STOCHASTIC MODELS FOR WIND SPEED TIME SERIES: A CASE STUDY*

Saverio Bivona, Giovanni Bonanno, Riccardo Burlon
Davide Gurrera†, Claudio Leone

Dipartimento di Fisica e Tecnologie Relative, University of Palermo
Viale Delle Scienze, Edificio 18, 90128 Palermo, Italy

(Received February 17, 2010)

The idea of using a mathematical model to describe the behaviour of a physical phenomenon is well established, but in many problems we have to consider a time-dependent phenomenon for which it is not possible to write a deterministic model; nevertheless, it may be possible to derive a stochastic model. The models for time series that are needed for example to achieve optimal forecasting and control are in fact stochastic models, but the choice of a proper model is never straightforward. In particular, this paper is concerned with the problem of forecasting a time series that possibly exhibits long-memory features. It results that the fractionally integrated ARMA processes may provide an adequate representation of the actual process, but do not yield a satisfactory forecasting performance.

PACS numbers: 05.45.Tp, 88.50.gg

1. Introduction

The autocorrelation function (ac.f.) ρ_k of an ARMA process converges rapidly (exponentially) to zero as the lag $k \to \infty$. Processes with this property are often referred to as short-memory processes. Stationary processes with much more slowly decreasing ac.f. do exist and they are known as long-memory processes [1]. More precisely, a process is said to possess a long memory if

$$\lim_{T\to\infty} \sum_{k=-T}^{T} |\rho_k| \tag{1}$$

* Presented at the XXII Marian Smoluchowski Symposium on Statistical Physics, Zakopane, Poland, September 12–17, 2009.

† davide.gurrera@difter.unipa.it

is nonfinite and that is tantamount to saying that the spectral density of a long-memory process becomes unbounded at low frequencies. There is empirical evidence that such long-memory processes occur frequently in fields as diverse as hydrology, geophysics, meteorology and economics [1–3]. From a practical modelling point of view, such processes may exhibit certain features that could give the impression of the need for differencing to achieve stationarity, although taking a first difference may be too extreme. A notable class of long-memory processes are the autoregressive fractionally integrated moving average (ARFIMA) processes which have recently proved to be adequate models in the analysis of time series with long-term dependence. However, several doubts have been raised as for their forecasting performance [4–6] and this paper, accordingly, tests their effectiveness upon a time series of hourly average wind speed (HAWS) recorded in Italy during one month. The series exhibits long-memory features but employing an ARFIMA model to compute the future values of the underlying stochastic process resulted in a poor prediction accuracy. The analysis of this kind of time series is of the utmost importance for the exploitation of wind energy, mainly hindered by stochastic wind speed fluctuations [7]. In order to compensate them and to take decisions in the context of the electricity market, a reliable weather forecast is necessary [8].

2. Time series analysis

Although it may be possible to increase the sample size by varying the length of the observed time series, there will only be a single outcome of the investigated stochastic process and time series analysis is essentially concerned with evaluating the properties of this underlying data generating process from the observed time series, even though this single realization is the only one we will ever observe. Once a hypothetical probability model to represent the data has been set up, it may be used to draw useful inferences from the time series. Different sources of variation may occur and accordingly different models have been developed in order to obtain an adequate representation [9].

Many time series contain a seasonal periodic component which repeats every s observations. We expect relationships to occur between adjacent observations and between observations separated by s units of time. A SARIMA$(p,d,q) \times (P,D,Q)_s$ process $\{Z_t\}$ is defined by the relation

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)W_t\,, \tag{2}$$

where B is the backward shift operator, $\phi_p(B)$ and $\theta_q(B)$ are polynomials in B of degree p and q, respectively, satisfying stationary and invertibility conditions, $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are polynomials in B^s of degree P

and Q, respectively, satisfying stationary and invertibility conditions, and $\{W_t\}$ are independent and identically distributed normal random variables having mean zero and variance σ_w^2.

If there is no cyclic variation, but the underlying data generating process is still supposed nonstationary, then the ARIMA(p,d,q) process $\{Z_t\}$, defined by

$$\phi(B)(1-B)^d Z_t = \theta(B) W_t \tag{3}$$

may represent an appropriate model, while the stationary, short-memory process provided by (3) by letting $d = 0$ (no trend) is called an ARMA(p,q) process.

An ARFIMA(p,d,q) process $\{Z_t\}$ is defined, for $0 < |d| < 0.5$, again by the relation (3), where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = 0$ and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q = 0$ have all roots greater than one in absolute value. For $d > -1$, the operator $(1-B)^d$ in (3) is defined by the binomial expansion

$$(1-B)^d = 1 + \sum_{j=1}^{\infty} \pi_j B^j \,, \tag{4}$$

where

$$\pi_j = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} \tag{5}$$

$\Gamma(x)$ being the gamma function. Hence the π_j follow the simple recursion

$$\pi_j = \left(\frac{j-1-d}{j}\right) \pi_{j-1} \tag{6}$$

with $\pi_0 = 1$.

A particular special case is the fractionally integrated white noise process $\{F_t\}$, defined by

$$(1-B)^d F_t = W_t \,. \tag{7}$$

From the above definitions, an ARFIMA process $\{Z_t\}$ can be interpreted as an ARMA process driven by fractionally integrated white noise or as a process whose fractional difference is an ARMA process.

The classical approach for detecting the presence of long memory in a time series $\{z_t\}$ traces back to Hurst's work [1], where the rescaled range statistic (R/S) is introduced:

$$\frac{R}{S_N} = \frac{\max\limits_{1\le k\le N} \sum_{t=1}^{k} (z_t - \bar{z}_N) - \min\limits_{1\le k\le N} \sum_{t=1}^{k} (z_t - \bar{z}_N)}{S_N} \tag{8}$$

with

$$\bar{z}_N = \frac{\sum_{t=1}^{N} z_t}{N} \tag{9}$$

and

$$S_N = \sqrt{\frac{\sum_{t=1}^{N}(z_t - \bar{z}_N)^2}{N}}\,. \tag{10}$$

Hurst showed that for large values of N

$$N^{-H} \times \frac{R}{S_N} \to \text{const.}\,, \tag{11}$$

where H is known as the Hurst exponent. It is also referred to as the scaling exponent because self-similar processes and long-memory processes are highly connected. For any stationary process with short-range dependence, H is expected to be $1/2$. Therefore in this case, for large values of N, $\log(R/S_N)$ should be scattered around a straight line with slope $1/2$. Instead, an estimated slope greater than $1/2$ is taken as an indication of long-term memory and in this last case the differencing parameter is estimated as

$$d = H - 1/2\,. \tag{12}$$

Forecasting for ARFIMA processes is not as straightforward as for non-fractionally integrated processes because forecasts cannot be obtained directly from a finite order difference equation form. For the ARFIMA model, forecasts are derived using the infinite AR form of the process, *i.e.*

$$\pi^*(B) Z_t = W_t\,, \tag{13}$$

where

$$\pi^*(B) = 1 - \sum_{j=1}^{\infty} \pi_j^* B^j = \theta^{-1}(B)\phi(B)(1-B)^d\,. \tag{14}$$

Multiplying both sides of this relation by $\theta(B)$, it is possible to obtain the π_j^* coefficients necessary for the AR form (13) of the general ARFIMA process (3). In fact,

$$\theta(B)\pi^*(B) = \phi(B)(1-B)^d \tag{15}$$

that is to say, using (14) and (4)

$$\theta(B)\left(1 - \sum_{j=1}^{\infty} \pi_j^* B^j\right) = \phi(B)\left(1 + \sum_{j=1}^{\infty} \pi_j B^j\right)\,. \tag{16}$$

Now, recalling the expressions for the polynomials $\phi(B)$ and $\theta(B)$, the π_j^* coefficients may be obtained recursively by

$$\pi_j^* - \theta_1 \pi_{j-1}^* - \cdots - \theta_q \pi_{j-q}^* = -\pi_j + \phi_1 \pi_{j-1} + \cdots + \phi_p \pi_{j-p}\,, \tag{17}$$

where

$$\begin{aligned} \pi_0 &= 1\,, \\ \pi_0^* &= -1\,. \end{aligned} \tag{18}$$

The l steps ahead forecast of Z_{t+l} based on the infinite past observations starting at time t is

$$\hat{z}_t(l) = \sum_{j=1}^{\infty} \pi_j^* \hat{z}_t(l-j)\,. \tag{19}$$

3. A comparative study

In order to make clear the difficulties encountered in setting up a proper class of models for the stochastic process considered in this paper, hourly average wind speed, a particular time series will be examined in detail. It has been selected because there seems to be no clear generating model for it and series with such a characteristic are not rare in the examined sample consisting of 96 time series. Fig. 1 shows the HAWS time series recorded in Cammarata (Italy) in February 2005, while Fig. 2 displays the series transformed in order to adjust for the non-Gaussian distribution, along with its ac.f. and smoothed spectrum, the latter obtained using an opportune autoregressive model fitted to the data. The estimated spectrum does not

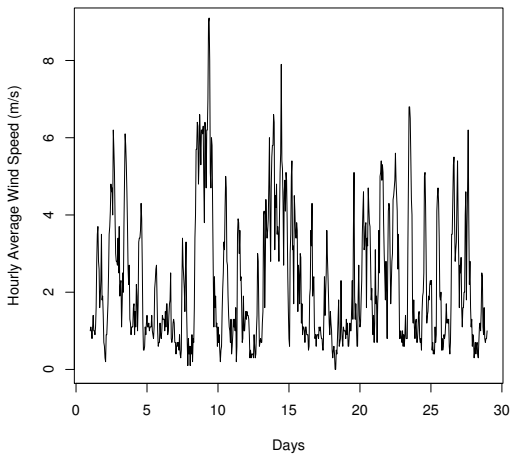

Fig. 1. HAWS (m/s) at 10 metres above ground in Cammarata, February 2005.

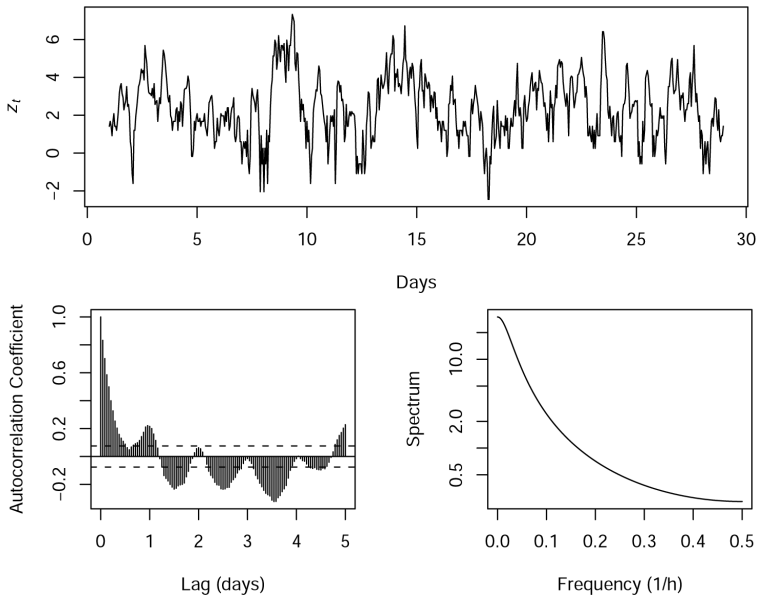

Fig. 2. Top — the time series shown in Fig. 1 after a nonlinear transformation. Bottom — autocorrelation function and smoothed spectrum of the series shown above.

reveal any periodicity. The autoregressive model selected for the estimation is simply an AR(1) model and this could be an indication that the investigated process is stationary. This hypothesis could be reinforced by looking at the time plot from which some qualitative features may be drawn. When one looks only at short time periods, then there seem to be cycles or local trends. However, looking at the whole series, there is no apparent persisting trend or cycle. It rather seems that cycles of different frequencies occur, superimposed and in random sequence. Overall, the series looks stationary. The interpretation of the ac.f. is much more difficult. The only clear feature is that the correlations do not die out quickly and this contradicts the hypothesis of stationarity unless a particular kind of stationary models — the long-memory models — is invoked. Moreover, looking at the correlations up to 2 days, a weak daily cycle could be detected, but for higher lags this possibility is not confirmed. Anticipating the results given below, it comes out that the examined process shares features common to different kinds of models. In particular, it could be considered a stationary process and a nonstationary process. Moreover, it could be a stationary short-memory process and a stationary long-memory process. Finally, it could be modelled by a nonstationary nonseasonal model and by a nonstationary seasonal model. Among all these fairly good candidates, there is only one clear winner, both in terms of modelling and forecasting.

This difficult selection of an appropriate data generating model for the observed time series must not surprise since not rarely it happens to be the case for other series. Support to this notion, that it is very difficult to determine the kind of stochastic process (stationary or not) on the basis of only one realization of its, is given in [9]. For a time series consisting of chemical process concentration readings, the Authors propose two possible ARMA/ARIMA models. Beran [1] also examined the same data and found that an ARFIMA model fits the series well. In the following, the four discussed kinds of models for the series shown in Fig. 2 will be given and used to forecast 24 hours ahead the original series in Fig. 1.

In order to assess the presence of long memory in the transformed series, the sample spectrum and the plot of the rescaled range statistic are displayed, respectively, in Fig. 3 and Fig. 4. Both tend to suggest that an ARFIMA model could be appropriate. More specifically, the spectrum grows at low frequencies and the values of $\log(R/S_N)$ versus $\log(N)$ are scattered around a straight line with slope greater than $1/2$, as expected

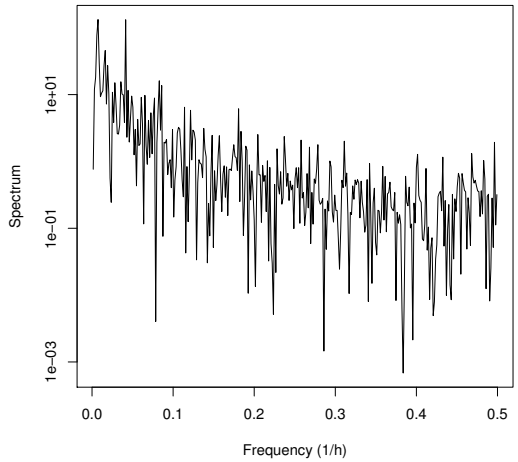

Fig. 3. Sample spectrum of the time series displayed in Fig. 2.

for long-memory processes (see Sec. 2). The slope of the estimated straight line is $H \approx 0.65$ and the derived differencing parameter is $d \approx 0.15$. The model selected by the AICc is ARFIMA(2,0.15,1) and the Ljung-Box-Pierce statistic reveals that the independence assumption for the residuals of the model fails beyond lag 35. Again upon a stationary assumption for the process, a class of short-memory models has been tested too, namely the class of ARMA models. The same information criterion provides in this case an ARMA(3,2) model as a possible candidate with the independence assumption for the residuals valid up to lag 30. Assuming instead that the underlying process is nonstationary, the following different kinds of models

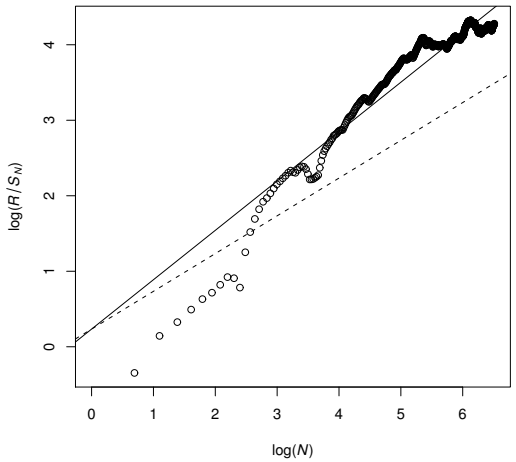

Fig. 4. Rescaled range analysis for the time series displayed in Fig. 2. The slope of the estimated straight line is $H \approx 0.65$, while the slope of the reference broken line is 1/2 (short-range dependence).

may be given, namely an ARIMA(1,1,1) and a SARIMA$(2,1,2)\times(0,1,2)_{24}$. In the first case the model fails the independence requirement for the residuals after lag 30, in the last case after lag 284. Fig. 5 displays the 24 hours ahead predictions obtained using each of the four developed models, while in Table I their quality is evaluated by four different indicators, namely

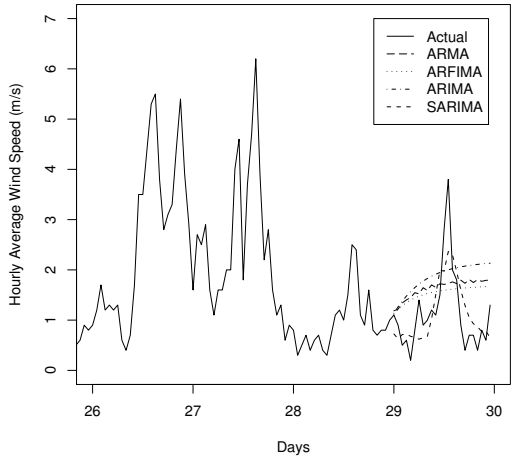

Fig. 5. 24 hours forecast of the time series shown in Fig. 1 by the models reported in Table I. Actual data displayed in the forecast window to assess forecasting accuracy have not been used to estimate the parameters of the models (out of sample forecasting).

TABLE I

The forecasting accuracy of the models developed for the considered case study.

Model	MAE (m/s)	RMSE (m/s)	MAPE	MASE
ARMA(3,2)	0.78	0.92	119.93	1.25
ARFIMA(2,0.15,1)	0.72	0.87	109.18	1.15
ARIMA(1,1,1)	0.93	1.06	144.43	1.49
SARIMA$(2,1,2)\times(0,1,2)_{24}$	0.43	0.53	52.42	0.69

the mean absolute error (MAE), the root mean square error (RMSE), the mean absolute percentage error (MAPE) and the mean absolute scaled error (MASE). The obtained results illustrate a fundamental point in time series analysis. The more dependence there is among past observations, the better a future observation can be predicted, provided that the existing dependence structure is exploited appropriately. In other words, to obtain good forecasts it is necessary to use good models. In the considered case study all four indicators agree on the ranking of the employed models. The winner is the SARIMA model, followed by the ARFIMA, then by the ARMA and finally by the ARIMA model. The predictions obtained by the last three look very similar, in that they all pass in the middle of the actual data, but they present different features. In fact, the forecast based on the ARIMA model diverges for greater lead times, while the prediction by the ARMA process converges very fast to the sample mean of the past observations and using the sample mean corresponds to total ignorance about the future observations (except for their unconditional expected value). On the other hand, the prediction based on the ARFIMA model converges more slowly and this means that past observations influence the forecasts even far into the future, as it is expected for a long-memory process. However, in view of the obtained results, it is clear that the analysed process is properly described only by the SARIMA model which has not been selected on a statistical basis, but upon the knowledge of the physical properties of the investigated phenomenon.

4. Conclusion

The predictions yielded by the ARFIMA models developed for the series analysed in this work clash with their modelling performance, as demonstrated by the reported case study. Given that predictions as that in Fig. 5 have been obtained also by Beran [1] and that many previous researches have raised doubts on the effectiveness of the ARFIMA specification as a forecasting tool, the validity of these models for HAWS should be further investigated.

REFERENCES

[1] J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, 1994.
[2] X. Zhang, J. Liu, J. Liu, B. Liu, *Math. Stat. Manage.* **20**, 1 (2001).
[3] R.G. Kavasseri, K. Seetharaman, *Renew. Energ.* **34**, 1388 (2009).
[4] V.A. Reisen, S. Lopes, *J. Statist. Plann. Inference* **80**, 269 (1999).
[5] C. Ellis, P. Wilson, *Int. Rev. Finan. Anal.* **13**, 63 (2004).
[6] J. Xiu, Y. Jin, *Physica A* **377**, 138 (2007).
[7] D. Lindley, *Nature* **463**, 18 (2010).
[8] M.S. Roulston, D.T. Kaplan, J. Hardenberg, L.A. Smith, *Renew. Energ.* **28**, 585 (2003).
[9] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, *Time Series Analysis*, Wiley, 2008.