57º SBPO
SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL
Gramado • RS
DE 5 A 9 DE OUTUBRO DE 2025

REPRESENTAÇÃO DE CORRELAÇÕES ENTRE FONTES RENOVÁVEIS DE ENERGIA NA AMOSTRAGEM POR ESPAÇO DE ESTADOS USANDO *KERNEL DENSITY ESTIMATION* E REDES BAYESIANAS

Mateus de Oliveira Vaz
Universidade Federal do Rio de Janeiro
Av. Athos da Silveira Ramos, 149. Cidade Universitária, Rio de Janeiro.
mateus.vaz@poli.ufrj.br

Carmen Lucia Tancredo Borges
Universidade Federal do Rio de Janeiro
Av. Athos da Silveira Ramos, 149. Cidade Universitária, Rio de Janeiro.
carmen@nacad.ufrj.br

RESUMO

O presente artigo explora um método para a preservação de correlações entre variáveis aleatórias na técnica de amostragem por espaço de estados adotada na Simulação Monte Carlo Não Sequencial. O método proposto combina a técnica *Kernel Density Estimation* com Redes Bayesianas para armazenar as relações de dependência entre variáveis. O modelo envolve a transformação dos dados de entrada históricos usando normalização e agregação. O algoritmo foi avaliado por meio de séries temporais reais de geração renovável (eólica e solar) e de carga da região Nordeste do Brasil. As séries originais e amostradas foram então comparadas usando gráficos de dispersão e métricas de correlação para validar a efetividade do método.

PALAVRAS CHAVE. ***Kernel Density Estimation*****, Redes Bayesianas, Simulação Monte Carlo Não Sequencial.**

Tópicos
OSP-Otimização de Sistemas de Potência
SIM-Simulação

ABSTRACT

This paper explores a method for preserving correlations between random variables in the state-space sampling technique that is adopted in Non-Sequential Monte Carlo Simulation. The proposed method combines the Kernel Density Estimation technique with Bayesian Networks to encode dependency relationships between variables. The model involves transforming historical input data through normalization and aggregation. The algorithm was evaluated using actual renewable generation (wind and solar) and load time series for the Northeast region of Brazil. The original and sampled series were then compared using scatter plots and correlation metrics to validate the effectiveness of the method.

KEYWORDS. Kernel Density Estimation. Bayesian Networks. Non-Sequential Monte Carlo Simulation.

Paper topics
OSP-Power Systems Optimization
SIM-Simulation

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

1. Introdução

A função primária dos Sistemas Elétricos de Potência é assegurar o fornecimento de energia aos consumidores da forma mais econômica possível, considerando ao mesmo tempo padrões estabelecidos de continuidade, qualidade e segurança. No entanto, falhas aleatórias em equipamentos e componentes podem comprometer tal objetivo, trazendo a necessidade de estudos de confiabilidade para o direcionamento de investimentos financeiros na rede, de modo a reduzir a probabilidade e impacto das falhas. Para este fim, técnicas probabilísticas como a Simulação Monte Carlo se sobressaem uma vez que incorporam probabilidades de ocorrências de eventos, diferentemente de métodos determinísticos [Li et al., 2013].

A Simulação Monte Carlo (SMC) é uma abordagem que pode ser aplicada à séries temporais cronológicas ou usando representação por espaço de estados. No primeiro caso é utilizada a SMC Sequencial, a qual pode representar adequadamente as dependências estatísticas entre as séries temporais mas possui um elevado custo computacional, tornando-a impraticável para sistemas de grandes dimensões. O segundo caso requer o uso da SMC Não Sequencial, que é bem menos custosa computacionalmente mas considera que as variáveis são estatisticamente independentes. Esta suposição pode levar a resultados imprecisos e discrepâncias consideráveis se comparados ao primeiro método, especialmente na presença de correlações espaciais ou temporais, comuns em séries históricas de vazões, velocidade do vento, irradiação solar e carga.

Em [Borges e Dias, 2016], um modelo não paramétrico baseado no método *Kernel Density Estimation* (KDE) foi usado para estimar as funções densidade de probabilidade (PDF) e funções de distribuição acumuladas (CDF) de variáveis aleatórias representadas por séries temporais. Tal modelo é então utilizado na SMC Não Sequencial para obter a mesma precisão da SMC Sequencial. Após transformações para normalizar e agregar os dados de entrada, as séries modificadas foram usadas para treinar uma Rede Bayesiana (RB), que armazena a informação da dependência entre as variáveis. A seguir, uma SMC Não Sequencial foi realizada, percorrendo a estrutura treinada da rede. Finalmente, depois de reverter as transformações nos dados, as amostras geradas - que preservam as correlações presentes nas séries históricas originais - foram usadas para calcular índices de confiabilidade composta no sistema RTS modificado, com resultados precisos e computacionalmente mais eficientes. Este método é aplicável a conjuntos de séries históricas com bases temporais diferentes, tornando-o mais genérico que técnicas de amostragem da mesma hora.

Em [Wang e Zou, 2020a], a RB foi treinada por meio de um procedimento similar ao descrito no trabalho anterior, e as correlações entre velocidade do vento, irradiação solar e carga - cujas séries temporais foram construídas usando um modelo de média móvel autoregressiva - foram analisadas. As amostras da RB foram então comparadas àquelas obtidas utilizando cópulas (*C-vine Copula*), sendo realizados a confrontação de ambas com as séries originais e o cálculo do fluxo de potência com os dados probabilísticos em um sistema IEEE 118 barras modificado.

Em [Wang e Zou, 2020b] é apresentada uma metodologia bastante parecida com a realizada em [Wang e Zou, 2020a]. Novamente é construída uma RB para verificação da dependência não linear entre dados de velocidade do vento, irradiação solar e carga, e suas amostras são comparadas àquelas obtidas por cópulas (*C-vine Copula*) e por uma variante da amostragem por hipercubo latino em termos de seus gráficos de dispersão, distribuições e resultados do cálculo do fluxo de potência probabilístico para um sistema de distribuição IEEE 69 barras.

Em [Wang e Zou, 2022] é considerada uma abordagem diferente das anteriores, com o treinamento de uma RB dinâmica. Este tipo de rede adiciona a informação de tempo e descreve o processo de evolução das variáveis, permitindo inclusive verificar de maneira explícita sua autocorrelação. Foi analisada a dependência entre séries temporais de geração eólica e carga, sendo realizada uma amostragem por Simulação Monte Carlo Sequencial na RB dinâmica. As carac-

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

terísticas estatísticas da amostra resultante foram comparadas com aquelas obtidas por um método alternativo que utiliza o modelo escondido de Markov.

O objetivo deste artigo é analisar as correlações de um conjunto de séries históricas de geração eólica, solar e curvas de carga da Região Nordeste do Brasil, seguindo o procedimento descrito em [Borges e Dias, 2016]. Os dados foram obtidos a partir do site do Operador Nacional do Sistema Elétrico [ONS, 2024a] [ONS, 2024b]. Também foram usadas medições da velocidade do vento de uma planta específica da região, convertidas em potência utilizando a curva de carga de um determinado modelo de aerogerador [Bauer e Matysik, 2024]. O propósito foi identificar o nível de correlação entre as medidas e os dados agregados de toda a região Nordeste.

O artigo é organizado da seguinte forma: a próxima seção faz um breve resumo dos principais conceitos teóricos que permeiam o estudo realizado e descreve a metodologia aplicada; a terceira seção apresenta e discute os resultados obtidos; a quarta seção contém as conclusões e considerações finais.

2. Metodologia

O processo seguido no estudo é mostrado no diagrama da Figura 1, sendo melhor detalhado em [Borges e Dias, 2016].

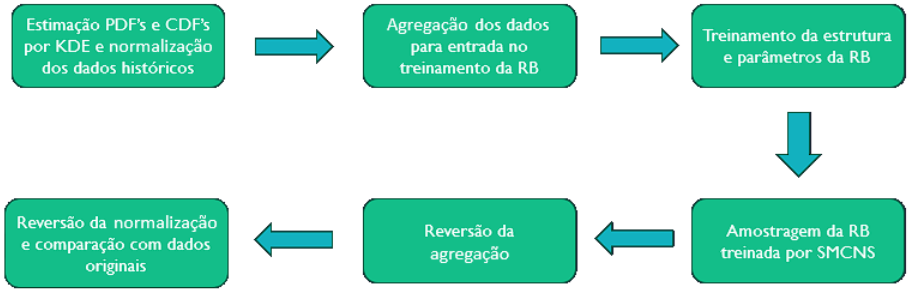


Figura 1: Sequência de passos realizados para geração das amostras e comparação com os dados históricos.

O primeiro passo consiste na normalização dos dados para uma faixa de [0,1] usando as distribuições estimadas pelo KDE. Inicialmente, as PDFs das variáveis são aproximadas por *kernels* gaussianos. Em seguida suas CDFs são obtidas por integração e aplicadas às séries históricas originais, o que resulta na sua normalização. Dado que as CDFs são funções monotonicamente crescentes e bijetivas, suas inversas existem e foram armazenadas para uso posterior.

O segundo passo refere-se à agregação das séries normalizadas de maneira tal que conserve a informação mútua no conjunto de variáveis cujas correlações serão preservadas. Para tanto, foi feito o cálculo de uma versão simplificada do Coeficiente de Informação Máximo (MIC) com o objetivo de maximizar a dependência entre as variáveis.

No terceiro passo, a RB é treinada usando as séries agregadas como entrada. Primeiro a estrutura da rede é aprendida através do método de otimização *Hill Climbing*, cuja métrica de escore é dada pelo algoritmo K2. Já a aprendizagem dos parâmetros é feita com a Estimação por Máxima Verossimilhança.

No quarto passo é feita a amostragem da RB treinada, percorrendo a estrutura dos nós pais aos nós filhos. Dez diferentes amostras foram obtidas, e sua média foi calculada para se chegar ao valor final de cada amostra.

No quinto passo é feita a reversão da agregação das variáveis amostradas por meio do sorteio de números reais, com distribuição uniforme, dentro dos limites dos intervalos entre os

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

agrupamentos.

No sexto passo, é feita a reversão da normalização aplicando às séries a inversa da CDF obtida na primeira etapa.

2.1. *Kernel Density Estimation*

KDE é uma técnica não paramétrica para estimação da PDF de variáveis aleatórias baseada num conjunto de dados observados. Ao contrário de métodos paramétricos, que assumem uma determinada distribuição de probabilidade para a variável, o KDE a estima por meio de funções de base (*kernels*) centradas em cada dado da amostra, o que confere à técnica uma maior capacidade de capturar padrões complexos.

A estimação para um ponto x é calculada como a média ponderada das contribuições dos *kernels* posicionados nos dados, de acordo com a fórmula [Bishop e Nasrabadi, 2006]:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{1}$$

onde n é o tamanho da amostra, h é o parâmetro de suavização ou de largura de banda, e $K(\cdot)$ é a função *kernel* - opções comumente escolhidas incluem o *kernel* Gaussiano e o uniforme.

O parâmetro h controla o grau de suavidade da densidade estimada e deve ser ajustado de forma cuidadosa para garantir uma maior qualidade na estimativa. Valores muito pequenos de h focam excessivamente em detalhes dos dados e são suscetíveis a ruídos (*overfitting*), enquanto valores muito grandes falham em capturar padrões importantes da PDF (*underfitting*).

2.2. Redes Bayesianas

As RBs são modelos probabilísticos gráficos que representam relações de dependência entre um conjunto de variáveis aleatórias por meio de um grafo acíclico direcionado. Cada nó no grafo corresponde a uma variável aleatória, enquanto as arestas direcionadas indicam relações de dependência condicional entre as variáveis.

Em termos matemáticos, uma RB modela a distribuição conjunta das variáveis como o produto das distribuições condicionais locais [Bishop e Nasrabadi, 2006]:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{Pais}(X_i)), \tag{2}$$

onde Pais(X_i) são as variáveis referentes aos nós pais de X_i no grafo. Comparando a Equação 2 com a aplicação da regra do produto para a mesma probabilidade conjunta, pode-se notar que a principal diferença entre elas é a ausência de ligações entre variáveis independentes na representação pela RB.

A construção de uma RB envolve dois componentes principais: a estrutura do grafo e seus parâmetros. A primeira especifica as relações de dependência entre as variáveis, que podem ser aprendidas a partir de dados ou definidas com base em conhecimento prévio. Já os parâmetros descrevem as distribuições condicionais $P(X_i|\text{Pais}(X_i))$, geralmente modeladas por distribuições discretas ou contínuas.

2.3. Métricas estatísticas

Para a validação dos resultados, gráficos de dispersão foram gerados de forma a comparar as amostras e suas correlações com os dados originais. Diferentes métricas estatísticas também foram calculadas para confirmar a equivalência entre as amostras e séries históricas. Elas são as seguintes:

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

1. Erro médio arcotangente absoluto proporcional (MAAPE): calcula o ângulo que mede a discrepância entre os valores históricos e previstos usando a função arcotangente, como representado na fórmula [Kim e Kim, 2016]:

$$\text{MAAPE} = \frac{1}{n}\sum_{i=1}^{n}\left[\arctan\left(\left|\frac{y_i - f_i}{y_i}\right|\right)\right], \tag{3}$$

onde y_i e f_i denotam os valores históricos e previstos para o dado i.

2. Coeficiente de determinação (R2): mede o grau de variação na variável dependente que é previsível pela variável independente. É dado pela seguinte expressão [Kambezidis et al., 2025]:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y})^2}, \tag{4}$$

onde $\overline{y}$ é a média dos dados históricos definida como:

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,. \tag{5}$$

3. MIC: mede a força da correlação linear ou não linear entre duas variáveis aleatórias. Seleciona um determinado particionamento dos dados que maximiza sua informação mútua. Uma possível formulação de otimização para a métrica é [Reshef et al., 2011]:

$$\text{MIC} = \max_{b_y, b_f} \frac{I(y, f, b_y, b_f)}{\log \min(b_y, b_f)}, \tag{6}$$

onde b_y e b_f são respectivamente o número de particionamentos dos dados históricos e previstos, e $I(y, f, b_y, b_f)$ é a informação mútua máxima como uma função dos dados históricos e previstos e seus particionamentos. A divisão pelo logaritmo assegura que a métrica permaneça na faixa [0,1].

4. Teste dos postos sinalizados de Wilcoxon: testa a hipótese nula de que duas amostras relacionadas vêm da mesma distribuição, verificando se a distribuição de suas diferenças é simétrica com relação à zero. Para a hipótese alternativa bicaudal (diferença entre amostras é simétrica em torno de um valor $\neq 0$, a estatística é calculada por [Conover, 1999] [Pratt e Gibbons, 2012]:

$$W = \min\left(\sum_{i:y_i - f_i > 0} R_i, \sum_{i:y_i - f_i < 0} R_i\right), \tag{7}$$

onde R_i são os postos definidos como os valores ordenados de $|y_i - f_i|$, preservando seus sinais originais. Considerando um grande número de pares de dados, W pode ser aproximado por uma distribuição normal. Portanto, a estatística Z e seu valor p correspondente (para uma hipótese alternativa bicaudal) são dados por [Conover, 1999] [Pratt e Gibbons, 2012]:

$$Z_W = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{8}$$

e

$$p = 2 \times P(Z > |Z_W|) \tag{9}$$

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

5. Teste de Kolmogorov-Smirnov (KS) para duas amostras: mede a diferença entre duas distribuições empíricas de probabilidade para testar a hipótese nula de que ambas vêm da mesma distribuição. As funções de distribuição empíricas para os valores históricos e previstos podem ser definidas como [Hollander et al., 2015]:

$$Y(x) = \frac{\text{número de amostras em y} \leq x}{n_y} \tag{10}$$

e

$$F(x) = \frac{\text{número de amostras em f} \leq x}{n_f}, \tag{11}$$

onde n_y e n_f são respectivamente o número de elementos nos dados históricos e previstos. A partir destas formulações, a estatística e seu valor p associado são calculados como [Hollander et al., 2015]:

$$D_{n_y,n_f} = \sqrt{\frac{n_y n_f}{n_y + n_f}} \sup_x |Y(x) - F(x)| \tag{12}$$

e

$$P(D_{n_y,n_f} > d) \approx \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 d^2}, \tag{13}$$

onde sup representa o supremo do conjunto de distâncias.

6. Teste de Anderson-Darling para K-amostras: testa se duas ou mais amostras vêm da mesma distribuição, e é mais sensível à discrepâncias nas caudas das distribuições. A seguinte expressão pode ser usada para calcular a estatística [Scholz e Stephens, 1987]:

$$A^2 = \frac{1}{N} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N-1} \frac{(NM_{ij} - jn_i)^2}{j(N-j)}, \tag{14}$$

onde $N = \sum_k n_k$ é a amostra agrupada e M_{ij} é o número de observações na i-ésima amostra que não são maiores do que Z_j, que é o j-ésimo elemento na amostra agrupada ordenada ($Z_1 < ... < Z_j < ... < Z_N$). Com respeito ao cálculo do valor p diferentes abordagens podem ser citadas, como o uso de aproximações empíricas e a Simulação Monte Carlo. No método `Anderson_ksamp` da biblioteca Scipy, aqui utilizada, o valor p é estimado baseado numa tabela de pontos percentuais da cauda superior. Recomenda-se a referência [Scholz e Stephens, 1987] para maiores informações.

7. Divergência Kullback-Leibler (KL): corresponde a uma medida do quanto uma distribuição de probabilidade difere de outra distribuição. Mais especificamente, é a quantidade média de informação adicional requerida para especificar o valor de uma variável aleatória como resultado de usar uma distribuição aproximada em vez da distribuição verdadeira. A divergência Kullback-Leibler entre as distribuições discretas históricas e previstas é dada por [Bishop e Nasrabadi, 2006]:

$$D_{KL}(Y||F) = \sum_x \left[Y(x) \log \left(\frac{Y(x)}{F(x)} \right) \right] . \tag{15}$$

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

8. Divergência Jensen-Shannon (JS): é uma versão melhorada da divergência Kullback-Leibler, uma vez que é simétrica e sempre retorna um número finito. Sua definição matemática é dada abaixo [Nielsen, 2019]:

$$D_{JS}(Y||F) = \frac{1}{2}D_{KL}(Y||M) + \frac{1}{2}D_{KL}(F||M), \tag{16}$$

onde M é a distribuição média definida como:

$$M(x) = \frac{Y(x) + F(x)}{2}\,. \tag{17}$$

3. Resultados e discussão

Os seguintes pares de conjuntos de dados históricos da região NE do Brasil foram avaliados e serão referidos pela numeração a seguir:

1. Geração eólica medida em uma planta específica e geração eólica do NE - 2011-2015;
2. Geração eólica medida em uma planta específica e geração eólica do NE - apenas 2015;
3. Geração eólica do NE e curva de carga do NE - 2011-2015;
4. Geração solar do NE e curva de carga do NE - apenas 2016.

As correlações entre as duas séries temporais em cada par de dados foram medidas pelos coeficientes de Pearson e τ de Kendall e são mostradas na Tabela 1.

Variáveis aleatórias	**Pearson**	**τ de Kendall**
Par de dados 1)	0,164	0,161
Par de dados 2)	0,488	0,324
Par de dados 3)	0,347	0,314
Par de dados 4)	0,234	0,118

Tabela 1: Correlação entre séries temporais em cada conjunto de dados.

Pode ser notado que as maiores e menores correlações vêm dos pares 2 e 1, respectivamente. Vale ressaltar que o tamanho das séries (período anual ou de cinco anos) pode exercer uma influência considerável.

3.1. Comparação entre variáveis originais e amostradas (monovariada)

Inicialmente, as variáveis aleatórias amostradas foram comparadas às suas correspondentes históricas. Um detalhe importante refere-se ao fato de que os dados originais são ordenados no tempo, enquanto as amostras são aleatórias e não possuem tal cronologia. Desse modo, para calcular as métricas MAAPE, R2 e MIC foi necessária a colocação das séries originais e amostradas em ordem crescente de valores. Para as métricas restantes, este ordenamento não foi necessário.

A Tabela 2 mostra os resultados para cada variável aleatória. As métricas que não são satisfeitas foram marcadas em vermelho. Nota-se que todos os valores da métrica R2 são próximos de 1, mostrando que as variáveis amostradas são altamente correlacionadas com as originais, como esperado. A métrica MIC também indica a forte correlação linear. Vale comentar ainda o valor dezenas de vezes maior do MAAPE para a geração solar do que para as outras variáveis.

Com respeito aos testes estatísticos de Wilcoxon, Kolmogorov-Smirnov e Anderson, a hipótese nula assume que os dados originais e amostrados vêm da mesma distribuição. Portanto,

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

Variável aleatória	MAAPE	R2	MIC	Valor p Wilcoxon	Valor p KS	Valor p Anderson
Geração eólica medida (2011-2015)	0,0065	0,9998	1,0000	0,7607	0	0,001
Geração eólica do NE (2011-2015)	0,0234	0,9993	1,0000	0,4827	0,3458	0,0618
Geração eólica medida (2015)	0,0078	0,9996	1,0000	0,2287	0	0,001
Geração eólica do NE (2015)	0,0071	0,9986	1,0000	0,9466	0,6780	0,25
Curva de carga do NE (2011-2015)	0,0025	0,9968	1,0000	0,3957	0,0395	0,003
Geração solar do NE (2016)	0,8211	0,9972	1,0000	0	0	0,001
Curva de carga do NE (2016)	0,0021	0,9986	1,0000	0,4985	0,7319	0,25

Tabela 2: Métricas de correlação entre as variáveis originais e amostradas.

espera-se que seus valores sejam maiores do que o nível de significância de 0,05, tendo em vista que valores menores representam forte evidência de rejeição à suposição de hipótese nula. Pode ser observado que as variáveis passam na maior parte dos testes. No entanto, a geração eólica medida para 2011-2015 e apenas 2015, bem como a curva de carga de 2011-2015, não passam em dois dos três testes, enquanto a geração solar falha em todos eles.

Investigações subsequentes são necessárias, mas possíveis explicações para os resultados obtidos incluem particularidades das distribuições das variáveis, em especial da geração solar onde um grande número de zeros está presente (devido ao período noturno).

3.2. Comparação dos pares de variáveis originais e amostradas (bivariada)

Figuras 2 a 5 mostram os gráficos de dispersão originais e amostrados para cada par de variáveis. Em todos os casos, distribuições similares para as variáveis originais e amostradas podem ser observadas.

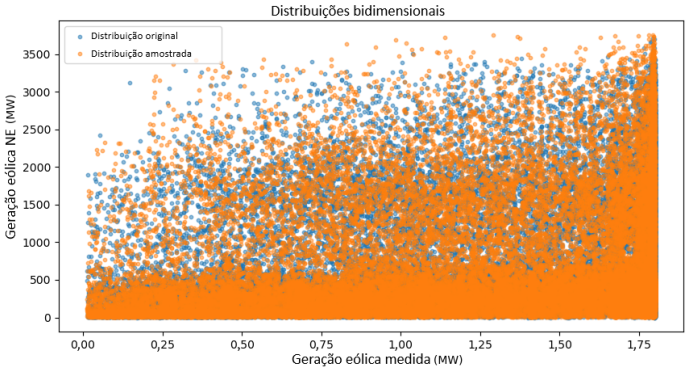


Figura 2: Gráficos de dispersão originais e amostrados - Par de dados 1.

Suas métricas de comparação quantitativas são apresentadas na Tabela 3. É possível observar que o MIC para o par de variáveis originais é bastante similar ao MIC do par de variáveis amostradas, com diferenças (Delta) próximas a zero, o que confirma a precisão da técnica de amostragem.

As divergências JS e KL referem-se à similaridade entre as distribuições bidimensionais originais e amostradas, sendo iguais a zero quando ambas são idênticas. Nota-se que os valores são muito próximos de zero no teste JS, significando que os dados originais e amostrados são próximos entre si. Já a divergência KL, apesar de numericamente mais elevada do que o teste anterior, é em

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

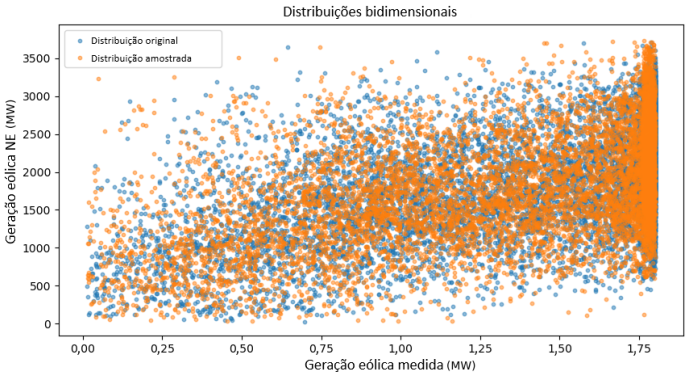


Figura 3: Gráficos de dispersão originais e amostrados - Par de dados 2.

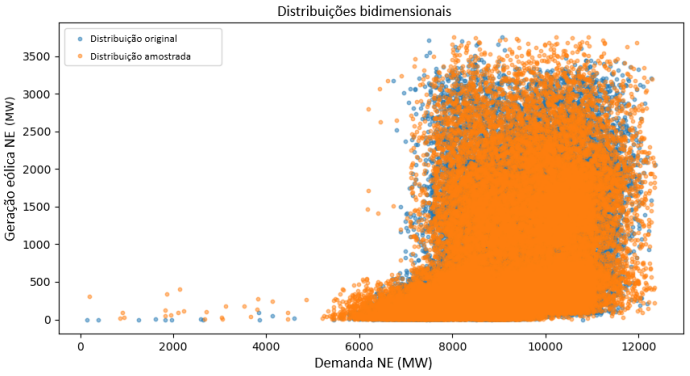


Figura 4: Gráficos de dispersão originais e amostrados - Par de dados 3.

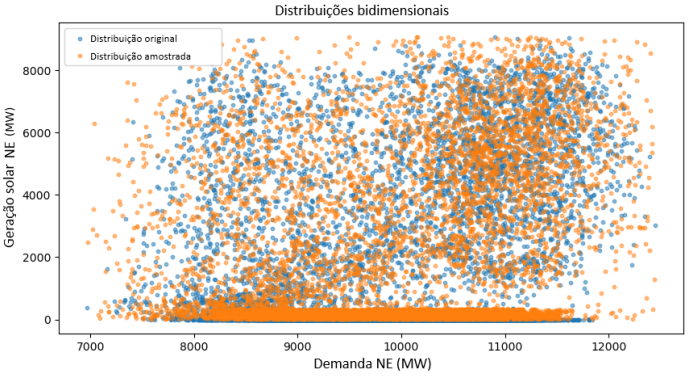


Figura 5: Gráficos de dispersão originais e amostrados - Par de dados 4.

Variáveis aleatórias	MIC			Divergência JS	Divergência KL
	Original	Amostrado	Delta		
Par de dados 1	0,1496	0,1030	0,0467	0,0200	0,1190
Par de dados 2	0,1988	0,1973	0,0015	0,0943	0,8590
Par de dados 3	0,1895	0,1857	0,0038	0,0146	0,0519
Par de dados 4	0,1529	0,1375	0,0154	0,2083	1,3017

Tabela 3: Métricas de correlação entre os pares de variáveis originais e amostradas.

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

geral menor que 1, com exceção do par de dados 4 que tem os maiores valores (menores similaridades) em ambos os testes. Este resultado provavelmente tem relação com o fato de que a geração solar falhou nos três testes estatísticos mostrados na Tabela 2.

3.3. Validação da amostragem bivariada

Para a validação da amostragem bivariada, séries temporais integradas foram calculadas para cada par de dados adicionando o par de variáveis originais a cada passo temporal, considerando sua natureza (geração e carga têm sinais diferentes). Dessa forma, nos pares de dados 1 e 2 a geração eólica local medida foi somada com a geração eólica total do NE, enquanto que nos pares de dados 3 e 4 a geração foi subtraída da carga (carga líquida). O propósito é utilizar a variável integrada para validar os resultados da amostragem bivariada sobre os pares de variáveis, usando as mesmas métricas aplicadas em 3.1.

A Tabela 4 apresenta as métricas comparando as variáveis integradas e a amostragem bivariada para cada par de dados. As métricas R2 e MIC são todas 1 ou muito próximas a este valor, indicando que as amostras bivariadas são quase idênticas às séries integradas. Os valores do MAAPE são muito próximos de zero para todas as variáveis integradas, o que também valida a técnica de amostragem. Com relação aos últimos três testes estatísticos, o par de dados 2 passa em todos eles, enquanto o par de dados 3 tem um valor p acima de 5% para o teste Wilcoxon. Entretanto, os pares de dados 1 e 4 falham em todos os testes, o que precisa ser investigado de forma mais profunda como ressaltado anteriormente.

Figuras 6 a 9 mostram os gráficos de cada variável amostrada, sua soma ou diferença e a variável original integrada, para cada par de dados, ordenadas dos valores menores aos maiores. As imagens demonstram que a soma das gerações ou a carga líquida obtidas a partir das amostras são praticamente iguais aos valores das séries integradas, novamente validando a técnica de amostragem bivariada.

4. Conclusão

Este artigo buscou demonstrar a aplicabilidade do KDE e RB para representação de dependências estatísticas entre séries temporais na abordagem que emprega a amostragem por espaço de estados. As amostras geradas foram validadas contra as séries históricas de geração renovável e demanda da região Nordeste do Brasil, considerando índices de dispersão e correlação. Os resultados indicam que as dependências estatísticas observadas nos dados históricos foram efetivamente preservadas nas amostras.

Tal aplicação é particularmente importante para os estudos de confiabilidade de Sistemas de Potência, uma vez que a técnica pode ser usada na SMC Não Sequencial para representar adequadamente as correlações entre geração renovável e carga com um esforço computacional muito menor do que o empregado na SMC Sequencial.

Referências

Bauer, L. e Matysik, S. (2024). Vestas v100-1.8 - 1,80 MW - Wind turbine. https://en.wind-turbine-models.com/turbines/1002-vestas-v100-1.8. Acessado: 2024-12-03.

Bishop, C. M. e Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

Variáveis aleatórias integradas	MAAPE	R2	MIC	Valor p Wilcoxon	Valor p KS	Valor p Anderson
Integradas x Amostras do par de dados 1	0,0391	0,9975	1,0000	0	0,0002	0,001
Integradas x Amostras do par de dados 2	0,0157	0,9989	1,0000	0,8584	0,8126	0,25
Integradas x Amostras do par de dados 3	0,0045	0,9960	1,0000	0,9773	0	0,001
Integradas x Amostras do par de dados 4	0,0147	0,9988	1,0000	0,0398	0,0001	0,001

Tabela 4: Métricas de correlação entre os pares de variáveis integradas originais e amostradas.

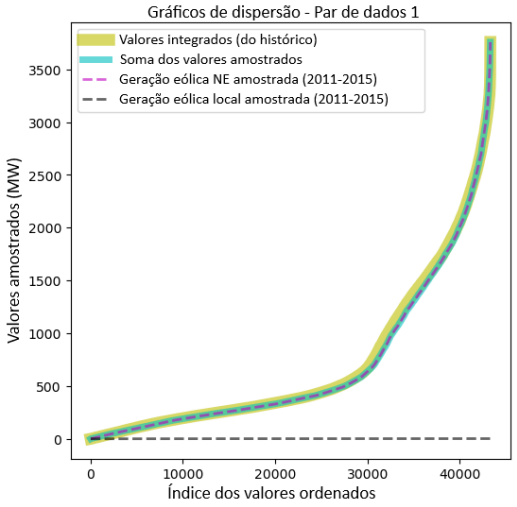


Figura 6: Variáveis integradas e soma das variáveis amostradas - Par de dados 1.

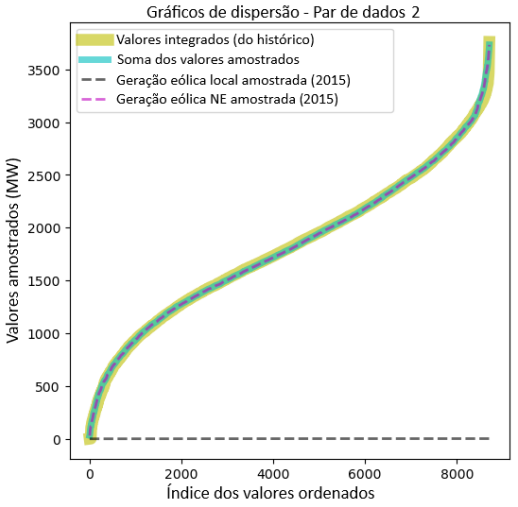


Figura 7: Variáveis integradas e soma das variáveis amostradas - Par de dados 2.

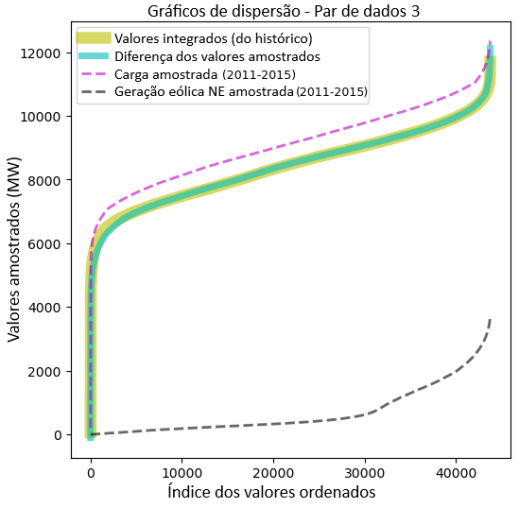


Figura 8: Variáveis integradas e diferença das variáveis amostradas - Par de dados 3.

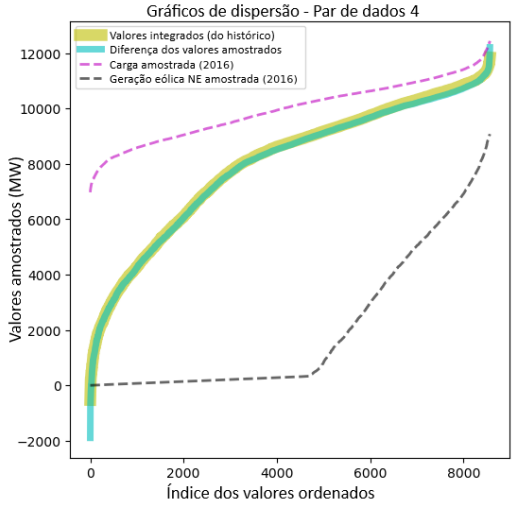


Figura 9: Variáveis integradas e diferença das variáveis amostradas - Par de dados 4.

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br

Borges, C. L. T. e Dias, J. A. (2016). A model to represent correlated time series in reliability evaluation by non-sequential monte carlo simulation. *IEEE Transactions on Power Systems*, 32 (2):1511–1519.

Conover, W. J. (1999). *Practical nonparametric statistics*. john wiley & sons.

Hollander, M., A. Wolfe, D., e Chicken, E. (2015). The two-sample dispersion problem and other two-sample problems. *Nonparametric statistical methods*, p. 151–201.

Kambezidis, H. D., Patelis, E., e Kavadias, K. A. (2025). An in-depth analysis of the ångström–prescott-type solar models: application for athens, greece. *Academia Environmental Sciences and Sustainability*, 2(1).

Kim, S. e Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, 32(3):669–679.

Li, W. et al. (2013). *Reliability assessment of electric power systems using Monte Carlo methods*. Springer Science & Business Media.

Nielsen, F. (2019). On the jensen–shannon symmetrization of distances relying on abstract means. *Entropy*, 21(5):485.

ONS (2024a). Curva de carga horária. `https://www.ons.org.br/Paginas/resultados-daoperacao/historico-da-operacao/curva`. Acessado: 2024-12-04.

ONS (2024b). Geração média diária e horária. `https://www.ons.org.br/Paginas/resultados-daoperacao/historico-da-operacao/geracao-media-diaria-horaria.aspx`. Acessado: 2024-12-04.

Pratt, J. W. e Gibbons, J. D. (2012). *Concepts of nonparametric theory*. Springer Science & Business Media.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., e Sabeti, P. C. (2011). Detecting novel associations in large data sets. *science*, 334(6062):1518–1524.

Scholz, F. W. e Stephens, M. A. (1987). K-sample anderson–darling tests. *Journal of the American Statistical Association*, 82(399):918–924.

Wang, H. e Zou, B. (2020a). Probabilistic computational model for correlated wind speed, solar irradiation, and load using bayesian network. *IEEE Access*, 8:51653–51663.

Wang, H. e Zou, B. (2020b). Probabilistic load flow calculation considering correlation based on bayesian network. *Electric Power Components and Systems*, 48(14-15):1571–1583.

Wang, H. e Zou, B. (2022). Wind power and load temporal dependence model based on dynamic bayesian network. *Electrical Engineering*, 104(3):1265–1276.

DOI: 10.59254/sbpo-2025-212158
https://proceedings.science/p/212158?lang=pt-br