it is natural for the accuracy to drop as we try to predict a value further in time. The second explanation comes from the way Box-Jenkins models are made. A predicted value is a linear combination of previous values; hence no sudden variation in the wind speed can be predicted. As the sampling rate increases, so are the variations of the wind speed, thus being harder to predict.

Regarding the capacity of AR, ARMA and ARIMA of modeling the wind speed, taking into account the fact that wind speed has a stochastic character and that it is best described in terms of statistical properties, the mean value and standard deviation of the real and modeled wind speeds were calculated. The results showed that these values are very close, so the main characteristics of wind speed are kept by the modeled speed. For exemplification, in Table 2 are presented the values for the Netherlands data set.

Table 2

Mean and standard deviation values for the measured and modeled wind speed

Unprojected wind speed				Projected wind speed			
real data	AR modeled	ARMA modeled	ARIMA modeled	real data	AR modeled	ARMA modeled	ARIMA modeled
Mean value in m/s for the real and modeled wind speed							
5,070	5,002	5,044	5,070	-0,524	-0,496	-0,515	-0,524
Standard deviation value in m/s for the real and modeled wind speed							
2,885	2,846	2,770	3,185	3,846	3,646	3,690	4,147

5. Conclusions

The undertaken studies showed that AR, ARMA and ARIMA models capacity of modeling and predicting wind speeds depend very much on the available recorded wind data.

It has been shown that although ARIMA is the best model to be used for predictions and modeling of the wind speed, AR and ARMA models can also be used because they can provide almost the same accuracy but with less computational time.

The modeled wind speed maintains the same major statistical characteristics so it can be used in simulations where wind data is needed but there are not enough measurements available and even for replacing missing or corrupted real wind measured data.

The results showed that the Box-Jenkins models could also be used for predictions with relatively high success, but only for short-term forecasts because high prediction horizons lead to low accuracy. Furthermore, by filtering these predicted wind speed values through the power curve of a wind turbine, wind power forecasts can be made thus providing a mean to "control" this turbine.

R E F E R E N C E S

[1]. ***, World Energy Outlook 2007, International Energy Agency.

[2]. ***, "A European Strategy for Sustainable, Competitive and Secure Energy", Green Paper, Commission of the European Communities, Brussels, March, 2006.

[3]. ***, Global Wind 2008 Report, Global Wind Energy Council, Brussels, 2008.

[4]. *G. Box, G. Jenkins,* Time series analysis – forecasting and control, 3rd edition, Holden-Day, Oakland, 1976.

[5]. Royal Netherlands Meteorological Institute (KNMI), KNMI HYDRA MetaDataConnection, available at http://www.knmi.nl/klimatologie/onderzoeksgegevens/potentiele_wind.

[6]. *C. Chatfield*, The analysis of time series: an introduction, 6th edition, Chapman & Hall/CRC, New York, 2004.