An Approach for Data Treatment of Solar Photovoltaic Generation

J. F. M. Pessanha, A. C. G. Melo, R. P. Caldas and D. M. Falcão

***Abstract*— A good solar power photovoltaic generation forecast depends on good quality time series data from measurements of Global Horizontal Irradiance and Solar Power Generation. However, measurement system failures and errors in data handling can corrupt data records with gaps and outliers that undermine forecasting accuracy. Therefore, it is important that the fitting of solar energy prediction models must be preceded by a data analysis in order to detect and correct measurement errors. Given that Global Horizontal Irradiance and Solar Power Generation are correlated variables, this paper aims to present the main characteristics of an offline approach developed for the joint treatment of hourly data values of both variables in a photovoltaic plant. In the proposed methodology, the measurements of Global Horizontal Irradiance and Solar Power Generation are analyzed by using reanalysis data and statistical and data mining techniques for the correction of outliers and the filling of data gaps. The application of the approach is illustrated by the analysis of measurements from a real solar PV system.**

***Index Terms*— Data Cleaning, Self-Organizing Map, Solar Power Generation, Statistical Methods, Time Series.**

I. Introdução

SEGUINDO a tendência mundial [1], verifica-se no Brasil, um grande impulso na implantação de novas energias renováveis, como a energia eólica, que vem crescendo exponencialmente na matriz elétrica brasileira, e a energia solar fotovoltaica (SFV), que tem apresentado relevante evolução tecnológica, traduzindo-se em maior eficiência na conversão para energia elétrica e custos decrescentes de produção de seus componentes. No entanto, em função da natureza intrínseca da variabilidade da irradiância solar, a intermitência da geração SFV impõe um desafio para sua integração aos sistemas de energia elétrica [2,3]. Uma capacidade inadequada de prever a produção de energia SFV pode afetar adversamente a estabilidade, a confiabilidade e a programação da operação do sistema de energia, e o seu benefício econômico.

J. F M. Pessanha, Centro de Pesquisas de Energia Elétrica-Cepel, Universidade do Estado do Rio de Janeiro-UERJ, Rio de Janeiro, Brasil (e-mail: francisc@cepel.br).

A. C. G. Melo, Centro de Pesquisas de Energia Elétrica-Cepel, Universidade do Estado do Rio de Janeiro-UERJ, Rio de Janeiro, Brasil, (e-mail: albert@cepel.br).

R. P. Caldas, Universidade Federal do Rio de Janeiro, Brasil (e-mail: robertopcaldas@gmail.com).

D. M. Falcão Universidade Federal do Rio de Janeiro, Brasil (falcao@nacad.ufrj.br).

A geração SFV depende fundamentalmente da irradiância solar ao nível do solo, (*Global Horizontal Irradiance* - GHI), que por sua vez é influenciada por fatores meteorológicos, como nebulosidade, precipitação pluviométrica, vento e umidade, entre outros. Estas características têm promovido um amplo espectro de estudos e desenvolvimentos de métodos, técnicas e modelos de previsão de geração SFV [4-6].

Os modelos baseiam-se no histórico de medições de grandezas como geração e radiação solar e, portanto, para o seu bom ajuste e desempenho é fundamental que esses dados tenham a maior qualidade possível. Erros e perturbações levam a dados discrepantes, com descontinuidades e lacunas. Assim, é essencial o tratamento prévio dos dados [7,8].

Este artigo visa apresentar as principais características de uma metodologia desenvolvida para o tratamento de medições horárias de GHI e geração SFV em uma planta fotovoltaica.

Na metodologia proposta as medições de geração SFV e GHI são analisadas de forma conjunta, empregando técnicas estatísticas [9-12], algoritmos de mineração de dados [13-15] e dados de reanálise [16], visando a correção de valores discrepantes (*outliers*), a substituição de valores errôneos (*bad data*) e preenchimento de lacunas de dados (*missing data*).

Sua aplicação é ilustrada por meio da análise de medições provenientes de sistema SFV localizado em Monteroni di Lecce, Puglia, Itália, com 4.710 m^2 de área efetiva, composto por 3.000 módulos de 320 Wp, totalizando 960 kWp e eficiência nominal de 19,6% [5,6].

II. Problemas Típicos nas Medições

Em um dia com céu limpo, o perfil diário da GHI corresponde ao ilustrado na Fig. 1. Contudo, em função das condições meteorológicas o perfil diário sofre variações naturais e pode apresentar padrões distintos do caso ideal ilustrado na Fig. 1. Adicionalmente, falhas no sistema de medição e no manuseio dos dados podem corromper os registros horários com lacunas de dados (Fig. 2), dados discrepantes (Fig. 3) e erros na datação dos registros (Fig. 4).

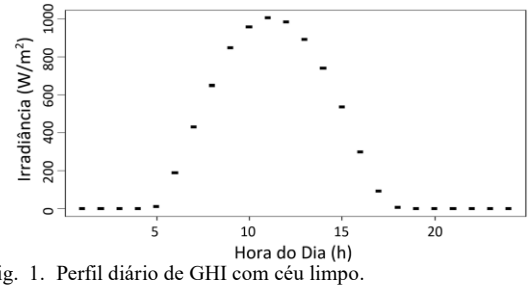


Fig. 1. Perfil diário de GHI com céu limpo.

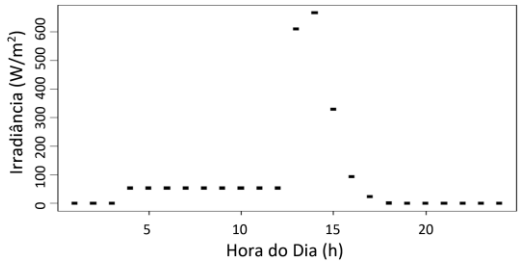


Fig. 2. Perfil diário de GHI com lacuna de dados.

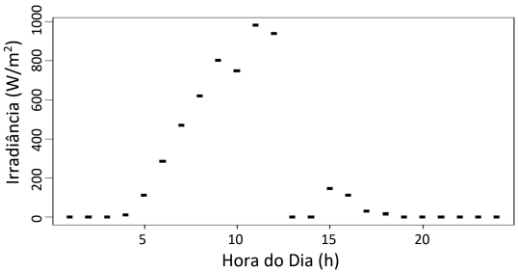


Fig. 3. Perfil diário de GHI com observações discrepantes.

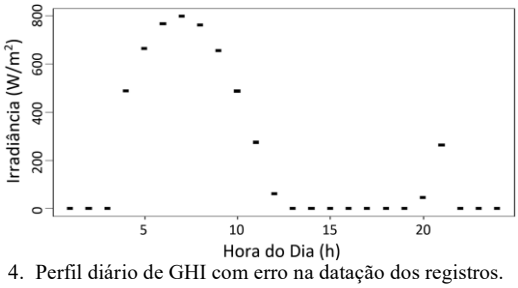


Fig. 4. Perfil diário de GHI com erro na datação dos registros.

Embora os perfis de GHI e geração SFV sejam afetados por causas naturais e não naturais, uma boa metodologia de tratamento de dados deve ser capaz de atuar apenas nas flutuações não naturais e evitar ou minimizar as atuações indevidas sobre as flutuações naturais. Para alcançar este objetivo, a metodologia proposta neste trabalho recorre ao uso combinado de métodos estatísticos [9,12], redes neurais artificiais [13-15] e dados de reanálise [16].

Dados de reanálises são produzidos por meio da assimilação de dados, uma técnica para construir conjuntos de dados de longo prazo, muito utilizada em estudos climáticos com base em modelos NWP (*Numerical Weather Prediction*) [22], num processo conhecido como análise retrospectiva — ou reanálise. A reanálise envolve a realização de assimilação de dados para períodos anteriores. Assim, sequências longas e abrangentes de valores (análises) de condições atmosféricas são produzidas, formando um conjunto de dados de reanálise.

III. Metodologia Proposta

Trabalhos anteriores abordaram alguns aspectos de tratamento de dados, como em [19], onde foram elaborados filtros de dados, constituídos de testes lógicos nos valores, verificando se uma medida estava entre um limite máximo e mínimo, retornando uma resposta booleana (dado aprovado ou reprovado, sendo que, neste último caso, ignorado na formação do conjunto de dados tratado). Os autores, entretanto, reconhecem que análises estatísticas suplementares poderiam colaborar significativamente para resultados mais completos. Em [20], são propostos dois métodos para detecção de *outliers* em dados de carga num sistema de distribuição de energia elétrica, um baseado em curvas representativas para cada dia da semana que é baseado em diferenças absolutas dos dados, *Boxplot* e testes com base em medidas de desvio padrão e absoluto médio. Adicionalmente, os dados detectados como *outliers*, são substituídos por dados do mesmo tipo, localizados em outros instantes de tempo, e que não foram identificados previamente como *outliers*.

O presente trabalho contribui com o uso de outras técnicas, além do *Boxplot*, como regressão não paramétrica (e.g., LOESS [7,12]) e mapa de Kohonen [13-15]. Também, dada a forte dependência entre a geração SFV e a GHI, a estratégia adotada consistiu no tratamento conjunto, e não separado, dos dados das medições de GHI e SFV. Adicionalmente, a metodologia proposta permite integrar dados de reanálise no tratamento das medições.

A Fig. 5 fornece uma visão geral das cinco etapas da metodologia proposta para o tratamento de dados.

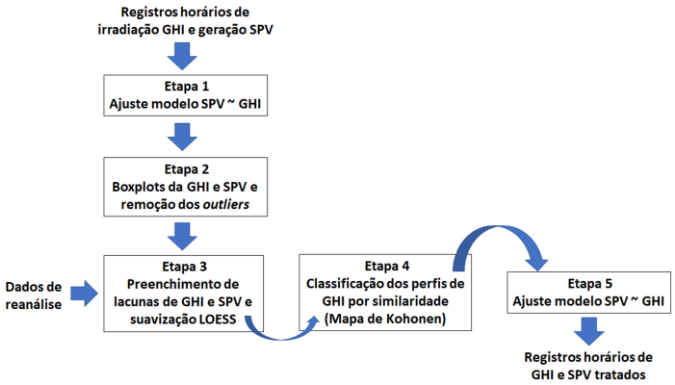
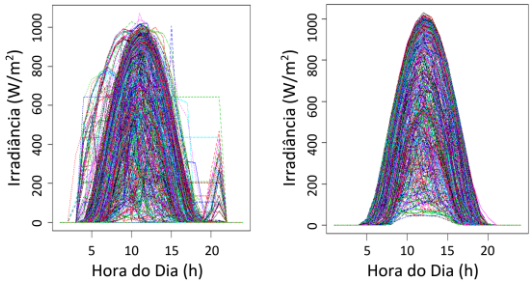


Fig. 5. Visão geral da metodologia proposta.

A. Etapa 1 Modelagem da Relação GHI x Geração SFV

Inicialmente, deve-se avaliar a compatibilidade dos perfis diários das medições de irradiância com os perfis de irradiância informados pelos dados de reanálises. Conforme ilustrado na Fig. 6, a comparação entre os perfis pode revelar algumas medições com valores de irradiância fora do intervalo definido entre o nascer e o pôr do sol. A Fig. 7 ilustra uma situação de incompatibilidade (erro por deslocamento temporal dos dados medidos) e a respectiva correção proposta.



(a) (b)

Fig. 6. Perfis diários de irradiância: (a) Medições; (b) Reanálises.

A incompatibilidade entre um perfil medido e a reanálise é evidenciada pelas diferenças observadas nas horas do nascer e do pôr do sol. A correção dos dados medidos consiste em

fazer a translação do perfil medido até que seja alcançada a minimização da soma dos desvios absolutos entre o perfil medido e o respectivo perfil na reanálise. A mesma translação aplicada a um perfil diário de GHI deve ser aplicada ao respectivo perfil de geração SFV.

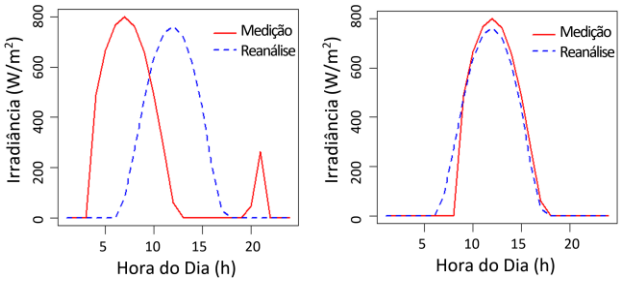


Fig. 7. Correção de erro (deslocamento temporal) da medição: (a) medição com erro; (b) medição corrigida.

Na sequência, ajuste o seguinte modelo de regressão linear simples aos dados brutos de geração SFV e GHI, em que ε_t denota o termo aleatório típico dos modelos de regressão linear, onde β_0 e β_1 são parâmetros estimados por mínimos quadrados [9]:

$$SFV_t = \beta_0 + \beta_1 GHI_t + \epsilon_t \ , \quad \forall\, t \in \{4\text{h}00, \ldots, 21\text{h}00\} \tag{1}$$

Substitua os registros nulos ou negativos nas medições de irradiância pelas respectivas estimativas oriundas das reanálises [16] e use a equação de regressão obtida em (1) para estimar as potências correspondentes.

B. *Etapa 2 Filtragem dos Registros Superestimados*

Faça *boxplots* [9,10] dos registros de GHI e de geração SFV para cada hora do dia; portanto, 24 *boxplots* para cada variável. Os *boxplots* permitem identificar os valores discrepantes em cada hora H ($\forall\ H$ =1,24), especialmente as medições superestimadas e que correspondem aos valores acima da cerca superior do *boxplot*, determinada pela seguinte expressão:

$$cerca_H = 3^o quartil_H + 1{,}5(3^o quartil_H - 3^o quartil_H) \tag{2}$$

Na sequência, em cada hora H, substitua os registros superestimados pelo respectivo valor horário da cerca superior.

C. *Etapa 3 Preenchimento de Lacunas nos Dados de GHI*

As lacunas de dados são sequências de valores repetidos; para os dias sem medições o mesmo valor é repetido nas 24 horas do dia. Inicialmente, calcule o desvio padrão das medições em cada dia - os dias com desvio padrão nulo correspondem aos dias sem medições ou com valores repetidos ao longo das 24 horas do dia. Na sequência, identifique os dias com lacunas de dados no horário entre 6h:00 e 18h00. Nos dias com lacunas que excedam 4 horas de duração, as lacunas devem ser preenchidas com dados provenientes da reanálise [16]. Já para os dias com lacunas de dados de até 4 horas, aplique um método de regressão não paramétrica (e.g., LOESS [7,12]) em cada perfil diário de GHI, com a finalidade de preencher as eventuais lacunas de curta duração por valores suavizados. Para um dado perfil diário de GHI, o algoritmo LOESS gera a sua versão suavizada, conforme o algoritmo a seguir, em que x é o vetor com as horas do dia, $x = \{1, \ldots, 24\}$ e y o vetor com o perfil horário de GHI. Para cada hora x_0 de 1 a 24 faça:

$$peso_i(x_0) = W\left[\frac{||x_0 - x_i||}{\Delta(x_0)}\right], \qquad onde\ W(u) = \begin{cases} (1-u^3)^3, & 0 \le u \le 1 \\ 0, & caso\ contrário \end{cases} \tag{3}$$

1) Pegue uma janela $\Delta(x_0)$ na vizinhança da hora x_0 e pondere os registros horários da seguinte forma:

2) Use o estimador de mínimo quadrados ponderados na regressão de y em x para obter estimativas suavizadas da GHI na hora x_0.

3) Calcule o desvio padrão (DP) dos desvios entre os valores medidos e suavizados para definir os limites de intervalos horários com 99,5% de confiança, conforme a seguir:

$$\begin{aligned} &\text{Limite superior } (x_0) = \text{valor suavizado } (x_0) + 3\ \text{DP}\ ; \\ &\text{Limite inferior } \ (x_0) = \text{valor suavizado } (x_0) - 3\ \text{DP} \end{aligned} \tag{4}$$

4) Substitua os valores fora dos intervalos de confiança pelos respectivos valores suavizados.

D. *Etapa 4 Análise de Agrupamentos dos Perfis de GHI*

Use um mapa de Kohonen [13-15] para formar grupos de perfis GHI similares (*clusters*), de tal forma que os perfis semelhantes sejam classificados em um mesmo grupo, enquanto perfis distintos sejam classificados em grupos diferentes.

O mapa de Kohonen ou mapa auto-organizável (SOM – *Self Organizing Map*) é um tipo de rede neural com treinamento competitivo e não supervisionado, aplicada em *data clustering* [13-15]. Por meio dela pode-se projetar um conjunto de dados com N padrões de entrada com dimensão p em um espaço de dimensão reduzida, usualmente duas dimensões, de tal forma que na projeção, a proximidade dos objetos no espaço de entrada é preservada no espaço de saída. Na rede SOM, os neurônios se organizam em uma grade ou reticulado, geralmente bidimensional, conforme mostra a Fig. 8, onde cada neurônio recebe todas as entradas.

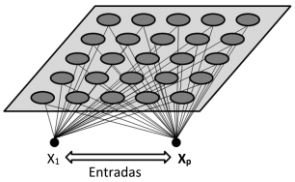


Fig. 8. Arquitetura da rede SOM.

No treinamento da rede, dado um padrão de entrada, os neurônios competem entre si para ver quem gera a maior saída (neurônio vencedor), i.e., quem é o neurônio com o vetor de pesos mais parecido com o padrão de entrada. Identificado o

neurônio vencedor, tem-se o início do processo de atualização dos pesos deste neurônio e de seus vizinhos. Durante a fase de aprendizado, os neurônios se especializam na detecção de um conjunto de padrões de entrada e se organizam topologicamente, fazendo com que os padrões detectados por um dado neurônio estejam relacionados com a posição do neurônio no reticulado. Assim, padrões de entrada semelhantes são detectados por neurônios próximos dentro do reticulado. O treinamento da rede SOM ocorre em duas fases [14]: fase de ordenação e fase de convergência. Na fase de ordenação, os vetores de pesos, inicialmente orientados de forma aleatória, são topologicamente ordenados de forma a agrupar os neurônios em *clusters* que reflitam a distribuição espacial dos padrões de entrada. Ao final desta fase a rede descobre quantos *clusters* deve identificar e quais as suas posições relativas no mapa. Durante o treinamento, a taxa de aprendizado é decrescente e a região de vizinhança se reduz gradualmente, A fase de convergência faz um ajuste fino do mapa obtido na fase ordenação, utilizando uma taxa de aprendizado baixa e um pequeno raio de vizinhança, que envolve um ou nenhum neurônio vizinho. Durante o treinamento a rede organiza topologicamente os neurônios de tal forma que os neurônios próximos respondem da mesma forma à padrões de entrada semelhantes. A seguir tem-se o algoritmo de treinamento da rede SOM [14,15]:

1) Inicialize os pesos (w_{ij}) e os parâmetros da rede SOM: taxa de aprendizado η e vizinhança. O peso w_{ij} é o peso da conexão entre o i-ésimo elemento do padrão de entrada (x_i) e o neurônio *j*.
2) Para cada padrão de treinamento *x*:
   Identifique o neurônio vencedor.
   Atualize os pesos deste neurônio e de seus vizinhos.
   Se o neurônio j ∈ Λ(t), a vizinhança do neurônio vencedor no instante de tempo t, faça

$$W_{ij}(t+1) = W_{ij}(t) + \eta(t)\left[x_i(t) - W_{ij}(t)\right] \quad (5)$$

   caso contrário, faça

$$W_{ij}(t+1) = W_{ij}(t) \quad (6)$$

   Reduza a taxa de aprendizado η e a área de vizinhança se o número de ciclos de treinamento for múltiplo de *N*.
3) Repita o passo 2 até que o mapa não mude.

A similaridade entre os perfis classificados em um mesmo *cluster* permite obter um perfil típico para cada *cluster*, bem como identificar perfis atípicos ou observações discrepantes que ainda possam estar presentes. A identificação de valores horários discrepantes em cada *cluster* pode ser realizada por meio de *boxplots* horários, como na etapa 2, porém desta vez com *boxplots* construídos apenas com base nos perfis em um mesmo *cluster* e os valores discrepantes em uma hora H são substituídos pela mediana da GHI na hora H.

E. Etapa 5 Ajuste de Novo Modelo de Regressão

Ajuste um modelo de regressão linear aos dados tratados, com a mesma especificação da equação (1). Na sequência, (a) calcule o intervalo de predição, com 97,5% de confiança, da geração SFV dada a radiação GHI, conforme indicado em (7), (b) identifique os *outliers*, i.e., os registros de SFV localizados fora dos limites dos intervalos de predição e (c) substitua os *outliers* pelas estimativas de geração SFV calculadas pelo modelo de regressão.

$$\beta_0 + \beta_1 GHI \pm t_{n-2}(2{,}5\%)\sqrt{\frac{SQRes}{n-2}\left[1+\frac{1}{n}+\frac{\left(GHI-\overline{GHI}\right)^2}{\sum_{i=1}^{n}\left(GHI_i-\overline{GHI}\right)^2}\right]} \quad (7)$$

Em (7) *n* é o número de observações, $t_{n-2}(2{,}5\%)$ é o quantil da distribuição t de Student com *n*-2 graus de liberdade, *SQRes* é a soma dos quadrados dos resíduos da regressão e $\overline{GHI}$ é a média amostral da irradiância.

IV. Experimento Computacional

O experimento computacional foi realizado em ambiente R [21] e baseou-se no sistema de geração SFV [5, 6], localizado no campus da Universidade de Salento, em Monteroni di Lecce (LE), Puglia, Itália (40°19'32"16N, 18°5'52"44E) desenvolvido no âmbito do projeto europeu "7th *Framework Programme Building Energy Advanced Management Systems* (BEAMS)". O sistema PV, com 4.710 m^2 de área efetiva, é composto por 3.000 módulos de 320 Wp. Com potência nominal total de 960 kWp e eficiência nominal de 19,6%, foi montado em estrutura metálica num estacionamento para veículos, segundo duas seções de inclinações diferentes (3° e 15°), e orientado para o sudeste (azimute de -10°). A primeira seção do sistema possui 1.104 módulos e uma área efetiva de 1.733,3 m^2, e a segunda, 1.896 módulos e área efetiva de 2.976,7 m^2. Conta ainda com inversores para injeção da energia gerada na rede, um sensor de temperatura ambiente, dois outros conjuntos de sensores (temperatura dos módulos e de irradiância solar global), um para cada seção do sistema, e com um sistema SCADA para coleta e armazenamento de dados. Foram utilizados os seguintes conjuntos de dados:

- medições do sistema SFV da Universidade de Salento [6], de período de 21 meses (de 1 de abril de 2012 até 30 de dezembro de 2013), abrangendo medições, em base horária, de GHI nas duas seções de painéis SFV e da geração total dos painéis PV ao longo de 639 dias ou 15.336 registros horários;
- dados horários de reanálise de irradiância solar direta e difusa, nas coordenadas geográficas do sistema SFV, disponibilizados pelo SARAH (*Surface Solar Radiation Data Set – Heliosat*) [16,18], na plataforma www.renewables.ninja.

A. Resultados da Etapa 1

Inicialmente foi verificado a compatibilidade dos perfis diários das medições de irradiância com os respectivos perfis oriundos da reanálise. Cerca de 15% dos perfis medidos apresentaram erros por deslocamento temporal dos dados, detectados pelas incompatibilidades com os dados da

reanálise, e assim foram transladados até alcançar a maior aderência ao respectivo perfil oriundo da reanálise. Na Fig. 9 apresentam-se os conjuntos de perfis diários das medições de GHI antes e após as translações dos perfis incompatíveis.

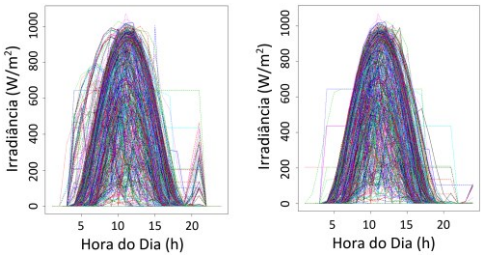


Fig. 9. Perfis de GHI: (a) com erros (deslocamento temporal); (b) após correção.

Na sequência, foi ajustado um modelo de regressão linear em que a geração SFV horária é explicada pela GHI (Fig. 10).

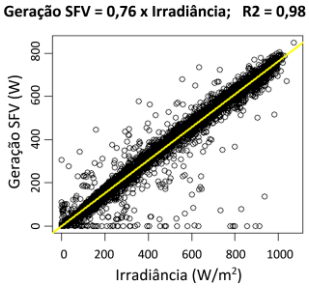


Fig. 10. Diagrama de dispersão e reta de regressão (sem erros de deslocamento temporal da medição).

B. *Resultados da Etapa 2*

Seguindo o algoritmo descrito na seção 3, foram construídos os *boxplots* horários de GHI (Fig. 11) e da geração SFV (Fig. 12). Nesta etapa, os *outliers* correspondem aos registros superestimados, i.e., os registros em cada hora situados acima das respectivas cercas superiores de GHI e SFV. Seguindo a metodologia proposta, os *outliers* foram substituídos pelas respectivas cercas superiores. Ao final, foram corrigidos 415 (2,7%) e 448 (2,9%) valores horários de irradiância e geração SFV respectivamente.

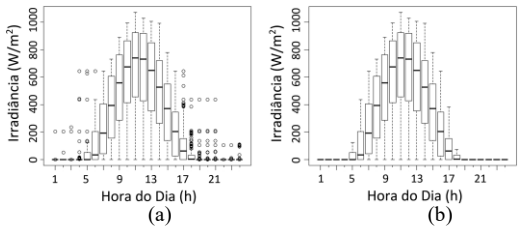


Fig. 11. *Boxplots* da GHI: (a) dados com *outliers*; (b) sem *outliers*.

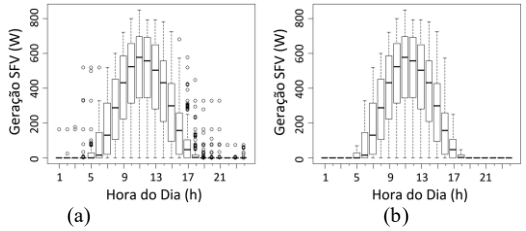


Fig. 12. *Boxplots* da geração SFV: (a) dados com *outliers*; (b) sem *outliers*.

C. *Resultados da Etapa 3*

No período entre 1 de abril de 2012 e 30 de dezembro de 2013 não foram identificados dias sem medições nas 24 horas do dia. Contudo, 175 (27%) dias possuem lacunas de dados entre 6h00 e 18h00. A distribuição de frequência das durações das lacunas nestes dias é ilustrada na Fig. 13. Nos dias em que a duração da lacuna (no intervalo entre 6h00 e 18h00) superar 4 horas, os dados de irradiância são substituídos pelos respectivos dados de reanálise. No total, em 6 dias as lacunas superam 4 horas. Nos demais 169 dias, as lacunas são menores que 4 horas. Na sequência, os perfis diários de irradiância são suavizados pelo LOESS e caso os registros horários situem-se fora do intervalo de confiança de 99,5%, eles são substituídos pelos valores suavizados e os respectivos registros de potência são corrigidos com base modelo de regressão previamente estimado na Etapa 1.

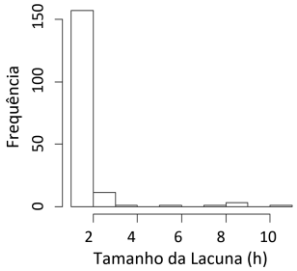


Fig. 13. Distribuição dos tamanhos das lacunas.

Na Fig. 14 apresentam-se alguns perfis de GHI com lacunas de dados e os respectivos perfis corrigidos (em vermelho o perfil original, em azul o perfil corrigido e em preto os limites dos intervalos de confiança LOESS). Ao todo, a correção atuou em 476 (3,1%) registros horários.

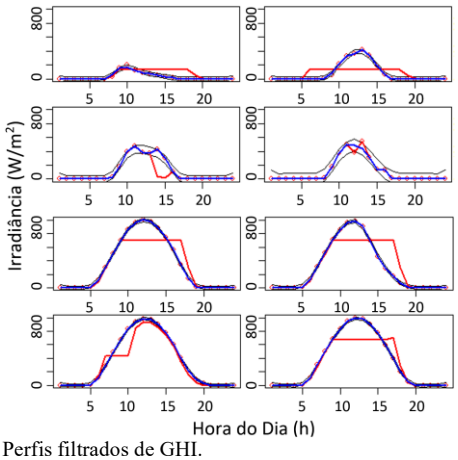


Fig. 14. Perfis filtrados de GHI.

D. *Resultados da Etapa 4*

Nesta etapa, os 639 perfis diários foram classificados em 25 *clusters* por meio do Mapa de Kohonen. O mapa resultante é ilustrado na Fig. 15, onde cada caixa é um cluster formado por perfis similares.

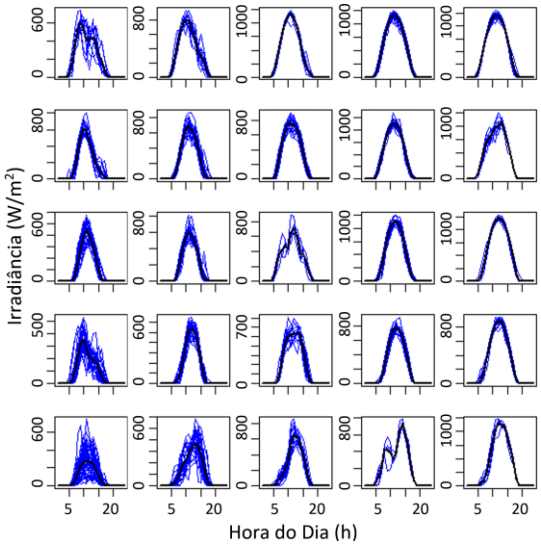


Fig. 15. Mapa dos perfis GHI.

Ressalte-se que o mapa obtido é topologicamente ordenado, i.e., os *clusters* semelhantes ocupam regiões vizinhas.

Na sequência, os *boxplots* em cada *cluster* (Fig. 16) permitem identificar as observações discrepantes com maior precisão. Os *outliers* correspondem aos registros de GHI localizados fora dos limites definidos pelas cercas inferior e superior dos *boxplots* horários. Os *outliers* são substituídos pelos respectivos valores medianos. No total foram corrigidos 284 (1,9%) registros horários de irradiância e de geração SFV.

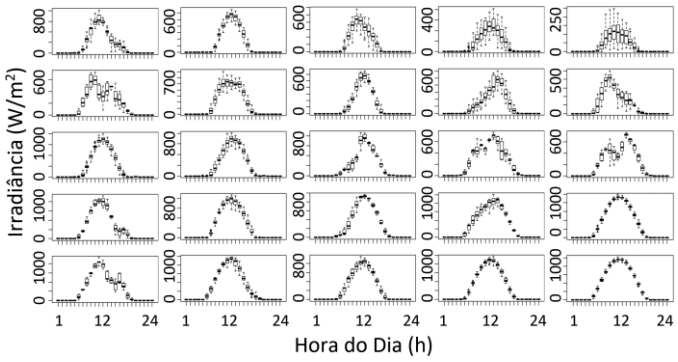


Fig. 16. *Boxplots* dos perfis GHI em cada *cluster*.

E. Resultados da Etapa 5

Com os registros de GHI corrigidos na Etapa 4, ajusta-se agora um novo modelo de regressão linear simples em que a geração SFV é explicada pelos valores tratados da GHI. O modelo resultante é apresentado na Fig. 17, juntamente com os *outliers* de geração SFV (em azul), i.e., valores fora dos limites dos intervalos de previsão com 95% de confiança.

O resultado final do tratamento de dados é ilustrado na Fig. 18. Nesta etapa foram modificados 552 (3,6%) registros horários de geração SFV.

Nas Fig. 19 e 20 apresentam-se os perfis diários de GHI e geração SFV após todas as etapas do tratamento de dados.

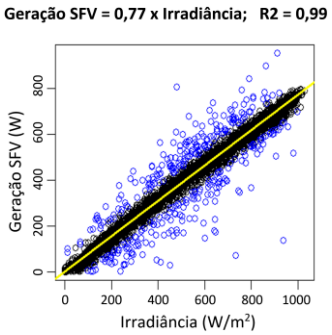


Fig. 17. Novo ajuste do modelo de regressão linear.

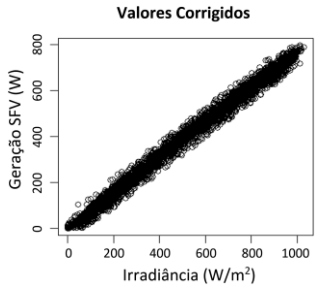


Fig. 18. Valores tratados da geração SFV e irradiância GHI.

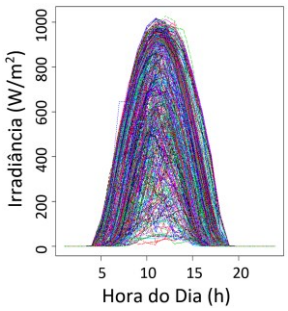


Fig. 19. Perfis diários de GHI após o tratamento de dados.

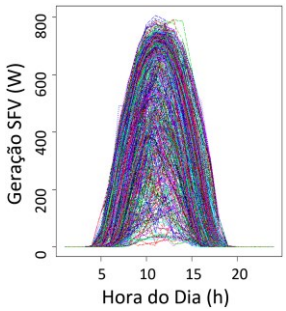


Fig. 20. Perfis de geração SFV após o tratamento de dados.

F. Desempenho do Processo de Tratamento de Dados

Conforme comentado anteriormente, o objetivo básico do processo de filtragem e tratamento de dados consiste na correção de valores discrepantes (*outliers*), na substituição de valores errôneos (*bad data*) e no preenchimento de lacunas de dados (*missing data*); porém sem alterar indevidamente o conjunto de dados ou a série temporal original.

Assim, neste trabalho foram comparadas as séries temporais de dados original e filtrada, tanto para a GHI quanto para a geração SFV, bem como as suas estruturas de autocorrelação.

Adicionalmente, também foram comparadas as distribuições acumuladas dos valores originais e filtrados da irradiância e da potência gerada. A título de ilustração, nas Fig. 21 e 22 apresentam-se as sequências de 168 horas em vermelho (dados originais) e em azul (dados filtrados), para um mesmo período. Conforme indicado, as correções realizadas conseguiram preencher corretamente as lacunas de dados, sem modificar de forma significativa os demais perfis.

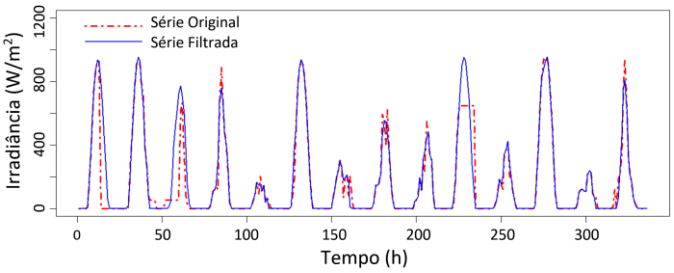


Fig. 21. Valores originais e filtrados de GHI em uma semana.

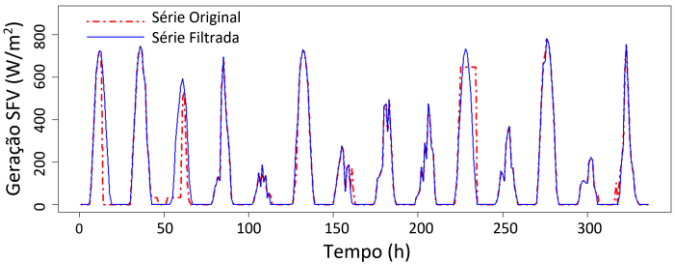


Fig. 22. Valores originais e filtrados de geração SFV.

A distribuição acumulada dos valores originais e filtrados da irradiância e da geração SFV são apresentados nas Fig. 23 e 24 respectivamente. Em cada caso, a semelhança entre as distribuições acumuladas indica que o tratamento de dados conseguiu realizar as correções necessárias, mas minimizando as mudanças indevidas nos dados.

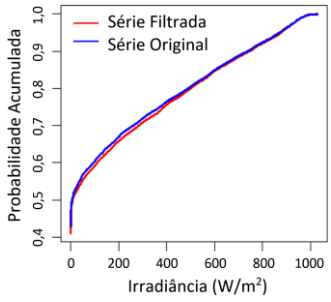


Fig. 23. Distribuições acumuladas dos valores de GHI.

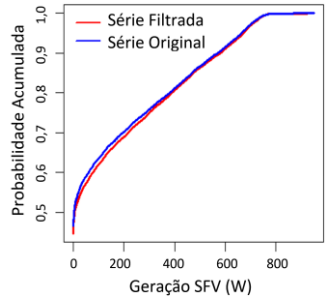


Fig. 24. Distribuições acumuladas da geração SFV.

Adicionalmente, o tratamento de dados não modificou as estruturas de autocorrelação das séries de GHI e geração SFV, conforme ilustrado nas Fig. 25 e 26 respectivamente.

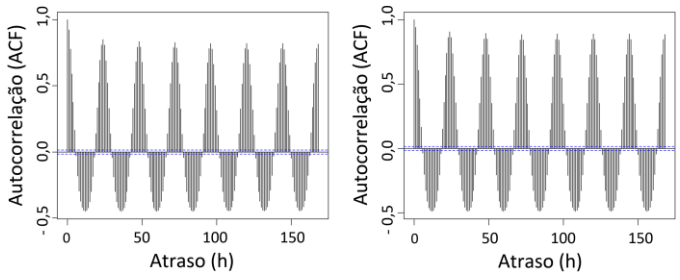


Fig. 25. Função de autocorrelação da GHI: (a) dados originais; (b) dados filtrados.

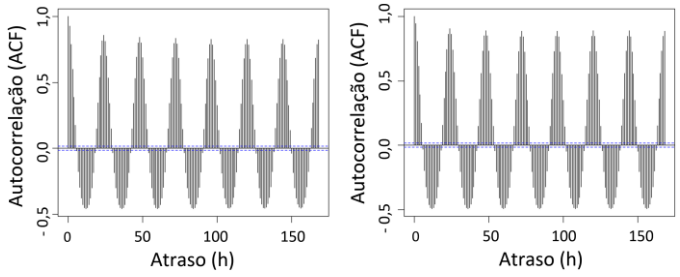


Fig. 26. Função de autocorrelação da geração SFV: (a) dados originais; (b) dados filtrados.

V. Conclusão

Este trabalho apresentou as principais características de uma metodologia desenvolvida para o tratamento conjunto de registros horários de geração solar fotovoltaica e irradiância solar (GHI). O tratamento de dados é uma etapa essencial para garantir a qualidade em qualquer análise preditiva, sobretudo quando esta se baseia em registros de medições. Neste sentido, o trabalho contribuiu com a introdução de técnicas, além do *Boxplot*, como métodos não paramétricos (LOESS) e de mineração de dados (Mapa de Kohonen) no tratamento de dados solarimétricos. Além disso, dada a forte dependência entre a geração SFV e a GHI, a estratégia adotada consistiu no tratamento conjunto, e não separado, dos dados das medições de GHI e SFV. Ressalta-se que os dados de reanálise são utilizados em situações extremas, por exemplo, no preenchimento de lacunas de dados de longa duração. A metodologia proposta foi aplicada a um sistema SFV localizado na região mediterrânea da Itália com capacidade total de 960 kWp e os resultados apresentados são satisfatórios. *Outliers* foram substituídos pelas respectivas cercas superiores, num total de 415 (2,7%) e 448 (2,9%) valores horários de irradiância e geração SFV respectivamente. Após verificação dos valores fora dos limites dos intervalos de previsão com 95% de confiança, a partir de um novo modelo de regressão linear simples, em que a geração SFV é explicada pelos valores tratados da GHI, foram modificados 552 (3,6%) registros horários de geração SFV. Os resultados obtidos encorajam a continuidade da pesquisa para

o aprimoramento da metodologia e apontam para a necessidade de disponibilização de uma base pública de dados solares, com discretização horária, e.g., a partir dos projetos vencedores nos leilões públicos de compra de energia elétrica.

REFERÊNCIAS

[1] International Energy Agency, *World Energy Outlook 2018*. [Online]. Available: www.iea.org.

[2] D.M. Falcão, G.N. Taranto, C.C.O. Hincapie, "Chronological Simulation of the Interaction between Intermittent Generation and Distribution Network,". in *2013 IEEE PES Innovative Smart Grid Technologies LATIN AMERICA* (ISGT LA 2013), São Paulo, 2013.

[3] G.B. Rosas, E.M. Lourenço, D.M. Falcão, T.S. Fernandes, "Superação de Equipamentos, Reserva de Energia e Controle de Tensão em Sistemas com Forte Penetração de Energias Eólica e Solar," in *XIV Symposium of Specialists in Operation Planning and Electrical Expansion*, Recife, 2018.

[4] J. Antonanzas, N. Osorio, R. Escobar, R. Urraca, F.J. Martinez-de-Pison, F. Antonanzas-Torres, "Review of photovoltaic power forecasting," *Solar Energy*, vol. 136, pp. 78–111, 2016.

[5] M.G. De Giorgi, P.M. Congedo, M. Malvoni, D. Laforgia, "Error analysis of hybrid photovoltaic power forecasting models: A case study of mediterranean climate," *Energy Conversion Management*, vol. 100, pp. 117–130, 2015.

[6] M. Malvoni, M.G. De Giorgi, P.M. Congedo, "Data on photovoltaic power forecasting models for Mediterranean climate," *Data in Brief*, vol. 7, no. May, pp 1639–1642, 2016.

[7] J.F.M. Pessanha, T.C. Justino, M.E.P. Maceira, "Metodologia para filtragem de registros de carga," XII *Symposium of Specialists in Operation Planning and Electrical Expansion*, Rio de Janeiro, 2012.

[8] J.F.M. Pessanha, V. Castellani, T.C. Justino, D.D.J. Penna, M.E.P. Maceira, "Uma metodologia para filtragem de medições anemométricas," *Learning and Nonlinear Models*, vol. 10, pp. 90-98, 2012.

[9] R.A. Johnson, D.W. Wichern, "Applied Multivariate Analysis", 4th ed., New Jersey: Prentice Hall, 1998.

[10] J.F.M. Pessanha, A.C.G. Melo, T.C. Justino, M.E.P. Maceira, "Combining Statistical Clustering techniques and Exploratory Data Analysis to Compute Typical Daily Load Profiles - Application to the Expansion and Operational Planning in Brazil," in *Probabilistic Methods Applied to Power Systems* (PMAPS), Boise, 2018.

[11] J.F.M. Pessanha, R.M. Velasquez, A.C.G. Melo, R.P. Caldas, "Técnicas de Cluster Analysis na Construção de Tipologias de Curva de Carga," in *XV Seminário Nacional de Distribuição de Energia Elétrica*, Salvador, 2002.

[12] S.C. William, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, 74.368, PP. 829-836, 1979.

[13] T. Kohonen, "Self-Organization and Associative Memory," Springer Berlin Heidelberg, vol. 8., 1989.

[14] S. Haykin, "Redes Neurais: princípios e prática", Porto Alegre, Bookman, 2001.

[15] A.C.P.L.F. Carvalho, A.P. Braga, T.B. Ludermir, "Fundamentos de Redes Neurais Artificiais", in *11ª Escola de Computação*, Rio de Janeiro, 1998.

[16] S. Pfenninger, I. Staffell, "Long-term patterns of European PV output using 30 years of validated hourly reanalysis and satellite data," *Energy - Elsevier*, vol. 114, pp. 1251–1265, 2016.

[17] W.S. Parker, "Reanalysis and observations: what´s the difference?," in *Bulletin of the American Meteorological Society*, vol. 97, no. 9, 2016.

[18] S. Pfenninger, I. Staffell, "Simulations Of Hourly Power Output From Wind And Solar PV Farms." [Online]. Available: https://www.renewables.ninja/.

[19] M. V. Contes Calça, M. R. Raniero, D. M. Zeca Fernando, S. A. Rodrigues and A. Dal Pai, "Outliers Detection in a Quality Control Procedure for Measurements of Solar Radiation," in *IEEE Latin America Transactions*, vol. 17, no. 11, pp. 1815-1822, November 2019.

[20] R. M. Salgado, T. C. Machado and T. Ohishi, "Intelligent Models to Identification and Treatment of Outliers in Electrical Load Data," in *IEEE Latin America Transactions*, vol. 14, no. 10, pp. 4279-4286, Oct. 2016.

[21] R Core Team (2018). *R: A language and environment for statistical computing*. [Online]. Available: https://www.R-project.org/.

[22] R. Kimura, " Numerical weather prediction," in Journal of Wind Engineering and Industrial Aerodynamics, vol. 90, no. 12–15, pp. 1403-1414, December 2002.

José Francisco Moreira Pessanha é Estatístico (ENCE, 1992) e engenheiro eletricista (UERJ, 1994). É mestre (COPPE/UFRJ, 1999) e doutor (PUC-Rio, 2006) em engenharia elétrica. Pesquisador do Centro de Pesquisas de Energia Elétrica - CEPEL desde 2002, tem atuado em estudos e no desenvolvimento de ferramentas computacionais para análise de confiabilidade de sistemas de potência, tarifação de sistemas de distribuição, previsão de mercado de longo prazo, previsão probabilística da geração eólica e previsão de carga para operação em tempo real e programação da operação. Também é professor adjunto da Universidade do Estado do Rio de Janeiro (UERJ). Em 2016 realizou pós doutorado no Inesc Tec Porto, Portugal, sobre previsão probabilística da geração eólica.

Albert Cordeiro Geber de Melo é graduado pela Universidade Federal de Pernambuco (1983), mestre (1986) e doutor (1990) pela PUC-Rio, em Engenharia Elétrica. Pesquisador do CEPEL desde 1985, onde exerceu vários cargos gerenciais, incluindo Diretor de Pesquisa, Desenvolvimento e Inovação (Jan 2005 – Jul 2008) e Diretor-Geral (Ago 2008 – Jan 2017), tendo ainda representado esta instituição em diversos fóruns nacionais e internacionais, como o Comitê de Monitoramento do Setor Elétrico – CMSE e a Comissão Permanente de Análise de Metodologias e Programas Computacionais do Setor Elétrico – CPAMP/CNPE, a Agência Internacional de Energia Elétrica e o Diálogo Estratégico em Energia Brasil - Estados Unidos da América. É Professor Adjunto da UERJ e membro titular da Academia Nacional de Engenharia

Roberto Pereira Caldas é Engenheiro de Eletrônica pelo Instituto Tecnológico de Aeronáutica (1978), Mestre (1990) e Doutorando em Engenharia Elétrica pela Universidade Federal do Rio de Janeiro (2018). No CEPEL, atuou como pesquisador (1983-2017) e Diretor de Pesquisa, Desenvolvimento e Inovação (2008-2016). Atuou na industrialização de dispositivos e sistemas de medição de energia elétrica e na introdução de redes elétricas inteligentes (*Smart Grid*) e de novas energias renováveis.

Djalma Mosquera Falcão é graduado em Engenharia Elétrica pela Universidade Federal do Paraná (1971), mestrado em Engenharia Elétrica pela COPPE / Universidade Federal do Rio de Janeiro (1973), doutorado em Engenharia Elétrica pela University of Manchester Institute of Science and Technology, Reino Unido (1981) e pós-doutorado pela University of California at Berkeley, USA (1993). Atualmente é professor titular da COPPE / Universidade Federal do Rio de Janeiro e membro titular da Academia Nacional de Engenharia.