A METHODOLOGY FOR DATA CLEANING OF WIND SPEED TIME SERIES

José F.M. Pessanha, Valk L.O. Castellani, Thatiana J. Conceição, Debora D.J. Penna, Maria E. P. Maceira

CEPEL – Centro de Pesquisas de Energia Elétrica
francisc@cepel.br vcastell@cepel.br thatiana@cepel.br debora@cepel.br elvira@cepel.br

Abstract – The prediction of wind resources is a key item for the safe and economic integration of wind farms in the operation of electrical systems. The accuracy of such predictions depends on the quality of the data. This article presents a methodology for filtering wind speed time series by using fuzzy clustering method (FCM) and local regression (LOESS). The goal is to improve the data quality for the time series modeling.

Keywords – Wind power, wind speed, data cleaning, smothing, cluster analysis.

1 Introducão

A integração segura e econômica dos aproveitamentos eólicos ao sistema elétrico requer previsões da disponibilidade dos recursos eólicos desde alguns minutos à frente até previsões horárias com horizontes que variam de uma hora até uma semana à frente. Na formulação de tais previsões pode-se contar com uma ampla variedade de metodologias, desde os tradicionais métodos estatísticos até métodos de inteligência computacional (Wu & Hong, 2007).
Em função das freqüentes falhas nos sistemas de medição, as séries de registros de velocidade de vento podem apresentar lacunas e observações aberrantes, conforme ilustrado na Figura 1. A presença de falhas nas medições compromete o desempenho dos modelos de previsão e, portanto, existe a necessidade de dispor de métodos para filtragem e imputação de dados (Wettayaprasit et al, 2007).

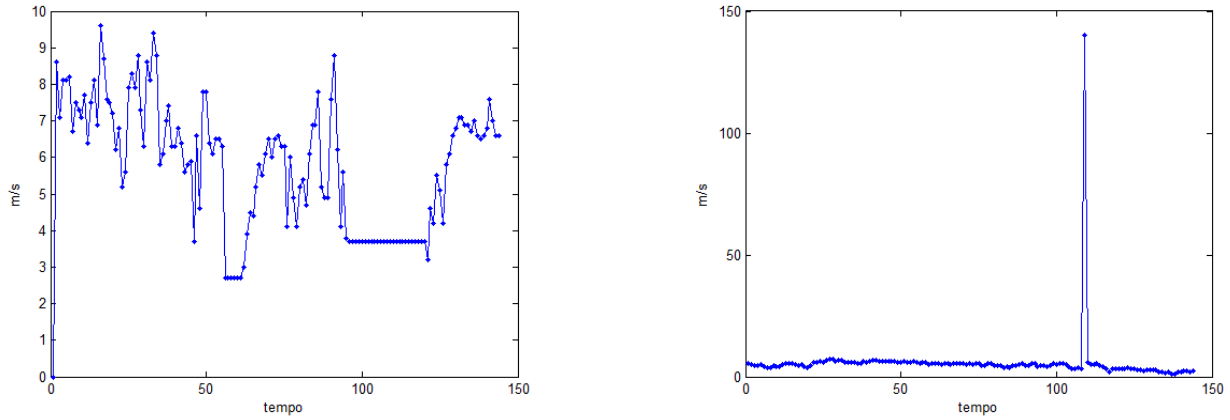


Figura 1 – Exemplos de problemas nos registros de velocidade de vento

Neste artigo apresenta-se uma metodologia para tratamento dos registros de velocidade do vento. A metodologia proposta baseia-se no uso combinado de técnicas para suavização de dados (LOESS) e técnicas de classificação não supervisionada (*cluster analysis*) , conforme ilustrado no diagrama da Figura 2.

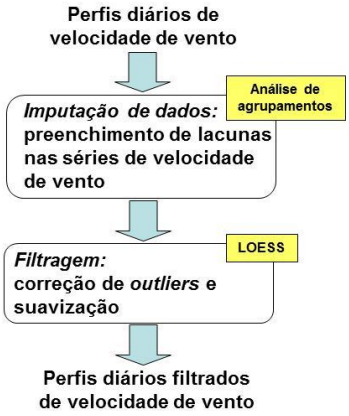


Figura 2 – Esquema da metodologia proposta

2 Imputação de dados de velocidade de vento

Os registros da velocidade de vento referem-se aos valores instantâneos tomados a cada 10 minutos e encontram-se organizados em matrizes, cujas colunas guardam os registros diários (144 pontos em cada coluna). Assim, as lacunas de dados foram classificadas em duas categorias: dias sem medição (colunas vazias) e dias com medições incompletas (colunas com lacunas de dados).

Para imputação dos valores de velocidade em um dia sem registros propõe-se identificar nos registros passados da velocidade do vento um dia com registros completos, cujos dias adjacentes apresentem perfis de velocidade similares aos verificados nos dias anterior e posterior ao dia sem registros. A seguir, na Figura 3, tem-se uma ilustração desta estratégia de imputação de dados.

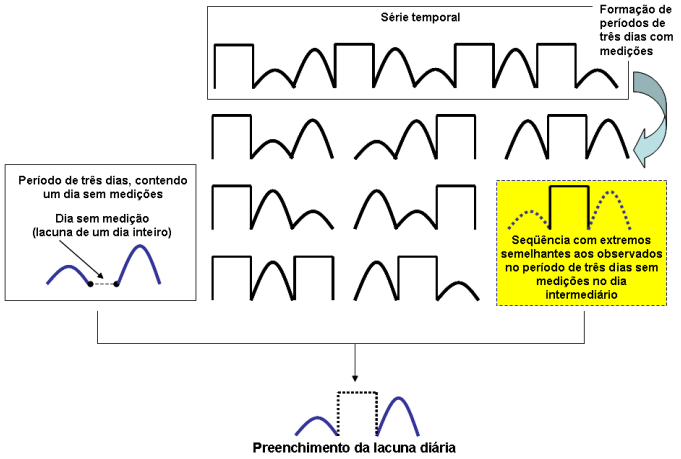


Figura 3 – Estratégia para imputação de dados em dias sem medições

A seguir, na Figura 4 tem-se uma ilustração do resultado da imputação de dados.

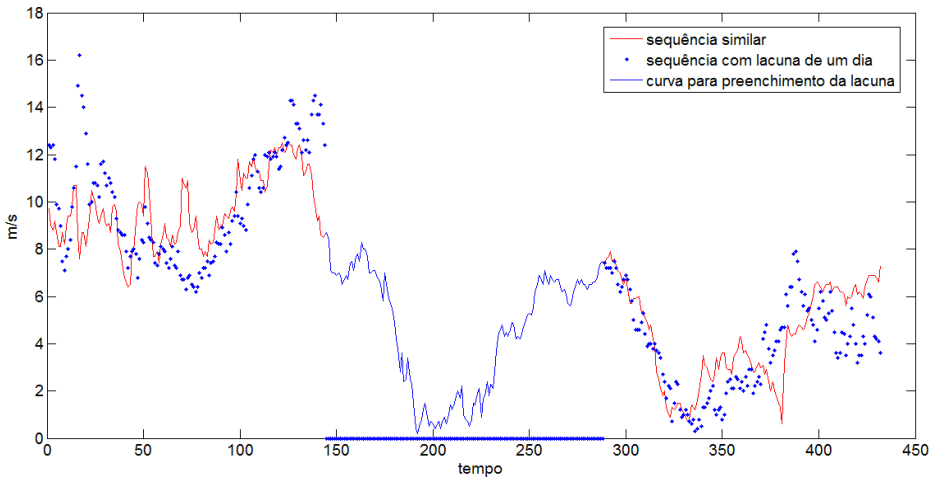


Figura 4 – Resultado da imputação de dados para um dia sem medições

3 Imputação de dados nos dias com medições incompletas

Na Figura 5 são apresentados dois exemplos de dias com medições incompletas. Naturalmente, antes de preencher as lacunas de dados presentes nestes dias é necessário identificá-las no conjunto de registros. As lacunas de dados são caracterizadas por sequências de valores constantes ou quase constantes.

A detecção de tais seqüências inicia-se com a normalização da curva diária da velocidade do vento pela respectiva média diária. Em seguida são calculadas as diferenças de primeira ordem em cada curva diária normalizada. Seja $v(t)$ a velocidade do vento no instante t, então a diferença de primeira ordem é dada por $\Delta v = v(t) - v(t\text{-}1)$.

Para detectar as seqüências de valores constantes ou quase constantes deve-se avaliar a magnitude das diferenças de primeira ordem Δv. Neste trabalho admitiu-se que diferenças menores que 0,001 são consideradas como sendo nulas. Uma sequência de diferenças nulas indica que os valores de velocidade subjacentes são constantes ou quase constantes. Assim, é possível identificar as sequências de velocidade constantes por meio da avaliação do comprimento das seqüências de diferenças consideradas nulas. Após alguns testes realizados com os registros históricos considera-se que sequências com 10 ou mais diferenças nulas são lacunas de dados, conforme indicado na Figura 5.

Como estratégia de imputação de dados propõe-se preencher as lacunas com segmentos extraídos de perfis típicos identificados por um algoritmo de análise de agrupamentos (*cluster analysis*) aplicado ao conjunto de perfis diários da velocidade de vento sem lacunas.

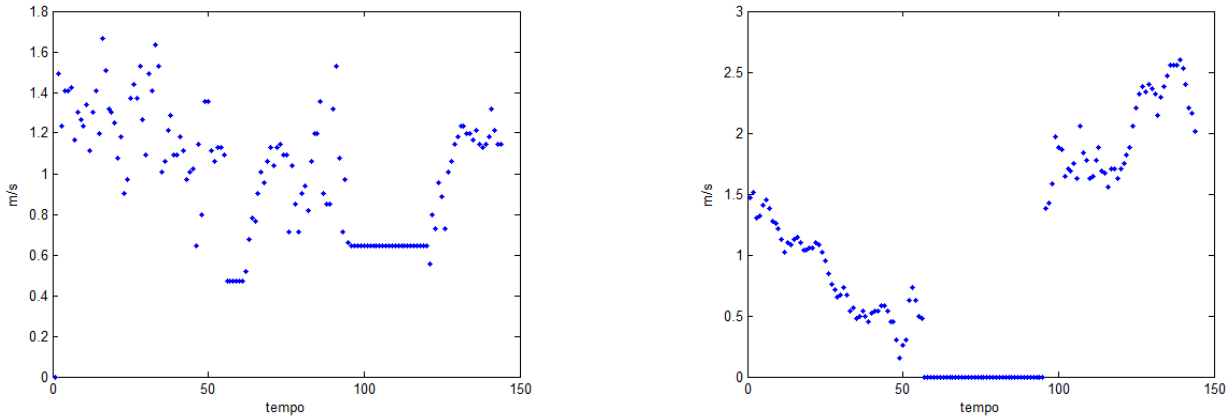


Figura 5 – Exemplos de lacunas nas medições de velocidade do vento

Neste trabalho foi utilizado o algoritmo *fuzzy clustering means* - FCM (Jang et al, 1997). A finalidade da análise de agrupamentos é identificar uma estrutura natural de agrupamento dos dados. No FCM a estrutura natural de agrupamentos é identifica pela solução do seguinte problema de programação não linear:

$$\underset{u_{ij},c_j}{Min} \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{m}\left\|x_i - c_j\right\|^2 \tag{1}$$

s.a.

$$\sum_{j=1}^{k} u_{1j} = 1$$

...

$$\sum_{j=1}^{k} u_{nj} = 1$$

onde n é o total de perfis na amostra, x_i denota o vetor contendo os 144 pontos do i-ésimo perfil diário de velocidade do vento (i=1,n), m é a constante de *fuzzyficação* (em geral m=1,25 ou m=2), k é o número de *clusters* em que os objetos serão agrupados, c_j é o centróide do j-ésimo *cluster* e u_{ij} é o grau de pertinência da i-ésima curva de carga no j-ésimo cluster ($0 \le u_{ij} \le 1$).

Com o auxílio da função Lagrangeana, o problema de otimização em (1) pode ser escrito como:

$$\underset{u_{ij},c_j,\lambda_i}{Min} \sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{m}\left\|x_i - c_j\right\| + \sum_{i=1}^{n}\lambda_i\left(\sum_{j=1}^{k} u_{ij} - 1\right) \tag{2}$$

Na função objetivo (2), λ_j, j=1,n são os multiplicadores de Lagrange para as n restrições de igualdade. A partir da equação (2) são obtidas as seguintes condições de otimalidade:

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \tag{3}$$

$$u_{ij} = \frac{1}{\sum_{t=1}^{k}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_t\right\|}\right)^{\frac{2}{m-1}}} \tag{4}$$

Usando as equações (3) e (4), pode-se programar o algoritmo a seguir, onde a matriz *U* formada pelos elementos u_{ij} (i=1,n e j=1,k) e os centroides dos k *clusters* (c_j, j=1,k) são obtidos iterativamente:

Passo 1 - Inicialize a matriz *U* com valores entre 0 e 1, observando que, em cada linha da matriz, a soma dos valores deve ser igual a unidade. Esta etapa é denominada por *fuzzyficação*.

Passo 2 - Use a equação (3) para calcular as coordenadas dos k centroides

Passo 3 – Calcule a função objetivo $\sum_{j=1}^{k}\sum_{i=1}^{n} u_{ij}^{m}\left\|x_i - c_j\right\|^2$. Pare se o valor da função objetivo estiver abaixo de uma tolerância ou se a melhoria em relação à iteração anterior for desprezível.

Passo 4 - Use a equação (4) para atualizar os elementos da matriz *U* e volte para o passo 2.

Após a convergência do algoritmo, os perfis de velocidade do vento são alocados nos *clusters* onde apresentam maior grau de pertinência (*defuzzyficação* pelo máximo). Como o objetivo é agrupar os perfis de velocidade semelhantes em um mesmo *cluster*, a análise de agrupamentos é aplicada nos perfis normalizados pelas respectivas médias. Os perfis típicos são os k centroides, ou seja, as médias dos perfis normalizadas em cada *cluster*.
A seguir, na Figura 6 são apresentados 2 dos 84 agrupamentos e respectivos perfis típicos (em negrito) identificados pelo método FCM.

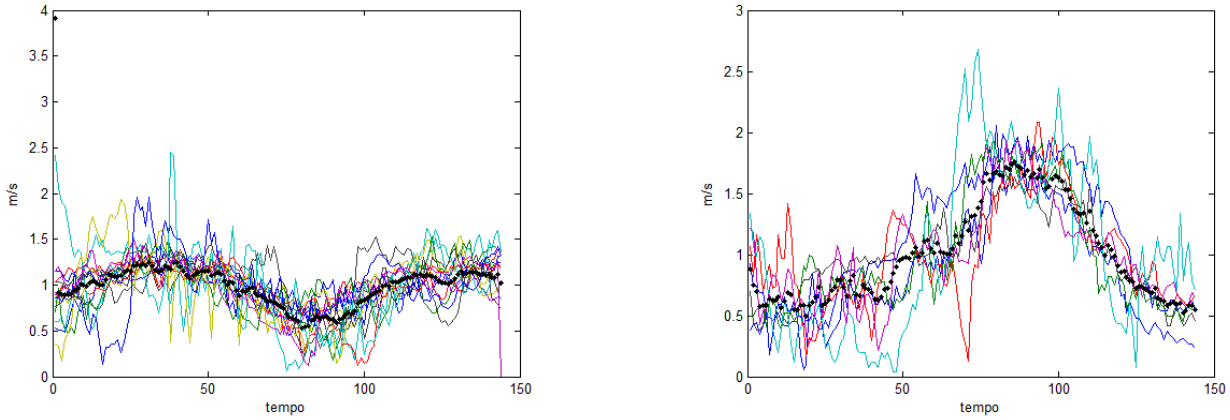


Figura 6 – Exemplos de agrupamentos e perfis típicos

A imputação de dados consiste em preencher as lacunas com estimativas fornecidas pelo perfil típico similar expresso em m/s. Os resultados deste procedimento são ilustrados na Figura 7.

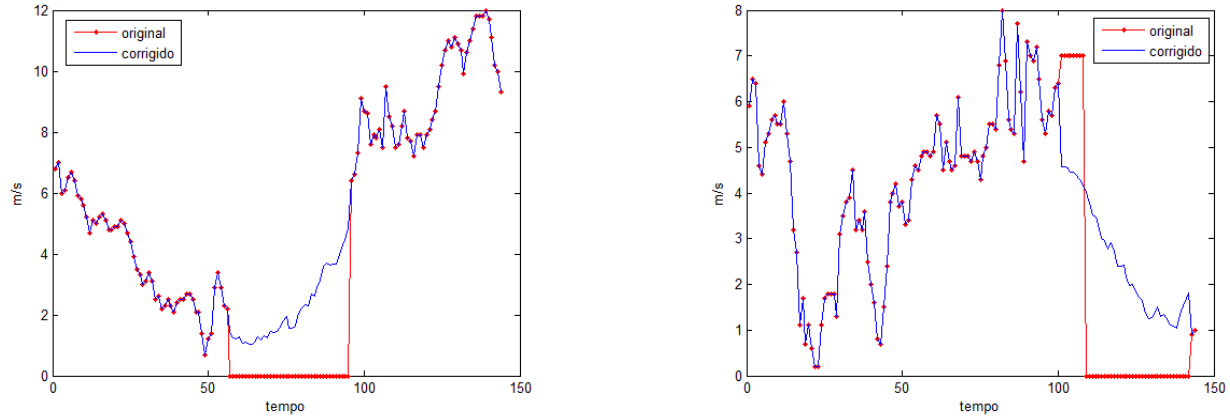


Figura 7 – Exemplos de agrupamentos e perfis típicos

4 Filtragem de dados de velocidade

A filtragem inicia-se com a aplicação do algoritmo de suavização LOESS (Martinez & Martinez, 2002) em cada perfil diário da velocidade do vento:

1) Seja x o vetor de variáveis explicativas e y o vetor com as observações da variável dependente. Nesta aplicação, o vetor y é o perfil diário da velocidade de vento com 144 pontos, enquanto o vetor x representa os instantes das observações.

2) Informe o tamanho k da janela de tempo.

3) Para cada instante x_0 identifique os k instantes x_i ($i=1,k$) na vizinhança de x_0 e denote este conjunto por $N(x_0)$.

$$\Delta(x_0) = máximo_{x_i \in N(x_0)} \|x_0 - x_i\|$$

4) Calcule a maior distância entre x_0 e o ponto x_i dentro da janela $N(x_0)$.

5) Pondere cada par (x_i,y_i), x_i em $N(x_0)$ com base na seguinte função:

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

6) Aplique mínimos quadrados ponderados para obter uma estimativa $\hat{y}$ para y no ponto x_0 ajustado ao conjunto de observações que pertencem à vizinhança $N(x_0)$.

7) Repita os passos de 3 a 6 para cada instante de tempo no vetor x.

A curva $Y(t)$ obtida pela aplicação da suavização LOESS é uma referência para comparação com o perfil de velocidade $v(t)$.

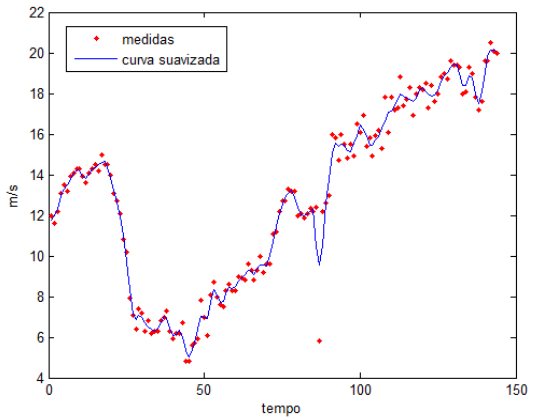
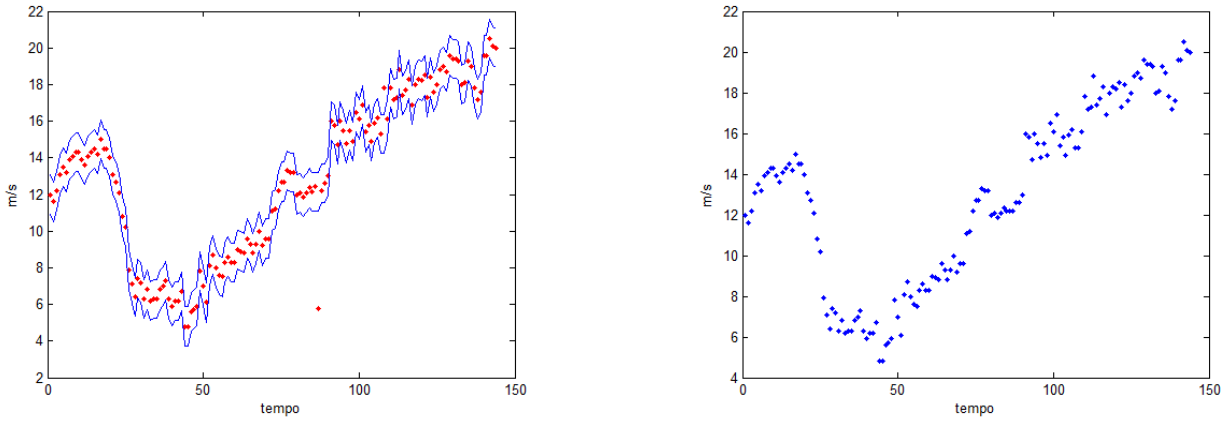


Figura 8 – Curva suavizada e curva medida

A partir do desvio padrão (DP) dos desvios entre as curvas suavizada e medida são calculados os limites dos intervalos de confiança em cada instante:

Limite superior: $LS(t) = Y(t) + 3.5\ DP$
Limite inferior: $LI(t) = Y(t) - 3.5\ DP$

Em cada instante de tempo verifica-se se o intervalo de confiança contém o valor medido. A seguir, na Figura 9 são apresentados alguns exemplos que ilustram observações fora dos limites dos intervalos de confiança e que, portanto, devem ser corrigidas. A correção consiste em substituir os valores fora dos intervalos por valores estimados pela suavização LOESS.



(a) curva medida e intervalos de confiança (b) curva filtrada

Figura 9 – Curvas medida e filtrada

5 Conclusões

Em função de falhas no sistema de medição os registros anemométricos apresentam erros como dados aberrantes, descontinuidades e lacunas. A introdução destes dados sem um tratamento estatístico prévio compromete o ajuste dos modelos de previsão e, portanto, implicam na perda da precisão das previsões de velocidade do vento. O presente artigo apresentou uma metodologia baseada em análise de agrupamentos e regressão local para a imputação e filtragem de registros de velocidade do vento. Os resultados obtidos são satisfatórios e mostram o potencial da metodologia proposta.

6 Referências bibliográficas

[1] Wu, Y.K. & Hong, J.S. A literature review of wind forecasting technology in the world, Power Tech, Lausanne, Switzerland, 1-5, July,2007.

[2] Wettayaprasit, W.; Laosen, N.; Chevakidagarn, S. Data filtering technique for neural network forecasting, 7th WSEAS International Conference on Simulation, Modeling and Optimization, Beijing, China, September, 15-17, 2007.

[3] Martinez, W.L.; Martinez, A.R. Computational statistics handbook with matlab, **Chapman & Hall/CRC**, 2002.

[4] Jang, J.S.R.; Sun, C.T.; Mizutani, E Neuro-Fuzzy and Soft Computing: A Computational Approach to Learming and Machine Intelligence, **Prentice Hall Inc**, 1997.