Confidence intervals for risk indicators in semi-Markov models: an application to wind energy production

Irene Votsi, Alexandre Brouste

Confidence intervals for risk indicators in semi-Markov models: an application to wind energy production

I. Votsi* and A. Brouste

Laboratoire Manceau de Mathématiques
Université du Maine

Abstract

Mean times to failure are fundamental indicators in reliability and related fields. Here we focus on the conditional mean time to failure defined in a semi-Markov context. A discrete time semi-Markov model with discrete state space is employed, which allows for realistic description of systems under risk. Our main objective is to estimate the conditional mean time to failure and provide asymptotic properties of its nonparametric estimator. Consistency and asymptotic normality results are provided. Our methodology is tested in a real wind dataset and indicators associated with the wind energy production are estimated.

1 Introduction

Semi-Markov models (SMMs) are stochastic models that are widely used in reliability, survival analysis, seismology and many other scientific fields. One of the main features of SMMs is that contrary to Markov models, they describe systems whose evolution is based not only on their last visited state but also on the elapsed time since this state. Due to this feature, popular

*Email: Eirini.Votsi@univ-lemans.fr; Corresponding author

"memory-full" lifetime distributions, such as the Weibull distribution, could be employed in a semi-Markov framework.

Many reliability indicators have been introduced for SMMs including hazard rates, availability and maintainability functions, mean times to failure *etc*. For advances in the topic concerning continuous time SMMs, see Limnios and Oprişan (2001) and Limnios and Ouhbi (2006). For discrete time SMMs, we address the interested reader to Votsi et al. (2014), Votsi and Limnios (2015), Barbu et al. (2016), Georgiadis et al. (2013) and Georgiadis (2017). Mean times to failure have been studied for continuous time SMMs by Limnios and Ouhbi (2006), among other reliability indicators. The authors introduced empirical, plug-in type estimators and studied their asymptotic properties. Here we focus on the discrete time case and study the consistency and asymptotic normality of the estimators of the conditional mean times to failure.

Markov models for wind modeling have been studied in numerous papers (see Tang et al., 2015, and references therein). The more general SMMs were first applied in the modeling of wind speed by D'Amico et al. (2013a) in order to generate synthetic wind speed data. Indexed SMMs were introduced by D'Amico et al. (2013c) to reproduce the autocorrelation of the speed data. However, the energy production of the wind farm deeply relies not only on the wind speed but also on the wind direction on the site through its power transfer function. There exists only very few models describing simultaneously the evolution of the wind speed and the wind direction (see Ailliot et al., 2015; Ettoumi et al., 2003; Masala (2014) for instance). There is thus a need for models which can reproduce both wind speed and wind direction data.

First, we aim to fill the aforementioned gap in a semi-Markov context. We propose SMMs as ready-to-use models guiding the decision-makers to identify and manage risks related to the wind energy production. Second, we focus on the operational management of a wind farm. To achieve that goal we extend previous reliability studies (see D'Amico et al., 2013b) by evaluating risk indicators that reflect the time-dependent nature of the risk. These indicators may help to signal a change in the level of risk exposure. We provide theoretical findings that enable us to construct confidence intervals for the respective estimators. Third, we make use of both real data coming from the private company EREN and simulated data coming from the National Renewable Energy Laboratory (NREL, USA). In the previous studies, the states of the SMMs correspond to wind speed data divided into a

fixed number of equal length. Here we propose a novel form of classification of states which is more coherent with the operational interest of the wind energy production.

The remainder of the paper is organized as follows: In Section 2 the notation and preliminaries of discrete time SMMs are presented. Section 3 describes risk indicators in a semi-Markov context. The relative statistical estimation aspects are specialized in Section 3.2 and Section 3.4 depicts a numerical example that enables us to validate our theoretical results. Section 4 concerns wind energy production management via SMMs. Section 4.1 discusses a goodness-of-fit test to provide evidence for the necessity of a more general model than the Markov model. In Section 4.2, the aforementioned risk indicators are evaluated to provide feedback for the management of the wind energy production. Finally, in Section 5, we give some concluding remarks.

2 Notation and preliminaries of SMMs

We briefly recall the main definitions from the theory of semi-Markov chains (see, e.g., Barbu and Limnios, 2008). Consider a random system with finite state space $E = \{1, 2, ..., s\}$. Let $\mathbb{N}$ be the set of nonnegative integers. We suppose that the evolution in time of the system is described by the following random sequences defined on a complete probability space:

1. The Markov chain $\mathbf{J} = (J_n)_{n\in\mathbb{N}}$ with state space E, where J_n is the system's state at the $n-$th jump time;

2. The $\mathbb{N}$-valued sequence $\mathbf{S} = (S_n)_{n\in\mathbb{N}}$, where S_n is the $n-$th jump time. We suppose that $S_0 = 0$ and $0 < S_1 < S_2 < \ldots < S_n < S_{n+1} < \ldots$ almost surely (a.s.);

3. The $\mathbb{N}$-valued sequence $\mathbf{X} = (X_n)_{n\in\mathbb{N}}$ defined by $X_0 = 0$ a.s. and $X_n = S_n - S_{n-1}$ for all $n \in \mathbb{N}$. Thus for all $n \in \mathbb{N}$, X_n is the sojourn time in state J_{n-1}, before the $n-$th jump.

The Markov renewal chain $(\mathbf{J}, \mathbf{S})$ (Fig. 1) is considered to be (time) homogeneous and is characterized by the semi-Markov kernel $\mathbf{q} = (q_{ij}(k);\ i, j \in E, k \in \mathbb{N})$ defined by

$$q_{ij}(k) = P(J_{n+1} = j, X_{n+1} = k | J_n = i),$$

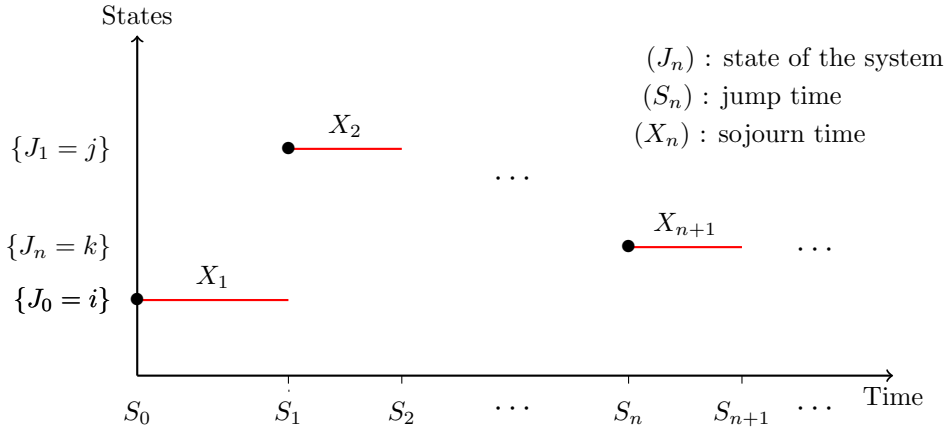

Figure 1: A typical trajectory of a Markov renewal chain.

where $i, j \in E$ and $k, n \in \mathbb{N}$. The process $\mathbf{J}$ is the embedded Markov chain (EMC) of the Markov renewal chain (MRC) with transition kernel $\mathbf{P} = (p_{ij};\ i, j \in E)$, where $p_{ij} = P(J_{n+1} = j | J_n = i)$, and initial distribution $\boldsymbol{\alpha} = (\alpha(i);\ i \in E)$, where $\alpha(i) = P(J_0 = i)$. Let us notice that we allow neither instantaneous transitions ($q_{ij}(0) = 0,\ i, j \in E$) nor self-transitions ($p_{ii} = 0,\ i \in E$).

We further denote the counting process of jumps $N(k) = \max\{n \geq 0 : S_n \leq k\}$, $k \in \mathbb{N}$. The semi-Markov chain $\mathbf{Z} = (Z_k)_{k \in \mathbb{N}}$ associated with the MRC $(\mathbf{J}, \mathbf{S})$ is defined by $Z_k = J_{N(k)}$, $k \in \mathbb{N}$. At the initial time, $t = 0$, the state of the semi-Markov chain coincides with the state of the EMC, *i.e.* $Z_0 = J_0$. We define the sojourn time distribution in state $i \in E$ by $h_i(k) = P(X_{n+1} = k | J_n = i) = \sum_{j \in E} q_{ij}(k)$, $k \in \mathbb{N}$, and the conditional distribution of the sojourn time in state $i \in E$ given that the next visited state is $j \in E$ by $f_{ij}(k) = P(X_{n+1} = k | J_n = i, J_{n+1} = j)$, $k \in \mathbb{N}$. We further define the cumulative sojourn time distribution in state $i \in E$, denoted $H_i(\cdot)$, by $H_i(\ell) = P(X_{n+1} \leq \ell | J_n = i)$, $\ell \in \mathbb{N}$, and the corresponding survival function, denoted $\overline{H}_i(\cdot)$, by $\overline{H}_i(\ell) = P(X_{n+1} > \ell | J_n = i)$, $\ell \in \mathbb{N}$. In addition we denote by $(S_n^i)_{n \in \mathbb{N}}$ the successive passage times in a fixed state $i \in E$ and by $\mu_{ii} = \mathbb{E}_i(S_1^i)$ the mean recurrence time in state i.

3 Risk indicators in SMMs

Our main objective is to provide reliability indicators associated with the wind energy production in order to identify opportunities for improving the reliability performance within the wind industry. These indicators can guide industry action to improve operating practices. In particular, we focus on some key risk indicators: the mean times to failure, the mean up and down times.

3.1 Conditional mean times to failure

First, we consider that the system of interest may be in either an operational (up) state, that belongs to $U = \{1, \ldots, s_1\}$, or an a failure (down) state belonging to $D = \{s_1 + 1, \ldots, s\}$, with $0 < s_1 < s$. Moreover, $E = U \cup D$ and $U \cap D = \varnothing$, $U, D \neq \varnothing$. One can think of the states of U as different operating modes or performance levels of the system. On the other hand the states of D can be seen as failures of the system with different modes. Up and down states associated to wind energy production are described in Section 4.

Second, we define by $\mathbf{m} = (m_1, \ldots, m_s)^\top$ the column vector of mean sojourn times, where $m_i = \mathbb{E}(S_1 | Z_0 = i)$ represents the mean sojourn time of the SMC in state $i \in E$. The next assumptions have to be fulfilled in the following:

A1 The SMC is irreducible and aperiodic (see Barbu and Limnios, 2008);

A2 The mean sojourn times are finite, *i.e.*, $m_i < \infty$, for any $i \in E$.

The two previous assumptions ensure the existence of the stationary distribution of the SMC, which is a function of the stationary distribution of the EMC denoted by $\boldsymbol{\nu} = (\nu(i);\ i \in E)$, s.t. $\nu(j) = \sum_{i \in E} \nu(i) p_{ij}$ and $\sum_{i \in E} \nu(i) = 1$.

In the sequel, we consider the partitions of $\mathbf{P}$, $\mathbf{m}$ and $\boldsymbol{\nu}$ to the subsets of states U and D

$$\mathbf{P} = \begin{array}{c} \begin{array}{cc} U & D \end{array} \\ \begin{pmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{pmatrix} \end{array} \begin{array}{c} U \\ D \end{array},$$

and

$$\mathbf{m} = \begin{array}{c} \begin{array}{cc} U & D \end{array} \\ \begin{pmatrix} \mathbf{m_1} & \mathbf{m_2} \end{pmatrix} \end{array}, \quad \boldsymbol{\nu} = \begin{array}{c} \begin{array}{cc} U & D \end{array} \\ \begin{pmatrix} \boldsymbol{\nu_1} & \boldsymbol{\nu_2} \end{pmatrix} \end{array}.$$

We recall that the system starts working at time $k = 0$. We denote by T_D the first passage time in subset D, *i.e.*

$$T_D := \inf\{k \in \mathbb{N} : Z_k \in D\} \quad \text{and} \quad \inf\{\varnothing\} = \infty.$$

We introduce the conditional mean time to failure, $MTTF_i = \mathbb{E}(T_D | J_0 = i)$, for any starting state $i \in E$. We further denote by

$$\mathbf{MTTF} = (MTTF_1, \ldots, MTTF_{s_1}),$$

the column vector of the conditional mean times to failure. The Markov renewal equation associated to the conditional mean time to failure is given by (Barbu and Limnios, 2008)

$$MTTF_i = m_i + \sum_{j \in U} p_{ij} MTTF_j.$$

or equivalently,

$$\mathbf{MTTF} = (\mathbf{I} - \mathbf{P}_{11})^{-1} \mathbf{m_1}.$$

Remark 1. *For the matrix* $(\mathbf{I} - \mathbf{P}_{11})$ *to be non-singular, the spectral radius of the matrix* $\mathbf{P}_{11}$ *should be smaller than* 1*, which means that there exists* $i \in \{1, \ldots, s_1\}$ *such that* $\sum_{j \in U} \mathbf{P}_{11}(i, j) < 1$*. In the following, the matrix* $(\mathbf{I} - \mathbf{P}_{11})$ *is assumed to be non-singular.*

3.2 Statistical estimation

In the literature concerning statistical inference of stochastic processes, there are two types of observational procedures. One either observes a single realization over the fixed time interval $[0, M]$, or observes $K > 1$ independent and identical copies of the process, each over the fixed duration $[0, M]$. In this study, we follow the first procedure and consider a single realization of the MRC $(\mathbf{J}, \mathbf{S})$, censored at fixed arbitrary time $M \in \mathbb{N}$,

$$\mathcal{H}(M) = (J_0, S_1, ..., J_{N(M)-1}, S_{N(M)}, J_{N(M)}, U_M),$$

where $U_M = M - S_{N(M)}$ is the censored sojourn time in the last visited state $J_{N(M)}$. For all $i, j \in E$ and $1 \leq k \leq M$, we define:

1. $N_{ij}(k,M) = \sum_{n=1}^{N(M)} \mathbf{1}_{\{J_{n-1}=i,J_n=j,X_n=k\}}$ is the number of transitions of the EMC from i to j, up to time M, with sojourn time in state i equal to k;

2. $N_{ij}(M) = \sum_{n=1}^{N(M)} \mathbf{1}_{\{J_{n-1}=i,J_n=j\}}$ is the number of transitions of the EMC from i to j, up to time M;

3. $N_i(M) = \sum_{n=1}^{N(M)-1} \mathbf{1}_{\{J_n=i\}}$ is the number of visits to state i of the EMC, before time M.

Given a sample path $\mathcal{H}(M)$ of the MRC, we define the empirical estimators

$$\widehat{p}_{ij}(M) = \frac{N_{ij}(M)}{N_i(M)} \quad \text{and} \quad \widehat{q}_{ij}(k,M) = \frac{N_{ij}(k,M)}{N_i(M)},$$

for all $i,j \in E$ and $k \in \mathbb{N}$, $k \leq M$. Then for all $i \in E$, $\ell \in \mathbb{N}$, the empirical estimators of $H_i(\ell)$ and $\overline{H}_i(\ell)$ are given by

$$\widehat{H}_i(\ell,M) = \sum_{j\in E}\sum_{k=0}^{\ell} \widehat{q}_{ij}(k,M) \quad \text{and} \quad \widehat{\overline{H}}_i(\ell,M) = 1 - \sum_{j\in E}\sum_{k=0}^{\ell} \widehat{q}_{ij}(k,M),$$

respectively. The previous estimators lead directly to the empirical, plug-in type estimator

$$\widehat{\mathbf{MTTF}}(M) = (\mathbf{I} - \widehat{\mathbf{P}}_{11}(M))^{-1}\widehat{\mathbf{m}}_1(M),$$

where $\widehat{\mathbf{P}}_{11}(M) = (\widehat{p}_{ij}(M);\ i,j \in U)$ and $\widehat{\mathbf{m}}_1(M) = (\widehat{m}_1(M),\ldots,\widehat{m}_{s_1}(M))^\top$ with $\widehat{m}_i(M) = \sum_{\ell\geq 0} \widehat{\overline{H}}_i(\ell,M)$, for any state $i \in U$.

The asymptotic properties of the empirical estimator of the mean time to failure were studied for continuous time SMMs by Limnios and Ouhbi (2006). Here we obtain the strong consistency and the asymptotic normality of the estimator of the conditional mean time to failure in the discrete time case. Namely, we state the following two theorems:

Theorem 1. *For any state $i \in U$, $\widehat{MTTF}_i(M)$ is strongly consistent, i.e.*

$$\widehat{MTTF}_i(M) \xrightarrow[M\to\infty]{a.s.} MTTF_i.$$

Proof. To simplify the notation, we omit M from the estimators. Then we have

$$\begin{aligned}\widehat{\mathbf{MTTF}} - \mathbf{MTTF} &= (\mathbf{I} - \widehat{\mathbf{P}}_{11})^{-1}\widehat{\mathbf{m}}_1 - (\mathbf{I} - \mathbf{P}_{11})^{-1}\mathbf{m}_1 \\ &= (\mathbf{I} - \widehat{\mathbf{P}}_{11})^{-1}\widehat{\mathbf{m}}_1 + (\mathbf{I} - \mathbf{P}_{11})^{-1}\widehat{\mathbf{m}}_1 \\ &\quad -(\mathbf{I} - \mathbf{P}_{11})^{-1}\widehat{\mathbf{m}}_1 - (\mathbf{I} - \mathbf{P}_{11})^{-1}\mathbf{m}_1 \\ &= \Big((\mathbf{I} - \widehat{\mathbf{P}}_{11})^{-1} - (\mathbf{I} - \mathbf{P}_{11})^{-1}\Big)\widehat{\mathbf{m}}_1 + (\mathbf{I} - \mathbf{P}_{11})^{-1}(\widehat{\mathbf{m}}_1 - \mathbf{m}_1).\end{aligned}$$

First, we know that $\widehat{\mathbf{P}}_{11}^n \xrightarrow[M\to\infty]{a.s.} \mathbf{P}_{11}^n$ and by means of the dominated convergence theorem (see Port (1994) for instance) we have that $\sum_{n=0}^{\infty} \widehat{\mathbf{P}}_{11}^n \xrightarrow[M\to\infty]{a.s.} \sum_{n=0}^{\infty} \mathbf{P}_{11}^n$. Then since $\sum_{n=0}^{\infty} \mathbf{P}_{11}^n = (\mathbf{I} - \widehat{\mathbf{P}}_{11})^{-1}$, we obtain directly that

$$||(\mathbf{I} - \widehat{\mathbf{P}}_{11})^{-1} - (\mathbf{I} - \mathbf{P}_{11})^{-1}|| \xrightarrow[M\to\infty]{a.s.} 0,$$

where $||\cdot||$ is a matrix norm. Using further the strong consistency of the empirical estimator of the mean sojourn time distribution (see Limnios and Ouhbi, 2006), we get the desired result. □

Theorem 2. *For any state $i \in U$, the random variable $\widehat{MTTF}_i(M)$, is asymptotically normal, in the sense that*

$$\sqrt{M}(\widehat{MTTF}_i(M) - MTTF_i) \xrightarrow[M\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2_{MTTF_i}),$$

with the asymptotic variance

$$\sigma^2_{MTTF_i} = \sum_{m\in E} a_{im}^2 \mu_{mm}\Big(\sigma_m^2 + \sum_{\ell\in E}(\eta_\ell - \widetilde{\eta}_m)^2 p_{m\ell} + 2\sum_{\ell\in E}\eta_\ell Q_{m\ell}\Big),$$

where $Q_{m\ell} = \sum_{u=1}^{+\infty}(u - m_m)q_{m\ell}(u)$, $a_{ij} = (\mathbf{I} - \mathbf{P}_{11})^{-1}_{ij}$, $\eta_\ell = \sum_{r\in U} m_r a_{r\ell}$, $\widetilde{\eta}_m = \sum_{j\in U} p_{mj}\eta_j$, and σ_m^2 is the variance of the sojourn time in state m.

Proof. First, for any fixed state $i \in E$, we notice that

$$\begin{aligned}\sqrt{M}(\widehat{MTTF}_i(M) - MTTF_i) &= \sqrt{M}\Big(\sum_{j\in U}(\widehat{a}_{ij}\widehat{m}_j - a_{ij}m_j)\Big) \\ &= \sqrt{M}\Big(\sum_{j\in U}(\widehat{a}_{ij} - a_{ij})(\widehat{m}_j - m_j) \\ &\quad + \sum_{j\in U} a_{ij}(\widehat{m}_j - m_j) + \sum_{j\in U}(\widehat{a}_{ij} - a_{ij})m_j\Big).\end{aligned}$$

Second, for any fixed state $j \in U$, and as M tends to infinity, the random variable $\sqrt{M}(\widehat{m}_j - m_j)$ converges in distribution to a zero mean normal random variable with asymptotic variance $\sigma^2_{m_j} = \mu_{jj}\sigma^2_j$ (see Georgiadis, 2014).

Then, using Slutsky theorem (see, e.g., van der Vaart, 1998), we obtain that

$$\sum_{j\in U}(\widehat{a}_{ij} - a_{ij})(\widehat{m}_j - m_j) \xrightarrow[M\to\infty]{\mathcal{L}} 0.$$

Thus, the random variable

$$\sqrt{M}(\widehat{MTTF}_i(M) - MTTF_i)$$

has the same limit in distribution as

$$\sqrt{M}\Big(\sum_{j\in U} a_{ij}(\widehat{m}_j - m_j) + \sum_{j\in U}(\widehat{a}_{ij} - a_{ij})m_j\Big).$$

For any $i, j \in U$, the random variable

$$\widehat{a}_{ij} - a_{ij} = (\mathbf{I} - \widehat{\mathbf{P}}_{11})^{-1}_{ij} - (\mathbf{I} - \mathbf{P}_{11})^{-1}_{ij}$$

has the same limit in distribution as the random variable

$$\sum_{r\in U}\sum_{\ell\in U} a_{ir}a_{\ell j}f_{r\ell},$$

where $f_{r\ell} = \widehat{p}_{r\ell} - p_{r\ell}$ (Sadek and Limnios, 2002, Theorem 4). Consecutively, we have that for any state $i \in U$, $\sqrt{M}(\widehat{MTTF}_i(M) - MTTF_i)$ has the same limit in distribution as

$$\begin{aligned}
&\sqrt{M}\Big(\sum_{j\in U} a_{ij}(\widehat{m}_j - m_j) + \sum_{j\in U}\Big(\sum_{r\in U}\sum_{\ell\in U} a_{ir}a_{\ell j}f_{r\ell}\Big)m_j\Big)\\
&= \sqrt{M}\sum_{j\in U}\left(a_{ij}(\widehat{m}_j - m_j) + \sum_{r\in U}\Big(\sum_{\ell\in U} a_{ir}a_{\ell j}f_{r\ell}\Big)m_j\right).
\end{aligned}$$

We further apply the central limit theorem for MRCs (Pyke and Schaufele, 1964). We define the function $f : E \times E \times \mathbb{N}$ by

$$\begin{aligned}
f(J_{n-1}, J_n, X_n) \;=\; &\sum_{j\in U} a_{ij}\mu_{jj}(X_n - m_j)1_{\{J_{n-1}=j\}}\\
&+\sum_{j\in U}\sum_{r\in U}\sum_{\ell'\in U} a_{ir}a_{\ell' j}\mu_{rr}\Big(1_{\{J_{n-1}=r, J_n=\ell'\}} - 1_{\{J_{n-1}=r\}}p_{r\ell'}\Big)m_j
\end{aligned}$$

and show that

$$A_m = \sum_{\ell \in E} A_{m\ell} = \sum_{\ell \in E} \sum_{u=1}^{+\infty} f(m,\ell,u)\, q_{m\ell}(u) = 0.$$

Second, we have that

$$\begin{aligned} B_{m\ell} &= \sum_{u=1}^{+\infty} f(m,\ell,u)^2\, q_{m\ell}(u) \\ &= a_{im}^2 \mu_{mm}^2 \sum_{u=1}^{\infty} (u-m_m)^2 q_{m\ell}(u) \\ &\quad + a_{im}^2 \mu_{mm}^2 \Big(\sum_{j \in U} \sum_{\ell' \in U} a_{\ell' j} (1_{\{\ell=\ell'\}} - p_{m\ell'}) m_j \Big)^2 p_{m\ell} \\ &\quad + 2 a_{im}^2 \mu_{mm}^2 \sum_{j \in U} \sum_{\ell' \in U} a_{\ell' j} (1_{\{\ell=\ell'\}} - p_{m\ell'}) m_j \sum_{u=1}^{\infty} (u-m_m) q_{m\ell}(u) \end{aligned}$$

and

$$\begin{aligned} B_m &= \sum_{\ell \in E} B_{m\ell} = a_{im}^2 \mu_{mm}^2 \sigma_m^2 + a_{im}^2 \mu_{mm}^2 \sum_{\ell \in E} (\eta_\ell - \widetilde{\eta}_m)^2 p_{m\ell} \\ &\qquad + 2 a_{im}^2 \mu_{mm}^2 \sum_{u=1}^{\infty} \sum_{\ell \in E} (\eta_\ell - \widetilde{\eta}_m) q_{m\ell}(u) (u - m_m), \end{aligned}$$

where $\eta_\ell = \sum_{r \in U} m_r a_{r\ell}$ and $\widetilde{\eta}_m = \sum_{j \in U} p_{mj} \eta_j$. We further obtain that

$$\sum_{m \in E} \frac{B_m}{\mu_{ii}} \frac{\mu_{ii}^*}{\mu_{mm}^*} = \sum_{m \in E} a_{im}^2 \mu_{mm} \Big(\sigma_m^2 + \sum_{\ell \in E} (\eta_\ell - \widetilde{\eta}_m)^2 p_{m\ell} + 2 \sum_{\ell \in E} (\eta_\ell - \widetilde{\eta}_m) Q_{m\ell} \Big),$$

where $Q_{m\ell} = \sum_{u=1}^{\infty} (u - m_m) q_{m\ell}(u)$ and $\frac{\mu_{ii}^*}{\mu_{ii}} = \frac{1}{\sum_{j \in E} \nu_j m_j}$, for any state $i \in E$. Finally we have that

$$\frac{\sum_{n=1}^{N(M)} f(J_{n-1}, J_n, X_n)}{\sqrt{M}} \xrightarrow[M \to \infty]{\mathcal{L}} \mathcal{N}\Big(0, \sigma_{MTTF_i}^2\Big),$$

where

$$\sigma_{MTTF_i}^2 = \sum_{m \in E} a_{im}^2 \mu_{mm} \Big(\sigma_m^2 + \sum_{\ell \in E} (\eta_\ell - \widetilde{\eta}_m)^2 p_{m\ell} + 2 \sum_{\ell \in E} \eta_\ell Q_{m\ell} \Big),$$

since $\sum_{\ell \in E} \widetilde{\eta}_m Q_{m\ell} = 0$.

□

The previous results enable us to construct asymptotic confidence intervals for the conditional mean times to failure. First, for a fixed state $i \in U$, we obtain the plug-in type estimator of the asymptotic variance

$$\widehat{\sigma}^2_{MTTF_i} = \sum_{m \in E} \widehat{a}^2_{im} \widehat{\mu}_{mm} \Big(\widehat{\sigma}^2_m + \sum_{\ell \in E} (\widehat{\eta}_\ell - \widehat{\widetilde{\eta}}_m)^2 \widehat{p}_{m\ell} + 2 \sum_{\ell \in E} \widehat{\eta}_\ell \widehat{Q}_{m\ell} \Big),$$

where $\widehat{Q}_{m\ell} = \sum_{u=1}^{+\infty} (u - \widehat{m}_m) \widehat{q}_{m\ell}(u)$, $\widehat{a}_{ij} = (\mathbf{I} - \widehat{\mathbf{P}}_{11})^{-1}_{ij}$, $\widehat{\eta}_\ell = \sum_{r \in U} \widehat{m}_r \widehat{a}_{r\ell}$, $\widehat{\widetilde{\eta}}_m = \sum_{j \in U} \widehat{p}_{mj} \widehat{\eta}_j$, and $\widehat{\sigma}^2_m$ is the empirical estimator of the variance of the sojourn time in state m. For notational simplicity we omit the M dependency on the estimators. Second, since the estimators that define $\widehat{\sigma}^2_{MTTF_i}(M)$ are strongly consistent, we use the dominated convergence theorem and obtain the strong consistency of $\widehat{\sigma}^2_{MTTF_i}(M)$. Thus, given the two estimators $\widehat{\sigma}^2_{MTTF_i}(M)$ and $\widehat{MTTF}_i(M)$, asymptotic confidence interval of $MTTF_i$ at level $100(1-\gamma)\%$, $\gamma \in (0,1)$ can be built, that is

$$\Big[\widehat{MTTF}_i(M) - u_{1-\gamma/2} \frac{\widehat{\sigma}_{MTTF_i}(M)}{\sqrt{M}},\ \widehat{MTTF}_i(M) + u_{1-\gamma/2} \frac{\widehat{\sigma}_{MTTF_i}(M)}{\sqrt{M}} \Big],$$

where u_γ is the $\gamma-$quantile of the $\mathcal{N}(0,1)$ distribution.

3.3 Additional risk indicators

In what follows, we evaluate additional risk indicators of repairable systems whose evolution in time is governed by an SMM. First, we define the row vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{s_1})$, where β_j is the probability that the system given that it has just entered U, it will be in state $j \in U$, when it achieved its stationarity. In particular, for all $j \in U$, we have that

$$\beta(j) = P_\nu(J_n = j | J_{n-1} \in D, J_n \in U) = \frac{\sum_{i \in D} p_{ij} \nu(i)}{\sum_{\ell \in U} \sum_{i \in D} p_{i\ell} \nu(i)}.$$

Then the mean up time is defined by $MUT = \mathbb{E}_{\boldsymbol{\beta}}(T_D)$.

Similarly, we define the row vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{s-s_1})$, where γ_j is the probability that the system given that it has just entered D, it will be in state $j \in D$, when it achieved its stationarity. For all $j \in D$, we have that

$$\gamma(j) = P_\nu(J_n = j | J_{n-1} \in U, J_n \in D) = \frac{\sum_{i \in U} p_{ij} \nu(i)}{\sum_{\ell \in D} \sum_{i \in U} p_{i\ell} \nu(i)}.$$

Symmetrically, the mean down time is defined by $MDT = \mathbb{E}_{\boldsymbol{\gamma}}(T_U)$, where $T_U := \inf\{k \in \mathbb{N} : Z_k \in U\}$. In what follows, we use the plug-in type, nonparametric, estimators

$$\widehat{MUT}(M) = \frac{\widehat{\boldsymbol{\nu}}_1(M)\widehat{\mathbf{m}}_1(M)}{\widehat{\boldsymbol{\nu}}_2(M)\widehat{\mathbf{P}}_{21}(M)\mathbf{1}_r} \qquad \text{and} \qquad \widehat{MDT}(M) = \frac{\widehat{\boldsymbol{\nu}}_2(M)\widehat{\mathbf{m}}_2(M)}{\widehat{\boldsymbol{\nu}}_1(M)\widehat{\mathbf{P}}_{12}(M)\mathbf{1}_{s-r}},$$

where $\widehat{\boldsymbol{\nu}}(M) = (\widehat{\nu}_1(M), \ldots, \widehat{\nu}_s(M))$ is a row vector with $\widehat{\nu}_i(M) = \frac{N_i(M)}{N(M)}$ and $\mathbf{1}_r$ is the $r-$dimensional column vector of ones.

3.4 Validation

To validate our results and see how well the model achieves to estimate the key indicator of the study, the conditional mean times to failure, we adopt simulation methods. In particular, we use a Monte Carlo algorithm to generate trajectories of an MRC in a fixed time interval $[0, M]$. The algorithm, in its general form, has as input the transition probability matrix of the EMC and the conditional sojourn time distributions.

We generate 5000 trajectories of the MRC with length equal to $M = 10000$. The initial law $\mathbf{a}$ and the transition kernel $\mathbf{P}$ are given by

$$\mathbf{a} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{P} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.6 & 0 & 0.4 \\ 0.3 & 0.7 & 0 \end{pmatrix},$$

respectively. The sojourn times are distributed according to the discrete Weibull distribution

$$f_{ij}(k) = \begin{cases} q^{k-1^b} - q^{k^b}, & \text{if} \quad k \geq 1, \\ 0, & \text{if} \quad k = 0, \end{cases}$$

with parameters $(q, b) = (0.1, 0.9)$ for the transitions $1 \to 2$ and $2 \to 1$, $(q, b) = (0.1, 2.0)$ for the transitions $2 \to 3$ and $3 \to 2$, and $(q, b) = (0.6, 0.9)$ for the transitions $3 \to 1$ and $1 \to 3$. In this case, the semi-Markov kernel is

$$\mathbf{q}(k) = \begin{pmatrix} 0 & 0.5f_{12}(k) & 0.5f_{13}(k) \\ 0.6f_{21}(k) & 0 & 0.4f_{23}(k) \\ 0.3f_{31}(k) & 0.7f_{32}(k) & 0 \end{pmatrix}, \quad k \in \mathbb{N}.$$

We denote by $U = \{1, 2\}$ the subset of up states and by $D = \{3\}$ the subset of down states. Our objective is to estimate the conditional mean

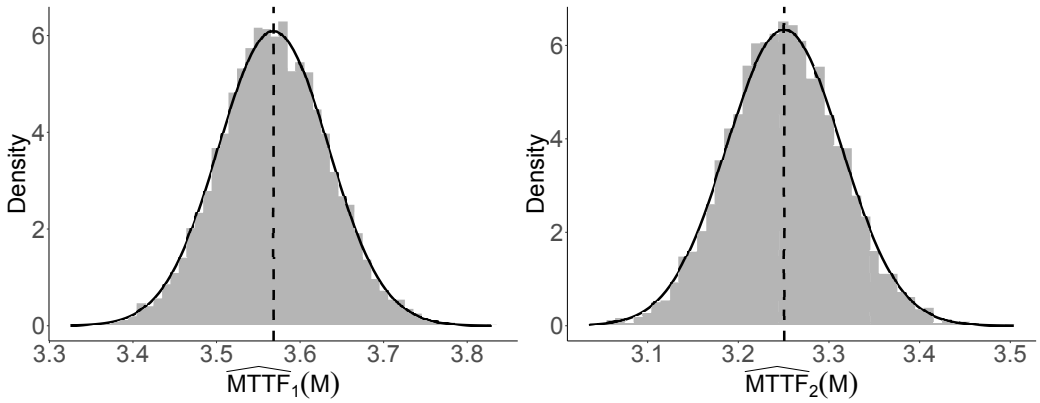

Figure 2: Histograms of the estimated conditional mean times to failure $\widehat{MTTF}_i(M)$, $i \in \{1, 2\}$, based on simulated data. The vertical dashed lines stand for the theoretical values, $MTTF_i$, $i \in \{1, 2\}$.

times to failure from the simulated data and quantify their discrepancies from the "true" (or theoretical) values. Figure 2 represents the empirical distribution of $\widehat{MTTF}_i(M)$, for $i = 1$ (right) and $i = 2$ (left). At a first sight, the empirical distribution of $\widehat{MTTF}_i(M)$ seems to be centered around $MTTF_i$ (black dashed line), for any state $i \in U$. The estimated means and variances are in good agreement with the theoretical values (Table 1).

State	$MTTF_i$	$\mathbb{E}(\widehat{MTTF}_i(M))$	$\sigma^2_{MTTF_i}$	$\widehat{\sigma}^2_{MTTF_i}$
$i = 1$	3.5683	3.5691	0.0043	0.0042
$i = 2$	3.2507	3.2508	0.0040	0.0038

Table 1: Estimated means and variances of the conditional mean times to failure and theoretical values.

4 Wind energy production management via SMMs

For operating a wind farm, the manager focuses on the wind energy production. The computation of the production takes into account the wind speed, the topography and the roughness of the site, the wake effect of the relative

positions of the turbines in each direction of the wind and the transfer power function of the turbines. It is then of paramount importance to put forward stochastic models that consider wind speeds, wind directions and production characteristics. In this work, for simplicity, wake effects are not taken into account and the transfer function of the wind farm does not depend on the wind direction. Furthermore the wind farm is reduced to a single wind turbine. For this turbine, the production at nominal power starts at 13 m/s. We distinguish low production from medium production at 8 m/s. To define the SMM,

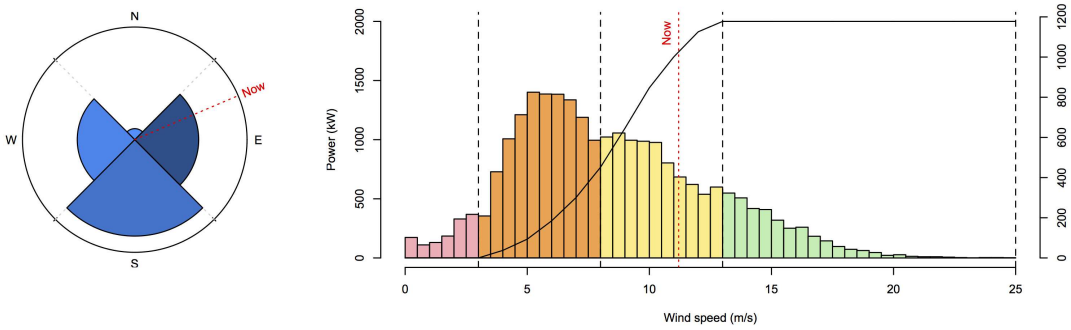

Figure 3: Wind directions and speeds states for operational management. Now the system is in the state (E, 8-13 m/s). The transfer power function of a 2MW wind turbine is superposed on the wind speeds histogram.

we consider the state space $E = E_1 \times E_2$, where E_1 is the set of direction states and E_2 is the set of wind speed states. For instance, the state spaces $E_1 = (N, W, S, E)$ and $E_2 = (0-3 \text{ m/s}, 3-8 \text{ m/s}, 8-13 \text{ m/s}, 13-25 \text{ m/s})$ could be considered for wind directions and speeds, respectively (see Figure 3 for an illustration). Breaks correspond to the cut-in (3 m/s) and the cut-off (25 m/s) wind speeds of a 2MW wind turbine. Table 2 in Appendix A presents the classification of the states and the estimated mean sojourn times along with the 95% asymptotic confidence intervals (see proof of Theorem 2). Our objective is to evaluate the aforementioned risk indicators on both simulated and real datasets. To the best of our knowledge, this is the first wind modeling in a semi-Markov context, that employs both speed and direction data to explore risk indicators. Before evaluating risk indicators for real and simulated data, we present the necessity of employing an SMM rather than a Markov model, by means of a goodness-of-fit test.

4.1 Goodness-of-fit test and sojourn time distribution

If the semi-Markov chain $\mathbf{Z}$ reduces to a (time) homogeneous Markov chain with transition probabilities $\widetilde{p}_{ij} = P(Z_{k+1} = j | Z_k = i)$ for any $i, j \in E$, $k \in \mathbb{N}$, then the semi-Markov kernel would be:

$$\begin{aligned} q_{ij}(k) &= P(J_{n+1} = j, X_{n+1} = k | J_n = i) \\ &= P(Z_{\ell+k} = j, Z_{\ell+k-1} = i, \ldots, Z_{\ell+1} = i | Z_\ell = i) \\ &= \widetilde{p}_{ij} \widetilde{p}_{ii}^{k-1}. \end{aligned}$$

If $\widetilde{p}_{ii} < 1$ and $i \neq j$ the transition probabilities of the EMC become

$$p_{ij} = \sum_{k=0}^{\infty} q_{ij}(k) = \widetilde{p}_{ij} \sum_{k=0}^{\infty} \widetilde{p}_{ii}^{k-1} = \frac{\widetilde{p}_{ij}}{1 - \widetilde{p}_{ii}},$$

whereas if $i = j$, then $q_{ij}(k) = 0$ for any $k \in \mathbb{N}$, which implies that $p_{ij} = 0$. The conditional sojourn time distribution does not depend on the next visited state j since

$$f_{ij}(k) = \frac{q_{ij}(k)}{p_{ij}} = (1 - \widetilde{p}_{ii}) \widetilde{p}_{ii}^{k-1},$$

for any $i, j \in E$, $k \in \mathbb{N}^*$. Concerning the sojourn time distribution, and for any $i \in E$, $k \in \mathbb{N}^*$, we have that

$$\begin{aligned} h_i(k) &= P(X_{n+1} = k | J_n = i) = \sum_{j \neq i} P(X_{n+1} = k, J_{n+1} = i | J_n = i) \\ &= \Big(\sum_{j \neq i} \widetilde{p}_{ij} \Big) \widetilde{p}_{ii}^{k-1} = (1 - \widetilde{p}_{ii}) \widetilde{p}_{ii}^{k-1}, \end{aligned}$$

which means that the sojourn times are geometrically distributed with parameter $\widetilde{p}_{ii}$. To justify the application of the SMM, it is sufficient to show that there exists a state whose the sojourn time distribution is significantly different from the geometric distribution. To verify the validity of geometrically distributed sojourn times, we use the chi-square goodness-of-fit test (see Delignette-Muller and Dutang, 2016). The null hypothesis is that the sojourn times are geometrically distributed. First, we fit the geometric distribution to the sojourn times and compute the maximum likelihood estimator

of the corresponding parameter $\widetilde{p}_{ii}$ for any state $i \in E$ and both real and simulated data (see Table 3 in Appendix B). Second, we compare the estimated sojourn time distribution and the geometric with pre-defined parameter $\widetilde{p}_{ii}$, for any state $i \in E$ (see Figures 6 and 7 in Appendix B). For any state $i \in E$ (apart from state 13 in simulated data), the p-value is smaller than 0.05, and therefore the null hypothesis is rejected against the alternative at 0.05 significance level.

4.2 Evaluation of risk indicators on real and simulated data

4.2.1 Real dataset

Here, we use a real dataset provided by the private company EREN. It concerns one production site with wind speed and direction measurements every ten minutes during one year (50958 records). To evaluate reliability indices, such as the aforementioned risk indicators, we focus on wind speeds that take their values over the interval $0-3$ m/s. Down states (D) are combinations of wind speeds belonging to $0-3$ m/s with any wind direction. In this case, $card(U) = 12$ and $card(D) = 4$. We pay special attention to the mean time to register a wind speed occupying the interval $0-3$ m/s, since it affirms no wind energy production. In this context, the risk indicators could provide feedback for the production management. We then present (Figure 4, upper panel) the estimates of $MTTF_i$ for any initial up state $i \in U$ along with its 95% confidence interval. For a given wind direction, the mean time to failure, *i.e.* the mean time that is required for the turbine to stop working, increases with the wind speed. For example, if the wind is blowing east and the speed is $3-8$ m/s, then the production will stop after 15.57 hours on average, instead of 18.55 hours that it would need if the current speed was $8-13$ m/s. For a given wind speed, we observe that the results vary with the wind direction. When the wind blows faster, the wind direction has a significant impact on the risk indicator. For example, it is worth mentioning that the mean time to failure is on average higher for wind blowing south with speed $8-13$ m/s than for wind speed blowing north with speed $13-25$ m/s. We can notice that we obtain more precise estimations if the wind is blowing north. We go one step further and concentrate on the mean time to register a wind speed higher than 3 m/s, given that the current wind speed is lower than 3 m/s. In other words, we are interested in the average time that the turbine needs

to restore operation, given that it is not operational. Thereafter, we will call this time conditional mean time to repair and denote it by $MTTR_i$, for any $i \in D = \{1, 2, 3, 4\}$. To estimate the conditional mean times to repair, we use the empirical, plug-in type estimator $\widehat{\mathbf{MTTR}}(M) = (\mathbf{I} - \widehat{\mathbf{P}}_{22}(M))^{-1}\widehat{\mathbf{m}}_2(M)$. We put emphasis on this indicator, since it highlights the activation of the wind energy production. Therefore it can serve as a key form of feedback for the production politics, by scheduling corrective interventions to reduce repair costs, *etc.* In Figure 5 the conditional mean times to repair are displayed along with their 95% confidence intervals. We conclude that if the current wind direction is N, the wind production needs less time to restart than if the wind is blowing west or east. The estimations are precise and the mean times to make the turbine functioning are quite short (less than one hour). Moreover, the expected time of stay in an up state (Section 3.3), *i.e.* the expected functioning time of the turbine after its reparation, is 16.31 hours. To be more precise, we are 95% confident that if the mean functioning time after a reparation was known, it would be between 16.09 and 16.53 hours. On the other side, the mean time that the turbine needs to become operational after it stops production (Section 3.3), turns out to be 1.58 hours with a 95% asymptotic confidence interval equal to $(1.54, 1.61)$.

4.2.2 Simulated dataset

The risk indicators are now evaluated based on simulated data. The dataset comes from the Wind Prospector of the National Renewable Energy Laboratory (NREL) and is available at `www.nrel.gov`. Wind directions and speeds are available every five minutes during one year on tens of thousands American sites (105120 records). Using the state space partition described previously, we estimate how long the turbine is expected to be operational. The operational efficiency of the turbine and therefore the production management could be improved by determining a preventive maintenance policy that includes risk indicators.

Figure 4 (lower panel) displays the estimated conditional mean times to failure along with their 95% confidence intervals. Concerning the results, the mean times to failure seem to have the same pattern for both simulated and real data, when the wind blows slowly. Narrower confidence intervals are obtained for simulated data w.r.t. the real data, which seems to advance the interruption of the production. Furthermore, high wind speeds delay the interruption of the wind energy production. The total time that

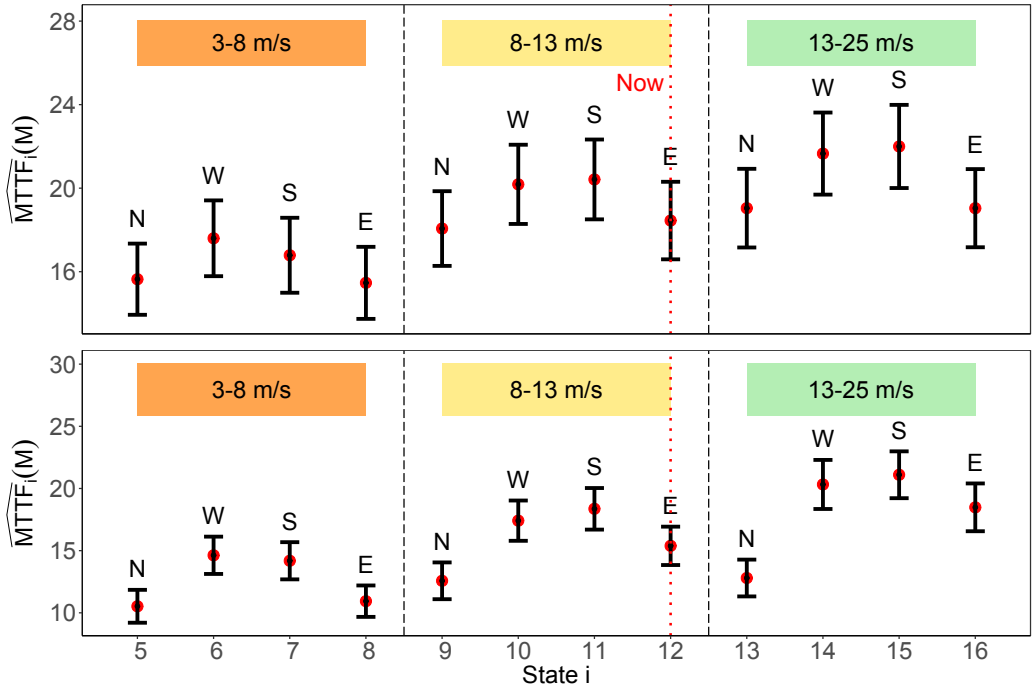

Figure 4: Empirical estimators of the conditional mean times to failure (in hours), $\widehat{MTTF}_i(M)$, $i \in U$, evaluated for real data (EREN) (upper panel) and simulated data (NREL) (lower panel).

the turbine is expected to function during an operating cycle, is estimated to be equal to 12.7 hours with a 95% asymptotic confidence interval equal to $(12.48, 12.92)$. On the other hand, the average time that the turbine is estimated to be non-operational, is equal to 1.53 hours and its confidence interval $(1.44, 1.61)$. Based on the previous results we conclude that the wind direction has an important effect on the wind energy production and thus it should be taken into account in the condition-based maintenance policy and operational management such as storage or trading.

5 Concluding Remarks

In the present paper we employ discrete time SMMs with discrete state space to evaluate risk indicators for wind modeling and energy production man-

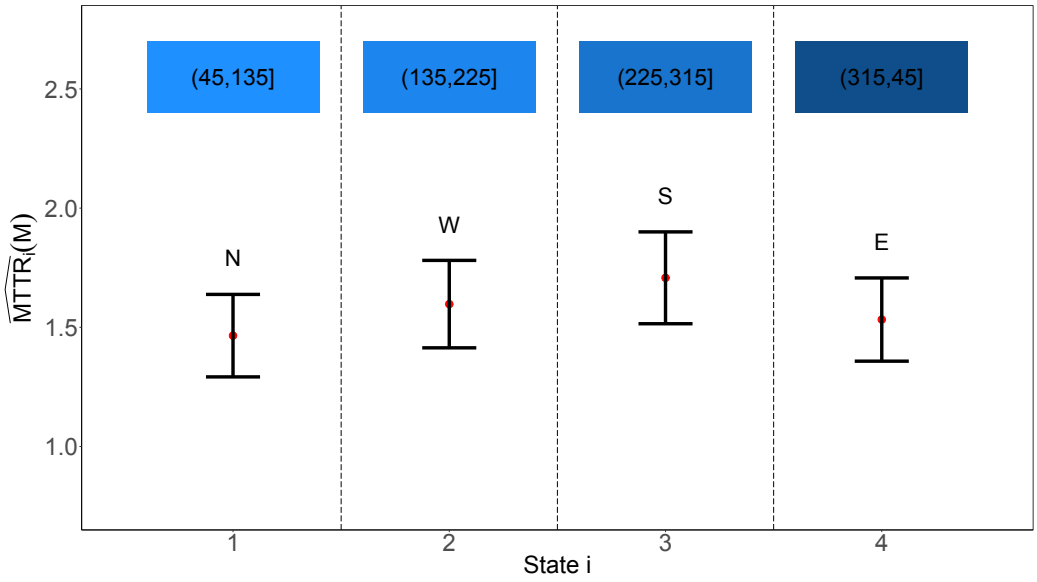

Figure 5: Empirical estimator of the conditional mean time to repair (in hours), $\widehat{MTTR}_i(M)$, $i \in D$.

agement. Among the numerous indicators in reliability, we focus on the conditional mean times to failure and show that their empirical estimators are strongly consistent and asymptotically normal. We use our results towards an application to real and simulated wind data. This is the first attempt to study SMMs that include both wind speed and direction data with a reliability orientation. To justify the necessity to employ an SMM versus a Markov model, we construct a goodness-of-fit test. We extend previous results on reliability indices for wind data modeled by SMMs (D'Amico et al., 2015a,b), by evaluating key indicators that could provide evidence in the management of a wind farm. The results highlight that the wind direction plays a critical role in a semi-Markov setting, and therefore it should be taken into account in maintenance and management decisions.

Due to the flexibility of the framework provided by SMMs, there exists a variety of extensions that could be adapted to obtain risk indicators. Conditional mean times to failure could be investigated when independent and identical copies of the semi-Markov chain are observed, each over a fixed duration, instead of one single copy over a fixed time interval. Further re-

search includes their extension to higher order SMMs. In this context, the stochastic evolution of the EMC depends on the jump time, *i.e.* the transition order. Finally, the employment of SMMs with a more refined structure, for example models that allow for many degradation states (Malefaki et al., 2014), could improve the understanding of wind data and eventually provide decision support tools in wind industry.

Acknowledgements The authors are partially supported by the project CAESARS grant ANR-15-CE05-0024. They would like to thank EREN group for providing the data used in this paper. They also thank M. Hamdaoui for fruitful discussions and assistance.

References

Ailliot, P., Bessac, J., Monbet, V., Pène, F. (2015). Non-homogeneous hidden Markov-switching models for wind time series. *Stat. Plan. Inference* 160: 75–88.

Barbu, V., Karagrigoriou, A., Makrides, A. (2016). Semi-Markov modeling for multi-state systems. *Methodol. Comput. Appl. Probab.* doi:10.1007/s11009-016-9510-y.

Barbu, V., Limnios, N. (2008). *Semi-Markov chains and hidden semi-Markov models toward applications: Their Use in Reliability and DNA Analysis. Lecture Notes in Statistics*. New York: Springer.

D'Amico, G., Petroni, F., Praticco, F. (2015a). Performance Analysis of Second Order Semi-Markov Chains: An Application to Wind Energy Production. *Methodol. Comput. Appl. Probab.* 17(3): 781-794.

D'Amico, G., Petroni, F., Praticco, F. (2015b). Reliability measures for indexed semi-Markov chains applied to wind energy production. *Reliab. Eng. Syst. Safe.* 144: 170–177.

D'Amico, G., Petroni, F., Praticco, F. (2013a). First and second order semi-Markov chains for wind speed modeling. *Physica* 392(5): 1194-1201.

D'Amico, G., Petroni, F., Praticco, F. (2013b). Reliability measures of second-order semi-Markov chain applied to wind energy production. *J. Renew. Energy.* Article ID 368940, 6 pages. doi:10.1155/2013/368940

D'Amico, G., Petroni, F., Praticco, F. (2013c). Wind speed modeled as an indexed semi-Markov process. *Environmetrics* 24: 367-376.

Delignette-Muller, M.-L., Dutang, C. (2016). Help to fit of a parametric distribution to non-censored or censored data. URL: https://cran.r-project.org/web/packages/fitdistrplus/index.html.

Ettoumi, F.Y., Sauvageot, H., Adane, A.-E.-H. (2003). Statistical bivariate modeling of wind using first-order Markov chain and Weibull distribution. *Renew. Energy* 28: 1787-1802.

Georgiadis, S. (2017). First hitting probabilities for semi-Markov chains and estimation. *Comm. Stat. Theory Methods* 46(5): 2435–2446.

Georgiadis, S. (2014). Estimation des systèmes semi-markoviens à temps discret avec applications, PhD Thesis, Université de Technologie de Compiègne, Compègne France.

Georgiadis, S., Limnios, N., Votsi, I. (2013). Reliability and probability of first occurred failure for discrete-time semi-Markov systems. In: Frenkel, I., Karagrigoriou, A., Lisnianski, A., Kleyner, A., eds. *Applied Reliability Engineering and Risk Analysis: Probabilistic Models and Statistical Inference* (chap. 12, 167179). New York: Wiley & Sons.

Limnios, N., Oprişan, G. (2001). *Convergence of Probability Measures*. New York: Wiley 2nd edition.

Limnios, N., Ouhbi, B. (2006). Nonparametric estimation of some important indicators in reliability for semi-Markov processes. *Stat. Methodol.* 3: 341–350.

Malefaki, S., Limnios, N., Dersin, P. (2014). Reliability of maintained system under a semi-Markov setting. *Reliab. Eng. Syst. Safe.* 131: 282–290.

Masala, G. (2014). Wind time series simulation with underlying semi-Markov model: an application to weather derivatives. *J. Stat. Manag. Syst.* 17(3): 285–300.

Port, S.C. (1994). *Theoretical probability for applications*. New York: Wiley.

Pyke, R., Schaufele, R. (1964). Limit theorems for Markov renewal processes. *Ann. Math. Statist.* 35: 1746–1764.

Sadek, A., Limnios, N. (2002). Asymptotic properties for maximum likelihood estimators for reliability and failure rates of Markov chains. *Comm. Stat. Theory Methods* 31(10): 1837–1861.

Tang, J., Brouste, A., Tsui, K. (2015). Some improvements of wind speed Markov chain modeling. *Renew. Energy* 81: 52-56.

van der Vaart, A.W. (1998). *Asymptotic Statistics*. New York: Cambridge University Press.

Votsi, I., Limnios, N. (2015). Estimation of the intensity of the hitting time for semi-Markov chains and hidden Markov renewal chains. *J. Nonparametr. Stat.* 27 (2): 149-166.

Votsi, I., Limnios, N., Tsaklidis, G., Papadimitriou, E. (2014). Hidden semi-Markov modeling for the estimation of earthquake occurrence rates. *Commun. Stat. Theory Methods* 43: 1484–1502.

A Data summary and estimations

Speed	Direction	State i	Frequency	$\widehat{m}_i(M)$	95%CI of m_i
$[0,3]$	$(45,135]$	1	1196	0.70	$[0.54,0.85]$
$[0,3]$	$(135,225]$	2	979	1.08	$[0.95,1.21]$
$[0,3]$	$(225,315]$	3	1342	1.16	$[0.99,1.32]$
$[0,3]$	$(315,45]$	4	980	1.51	$[1.18,1.85]$
$(3,8]$	$(45,135]$	5	6041	0.32	$[0.26,0.39]$
$(3,8]$	$(135,225]$	6	9339	0.81	$[0.62,1.00]$
$(3,8]$	$(225,315]$	7	7453	1.17	$[0.90,1.44]$
$(3,8]$	$(315,45]$	8	4813	0.58	$[0.39,0.77]$
$(8,13]$	$(45,135]$	9	1219	0.62	$[0.54,0.70]$
$(8,13]$	$(135,225]$	10	6240	0.56	$[0.48,0.65]$
$(8,13]$	$(225,315]$	11	4999	0.71	$[0.61,0.80]$
$(8,13]$	$(315,45]$	12	2605	0.58	$[0.49,0.67]$
$(13,25]$	$(45,135]$	13	37	1.26	$[1.11,1.41]$
$(13,25]$	$(135,225]$	14	1466	1.29	$[1.16,1.42]$
$(13,25]$	$(225,315]$	15	2057	1.45	$[1.26,1.63]$
$(13,25]$	$(315,45]$	16	192	1.37	$[1.19,1.56]$

Table 2: States classification, frequencies and estimated mean sojourn times (in hours) based on the real dataset (EREN).

B Goodness-of-fit test

State	*Real data (EREN)*		*Simulated data (NREL)*	
	$\widetilde{p}_{ii}$	p-value	$\widetilde{p}_{ii}$	p-value
1	0.27	$3.52e-09$	0.12	$2.96e-06$
2	0.30	$3.59e-12$	0.15	$1.80e-04$
3	0.24	$1.66e-13$	0.14	$1.22e-02$
4	0.29	$1.40e-10$	0.10	$1.18e-15$
5	0.13	$1.22e-120$	0.10	$6.75e-29$
6	0.13	$1.93e-195$	0.05	$1.43e-83$
7	0.12	$2.32e-259$	0.05	$6.14e-55$
8	0.12	$1.41e-99$	0.04	$4.38e-32$
9	0.24	$1.14e-16$	0.13	$1.47e-05$
10	0.15	$9.37e-125$	0.06	$4.62e-52$
11	0.14	$9.32e-142$	0.04	$8.70e-44$
12	0.11	$7.25e-55$	0.05	$4.02e-29$
13	0.51	$2.30e-03$	0.37	$5.47e-01$
14	0.21	$7.96e-42$	0.06	$3.11e-27$
15	0.14	$5.45e-38$	0.03	$4.80e-36$
16	0.29	$4.32e-04$	0.05	$1.31e-07$

Table 3: MLE of the parameter of the geometric distribution, $\widetilde{p}_{ii}$, $i \in E$, and summary of the goodness-of-fit test for real and simulated data.

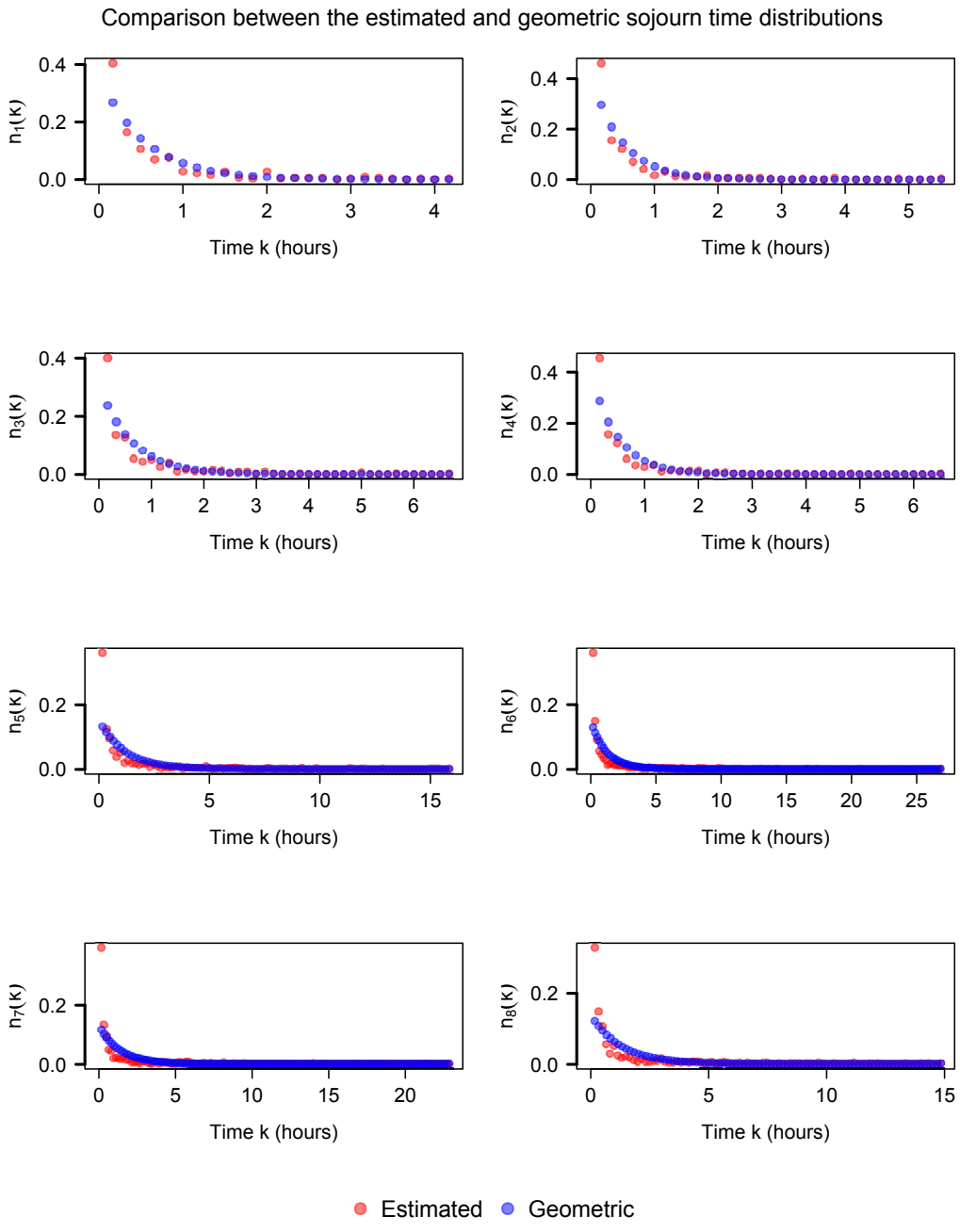

Figure 6: Estimated and geometric sojourn time distributions for any state $i \in \{1, \ldots, 8\}$.

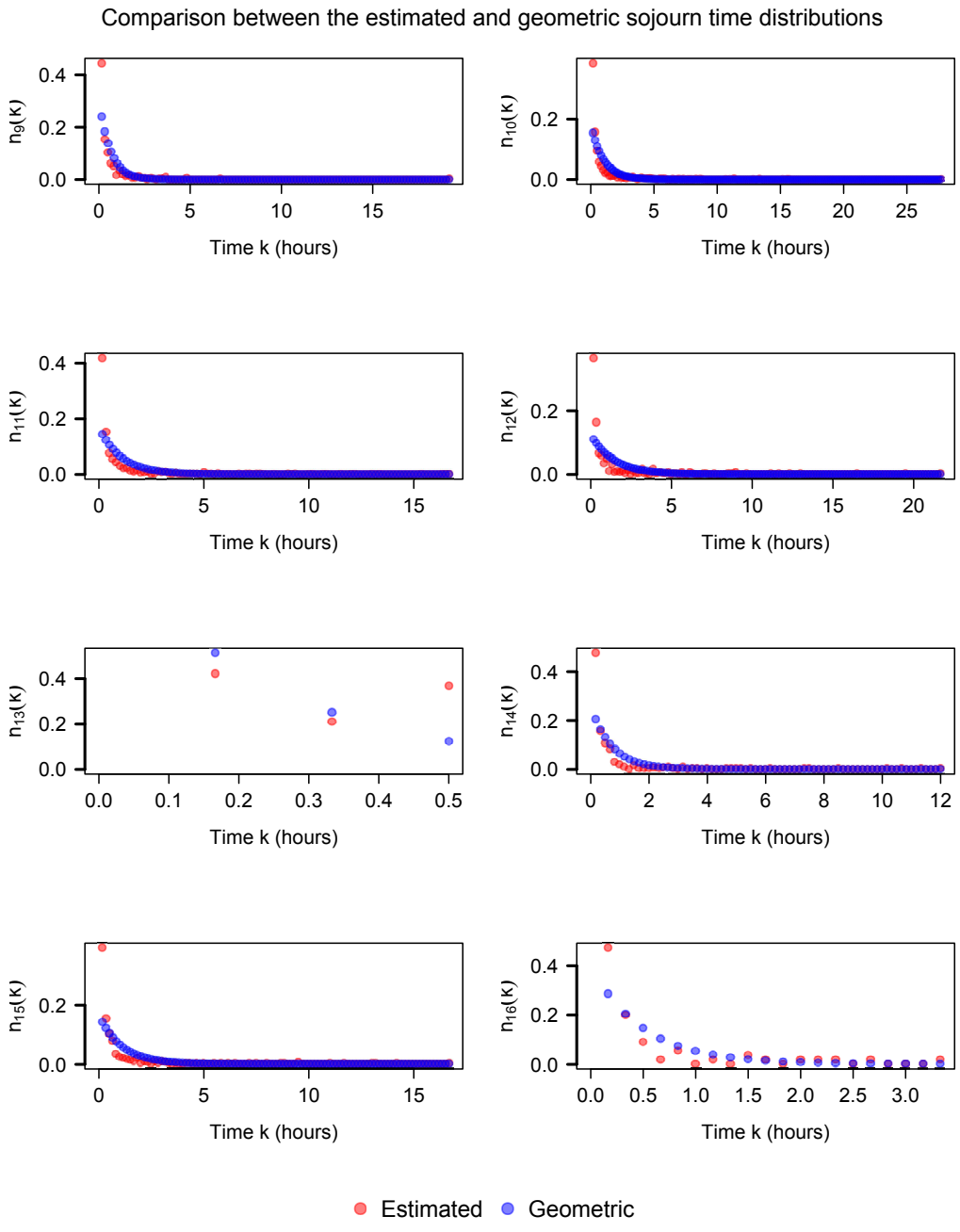

Figure 7: Estimated and geometric sojourn time distributions for any state $i \in \{9, \ldots, 16\}$.