First and second order semi-Markov chains for wind speed modeling

Guglielmo D'Amico[1], Filippo Petroni[2], and Flavio Prattico[3]

[1] Dipartimento di Scienze del Farmaco, Università G. d'Annunzio, 66013 Chieti, Italy, (E-mail: g.damico@unich.it))
[2] Dipartimento di Scienze Economiche e Aziendali, Università degli studi di Cagliari, 09123 Cagliari, Italy, (E-mail: fpetroni@unica.it))
[3] Dipartimento di Ingegneria Meccanica, Energetica e Gestionale, Università degli studi dell'Aquila, 67100 L'Aquila, Italy, (E-mail: flavioprattico@gmail.com)

Abstract. The increasing interest in renewable energy, particularly in wind, has given rise to the necessity of accurate models for the generation of good synthetic wind speed data. Markov chains are often used with this purpose but better models are needed to reproduce the statistical properties of wind speed data. We downloaded a database, freely available from the web, in which are included wind speed data taken from L.S.I. -Lastem station (Italy) and sampled every 10 minutes. With the aim of reproducing the statistical properties of this data we propose the use of three semi-Markov models. We generate synthetic time series for wind speed by means of Monte Carlo simulations. The time lagged autocorrelation is then used to compare statistical properties of the proposed models with those of real data and also with a synthetic time series generated though a simple Markov chain.

Keywords: semi-Markov chains, synthetic time series, autocorrelation.

1 Introduction

The increasing interest in renewable energy leads scientific research to find a better way to recover most of the available energy. Particularly, the maximum energy recoverable from wind is equal to 59.3% of that available (Betz law) at a specific pitch angle and when the ratio between the wind speed in output and in input is equal to 1/3. A powerful system control of the pitch angle allows the wind turbine to recover better the energy in transient regime. A good stochastic model for wind speed is then needed to help both the optimization of turbine design and to assist the system control to predict the value of the wind speed to positioning the blades quickly and correctly. The possibility to have synthetic data of wind speed is a powerful instrument to assist designer to verify the structures of the wind turbines or to estimate the energy recoverable from a specific site. To generate synthetic data, Markov chains of first or higher order are often used [1–3] but the search for a better model is still open. Approaching this issue, we applied new models which are generalization of Markov models. More precisely we applied first and second order semi-Markov models to generate synthetic wind speed time series.

The models are used to generate synthetic time series for wind speed by means of Monte Carlo simulations and the time lagged autocorrelation is used to compare statistical properties of the proposed models with those of real data and also with a time series generated though a simple Markov chain.

2 Wind speed modeling with semi-Markov chains

Semi-Markov chains are a generalization of Markov chains allowing the times between transitions to occur at random times according to any kind of distribution functions which may depend on the current and the next visited state. As it is well known, Markov chains have sojourn times between transitions geometrically distributed, for this reason the memoryless property is preserved and no duration effect is observed. The more general semi-Markov environment allows the possibility to use also no memoryless distributions and then can reproduce a duration effect. The duration effect affirms that the time the system is in a state influences its transition probabilities. The states of the process in our data are represented by different speed data, then in this paper we detect the presence of a duration effect in wind speed modeling and forecasting.

Here below we propose a semi-Markov model of order two in state and duration and we compare its performance with the Markov chain models often used to describe wind speed, see [1–3] and with some particular cases of semi-Markov chain models.
Let consider a finite set of states $E = \{1, 2, ..., S\}$ in which the system can be into and a complete probability space (Ω, F, P) on which we define the following random variables:

$$J_n : \Omega \to E, \;\; T_n : \Omega \to \mathbb{N}. \tag{1}$$

They denote the state occupied at the n-th transition and the time of the n-th transition, respectively. To be more concrete, by J_n we denote the wind speed at the nth transition and by T_n the time of the nth transition in the wind speed. We do the following conditional independence assumption:

$$\begin{aligned} &P[J_{n+1} = j, T_{n+1} - T_n = t | \sigma(J_s, T_s), J_n = k, J_{n-1} = i, T_n - T_{n-1} = x, 0 \le s \le n] \\ &\quad = P[J_{n+1} = j, T_{n+1} - T_n = t | J_n = k, J_{n-1} = i, T_n - T_{n-1} = x] := {}_x q_{i.k,j}(t). \end{aligned} \tag{2}$$

Relation (2) asserts that, the knowledge of the values $J_n, J_{n-1}, T_n - T_{n-1}$ suffices to give the conditional distribution of the couple $J_{n+1}, T_{n+1} - T_n$ whatever the values of the past variables might be. Therefore to make probabilistic forecasting we need the knowledge of the last two visited state and the duration time of the transition between them. For this reason we called this model a second order semi-Markov chains in state and duration.

The conditional probabilities

$${}_x q_{i.k,j}(t) = P[J_{n+1} = j, T_{n+1} - T_n = t | J_n = k, J_{n-1} = i, T_n - T_{n-1} = x]$$

are stored in a matrix of functions $\mathbf{q} = ({}_{x}q_{i.k,j}(t))$ called the second order kernel (in state and duration). The element ${}_{x}q_{i.k,j}(t)$ represents the probability that next wind speed will be in speed j at time t given that the current wind speed is k and the previous wind speed state was i and the duration in wind speed i before of reaching wind speed k was equal to x units of time.

We can define the cumulated second order kernel:

$$\begin{aligned} {}_{x}Q_{i.k,j}(t) &:= P[J_{n+1} = j, T_{n+1} - T_n \leq t | J_n = k, J_{n-1} = i, T_n - T_{n-1} = x] \\ &= \sum_{s=1}^{t} {}_{x}q_{i.k,j}(s). \end{aligned} \tag{3}$$

The process $\{J_n\}$ is a second order Markov chain with state space E and transition probability matrix ${}_{x}\mathbf{P} = {}_{x}\mathbf{Q}(\infty)$. We shall refer to it as the embedded Markov chain.

Define the unconditional waiting time distribution function in states i and k with duration x as

$${}_{x}H_{i.k}(t) := P[T_{n+1} - T_n \leq t | J_n = k, J_{n-1} = i, T_n - T_{n-1} = x] = \sum_{j \in E} {}_{x}Q_{i.k,j}(t). \tag{4}$$

The conditional cumulative distribution functions of the waiting time in each state, given the state subsequently occupied is defined as

$$\begin{aligned} {}_{x}G_{i.k,j}(t) &= P[T_{n+1} - T_n \leq t | J_n = k, J_{n-1} = i, J_{n+1} = j, T_n - T_{n-1} = x] \\ &= \frac{1}{{}_{x}p_{i.k,j}} \sum_{s=1}^{t} {}_{x}q_{i.k,j}(s) \cdot 1_{\{{}_{x}p_{i.k,j} \neq 0\}} + 1_{\{{}_{x}p_{i.k,j} = 0\}} \end{aligned} \tag{5}$$

Define by $N(t) = \sup\{n : T_n \leq t\}\ \forall t \in \mathbb{N}$. We define the second order (in state and duration) semi-Markov chain as $Z(t) = (Z^1(t), Z^2(t)) = (J_{N(t)-1}, J_{N(t)}))$.

If we define, $\forall i, k, j \in E$, and $t \in \mathbb{N}$, the semi-Markov transition probabilities:

$${}_{x}\phi_{i.k,h.j}(t) := P[J_{N(t)} = j, J_{N(t)-1} = h | J_0 = k, J_{-1} = i, T_0 = 0, T_0 - T_{-1} = x], \tag{6}$$

then the following system of equations is verified:

$${}_{x}\phi_{i.k,h.j}(t) = 1_{\{i=h, k=j\}}(1 - {}_{x}H_{i.k}(t)) + \sum_{r \in E} \sum_{s=1}^{t} {}_{x}q_{i.k,r}(s)\, {}_{s}\phi_{k.r,h.j}(t-s). \tag{7}$$

The proof of equation (7) is not given here because it is a particular case of the equation established and proved later.

To detect the duration effects let us introduce the backward recurrence time process defined for each time $t \in \mathbb{N}$ by:

$$B(t) = t - T_{N(t)}. \tag{8}$$

If the semi-Markov process $Z(t)$ indicates the wind speed at time t, the backward process $B(t)$ indicates the time since the last transition, that is from how long time the wind speed is at the value $Z(t)$.

To quantify the duration effect in our second order semi-Markov model, let us define the following probabilities:

$$
\begin{aligned}
&{}_{x}^{b}\phi_{i.k,h.j}^{b}(v;v',t) := \\
&P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h | J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x].
\end{aligned}
\tag{9}
$$

Expression (10) gives the probability that the wind speed will enter in the state j at time $t-v'$ coming from state h and will remains inside the state j without any other transition up to the time t given that at the present the wind speed is k and it entered into this state with the last transition v periods before coming from a wind speed equal to i with a duration in i of x periods.

Proposition 1. *The relation* (10) *represents the evolution equation of* (9)*:*

$$
\begin{aligned}
{}_{x}^{b}\phi_{i.k,h,j}^{b}(v;v',t) &= 1_{\{i=h,k=j,v'=t+v\}} \frac{[1-{}_{x}H_{i.k}(t+v)]}{[1-{}_{x}H_{i.k}(v)]} \\
&+ \sum_{r\in E}\sum_{s=1}^{t-v'} \frac{{}_{x}q_{i.k,r}(s+v)}{[1-{}_{x}H_{i.k}(v)]}\ {}_{s+v}^{b}\phi_{k.r,h.j}^{b}(0;v',t-s).
\end{aligned}
\tag{10}
$$

Proof: We have

$$
\begin{aligned}
&{}_{x}^{b}\phi_{i.k,h,j}^{b}(v;v',t) \\
&= P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h, T_1>t | J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x] \\
&+ P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h, T_1\le t | J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x].
\end{aligned}
\tag{11}
$$

Observe that

$$
\begin{aligned}
&P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h, T_1>t | J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x] \\
&= P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h | T_1>t, J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x] \\
&\cdot P[T_1>t | J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x].
\end{aligned}
\tag{12}
$$

If $T_1>t$ then $J_{N(t)}=J_0$, $J_{N(t)-1}=J_{-1}$ and $B(t)=v'=v+t$. This gives:

$$
\begin{aligned}
&P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h, T_1>t | J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x] \\
&= P[k=j, v'=v+t, i=h | T_1>t, J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x] \\
&\cdot P[T_1>t | J_0=k, J_{-1}=i, T_0-T_{-1}=x, T_0=-v, T_1>0] \\
&= 1_{\{k=j,i=h,v'=v+t\}} \frac{P[T_1>t | J_0=k, J_{-1}=i, T_0-T_{-1}=x, T_0=-v]}{P[T_1>0 | J_0=k, J_{-1}=i, T_0-T_{-1}=x, T_0=-v]} \\
&= 1_{\{k=j,i=h,v'=v+t\}} \frac{1-{}_{x}H_{i.k}(t+v)}{1-{}_{x}H_{i.k}(v)}.
\end{aligned}
\tag{13}
$$

The second addend on the right hand side of (11) can be represented as follows:

$$\sum_{r\in E}\sum_{s=1}^{t-v'} P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h, J_1=r, T_1=s | J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x]$$

$$\sum_{r\in E}\sum_{s=1}^{t-v'} P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h | J_1=r, T_1=s, J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x]$$

$$\cdot P[J_1=r, T_1=s | J_0=k, J_{-1}=i, B(0)=v, T_0-T_{-1}=x]$$

$$= \sum_{r\in E}\sum_{s=1}^{t-v'} P[J_{N(t)}=j, B(t)=v', J_{N(t)-1}=h | J_1=r, J_0=k, B(s)=0, T_1-T_0=s+v]$$

$$\cdot P[J_1=r, T_1-T_0=s+v | J_0=k, J_{-1}=i, T_0=-v, T_1>0, T_0-T_{-1}=x]$$

$$= \sum_{r\in E}\sum_{s=1}^{t-v'} {}_{s+v}^{b}\phi_{k.r,h.j}^{b}(0;v',t-s)\frac{P[J_1=r, T_1-T_0=s+v | J_0=k, J_{-1}=i, T_0-T_{-1}=x]}{P[T_1-T_0>v | J_0=k, J_{-1}=i, T_0-T_{-1}=x]}$$

$$\sum_{r\in E}\sum_{s=1}^{t-v'} \frac{{}_{x}q_{i.k,r}(s+v)}{[1-{}_{x}H_{i.k}(v)]}\ {}_{s+v}^{b}\phi_{k.r,h.j}^{b}(0;v',t-s). \tag{14}$$

A substituion of (13) and (14) in (11) concludes the proof.

Obviously we have that

$$\begin{aligned} &{}_{x}^{b}\phi_{i.k,h.j}(v;t) \\ &:= P[J_{N(t)}=j, J_{N(t)-1}=h | J_0=k, J_{-1}=i, B(0)=v, T_0=0, T_0-T_{-1}=x] \\ &= \sum_{v'\geq 0} {}_{x}^{b}\phi_{i.k,h.j}^{b}(v;v',t). \end{aligned} \tag{15}$$

It expresses the probability that the wind speed will be in the state j at time t coming from a wind speed equal to h given that at present the wind speed is k and it entered into this state with the last transition v periods before coming from a wind speed equal to i with a duration in i of x periods.

Moreover if $v=0$ we obtain the equation (7).

It should be noted that our semi-Markov model of order two in state and duration contains as a special case the semi-Markov model of order two in state. The paper [4] proposed a n-order semi-Markov process (in state) in continuous time. The discrete time counterpart of order two (in state) is obtained through the following assumption:

$$\begin{aligned} &P[J_{n+1}=j, T_{n+1}-T_n=t | \sigma(J_s,T_s), J_n=k, J_{n-1}=i, 0\leq s\leq n] \\ &\quad = P[J_{n+1}=j, T_{n+1}-T_n=t | J_n=k, J_{n-1}=i] := q_{i.k,j}(t). \end{aligned} \tag{16}$$

Relation (16) asserts that, the knowledge of the values J_n, J_{n-1} suffices to give the conditional distribution of the couple $J_{n+1}, T_{n+1}-T_n$ whatever

the values of the past variables might be. Therefore to make probabilistic forecasting we need the knowledge of the last two visited state. In the application we will refer to this model as the model named semi-Markov II.

Finally remark that if we assume that

$$\begin{aligned} &P[J_{n+1}=j, T_{n+1}-T_n=t|\sigma(J_s,T_s), J_n=i, 0\leq s\leq n] \\ &\quad = P[J_{n+1}=j, T_{n+1}-T_n=t|J_n=i] := q_{ij}(t). \end{aligned} \tag{17}$$

then we recover the classical semi-Markov chain model.

3 Application to real data

To check the validity of our model we perform a comparison of the behavior of real data and wind speeds generated through Monte Carlo simulations based on the models described above. In this section we describe the database of real data used for the analysis, the method used to simulate synthetic wind speed time series and, at the end, we compare results from real and simulated data.

The data used in this analysis are freely available from $http://www.lsi-lastem.it/meteo/page/dwnldata.aspx$. The station of L.S.I. -Lastem is situated in Italy at N 45 28' 14,9" − E 9 22' 19,9" and at 107 m of altitude. The station use a combined speed-direction anemometer at 22 m above the ground. It has a measurement range that goes from 0 to 60 m/s, a threshold of 0,38 m/s and a resolution of 0,05 m/s. The station processes the speed every 10 minute in a time interval ranging from 25/10/2006 to 28/06/2011. During the 10 minutes are performed 31 sampling which are then averaged in the time interval. In this work, we use the sampled data that represents the average of the modulus of the wind speed (m/s) without considered a specific direction. The database is then composed of about 230thousands wind speed measures ranging from 0 to 16 m/s. The time series, together with its empirical probability density distribution are represented in Figure 1.

To be able to model the wind speed as a semi-Markov process the state space of wind speed has been discretized. In the example shown in this work we discretized wind speed into 7 states chosen to cover all the wind speed distribution.

From the discretized wind speeds we estimated the probabilities $\mathbf{P}$ and G to generate synthetic trajectories for three semi-Markov models: a simple semi-Markov model of the first order named semi-Markov I, semi-Markov II a second order semi-Markov model in state (as described in section 2) and the second order semi-Markov model in state and duration is named semi-Markov III.

For comparison reason, we also generated a synthetic trajectory which follows a simple Markov model with transition probability matrix estimated

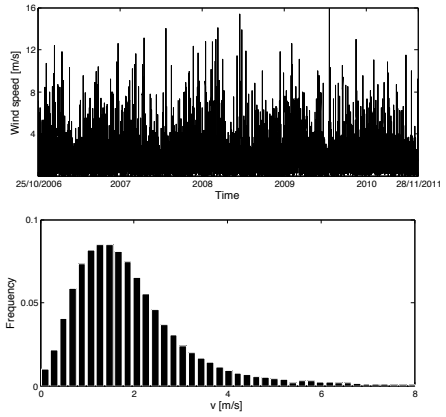

Fig. 1. Time series of wind speed and its empirical distribution.

from the real data. We then ended up with five trajectories: one representing real data, three representing the three trajectories according to the three semi-Markov models and the last one a Markov chain. The three time series are used in the following to compare results on the time lagged autocorrelation function. Real data are, in fact, strongly autocorrelated and the autocorrelation function decreases rapidly with time. We then tested our models to check whether they are able to reproduce such behavior.

If w indicates wind speed, the time lagged (τ) autocorrelation of wind speed is defined as

$$\Sigma(\tau) = \frac{Cov(w(t+\tau), w(t))}{Var(w(t))} \tag{18}$$

The time lag τ was made to run from 1 minute up to 1000 minutes. Note that to be able to compare results for $\Sigma(\tau)$ each simulated time series was generated with the same length as real data. Results shown in Figure 2 indicate that semi-Markov models reproduce better the autocorrelation present in real data especially if a second order semi-Markov chain is used to generate synthetic data. We can also notice that the second order semi-Markov model in state and duration (Model III) has to be preferred to the second order semi-Markov model in state (Model II) because it exhibits a slightly higher autocorrelation uniformly in time and also because, asymptotically, its autocorrelation has the same slope of that observed in real data whereas in the Model II, the autocorrelation drops to zero. Although the evidence shows the semi-Markovian nature of the studied phenomenon, probably a third/fourth order semi-Markov chain would be needed to decrease the difference between autocorrelation of real and simulated data. In our view this approach would be too much computationally and data consuming and further research on a simplified, but still with longer memory, model is needed.

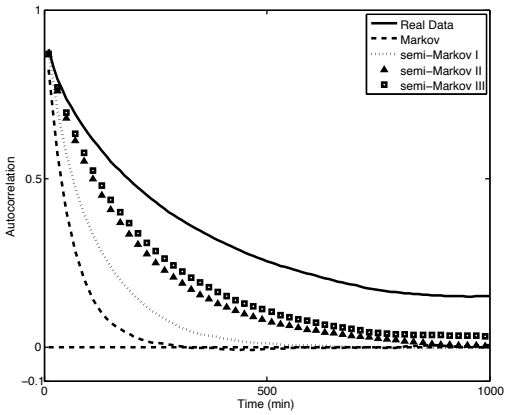

Fig. 2. Time lagged autocorrelation function.

4 Conclusions

Wind speed is a stochastic process for which a completely satisfactory model is still lacking. Many authors have used Markov chain to model the process but this approach does not give the same persistence present in real data. We presented in this work three semi-Markov models with the aim of generate synthetic wind speed data. We have showed that all our models perform better than a simple Markov chain in reproducing the statistical properties of wind speed data. In particular, the model that we recognized as being the more suitable is our third model which is a second order semi-Markov process in state and duration. We conclude that semi-Markov models should be used when dealing with wind speed data.

References

1. A. Shamshad, M.A. Bawadi, W.M.W. Wan Hussin, T.A. Majid, S.A.M. Sanusi, First and second order Markov chain models for synthetic generation of wind speed time series, *Energy* **30** (2005) 693-708.
2. H. Nfaoui, H. Essiarab, A.A.M. Sayigh, A stochastic Markov chain model for simulating wind speed time series at Tangiers, Morocco, *Renewable Energy* **29** (2004) 1407-1418.
3. F. Youcef Ettoumi, H. Sauvageot, A.-E.-H. Adane, Statistical bivariate modeling of wind using first-order Markov chain and Weibull distribution, *Renewable Energy* **28** (2003) 1787-1802.
4. N. Limnios, G. Oprişan. *An introduction to Semi-Markov Processes with Application to Reliability*, In D.N. Shanbhag and C.R. Rao, eds., Handbook of Statistics, **21**, (2003), 515-556.