20th International Conference on Renewable Energies and Power Quality (ICREPQ'22)
Vigo (Spain), 27th to 29th July 2022
Renewable Energy and Power Quality Journal (RE&PQJ)
ISSN 2172-038 X, Volume No.20, September 2022

Modeling and Simulation of Renewable Energy Sources by Markov Chains

M. A. S. G. Araujo[1], G. A. Melo[2], S. C. A. Ferreira[3], R. C. Souza[4], F. L. Cyrino Oliveira[5] and P. M. M. Louro[6]

Department of Industrial Engineering, Pontifical Catholic University of Rio de Janeiro, Brazil
[1]e-mail: **margarete.afonso.sousa@gmail.com**
[2]e-mail: **gustavo.melo.rio@gmail.com**
[3]e-mail: **saulocustodio@tecgraf.puc-rio.br**
[4]e-mail: **reinaldo@puc-rio.br**
[5]e-mail: **cyrino@puc-rio.br**
[6]e-mail: **paulamacaira@puc-rio.br**

Abstract. The increased participation of renewable variable energy sources (RVS) in the Brazilian electricity matrix brings several challenges to the planning and operation of the Brazilian Electricity System (BES) due to the stochasticity present in RVS. Such challenges involve the modeling and simulation of intermittent generation processes. In this context, this work aims to simulate power generation scenarios of three Brazilian plants, each based on three distinct renewable sources: wind power, solar, and biomass. The methodology used is based on the modeling of historical time series by Markov Chains, and the generation of scenarios is performed by Monte Carlo simulation. The results obtained are promising: the simulated scenarios satisfactorily reproduced the characteristics of the historical generation data of the plants.

Keywords. Wind Power, Solar Photovoltaic, Biomass, Markov Chains, Monte Carlo Simulation.

1. Introduction

In recent years, renewable variable energy sources (RVS) have become a cost-effective and environmentally friendly alternative to supply power to isolated and integrated electrical grids worldwide. According to data from the International Renewable Energy Agency (IRENA) [1], the world added more than 260 gigawatts (GW) of renewable energy capacity in 2020, outpacing the expansion in 2019 by about 50%.

The growing share of renewable sources is partly attributable to the net dismantling of power generation from fossil fuels, whose total world additions fell from 64 GW in 2019 to 60 GW in 2020 [1], demonstrating a continued downward trend of expansion of fossil fuels. Thus, it is noteworthy that more than 80% of the world's electricity generation capacity added in 2020 was RVS, highlighting wind, solar, and biomass sources.

In Brazil, following the global trend, investments in renewable sources have increased significantly over the last decades. It is worth noting that the share of renewables in the Brazilian electricity matrix is substantially higher when compared to the world matrix [2], as shown in Figure 1 based on 2019 data information.

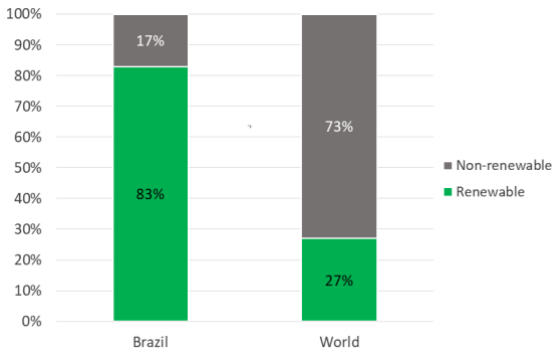

Figure 1. The relative share of renewable and non-renewable sources in the electricity matrices of Brazil and the world in 2019. Source: Adapted from Empresa de Pesquisa Energética (EPE) [2].

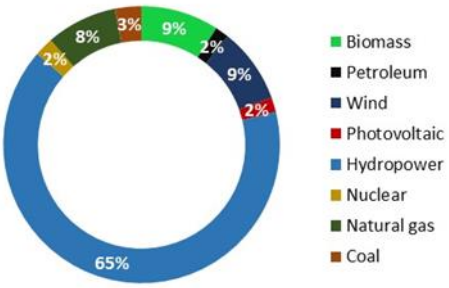

Figure 2. Share electricity generation by source, 2020. Source: Adapted from Empresa de Pesquisa Energética (EPE) [2].

Figure 2 presents the Brazilian electricity matrix by source in 2020. Although the hydroelectric source accounts for 65% of the total installed capacity, the joint participation of wind, photovoltaic, and biomass RVS is already equivalent to about 20% of the country's electricity matrix and is increasing substantially every year.

However, it is essential to mention that the large-scale integration of RVS in electrical systems strongly impacts their planning and operation due to the uncertainties

inherent to such sources, bringing challenges to the dispatch of energy generation worldwide [3]. According to [4], such challenges require the simulation of stochastic processes of renewable energy generation on temporal and spatial scales to support decision-making in the public and private sectors.

According to [5], extensive historical measurements of variables linked to intermittent sources are necessary to model the random behavior of these variables at a given location. However, the availability of such data is often insufficient, emerging the need to develop simulation techniques that capture and satisfactorily reproduce the random behavior of RVS.

In this sense, the objective of this work is to simulate electricity generation scenarios for three different RVS: wind, solar, and biomass. Data from three Brazilian plants is based on one of the mentioned sources. The methodology used is based on the works of [5] and [6], which apply a technique called Monte Carlo via Markov Chains to simulate wind power generation scenarios in Canada and Brazil, respectively. Furthermore, as the objective is to simulate reliable scenarios compatible with reality, statistical characteristics of the synthetic series are compared with the historical ones to evaluate the simulations.

The remainder of this article is divided as follows: Chapter 2 presents the methodology used to generate the scenarios; in chapter 3, a brief description of the data from each simulated source is made; chapter 4 brings the results obtained; finally, in chapter 5, the conclusions of this research and some proposals for future work are presented.

2. Methodology

In this paper, the authors propose the methodology shown in Figure 3, with four steps: first, if necessary, the historical data is divided into subsets, then clustered, followed by the construction of a transition matrix, and finally, simulated based on the matrix created on step 3.

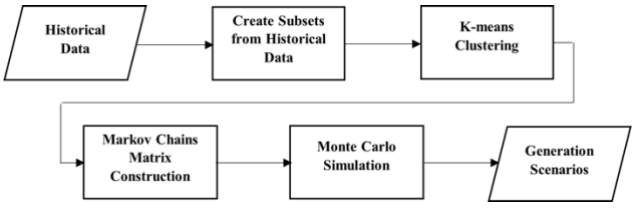

Figure 3. Methodology steps for scenarios generation.

A. Create subsets from historical data

For each renewable energy source, if necessary, subsets have to be created based on the characteristic of the historical input data. This phase needs a descriptive data analysis to identify trends and seasonality to help decide the best subset of data.

B. K-means clustering

The objective of this phase is to define a finite amount of values, discretizing the historical generation data. The limited amount of values enables the construction of a transition matrix of K x K dimensions, where K is the number of clusters that represent this amount of discrete values.

The k-means clustering [7] groups the data, which are continuous, in a number k of clusters randomly selected with an initial centroid (cluster's center). The distance between each historical observation to this initial centroid is calculated, and the data value (generation) is assigned to the nearest cluster. Then, new centroids are defined, the mean of all data values of each cluster. This process is repeated until the centroids remain fixed after multiple iterations. The historical generation values are replaced by the cluster centroids to which they belong [8].

C. Markov chains matrix construction

Making a link of k-means with Markov Chains implementation, each cluster is interpreted as a state of a stochastic process. The stochastic process represents an evolution of a random variable over time; precisely, the variable corresponds to a discrete generation. The process is called a Markov chain when there is a finite number of states, which means when the variable is discrete [9].

$\boldsymbol{P_{a,b}}$ is the transition probability, the probability of the stochastic process, assuming the value of state $\boldsymbol{b}$, starting from a state $\boldsymbol{a}$. Each probability is an element of the matrix and is calculated by equation 1.

$$\boldsymbol{P_{a,b}} = \frac{\boldsymbol{n_{a,b}}}{\sum_{\boldsymbol{k}} \boldsymbol{n_{a,k}}} \tag{1}$$

Where $\boldsymbol{k} \in \boldsymbol{K}$, set of all states, $\boldsymbol{n_{a,b}}$ is the number of times the process assumed value $\boldsymbol{b}$, starting from the state $\boldsymbol{a}$ and $\boldsymbol{n_{a,k}}$ the number of transitions in which the process assumed any state $\boldsymbol{k}$ from $\boldsymbol{a}$.
Equation 1 results in the empirical probability of the state transition. There is the transition probability from each state to any other state, thus constructing the transition probability matrix, $\boldsymbol{P}$, presented by Equation 2.

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{P_{1,1}} & \cdots & \boldsymbol{P_{1,k}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{P_{k,1}} & \cdots & \boldsymbol{P_{k.k}} \end{pmatrix} \tag{2}$$

D. Monte Carlo Simulation

Each matrix row corresponds to the initial state at instant $\boldsymbol{t}$, and the columns refer to the next state assumed at $\boldsymbol{t+1}$ [10]. Each line can represent a discrete probabilistic density conditioned to the initial state $\boldsymbol{t}$. A random value is drawn for the instant $\boldsymbol{t+1}$ and follows the respective density. Equation 3 represents the iterative simulation process.

$$\boldsymbol{Sim_{t+1}} = \boldsymbol{M_k} \sim \boldsymbol{P_{Sim_t,k}} \tag{3}$$

Where $\boldsymbol{k} \in \boldsymbol{K}$, for each instant $\boldsymbol{t}$.
Equation 3 is already contemplating a methodology peculiarity. The value drawn is a centroid following the initial state probabilistic density at instant $\boldsymbol{t}$, $\boldsymbol{P_{Sim_t,k}}$. Simulation finish when the iteration process reaches the previously determined horizon. The methodology was implemented using R software [11].

3. Data description

This section summarizes the generation series of Brazilian plants descriptions concerning the three energy sources treated in this work: wind, solar, and biomass.

The seasonality existing in the three categories of generation were considered separately for each one of them, as they have different behavior. This is described in the presentation of each one of them.

A. Wind power

For the development of wind generation scenarios, we use active power time series between April 2018 and February 2021, and they are supplied every 15 minutes. These data belong to a wind power plant located on the coast of northeastern Brazil, a region with a large capacity for wind power generation.

Before preparing the scenarios, the data underwent a pre-processing and statistical analysis. The pre-processing consisted of processing and validation of the data: (i) no missing data identified; (ii) the data above the installed capacity were corrected to maximum capacity; and (iii) no negative values were identified, which would be substituted by zero. Corrections accounted for less than 0.001% of the data.

The monthly boxplot data and the density curve were generated in the descriptive data stage of the modeling process. For all the energy sources, the data available by the Brazilian electricity sector utilities are shown per unit of installed capacity (p.u.) to preserve confidentiality.

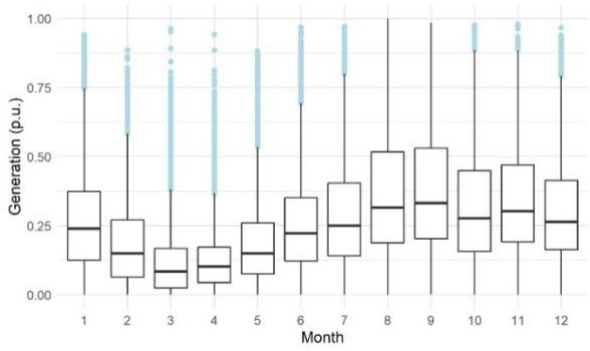

Figure 4. Power generation boxplot from the Wind Farm.

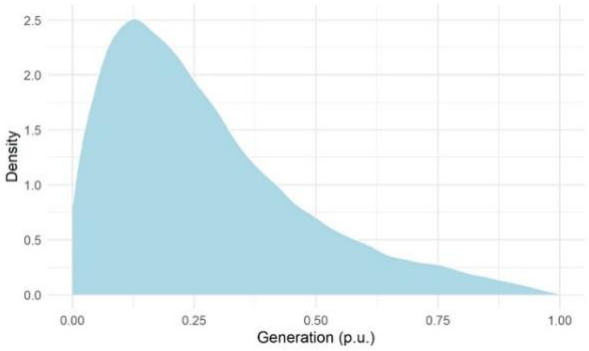

Figure 5. Power generation probabilistic density from the Wind Farm.

Figure 4 presents the monthly boxplot of wind generation data, a seasonal movement of the month's medians is noted, a peak in September, and a valley in March. The months belonging to the second half of the year tend to present higher generations than the first semester; besides, the data shows a significant dispersion.

The density curve of the data is exposed in Figure 5. It displays a single peak closer to the beginning of the data range, indicating that a Gama or even a Weibull distribution would fit the data.

B. Solar photovoltaic

In the case of the Solar Photovoltaic source, the input data period comprises the months from October 2018 to February 2021, with observations every 15 minutes. The first three months (2018) show a generation growth due to the start of operation of the plant. The authors decided not to consider this period to avoid a model misinterpretation because this trend does not exist.

In addition, all missing values were completed with the mean of the last seven days before the day/hour with no generation. The missing values were interpreted as a measurement error.

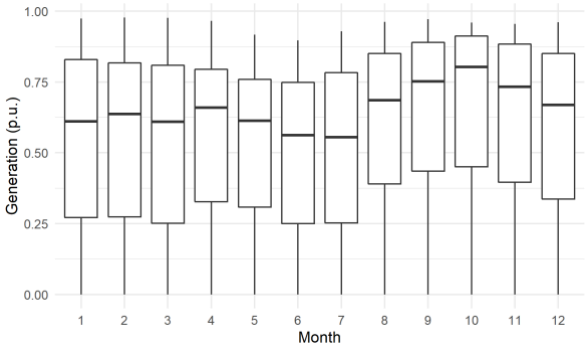

Figure 6. Power generation boxplot from the solar photovoltaic plant.

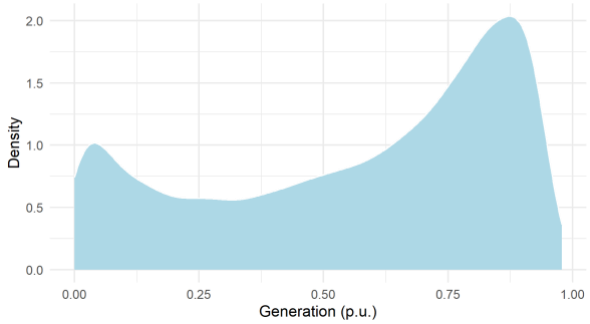

Figure 7. Power generation probabilistic density from the solar photovoltaic plant.

The boxplot, in Figure 6, shows that from January to May, the solar power generation is almost the same, with a decrease in generation during June and July. On the other hand, solar power generation grows in August, September, and October, returning to the same generation level of December and January. This behavior justifies dividing the data in a monthly subset before the clustering phase.

Solar Photovoltaic source differs from wind and biomass because power generation is observed only during the day (from 6 am to 6 pm). Thus, for simulation, the night period is not considered. After the process is finished, the hours from 6 am to 6 pm are completed with zeros.

The density curve (Figure 7) shows a bi-modal curve for low values near zero (first hours of each day and near 6 pm) and a significant number of high generation between 10 am and 3 pm.

C. Biomass

The historical biomass plant series goes from August 2019 to February 2021, with observations every 15 minutes. On the power generation boxplot, Figure 8, it is possible to observe an annual intermittence period between March and July, explained by the off-season production period of the primary raw material used in the plant, in this case, the sugarcane.

In addition, there are three well-defined operation phases: an initial period between August and September, a somehow steady generation period between October and December, and a final period between January and February.

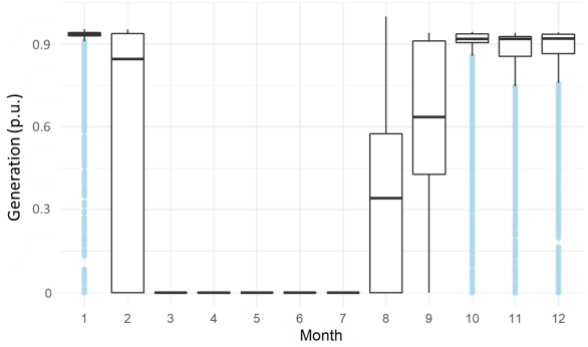

Figure 8. Power generation boxplot from the biomass plant.

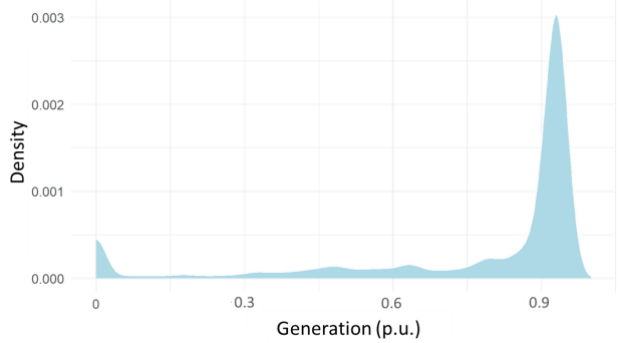

Figure 9. Power generation probabilistic density from the biomass plant.

Figure 9 presents the generation density curve of the plant in its period of operation, that is, excluding the intermittency period. The curve is bimodal, with a peak close to its maximum generation and small concentrations at lower and null values.

4. Results

In this section, the results of the simulations obtained from the data of the three plants are presented.

A. Wind power

From the analysis of wind generation data presented in section 3.1, different behavior was noted between months. In this way, the authors decided to generate monthly scenarios. The data is divided into month subsets and for each month, the methodology proposed was followed, creating 100 wind power scenarios. The number of clusters selected per month ranged from five to eleven; the most recurrent amount was seven clusters.

Figure 10 compares the proportion of values generated through simulations and those from the historical series for January as an example. For all the histograms, the number of values belonging to a state in the simulated series is approximately equal to that of the historical series for all cases, pointing out that the method to generate synthetic scenarios used satisfactorily reproduces the observed data behavior of the wind power plant.

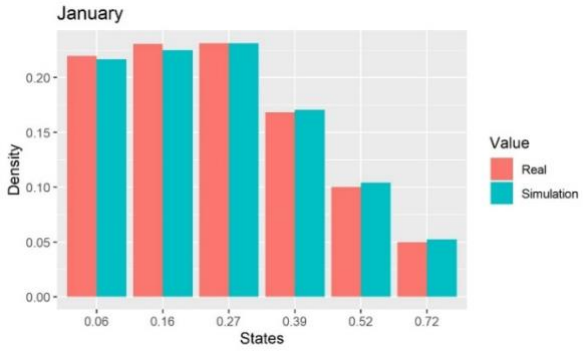

Figure 10. Historical clusters and simulations' distributions of Janeiro of the Wind Farm.

Table I compares the means and monthly standard deviations between the measured and synthetic series. It is possible to see that the percentage errors between values are minimal. The mean percentage error ranges from 0.14% to 2.09% and between 0.49% and 5.47% for the standard deviation. Those ranges indicate that the proposed method replicates the annual seasonality of wind generations very well.

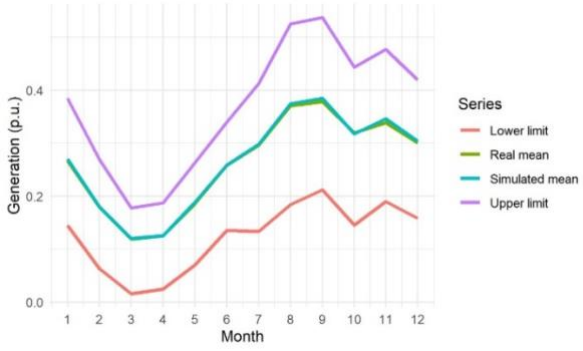

Figure 11. Comparison of the simulated scenarios with the means of the monthly historical generations.

Figure 11 compares the simulated scenarios calculated means with the actual generations' mean per month—the measured power generation average in red, the mean of the simulated scenarios in orange, the first (light blue), and the third (dark blue) means of the set of scenarios quartiles. It is noteworthy that the series corresponding to the actual averages is between the first and the third quartiles each

month. In addition, it is noted that the synthetic mean accompanies the seasonal movement of the real average generation. Thus, we can conclude that the simulated series satisfactorily reproduce the behavior of the historical observed time series.

Table I. Comparison between the measured and synthetic series – mean and standard deviation for Wind Farm.

Month	Mean (p.u.)		Standard deviation (p.u.)	
	Historic	Simulation	Historic	Simulation
Jan	0.2670	0.2707	0.1803	0.1770
Feb	0.1800	0.1804	0.1496	0.1466
Mar	0.1192	0.1206	0.1306	0.1277
Apr	0.1254	0.1255	0.1148	0.1085
May	0.1862	0.1888	0.1519	0.1498
June	0.2585	0.2592	0.1846	0.1810
July	0.2963	0.2983	0.2129	0.2115
Aug	0.3712	0.3749	0.2357	0.2340
Sept	0.3787	0.3848	0.2223	0.2200
Oct	0.3198	0.3180	0.2085	0.2074
Nov	0.3389	0.3460	0.1930	0.1909
Dec	0.3009	0.3042	0.1890	0.1872

B. Solar photovoltaic

The data were also divided into months for the solar photovoltaic source to better reproduce the time series behavior throughout the year. However, the number of clusters per group ranged from two to six clusters. Figure 12 shows the histogram for January to exemplify the clustering process. In most of the cases, the clustering method resulted in 7 clusters. As well as for wind and biomass cases, simulated solar photovoltaic series presents almost the same number of values of the historical series.

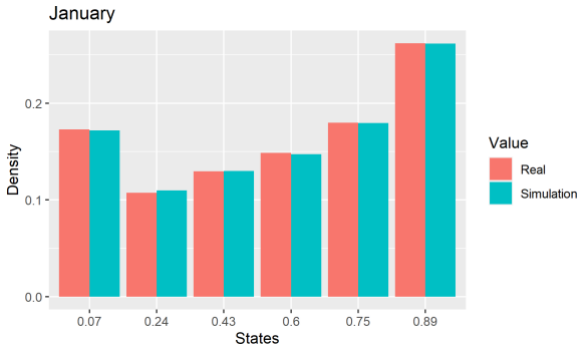

Figure 12. Historical clusters and simulations' distributions of January of the solar photovoltaic plant.

Finally, after the Monte Carlo simulation, comparing the mean from the simulated series with the historical data available, it is possible, with the line graph (Figure 13), to visualize that this methodology reproduces the aggregated monthly behavior of the solar photovoltaic generation. As for wind power, 100 scenarios were generated for photovoltaic solar generation.

Table II presents the means and the standard deviations. The worst result was obtained in August, where the percentage error for the mean was 1.17%. Despite this result, the mean bias error found is 0.21% for the means and 1.49% for the standard deviation, confirming the visual perception shown in Figure 13.

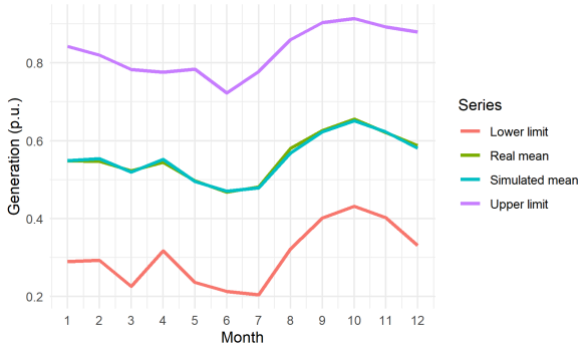

Figure 13. Simulated limits and means of the historic plant and solar photovoltaic simulations.

Table II. Comparison between the measured and synthetic series – mean and standard deviation for the solar photovoltaic plant.

Month	Mean (p.u.)		Standard deviation (p.u.)	
	Historic	Simulation	Historic	Simulation
Jan	0.5490	0.5487	0.3052	0.2993
Feb	0.5469	0.5536	0.3045	0.2978
Mar	0.5226	0.5198	0.3109	0.3057
Apr	0.5448	0.5525	0.2939	0.2865
May	0.4971	0.4952	0.2921	0.2864
June	0.4680	0.4706	0.2896	0.2846
July	0.4818	0.4787	0.2952	0.2916
Aug	0.5805	0.5688	0.3011	0.3017
Sept	0.6260	0.6232	0.3056	0.3016
Oct	0.6554	0.6513	0.3118	0.3075
Nov	0.6210	0.6233	0.3005	0.2959
Dec	0.5872	0.5804	0.2932	0.2914

C. Biomass

As highlighted in section 3.3, three distinct phases were identified in the biomass plant's operation period: beginning, intermediate or entire generation phase, and a final period. Therefore, the clustering of the historical series was carried out separately for each of the three periods mentioned so that the simulated scenarios could well represent the individual characteristics of each stage. Thus, the K-means method identified 12, 9, and 4 clusters for each phase.

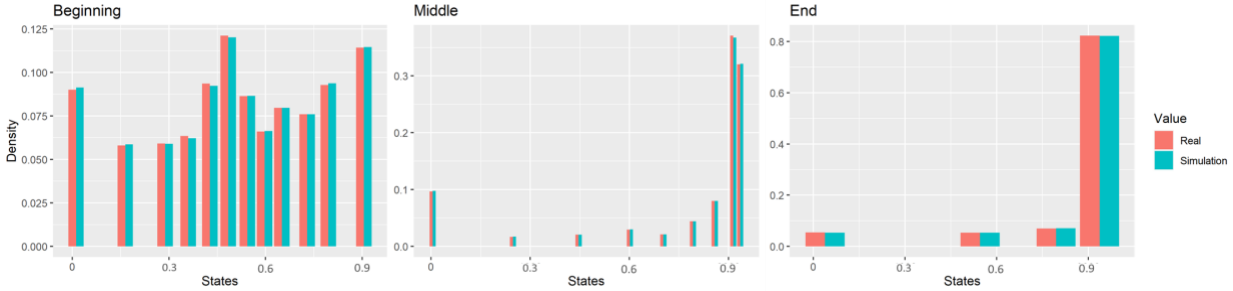

Figure 14. Historical clusters and simulations' distributions of the biomass plant's initial, intermediate, and final stages.

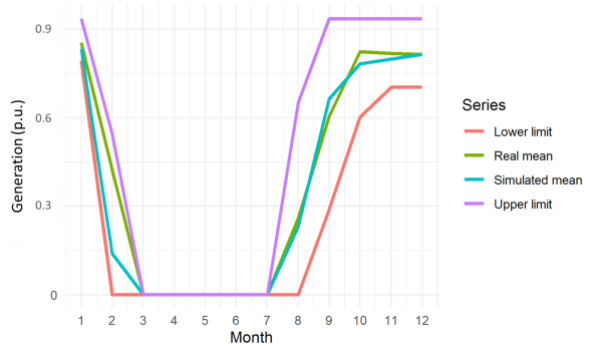

Figure 15. Simulated limits and means of the historic plant and biomass simulations.

Figure 14 compares the distribution of clusters, represented by the values of their centroids, between the observed series and the simulated three stages. Note that the historical distribution was reproduced satisfactorily by the simulations. Therefore, it is possible to affirm that the methodology adheres to the time series trends.

Figure 15 compares the simulated scenarios' averages with the observed generations' monthly averages. The simulation averages followed the generation trends, while the simulated upper and lower bounds comprised the generation capacity. In parallel with the cluster distribution graphs, it can be said that the 100 scenarios reproduced the statistical characteristics of the generation history.

Table III compares the means and monthly standard deviations between the measured and synthetic series. It is possible to see that the percentage errors between values are minimal. The mean percentage error ranges from 0.00% to 0.16% and between 0.18% and 2.55% for the standard deviation. Those ranges indicate that the proposed method replicates the annual seasonality of biomass generations very well.

Table III. Comparison between the measured and synthetic series – mean and standard deviation for the biomass plant.

Phase	Mean (p.u.)		Standard deviation (p.u.)	
	Historic	Simulation	Historic	Simulation
Beginning	0.5196	0.5195	0.2567	0.2556
Middle	0.7887	0.7874	0.2854	0.2849
End	0.8579	0.8579	0.2192	0.2136

5. Conclusion

Given that the generation of renewable energy scenarios is essential for the planning and operation of sustainable electrical networks, this work aimed to simulate synthetic energy series from three plants with different renewable sources - wind, solar, and biomass. From the application of the simulation methodology, based on the modeling of time series by Markov Chains, the scenarios obtained reproduced well the statistical characteristics of the historical data of the three RVS. As the scenarios were generated for three plants, it is suggested to investigate the robustness of the methodology, applying it to a more significant number of plants.

Acknowledgment

This work was supported by: the Brazilian Coordination for the Improvement of Higher Level Personnel (CAPES) under Grant [number 001]; the Brazilian National Council for Scientific and Technological Development (CNPq) under Grants [numbers 307403/2019-0 and 422470/2021-0], the Carlos Chagas Filho Research Support Foundation of the State of Rio de Janeiro (FAPERJ) under Grants [numbers 202.673/2018 and 211.086/2019] and ANEEL R&D Program and Grupo Energisa under Grant number 06585-1802/2018 (Análise Estocástica de Perdas Técnicas em Sistemas de Energia).

Appendix A. Supplementary data

Supplementary data (table of centroids and all the months histograms) to this article can be found at https://github.com/paulamacaira/Araujo_et_al_ICREPQ_2022.

References

[1] International Renewable Energy Agency 2021, accessed 5 March 2022, < https://www.irena.org/>
[2] Empresa de Pesquisa Energética 2021, accessed 6 March 2022, <https://www.epe.gov.br/pt/abcdenergia/matriz-energetica-e-eletrica#ELETRICA>
[3] Esteves, G. R. T., Maçaira, P. M., Oliveira. F. L. C., Amador, G. and Souza, R. C., "Improvements in the current Brazil's energy dispatch optimization: Load forecast and wind power", ICORES 2019 - Proceedings of the 8th International Conference on Operations Research and Enterprise Systems (2019). Pp. 398-405.
[4] Pinson, P., "Wind energy: Forecasting challenges for its operational management", Statistical Science (2013). Vol. 28, pp. 564–585.
[5] Almutairi, A., Hassan Ahmed, M. and Salama, M. M. A., "Use of MCMC to incorporate a wind power model to evaluate generating capacity adequacy", Electric Power Systems Research (2016). Vol. 133, pp. 63–70.
[6] Maçaira, P. M., Cyrillo, Y. M., Oliveira, F. L. C. and Souza, R. C., "Including wind power generation in Brazil's long-term optimization model for energy planning", Energies (2019). Vol. 12, n. 826.
[7] MacQueen, J., "Some methods for classification and analysis of multivariate observations". In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (1967). Oakland, CA, USA, p. 14.
[8] Celebi, H. A. V. P. A., M. E.; Kingravi., "A comparative study of efficient initialization methods for the k-means clustering algorithm", Expert Systems with Applications (2013). 40:200–210.
[9] Gagniuc, P. A., "Markov Chains: From Theory to Implementation and Experimentation", USA, NJ: John Wiley Sons (2017).
[10] van Ravenzwaaij, P. B. S. D., Don; Cassey, "A Simple Introduction to Markov Chain Monte–Carlo Sampling". Psychonomic Bulletin Review (2016). 25:143–154. ISSN 0167-6105.
[11] R Core Team 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.