A REVIEW OF TIME SERIES MODELS FOR THE LONG TERM MODELING OF WIND SPEED IN DISTRIBUTION NETWORK PLANNING

Zacharie DE GREVE
Power Elec. Eng. Dept.
Univ. of Mons – Belgium
zacharie.degreve@umons.ac.be

Cyril DANIELS
Power Elec. Eng. Dept.
Univ. of Mons – Belgium
cyril.daniels@student.umons.ac.be

Lazaros EXIZIDIS
Power Elec. Eng. Dept.
Univ. of Mons – Belgium
lazaros.exizidis@umons.ac.be

François VALLEE
General Physics Dept.
University of Mons – Belgium
francois.vallee@umons.ac.be

Jacques LOBRY
Power Elec. Eng. Dept.
Univ. of Mons– Belgium
Jacques.lobry@umons.ac.be

ABSTRACT

In this work, a review of univariate time series models for the long term modeling of wind speed is performed. These models intend to help the distribution system operators to account for the stochastic nature of wind during the network planning phase. Two groups of criteria are proposed for comparing the performance of the models. The first one is focused on the representativeness (i.e. the ability of the model to reproduce the statistical properties of the real data), whereas the second one is linked to computational issues, such as the size of the historical dataset which is needed to obtain a given level of representativeness, or the computational burden associated to the training procedure.

INTRODUCTION

The increasing penetration of wind power has raised new challenges concerning the operation and the long term planning of distribution grids. Indeed, the stochastic nature of wind production significantly complicates the formulation and the resolution of the underlying optimization problem (see *e.g.* [1] for a recent formulation of the day-ahead operational management of distribution networks, and [2] for the long term planning). In the case of distribution network planning more precisely, the system operator is particularly interested in obtaining future trajectories of wind speed (*e.g.* for the upcoming year) at different wind sites. Indeed, this information is crucial for the determination of optimal investment plans (after a transformation in the power domain, *e.g.* using power curve functions).

Time series models are powerful solutions for that purpose. They permit to generate synthetic wind data which mimic the statistical properties of real measurements. Moreover, they can be stored in a compact form, as they are usually represented by mathematical formulae with a small number of parameters. There is therefore no need to embed the entire historical dataset in the optimization tool for sampling.

Over the past years, the literature has been abundant on the topic of long term time series modeling of wind speed. References [3-4] are for instance focused on the use of AutoRegressive Moving Average (ARMA) models, whereas in [5], the authors propose a method based on an ARMA-GARCH approach (with GARCH standing for Generalized AutoRegressive Conditionnal Heteroscedasticity). A recent contribution [6] investigates the class of AutoRegressive Integrated Moving Average models (ARIMA). Several attempts articulated around the ARMA family can also be noticed, namely the ARMA-GARCH-in-Mean model [7], and Fractional ARMA models [8]. The list is non exhaustive, but the section “Models” will give a detailed description of some of the most significant approaches.

Nevertheless, to the best of the authors’ knowledge, no comparative study has been published so far. This work intends therefore to provide an objective comparison between the available approaches, by testing them on the same dataset. A particular attention will be paid to the definition of two groups of comparison criteria, focused on the representativeness of the model on the one hand, and on computational issues on the other hand. The most significant contributions will be described in section “Models”, and tested on a common dataset, obtained from the Royal Netherlands Meteorological Institute (KNMI [9]), in section “Comparative study”. The issue of the preprocessing of the raw data (*i.e.* the treatment to apply to wind speed data before using it in a time series mathematical model) will also be discussed.

This contribution is focused on time series models only: other approaches, making use of Artificial Neural Networks [10] or Markov chains [11] for instance, are available in the literature. These will be tested in the future, as it will be discussed in section “Conclusions and Perspectives”.

MODELS

Data pre-processing

The ARMA and ARMA-GARCH classes require to work on weak-sense stationary processes, *i.e.* stochastic processes for which the mean is constant over time, the variance is finite at each time *t*, and for which the covariance function is independent of the time lag [12]. In practice, collected wind speed data does not verify these properties: it naturally shows seasonal patterns (day/night cycles, seasons), and may contain a trend. Therefore, a pre-processing must be applied to the raw data in order to remove such effects. Two approaches are implemented and compared in this paper.

Centralization-reduction [3]

An elegant procedure, based on a centralization-reduction operation, is proposed in [3]. The idea is to work on a standardized version X_t of the initial wind speed time series W_t, obtained using the following equations:

$$X_t = (W_t - \mu_t)/\sigma_t \ , \qquad (1)$$

with μ_t and σ_t respectively the mean and the standard deviation of observed wind speed at time *t*.

Inversion of Cumulative Distribution Functions [12]

In that case, the standardized time series X_t are obtained by following three steps:

1. Compute the cumulative distribution functions (cdfs) of observed wind speed at each time *t*.
2. Transform each observed wind speed to uniformity using the computed cdfs.
3. Transform the uniformly distributed data to normality, using the inverse cdf of a normal with zero mean and unit variance.

ARMA

A zero mean ARMA process $\{X_t\}$ of order (p,q) can be defined as follows [13]:

$$X_t = \sum_{k=1}^{p} a_k X_{t-k} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \ , \qquad (2)$$

with $\{\varepsilon_t\}$ the process of innovations (a gaussian white noise $N(0,\sigma_\varepsilon^2)$ of variance σ_ε^2), and with a_k and θ_j non zero constants. In this work, the AR order *p* and MA order *q* are estimated using the Bayesian Information Criterion (BIC) [14], knowing that any stationary process can be approximated as closely as required by an ARMA(*n*,*n*-1) model [15]. Then, the coefficients a_k,θ_j and σ_ε^2 are computed using a conditional MLE (Maximum Likelihood Estimation) procedure [13].

ARMA-GARCH

An ARMA(*p*,*q*)-GARCH(*a*,*b*) model consists in an ARMA(*p*,*q*) process for which the process of innovations $\{\varepsilon_t\}$, is written as follows [16]:

$$\varepsilon_t^2 = z_t\sigma_t \ , \ z_t \sim D(0,1) \ , \qquad (3)$$

where the conditional variance σ_t^2 reads:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{a} \alpha_i \sigma_{t-i}^2 + \sum_{l=1}^{b} \beta_l \varepsilon_{t-l}^2 \ , \qquad (4)$$

with α_i and β_l positive coefficients and z_t the standardized residuals (independent and identically distributed). Their probability density function is *D* (with zero mean and unit variance), which may be normal or Student-T. Such a model allows to take into account a temporal variability in terms of variance, coming from the past evolution of the process. It is for instance well suited for the representation of series which tend to group the variability into clusters, *i.e.* which separates periods of high and low variability. Again, in this work, the optimal orders of the ARMA-GARCH model are estimated using the BIC, and the coefficients are computed using the MLE approach.

ARIMA

In the “Data pre-processing” subsection, two procedures have been exposed for ensuring that the time series are stationary. Another typical solution consists in applying multiple numerical differentiations to the time series. This is the approach retained for ARIMA and Seasonal ARIMA (SARIMA) models [6]. These are ARMA models for which the series have been subject to differentiations of appropriate orders so as to remove trend and seasonality effects. However, in this work, the simplicity of implementation of [3] and [12] has been favoured against that class of models.

COMPARATIVE ANALYSIS

Two groups of comparison criteria are proposed, linked to the notion of representativeness of a model on the one hand, and on computational issues on the other hand. They are tested on wind speed data from the Schiphol wind site, in the Netherlands [9]. The data has been collected on an hourly basis, for 54 years (between 1951 and 2005), and has been classified month by month in order to take seasonal effects into account.

In the “Representativeness” subsection, the whole set of 54 years has been employed for the data pre-processing (centralization-reduction method of [3], or inverse cdfs method of [12]). In the case of June for instance, it means that the μ_t and σ_t are computed at each hour of the month, basing on 54 months of data. On the other hand, three years of data (namely June 2003, 2004 and 2005) has been employed for estimating the parameters of the ARMA and ARMA-GARCH models. The size of the database will vary in the “Computational issues” subsection, as it will be discussed.

Representativeness

The ARMA and ARMA-GARCH models obtained using the inverse cdfs and centralization-reduction methods are listed in Table 1 and 2 respectively, for the dataset exposed above (June, Gaussian innovations). The estimated models are nearly identical, for the two data pre-processing methods. The reason for choosing one of the two approaches is linked with computational issues, as it will be discussed in the next section.

Table 1: parameters of ARMA and ARMA-GARCH models (three years database), using inverse cdfs method (54 years database), for June.

	α_1	α_2	θ_1	σ_t^2
ARMA(2,1)	1.14	-0.23	-0.51	0.35
ARMA(2,1)-GARCH(1,1)	1.11	-0.21	-0.42	$0.043 + 0.63\sigma_{t-1}^2$ $+ 0.26\varepsilon_{t-1}^2$

Table 2: parameters of ARMA and ARMA-GARCH models (three years database), using centralization-reduction method (54 years database), for June.

	α_1	α_2	θ_1	σ_t^2
ARMA(2,1)	1.16	-0.24	-0.5	0.34
ARMA(2,1)-GARCH(1,1)	1.1	-0.21	-0.37	$0.042 + 0.63\sigma_{t-1}^2$ $+ 0.26\varepsilon_{t-1}^2$

The purpose of the models is to generate future trajectories of the wind speed at a given site, *e.g.* for the upcoming year, which are representative of the behavior of the measured speed. This first category of criteria is devoted to the verification of the correspondence between the statistical properties of real and simulated data. Discussions with system planners lead to the definition of three indicators regarding the representativeness, which are exposed below.

Energy criterion

The first criterion is focused on the energy content of the simulated time series. This can be analyzed by plotting the histograms of simulated and observed wind speeds for a given month (June in our case). Figure 1 shows for instance such histograms for real observed data (blue and black curves), as well as the 95% confidence bounds of simulated data using the ARMA (red curves) and ARMA-GARCH (green curves) models. It can be observed that the 95% confidence bounds of the two models encompass the observed histograms. Moreover, their performance is similar regarding the energy criterion.

Autocorrelation criterion

The autocorrelation function of the simulated series must reproduce faithfully the temporal correlation of the measured data. Figure 2 depicts the autocorrelation plots of observed (blue curves) and simulated data (red curves for ARMA and green curves for ARMA-GARCH), for June. Again, the performance of the two models is similar, even if the ARMA-GARCH seems to behave slightly better for high lags.

Variability criterion

The energy criterion evaluates the ability of the model to simulate data which mimics the mean behavior of real wind speed. However, no information is given regarding the wind speed variations between successive hours (sign and amplitude), which need to be correctly reproduced during the system planning phase. Therefore, a third criterion has been proposed, which consists in comparing the distributions of wind speed variations between real and simulated data. Figure 3 shows for instance such distributions for the month of June, the black lines corresponding to observed data, the red (green) ones to the 95% confidence bounds of simulated data using the ARMA (ARMA-GARCH) model. It can be seen that the two models capture well the variability of the real data, excepted for very small variations (*e.g.* 1m/s). The ARMA-GARCH model performs however slightly better than the ARMA in these conditions.

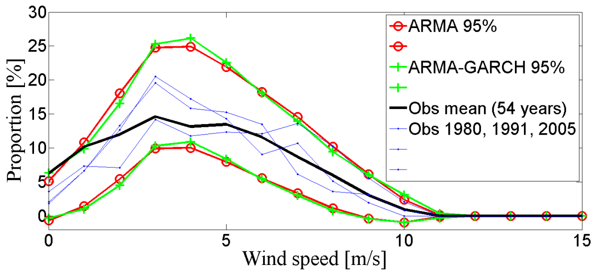

Fig.1: histograms of observed data for June (blue and black curves), and 95% confidence bounds of simulated data (red curves for ARMA and green curves for ARMA-GARCH).

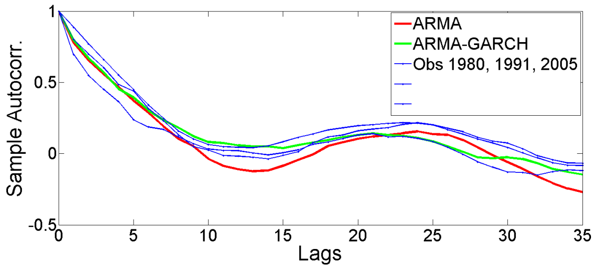

Fig.2: autocorrelation plots of observed (blue curves) and simulated data (red curves for ARMA and green curves for ARMA-GARCH), for June.

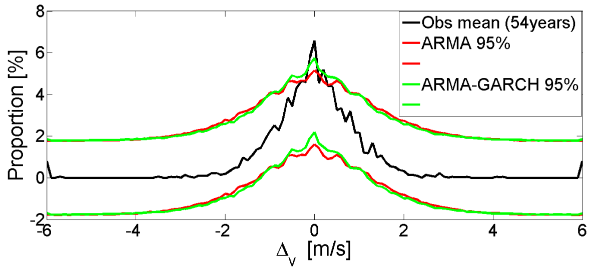

Fig.3: distribution of the wind speed variations for observed data (in black), and 95% confidence bounds for simulated data (ARMA in red and ARMA-GARCH in green).

Based on the three representativeness criterion, and provided that the size of the database is large (here 54 years of wind data for the pre-processing, and three years for model estimation), no significant differences appear between the ARMA and ARMA-GARCH class of models (even if the variability of real data is captured by the ARMA-GARCH family slightly better).

Computational issues

Size of the dataset

The influence of the size of the database on the representativeness performance of the models is studied. A first analysis has been conducted by reducing the amount of data used during the preprocessing phase (from 54 years to 3 years), keeping a fixed amount of three years of data for the estimation of the times series models. The main conclusion was that the representativeness of the models was good, even with the reduced set of 3 years.

However, further comments need to be made when only one year of data is available. In that case, μ_t and σ_t for the centralization-reduction method [3] (or the hourly cdfs for the method of [12]) must be computed for a typical day of the studied month. Indeed, it is not possible to calculate them for each hour of the month, because of the reduced amount of available data. Two ARMA-GARCH models have been estimated, starting from the data collected in June 2005 only, for the centralization-reduction procedure (ARMA(3,2)-GARCH(2,1)) and for the inverse cdfs method (ARMA(1,1)-GARCH(1,1)). Figures 4, 5 and 6 compare the performances of the two models regarding the three representativeness criteria. It can be seen that the model based on inverse cdfs performs better than the other, especially for the energy and variability criteria. This is an important result when a reduced amount of data is at hand (typically one year). Results for simple ARMA models are not presented here, since they do not bring anything new to the discussion

Computational burden

The two main steps which influence the overall computational burden are the time needed to estimate the models on the one hand, and the time required for simulating synthetic data once the model is known on the other hand, including the de-standardization procedure.

Regardless of the type of model (ARMA or ARMA-GARCH), the whole estimation procedure last around 60s of CPU time for three years of data, for a given month (three years appeared as an optimal trade-off between representativeness and the computational burden linked to the estimation phase). This is not a constraint since that procedure is performed only one time, at the construction of the model.

The issue of generating synthetic data may be more time consuming. It is really fast in the case of the centralization-reduction pre-processing (less than 1s of CPU time for generating 100 series, including de-standardization). However, the inverse cdfs approach has logically required a CPU time of 133s in our case, for the same number of series, since it implies to construct the cdfs as well as their inverse. This may be problematic within the framework of a system planning software, for which a lot of synthetic series need to be produced in a Monte Carlo framework, for a lot of different wind sites. Moreover, the latter approach requires to store the cdfs in the tool, which is more memory consuming than storing the scalars μ_t and σ_t of the centralization-reduction method.

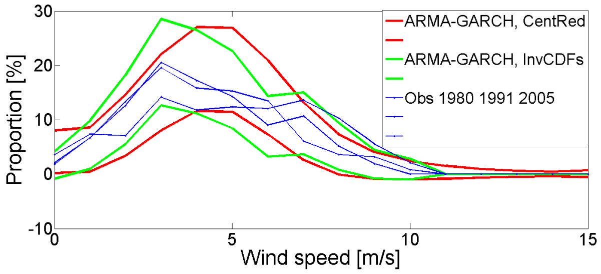

Fig.4: energy criterion, database of June 2005 only. In blue are depicted observations for years 1980, 1991, 2005, in red (green) the 95% confidence bounds for data simulated by the ARMA-GARCH model with the centralization-reduction (inverse cdfs) method.

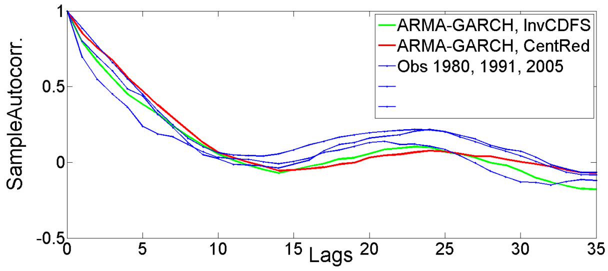

Fig.5: autocorrelation criterion, database of June 2005 only.

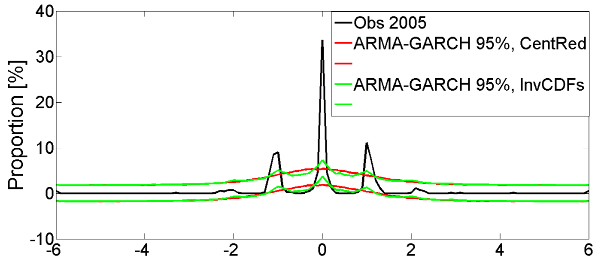

Fig.6: variability criterion, database of June 2005 only.

CONCLUSIONS AND PERSPECTIVES

Various time series models (mainly ARMA [3] and ARMA-GARCH [5]) have been reviewed for the generation of synthetic wind speed data in the framework of distribution network planning, with an emphasis on the pre-processing which has to be applied on the raw data before fitting the models (centralization-reduction method of [3] and inverse cdfs transformation of [12]). Two class of comparison criteria have been proposed, linked to the notion of representativeness of a model on the one hand, and on computational issues on the other hand.

It has been shown that the performance of the ARMA and ARMA-GARCH models were quite similar regarding the representativity as well as the computational burden (even if the ARMA-GARCH class captures the variability of the observed wind data slightly better, at the expense of a slightly higher computational burden during the estimation phase).

A critical study has been conducted on the size of the training datasets. It has been shown that the representativeness of the models remain acceptable even if only a few years of data are available. On the other hand, when only one year is at hand, a time series model based on the inverse cdfs transform performs significantly better. Unfortunately, the CPU time required for the generation of synthetic data drastically increases in that case. Therefore, the choice of the method will depend on the available data. If the dataset contains a few years or more, the centralization-reduction method prevails as it is fast and light, whereas the inverse cdfs approach may be employed for small dataset for a better representativeness.

The literature is very abundant on the topic of statistical modelling of wind data. This work is a first attempt to clarify the problem among the power electrical engineers community. It will be completed in the future, with the objective to propose an exhaustive state-of the-art in the field of long term wind speed modelling (use of Artificial Neural Networks, Markov chains, Kalmann filters, etc.).

The important issues of geographical correlation, and of the generation of synthetic series in the presence of incomplete data (missing samples, etc.) will also need to be investigated. The set of proposed comparison criteria will be extended, as soon as these aspects will be taken into account in the study.

REFERENCES

[1] Q. Gemine, E. Karangelos, D. Ernst, B. Cornélusse, 2013, "Active network management: planning under uncertainty for exploiting load modulation", *Proceedings of the Bulk Power System Dynamics and Control – IX Optimization, Security and Control of the Emerging power Grid, IREP2013 Symposium*, Greece, 9 pages.

[2] V.F. Martins, C.L.T. Borges, 2011, "Active distribution network integrated planning incorporating distributed generation and load response uncertainties", *IEEE Transactions on Power Systems*, vol. 26, no. 4, 2164-2172

[3] R. Billiton, H. Chen, R. Ghajar, 1996, "Time-series models for reliability evaluation of power systems including wind energy", *Micorelectronics Reliability*, vol. 36, no. 9, 1253-1261.

[4] J.L. Torres, A. Garcia, M. De Blas and A. De Francisco, 2005, "Forecast of hourly average wind speed with ARMA models in Navarre (Spain)", *Solar Energy*, vol. 79, 65-77.

[5] A. Lojowska, D. Kurowicka, G. Papaefthymiou and L. Van Der Sluis, 2010, "Advantages of ARMA-GARCH wind speed time series modeling", *Proc. of the 11th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS2010)*, Singapore, 83-88.

[6] P. Chen, T. Petersen, B. Bak-Jensen, Z. Chen, 2010, "ARIMA-based time series model of stochastic wind power generation", *IEEE Transactions on Power Systems*, vol. 25, no. 2, 667-676.

[7] B.T. Ewing, J.B. Kruse, J.L. Schroeder, 2006, "Time series analysis of wind speed with time-varying turbulence", *Environmetrics*, vol.17, 119-127.

[8] S. Hussain, A. Elbergali, A. Al-Masri, G. Shukur, 2004 "Parsimonious modelling, testing and forecasting of long-range dependence in wind speed", *Environmetrics*, vol. 15, 155-171.

[9] Koninklijk Nederlands Meteorologisch Instituut website (KNMI), http://www.knmi.nl/samenw/hydra/

[10] M. Monfared, H. Rastegar, H.M. Kojabadi, 2009, "A new strategy for wind speed forecasting using artificial intelligent methods", *Renewable Energy*, vol.34, 845-848.

[11] A. Shamshad, W.M.A. Wan Hussin, M.A. Bawadi, S.A. Mohd. Sanusi, 2005, "First and second order Markov chain models for synthetic generation of wind speed time series", *Energy*, vol. 30, 693-708.

[12] B. Klöckl, G. Papaefthymiou, 2010, "Multivariate time series models for studies on stochastic generators in power systems", *Electric Power Systems Research*, vol.80, 265-276.

[13] G.E. Box, G.M. Jenkins, G.C. Reinsel, 1994, *Time series analysis: forecasting and control (3rd ed.)*, Prentice-Hall, Upper Saddle River, New Jersey, USA.

[14] G. Schwarz, 1978, "Estimating the dimension of a model", *The Annals of Statistics*, vol.6, no.2, 461-464.

[15] S.M. Pandit, S. Wu, 1983, *Time series and system analysis with application*, John Wiley, New York, USA.

[16] A.A. Weiss, 1984, "ARMA models with ARCH errors", *Journal of Time Series Analysis*, vol.5, no.2, 129-143.