ARIMA Modeling Applied to Wind Speed Records from SONDA Database

Danilo Zucolli Figueiredo
Departamento de Engenharia de Telecomunicações e Controle
Escola Politécnica da Universidade de São Paulo
CEP: 05508-010 - São Paulo/SP, Brasil
d.z.figueiredo@ieee.org

Oswaldo Luiz do Valle Costa
Departamento de Engenharia de Telecomunicações e Controle
Escola Politécnica da Universidade de São Paulo
CEP: 05508-010 - São Paulo/SP, Brasil
oswaldo@lac.usp.br

RESUMO

Neste artigo, um procedimento de identificação é apresentado para a obtenção de um modelo ARIMA a partir de dados de velocidade do vento da base de dados SONDA (Sistema de Organização Nacional de Dados Ambientais). Este procedimento é ilustrado por um exemplo numérico considerando registros de velocidade de vento de uma estação anemométrica localizada em Triunfo, Pernambuco. Espera-se que o presente artigo possa motivar a utilização dos dados do projeto SONDA em outros trabalhos que tratem da previsão da velocidade do vento a curto prazo.

PALAVRAS CHAVE. Identificação de sistemas, Séries temporais, Velocidade do vento.

EN - PO na Área de Energia

ABSTRACT

In this paper, an identification procedure to derive an ARIMA model from wind speed data records of the SONDA (*Sistema de Organização Nacional de Dados Ambientais* - National Organization System of Environment Data) database is presented. This procedure is illustrated by a numerical example considering wind speed data records of an anemometric station located at Triunfo, Pernambuco. It is expected that the present article may motivate the use of SONDA project data in other works in the field of short-time wind speed forecast.

KEYWORDS. System identification. Time series. Wind speed.

EN - OR in Energy

1. Introduction

It is a well known fact that over the last decade there has been a substantial growth of renewable energy participation in the energy matrix of many countries. An example of this is the wind power generation in Brazil that from 2014 to 2015 has expanded 77% of its capacity (going from 12.210GWh to 21.626GWh) [Ministério de Minas e Energia, 2016]. As Brazil has several locations with valuable wind energy potential (Brazilian Northeast and South are the regions that offer the highest wind energy resources) [Pereira and Lima, 2008], it is expected that a great growth of wind power penetration may continue to occur in the country.

In spite of the benefits of wind energy, the volatility of wind presents new challenges when compared to conventionally (dispatchable) power sources. Due to its stochastic nature, wind generation is considered a non-dispatchable source of electricity and, in this sense, models and forecasts of wind speed and wind power bring a valuable information for system operators and for the players in the electricity market.

The modeling of wind speed is motivated by many reasons, but at least two of them should be highlighted. The first is that, in many cases, when one wishes to evaluate the potential contribution or the impact of a wind power site to the electrical grid, there are not enough data records. In this case, simulations are employed in order to build time series with a longer time span or to synthesize new wind speed sample paths. The other reason is that wind speed models can be used to forecast the wind power output of a wind power generator or wind farm. The obtained forecasts can then be used, for instance, to optimize the value of the produced electricity in the market, to optimize the dispatch of conventional power plants or to schedule some maintenance tasks [Brown et al., 1984; Costa et al., 2008; Giebel et al., 2011].

This paper is concerned with short-time modeling of wind speed. The time scales adopted in short-time prediction are in the order of some days (for the forecast horizon) and from minutes to hours (for the time-step) [Costa et al., 2008]. Different approaches are used in wind energy modeling, depending on the desired forecast horizon. For short horizons (up to 3 to 6 hours) statistical approaches using only data available on-line as input are generally employed. For longer time horizons, prediction models usually make use of a numerical weather prediction model [see Giebel et al., 2011, for more details].

There is already a rich literature regarding wind speed and wind power models for short time scales. In some articles, the model for wind speed is obtained by fitting distributions to wind speed data records, that is, by estimating the parameters of the marginal distribution of wind speed distribution. On this approach, the positive correlation between consecutive observations of wind speed is ignored. Several works, on the other hand, try to incorporate the autocorrelation into wind speed models mainly by using time series or artificial neural networks techniques, and one can even find papers that propose ensembles of several models. For detailed reviews on this subject, refer to [Costa et al., 2008; Giebel et al., 2011; Foley et al., 2012; Zhu and Genton, 2012].

The goal of this paper is to present an application of the Box-Jenkins model identification procedure to identify an univariate ARIMA model from wind speed data records of an anemometric station of the SONDA network. The SONDA (*Sistema de Organização Nacional de Dados Ambientais* - National Organization System of Environment Data) project aims at providing high confidence data of interest to the Brazilian energy sector, and it has established a network of ground data acquisition sites for solar and wind data throughout the Brazilian territory [Martins et al., 2004]. Although several studies use data from the SONDA database, there are few papers dealing with the modeling and forecasting of wind speed that make use of this database [Oliveira et al., 2016, is an example]. It is expected that the present article may motivate the use of SONDA project data in other works in this field. Furthermore, as highlighted by [Meyn and Tweedie, 2009, Section 2.1.2], linear models such as ARIMA models are suitable for a Markovian description and, thus, may be of interest in the context of discrete-time Markov jump linear systems [see Costa et al., 2005; Costa

and Figueiredo, 2015, 2016; Figueiredo and Costa, 2016].

The rest of the paper is organized as follows. In Section 2 the SONDA database is described. In Section 3 the ARIMA models are presented and in Section 4 a procedure is introduced in order to identify an ARIMA model from wind speed data. Next, in Section 5 a numerical example of the proposed identification procedure, considering data from an anemometric station located at Triunfo, Pernambuco, is presented. Finally, some final comments are discussed in Section 6.

2. SONDA Database

The SONDA database is public available at the SONDA project website [Sistema de Organização Nacional de Dados Ambientais - SONDA, 2017]. Data is organized by measurement station, type of dataset (environmental data or anemometric data), and month and year. Besides the data records of measurements, complementary information about data quality is also available. The data validation process employed by the SONDA network is based on the quality control policies proposed by the Baseline Surface Radiation Network and Webmet.com.

A collection of text files (semicolon separated csv files) composes the SONDA database. The SONDA wind towers have wind sensors installed at 25m and 50m heights and measurements are performed every 10 minutes. For a given anemometric data file, content is organized as presented in Table 1.

Table 1: Anemometric data file content

Column	ID	Data
1	id	Station ID
2	year	Current year of data collection
3	day	Julian calendar day
4	datetm	Date in Gregorian calendar and time:minute:second in UTC time
5	min	Minute (indicates the average of the subsequent 10 minutes)
6	ws_25	Average wind speed (ms^{-1}) at 25m
7	wd_25	Average wind direction at 25m (from 0° - North to 360° clockwise)
8	tp_25	Air temperature (°C) at 25m
9	ws_50	Average wind speed (ms^{-1}) at 50m
10	wd_50	Average wind direction at 50m (from 0° - North to 360° clockwise)
11	tp_50	Air temperature (°C) at 50m

Analogously to anemometric data, the SONDA database provides csv files with environmental data (such as surface air temperature, relative humidity, atmospheric pressure, etc).

3. ARIMA Models

In this section, the autoregressive integrated moving average (ARIMA) models are introduced. An ARIMA process is such that its dth difference is a stationary, invertible mixed autoregressive-moving average (ARMA) process, thus this class of models describes homogeneous nonstationary behavior. For more details on this topic, refer to [Box et al., 2015, Chapter 4].

Some notation must be introduced. Let $\mathbb{Z}$, $E(\cdot)$, and var$(\cdot)$ denote the set of nonnegative integers, the expectation operator, and the variance operator, respectively. As usual, for a time series $\{z_t, t \in \mathbb{Z}\}$, the backward shift operator is denoted by B and is such that $Bz_t = z_{t-1}$, and the difference operator is denoted by Δ and is such that $\Delta = (1 - B)$, that is, $\Delta z_t = (1 - B)z_t = z_t - z_{t-1}$. Moreover, throughout the paper $\{a_t, t \in \mathbb{Z}\}$ denotes a white noise process, that is, $\{a_t, t \in \mathbb{Z}\}$ is a sequence of independent identically distributed random variables with $E(a_t) = 0$ and var$(a_t) = \sigma_a^2$.

An ARIMA model of order (p, d, q), or simply an ARIMA(p, d, q), can be written in the form

$$\phi(B)\Delta^d z_t = \theta_0 + \theta(B)a_t, \tag{1}$$

where $\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$ is a stationary autoregressive operator of order p, $\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)$ is an invertible moving average operator of order q, and θ_0 is a constant term.

Some comments on the model (1) should be pointed out. It is immediate to see that if $d = 0$, the model (1) is a stationary process. The constant term θ_0 allows that model (1) represents series that have stochastic ($\theta_0 = 0$) or deterministic ($\theta_0 \neq 0$) trends. It should also be mentioned that, for the representation of nonseasonal time series, parameters p, d, or q rarely need to be greater than 2.

4. Wind Speed Modeling

In this section, an identification procedure based on the Box-Jenkins method [Box and Jenkins, 1970] is presented in order to identify an ARIMA model for wind speed from data records of SONDA database.

4.1. Data Preparation

The SONDA stations record average wind speed every ten minutes. From this records, other time scales may be considered. For instance, the value of the hourly average wind speed can be obtained averaging the six values measured within each hour (Torres et al. [2005] performed this procedure to analyze hourly wind speeds at Navarre, Spain).

Depending on the time step of the time series, wind speed series may present non-Gaussian distributions, diurnal nonstationary, and seasonal nonstationary. An approach to deal with it is by applying an adequate transformation to adjust the original time series, thus z_t is substituted by z_t^* in (1), where z_t^* is a transformation of z_t involving one or more transformation parameters λ:

i. A power transformation $z_t' = z_t^\lambda$ may be applied to the wind speed time series, so that the distribution becomes approximately Gaussian [see Brown et al., 1984; Nfaoui et al., 1996; Torres et al., 2005, for instance].

ii. Regarding the diurnal nonstationary of wind speed data, many authors have employed the transformation given by

$$z_t^* = \frac{z_t' - \mu(t)}{\sigma(t)}, \tag{2}$$

in order to remove it. For instance, Brown et al. [1984]; Nfaoui et al. [1996]; Torres et al. [2005] considered ARMA models to analyze hourly wind speed time series transformed by (2), that is, they removed the diurnal nonstationary by subtracting the hourly expected wind speed values $\mu(k)$ and dividing by the hourly standard deviations $\sigma(k)$, $k \in \{0, 1, \ldots, 23\}$. In (2) it is assumed that μ and σ are periodic with $\mu(k+24\,i) = \mu(k)$ and $\sigma(k+24\,i) = \sigma(k)$, for $k \in \{0, 1, \ldots, 23\}$, $i \in \mathbb{Z}$.

iii. Seasonal nonstationarity can be removed by fitting a separate model for each month [see Brown et al., 1984; Nfaoui et al., 1996; Torres et al., 2005, for instance]. Another approach to deal with seasonality is to analyze seasonal ARIMA (SARIMA) models [see Kamal and Jafri, 1997, for instance]. As only ARIMA models, given by (1), are considered in this paper, this alternative approach will not be further detailed.

4.2. Model Structure Selection and Parameters Estimation

During an ARIMA model identification, usually an ARIMA model structure (p, d, q) is defined by analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the time series. For a chosen structure (p, d, q), the parameter's estimation of the associated ARIMA model can be performed by maximum likelihood estimation methods.

In a more pragmatical approach, one can estimate several models and, after that, select a model by applying a model selection criteria such as minimizing the Akaike information criterion (AIC), the Akaike information criterion corrected for small sample sizes (AICc) or the Bayesian information criterion (BIC). The information criteria AIC is given by

$$\text{AIC}_{p,q} = -2\ln(\text{maximized likelihood}) + 2r, \tag{3}$$

the AICc by

$$\text{AICc}_{p,q} = \text{AIC} + \frac{2r(r+1)}{n-r-1}, \tag{4}$$

and the BIC by

$$\text{BIC}_{p,q} = -2\ln(\text{maximized likelihood}) + r\ln(n), \tag{5}$$

where n is the sample size, $r = p + q + 1$ is the number of parameters estimated in the model, including constant term. For more details on the use of model selection criteria and on parameter estimation, refer to [Box et al., 2015, Section 6.2 and Chapter 7].

Regarding wind series, no deterministic trend is expected in a wind speed time series. Thus θ_0 can be omitted in (1) and only models given by

$$\phi(B)\Delta^d z_t = \theta(B)a_t \tag{6}$$

should be considered.

4.3. Model Validation

The model validation requires confirming that the hypotheses made with respect to the model residuals hold true. Recall that it was assumed that the residuals follow a white noise process. Thus, in this step of the model identification procedure, the main concern is checking if the residuals are independent, and distributed with zero mean and constant variance. For more details on model validation, refer to [Box et al., 2015, Chapter 8].

4.4. Model Performance

In the literature on modeling and forecasting of wind speed, several different measures of performance have been adopted for model performance assessment, e.g. mean error, mean absolute error, root mean square error, etc [see Foley et al., 2012, Table 2]. Many authors consider the persistence model (naive predictor) as a reference model and evaluate more advanced models in terms of improvement over persistence. The persistence forecasting assumes that the future values of wind speed are the same as the current one, that is,

$$\hat{z}_t(k) = z_t, \tag{7}$$

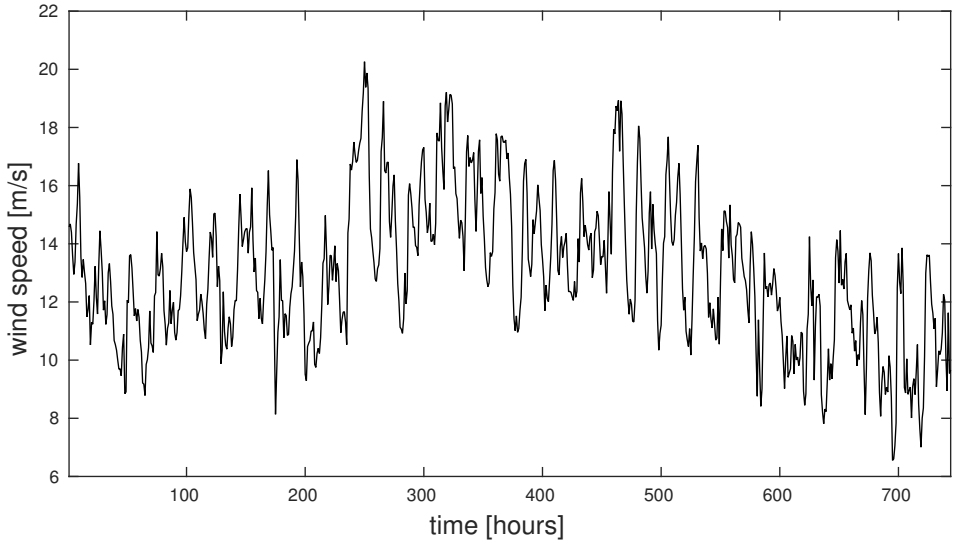
where $\hat{z}_t(k)$ denotes the forecast of z_{t+k}, i.e. the wind speed at some future time instant $t+k$ when we are currently at time t. In order to have best validation results and avoid overfitting, it is usual to consider different data sets for model identification and validation/performance assessment.

5. Numerical Example

In this section, a numerical example of the procedure introduced in Section 4 is presented. The computations and graph visualizations were done using the statistical software package R and the numerical computing software Matlab.

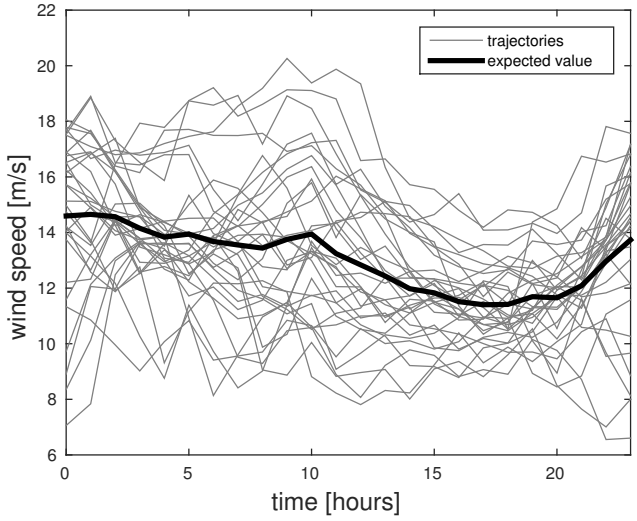
Wind speed data for January 2006 obtained from an anemometric station located at Triunfo, Pernambuco, is presented in Figure 1.

Figure 1: Wind speed time series (January 2006)



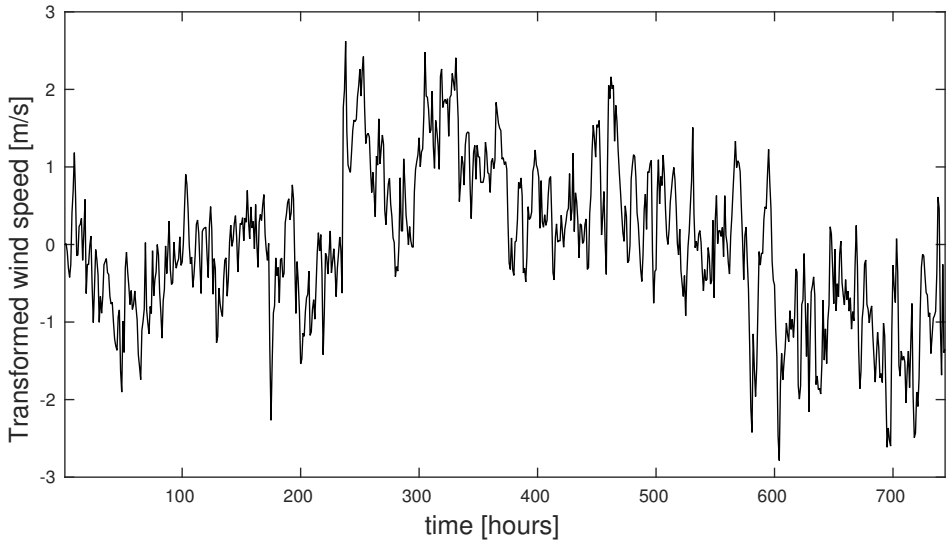
The daily wind speed time series for January 2006 is presented in Figure 2. That is, this figure contains the 31 trajectories of hourly average wind speed, each trajectory representing the time series of one of the days of January 2006, and the thick line in the figure is the expected value trajectory. The value of the hourly average wind speed was obtained by averaging the six values measured within each hour (recall that the SONDA measurements are performed every 10 min).

Figure 2: Daily wind speed time series (January 2006)



It is easy to see from Figure 2 that the wind speed presents diurnal nonstationary. Thus the transformation (2) may be applied to adjust data. Figure 3 presents the time series of 744 transformed hourly wind speed data, each point corresponding to a data record of one of the 24 hours of the 31 days of January 2006. In what follows the time series of transformed hourly wind speed is divided into two data sets: the first one is used for model identification (the first 520 points) and the other for performance assessment (the last 224 points).

Figure 3: Transformed wind speed time series (January 2006)



By applying the pragmatical approach proposed in Section 4.2, an ARIMA model can be obtained to model the transformed time series (considering only the data set for model identification). This step has been done by employing the auto.arima() function from the forecast package of R, considering the AIC information criterion. The output from the identification process is presented in Table 2.

Table 2: Estimated ARIMA(1,1,2) model

Parameter	**Value**	**Standard Error**
ϕ_1	0.6157	0.0921
θ_1	0.7150	0.0981
θ_2	0.1671	0.0597

Let us denote by y_t the transformed hourly wind speed time series. From the estimated parameters (Table 2), the obtained ARIMA(1,1,2) model is given by

$$(1 - 0.6157B)(1 - B)y_t = (1 - 0.7150B - 0.1671B^2)a_t, \tag{8}$$

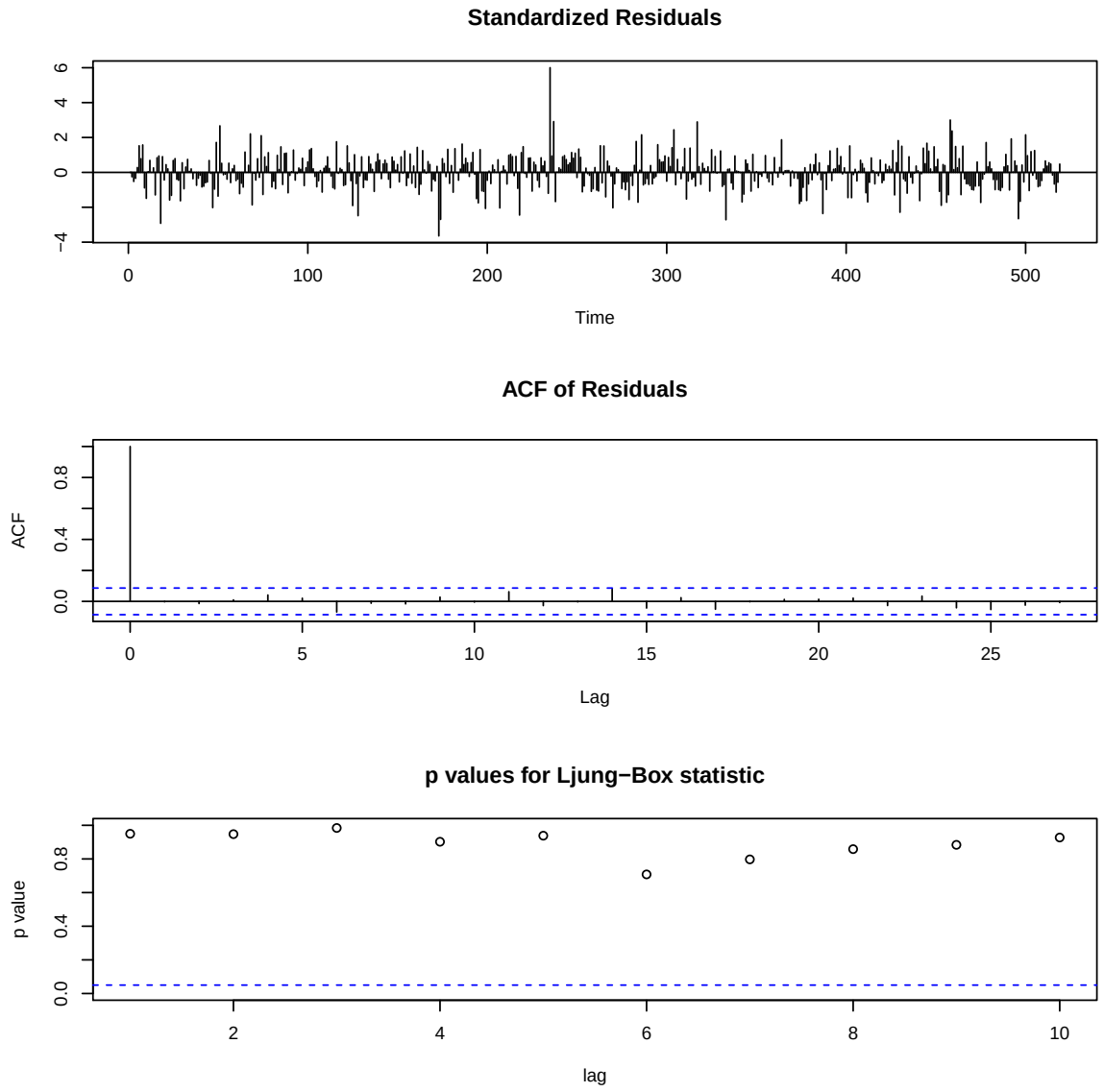
that can also be written as

$$y_t = 1.6157y_{t-1} - 0.6157y_{t-2} + a_t - 0.7150a_{t-1} - 0.1671a_{t-2}. \tag{9}$$

The model residuals, the autocorrelation function of the residuals, and the p-values of a Portmanteau test (Ljung-Box test) for several lags are plotted in Figure 4. As expected, this figure shows a small error range, centered around zero (the mean of the residuals is approximately zero),

and the variation of the residuals seems the same across the time series. Checking the autocorrelation plot, the residuals have no significant autocorrelations. Also, by applying the Ljung-Box test [see Box et al., 2015, Section 8.2] it follows that the residuals from the estimated ARIMA model are independently distributed. From this analysis, one can conclude that (8)-(9) adequately model the considered time series.

Figure 4: Residuals diagnostics



The difference equations for computing forecasts k-hours ahead for the model (8)-(9) are given by

$$\hat{y}_t(k) = \begin{cases} 1.6157y_t - 0.6157y_{t-1} - 0.7150a_t - 0.1671a_{t-1}, & \text{if } k = 1 \\ 1.6157\hat{y}_t(1) - 0.6157y_t - 0.1671a_t, & \text{if } k = 2 \\ 1.6157\hat{y}_t(k-1) - 0.6157y_t(k-2), & \text{if } k \geq 3. \end{cases} \tag{10}$$

To convert $\hat{y}_t(k)$ obtained in (10) into a wind speed forecast k-hours ahead $\hat{z}_t(k)$, the relationship

$$\hat{z}_t(k) = \hat{y}_t(k) \cdot \hat{\sigma}(t+k) + \hat{\mu}(t+k) \tag{11}$$

is used. Recall that $\hat{\mu}(t)$ and $\hat{\sigma}(t)$ have been defined when applying transformation (2).

The results in terms of root mean square error (RMSE) of forecasts obtained by applying model (11) and by applying the persistence model (7) are presented in Table 3, where the improvement is given by

$$\frac{\text{RMSE}_{\text{persistence}} - \text{RMSE}_{\text{ARIMA}}}{\text{RMSE}_{\text{persistence}}}.$$

It is worth pointing out that in order to avoid overfitting in this analysis only the data set for performance assessment have been used. In summary, the ARIMA approach yielded an improvement in the RMSE compared to persistence for all the considered forecast horizons. The results presented in Table 3 are similar to those found in the literature [see Foley et al., 2012, Section 2.4].

Table 3: Root mean square error (RMSE) in m/s by model

Hours ahead	ARIMA	Persistence	Improvement
1	1.0374	1.1702	11.35%
2	1.4507	1.7051	14.92%
3	1.6501	2.0388	19.06%
4	1.7734	2.2914	22.60%
5	1.8442	2.4561	24.91%
6	1.8730	2.5473	26.47%
7	1.8908	2.5762	26.60%
8	1.9293	2.6333	26.73%
9	1.9681	2.6861	26.73%
10	1.9895	2.7164	26.76%

6. Final Comments

In this paper, an identification procedure to derive an ARIMA model from wind speed data records of the SONDA database was presented. This identification procedure was illustrated in a numerical example in which a model was obtained from wind speed data records of a SONDA anemometric station located at Triunfo, Pernambuco. In the numerical example, the ARIMA approach yielded an improvement in the RMSE of forecasts compared to the persistence model for all the considered forecast horizons. In comparison with the present paper, future works on short-time modeling of wind speed may deal with different data records from SONDA project, employ different time scales and forecast horizons, or adopt other modeling techniques (such as multivariate time series analysis, artificial neural networks techniques, wavelets techniques, etc) and other input data besides wind speed (such as wind direction, air temperature, relative humidity, atmospheric pressure, etc).

Acknowledgments

The authors are grateful to the anonymous referee for the suggestion that led to a more detailed numerical example. This work was supported in part by the project INCT (National Institute of Science and Technology) under the grant CNPq (Brazilian National Research Council) 465755/2014-3, FAPESP (São Paulo Research Foundation) 2014/50851-0, FAPESP/Shell, through the Research Center for Gas Innovation, grant FAPESP 2014/50279-4, and FUSP (Fundação de Apoio à Universidade de São Paulo). D.Z. Figueiredo was supported by the grant CNPq 380610/2017-5 and O.L.V. Costa was supported in part by the grant CNPq 304091/2014-6.

References

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. Wiley.

Box, G. E. P. and Jenkins, G. (1970). *Time series analysis: forecasting and control*. Holden-Day.

Brown, B. G., Katzb, R. W., and Murphyc, A. H. (1984). Time series models to simulate and forecast wind speed and wind power. *Journal of Climate and Applied Meteorology*, 23(8):1184–1195.

Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., and Feitosa, E. (2008). A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, 12(6):1725–1744.

Costa, O. L. V. and Figueiredo, D. Z. (2015). LQ control of discrete-time jump systems with Markov chain in a general Borel space. *IEEE Transactions on Automatic Control*, 60(9):2530–2535.

Costa, O. L. V. and Figueiredo, D. Z. (2016). Quadratic control with partial information for discrete-time jump systems with the Markov chain in a general Borel space. *Automatica*, 66:73–84.

Costa, O. L. V., Fragoso, M. D., and Marques, R. P. (2005). *Discrete-Time Markov Jump Linear Systems*. Probability and Its Applications. Springer.

Figueiredo, D. Z. and Costa, O. L. V. (2016). Jump linear quadratic optimal control applied to a wind turbine generator system. In *XXI Congresso Brasileiro de Automática*, p. 887–892.

Foley, A. M., Leahy, P. G., Marvuglia, A., and McKeogh, E. J. (2012). Current methods and advances in forecasting of wind power generation. *Renewable Energy*, 37(1):1–8.

Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., and Draxl, C. (2011). *The State-Of-The-Art in Short-Term Prediction of Wind Power: A Literature Overview, 2nd edition*. ANEMOS.plus. Project funded by the European Commission under the 6th Framework Program, Priority 6.1: Sustainable Energy Systems.

Kamal, L. and Jafri, Y. Z. (1997). Time series models to simulate and forecast hourly averaged wind speed in Quetta, Pakistan. *Solar Energy*, 61(1):23–32.

Martins, F. R., Pereira, E. B., Neto, S. L. M., Abreu, S. L., Colle, S., and Beyer, H. G. (2004). Solar and wind resources database to support energy policy and investiments in South America. In Ortega, E. and Ulgiati, S., editors, *Proceedings of IV Biennial International Workshop - Advances in Energy Studies*, p. 419–427, Campinas. Unicamp.

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition.

Ministério de Minas e Energia (2016). *Resenha Energética Brasileira - Exercício de 2015 (Edição de Maio de 2016)*. Ministério de Minas e Energia.

Nfaoui, H., Buret, J., and Sayigh, A. A. M. (1996). Stochastic simulation of hourly average wind speed sequences in Tangiers (Morocco). *Solar Energy*, 56(3):301–314.

Oliveira, G. S. d., Barros, M. F. d., and Oliveira, F. L. C. (2016). Aplicação de cadeias de Markov para análise de séries temporais de energia eólica. In *Anais do XLVIII SBPO - Simpósio Brasileiro de Pesquisa Operacional*, p. 818–829, Vitória.

Pereira, E. B. and Lima, J. H. G. (2008). *Solar and wind energy resource assessment in Brazil*. National Institute for Space Research.

Sistema de Organização Nacional de Dados Ambientais - SONDA (2017). Sonda project website. URL `http://sonda.ccst.inpe.br/`.

Torres, J. L., Garcia, A., De Blas, M., and De Francisco, A. (2005). Forecast of hourly average wind speed with ARMA models in Navarre (Spain). *Solar Energy*, 79(1):65–77.

Zhu, X. and Genton, M. G. (2012). Short-term wind speed forecasting for power system operations. *International Statistical Review*, 80(1):2–23.