Uso de árvore de decisão para escolha de método de preenchimento de falhas em dados meteorológicos

Pedro Vinicius S. Zenere[1], **Thiago M. Ventura**[1], **Raphael S. R. Gomes**[1], **Thiago R. Rodrigues**[2]

[1]Instituto da Computação – Universidade Federal de Mato Grosso (UFMT)
Av. Fernando Corrêa da Costa, nº 2367 - Boa Esperança. Cuiabá - MT - 78060-900

[2]Instituto de Física – Universidade Federal de Mato Grosso do Sul (UFMS)
Av. Costa e Silva, Cidade Universitária. Pioneiros - MS - 79070-900

pedro.zenere@aluno.ic.ufmt.br, {thiago,raphael}@ic.ufmt.br,

thiagorangel@pgfa.ufmt.br

Abstract. *Traditional meteorological stations are being replaced by automated ones, creating a greater amount of information. However, there may be gaps on the data series, which creates difficulties in the analysis and in decision-making processes. Such gaps must be corrected in order to obtain a quality dataset and enable more detailed analyzes. There are gap filling methods that do this task. Each method has different performances depending on the aspects of the gaps. This paper proposes a methodology with a decision tree to determine which gap filling method should be used by analyzing a meteorological dataset.*

Resumo. *As estações tradicionais de coleta de dados meteorológicos estão sendo substituídas por estações automatizadas, gerando uma maior quantidade de informações. Entretanto, podem ocorrer ausência de dados, o que gera dificuldades nas análises e na tomada de decisão dos processos. Tais falhas nos dados devem ser corrigidas para se obter uma base de dados de qualidade e propiciar análises mais detalhadas. Existem métodos de preenchimento de falhas que realizam este trabalho. No entanto, cada método possui desempenhos diferentes dependendo dos aspectos das falhas. Este trabalho propõe uma metodologia com árvore de decisão para determinar qual método de preenchimento de falhas deve ser utilizado analisando uma série de dados meteorológicos.*

1. Introdução

No estudo da interação entre a superfície terrestre e a atmosfera é necessário dados e a análise das dinâmicas climáticas. Sendo que, através da coleta e da interpretação dos dados e fenômenos climáticos, podemos chegar a previsões para a agricultura, saúde, catástrofes, entre outras atividades [Giroto et al. 2015]. Assim, seria difícil, sem o estudo da meteorologia a prática dessas atividades que possuem como uma das principais premissas o estado do tempo, bem como sua compreensão, para alcançar resultados satisfatórios.

Para realizar estas análises e obter os resultados desejados é necessária a utilização de uma base de dados que contenha informações sobre a região em questão. Em

[de Oliveira 2009] é explicado a importância de se ter uma boa base de dados. Notadamente essas pesquisas devem ser embasadas em um minucioso, bem elaborado e completo banco de dados sobre o tempo, que contenham as anotações das diversas variáveis climáticas (elementos meteorológicos) em uma escala espaço-temporal definida. Os equipamentos responsáveis pela coleta desses dados estão instalados em estações meteorológicas.

Normalmente, vários equipamentos são instalados nas estações meteorológicas para coletar e armazenar dados para análise, de modo que cada equipamento é responsável por mensurar uma ou mais variáveis climáticas. Como todo aparelho eletrônico, os equipamentos de medições meteorológicas estão sujeitos a falha, e uma falha pode ser resultado de um erro técnico ou até mesmo de fenômenos naturais, atrapalhando a leitura dos dados. Essas falhas comprometem as análises realizadas com base nos dados das estações meteorológicas [Ventura et al. 2013]. Neste cenário, torna-se fundamental aplicar métodos de preenchimento de falhas no intuito de fortalecer os bancos de dados meteorológicos e melhorar a qualidade dos dados que são fontes para diversas análises e estudos que irão impactar diretamente os resultados obtidos através destas informações.

Entretanto, por haver diversos métodos de preenchimento de falhas, cada um com suas próprias características, há a dificuldade de escolher o melhor método para a série de dados que está sendo tratada. Este artigo tem como objetivo propor uma metodologia capaz de extrair características específicas da base de dados e mapear os diversos cenários existentes dando como resposta o método mais preciso para corrigir as falhas.

2. Trabalhos relacionados

Diversos trabalhos demonstram a aplicação de métodos para realizar o preenchimento de falhas. Em [da Silva Júnior et al. 2019] é mostrado os métodos de preenchimento de falhas mais populares utilizados pelos pesquisadores brasileiros. No trabalho de [Mello et al. 2017] são utilizados métodos estatísticos para preencher falhas em estações pluviométricas, assim como em [Ventura et al. 2016] que apresentam aplicações de métodos de regressão para preenchimento de falhas em dados meteorológicos. Em [Wanderley et al. 2012] é detalhado a utilização de técnicas geoestatísticas para preenchimento de falhas de dados pluviométricos. Em [Correia et al. 2016] é descrito o uso de redes neurais artificiais no preenchimento de falhas em séries de precipitação mensal. Já em [Ventura et al. 2019] é apresentado um método de Inteligencia Artificial para preencher falhas em dados meteorológicos, do mesmo modo que [Júnior et al. 2020] utiliza Inteligencia Artificial para modelar dados meteorológicos. O trabalho de [Rangel et al. 2018] faz uso de transformadas de Fourier no preenchimento de falhas em dados de intensidade do vento.

Como visto nos trabalhos citados, diferentes métodos são aplicados para preencher falhas em diferentes variáveis. Desse modo há uma dificuldade para avaliar qual método comporta-se melhor em determinadas variáveis meteorológicas [Ventura et al. 2016]. Além disso, os dados meteorológicos possuem diversas características diferentes entre si e, como consequência, métodos de preenchimento de falhas podem ter desempenhos variados dependendo do cenário encontrado. Portanto, a primeira etapa consiste em reconhecer quais características estão presentes na série de dados, e para testar esse reconhecimento, foi criada uma base de dados com diferentes cenários de falhas.

3. Materiais e métodos

3.1. Cenários de falhas de dados meteorológicos

Por meio de uma estação meteorológica, foi obtida uma série contendo 1 ano de dados. Nesta série, os dados foram armazenados a cada 30 minutos referente aos valores de temperatura do ar, temperatura do solo, umidade relativa do ar, umidade do solo, velocidade do vento, precipitação, saldo de radiação, radiação solar, fluxo de calor latente e fluxo de calor sensível, além do dado temporal (data e hora).

Diversos cenários de falhas foram simulados por meio desta série de dados, e os cenários tentaram simular situações reais encontradas pelos pesquisadores. Foram geradas falhas de 1% a 50% para cada variável da série de dados, e, assim, foi possível a avaliação dos métodos de correção desde, poucas falhas até quantidades consideráveis sem informação. Foram consideradas também falhas aleatórias ou sequenciais, pois é possível que, por falhas nos equipamentos, haja a ausência de dados por um longo período de tempo, assim como são possíveis falhas pontuais. Para a simulação de falhas de um equipamento que mede mais de uma variável, foram gerados 5 conjuntos diferentes no qual certas variáveis são falhadas em grupos, da seguinte maneira:

- Falhas nas variáveis relacionadas à radiação;
- Falhas em todas as variáveis, restando apenas o dado temporal;
- Falhas nas variáveis de temperatura e umidade relativa do ar;
- Falhas nas variáveis relacionadas ao solo;
- Falha em cada variável independentemente das demais.

Os cenários foram gerados baseados na combinação dessas variações, reproduzindo milhares de cenários diferentes.

3.2. Métodos de preenchimento de falhas testados

Quatro métodos de preenchimento de falhas foram testados com os cenários criados: média aritmética simples, regressão linear simples (RLS) e múltipla (RLM), e *Support Vector Machines* (SVM). Portanto, foram selecionados métodos univariados, multivariados, e da área de Inteligência Artificial (IA). Os métodos estatísticos foram escolhidos por serem populares entre pesquisadores que trabalham no âmbito da climatologia, além de já ter sido avaliado sua eficácia no trabalho de [Ventura et al. 2016] no qual é realizado uma análise da aplicabilidade destes métodos em falhas de dados meteorológicos.

A média aritmética simples consiste em somar os valores anteriores e posteriores à falha e dividi-los por 2. É o método mais popularmente utilizado entre pesquisadores da área para corrigir seus dados de acordo com [da Silva Júnior et al. 2019], provavelmente pela facilidade em calculá-lo. Segundo [Ventura et al. 2016], apresenta boa performance em variáveis que tem baixo desvio padrão. Apresenta boa precisão também em outras variáveis porém em cenários de poucas falhas. A equação (1) demonstra como é feito o calculo da média.

$$\bar{X}_i = \frac{x_{i-1} + x_{i+1}}{2} \tag{1}$$

A regressão linear é uma tentativa de modelar uma equação matemática linear que descreva o relacionamento entre duas variáveis [Curral 1994]. Este método obtém bons resultados em variáveis com cenários de muitas falhas. Nas análises realizadas por

[Ventura et al. 2016] ressalta que os métodos de regressão linear foram mais robustos com relação ao aumento das falhas. A equação (2) define a regressão linear simples.

$$y = \alpha + \beta x \tag{2}$$

Onde y é a variável dependente ou regressando, α é o intercepto ou constante do modelo, β é o coeficiente angular e x é a variável independente ou explanatória. Para o cálculo de α e β é dado pelas equações (3) e (4).

$$\hat{\alpha} = \bar{y} - \hat{\beta}_1 \hat{x} \tag{3}$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{4}$$

Para a regressão linear múltipla a variável dependente (y) será determinada por mais de uma variável independente (x). É definida pela equação (5).

$$y_i = \alpha + \sum_{i=n}^{n} \beta_i x_i \tag{5}$$

Por fim, o método SVM foi selecionado da área de IA. É um método robusto, utiliza do aprendizado supervisionado e tem como principal objetivo a determinação de limites de decisão que permitam a separação ótima entre classes a partir da minimização dos erros [Nascimento et al. 2009]. Para determinar os hiperparâmetros do SVM foi utilizado a técnica de *grid search*. Desta forma foi definido para *C* e *gamma* subconjuntos crescentes, sendo que para o parâmetro *C* foi definido um subconjunto de 1 a 15, e para *gamma* de 1 a 10.

3.3. Metodologia de escolha do método de preenchimento de falhas

Há dificuldades em determinar o melhor método de preenchimento de falhas, já que o desempenho de cada um varia dependendo das características das falhas existentes na série de dados. Entretanto, pode ser avaliado entre diversas situações, o método que costuma ter melhor desempenho. Assim, em um ambiente real, pode ser utilizado o método de preenchimento de falha mais indicado, bastando reconhecer as características das falhas presentes na série de dados a ser tratada.

Neste trabalho, a determinação do melhor método para determinada situação foi feita avaliando o desempenho de cada método em cada cenário criado. O desempenho dos métodos foi mensurado utilizando Erro Médio Absoluto (EMA). De acordo com [Mentzer and Bienstock 1998], o EMA mede o afastamento médio das previsões em relação aos valores observados, constituindo na média dos erros da previsão (e_t). Desta forma um valor próximo a 0 é o ideal. A equação (6) descreve como é realizado o cálculo.

$$EMA = \frac{\sum_{t=1}^{n} |e_t|}{n} \tag{6}$$

Após executar o preenchimento de falhas de todos os cenários com todos os métodos, uma árvore de decisão foi montada. Segundo [Fayyad et al. 1996], as árvores de decisão são de interesse especial para a mineração de dados ou descoberta de conhecimento em bases de dados, visto que utiliza-se de símbolos e representações gráficas para reproduzir o conhecimento. A árvore de decisão é um modelo representado graficamente por nós e ramos, parecido com uma árvore, mas no sentido invertido [Han and Kamber 2001, Witten and Frank 2005] e fornece uma visão ampla e de fácil compreensão dos dados. A estrutura de uma árvore de decisão pode ser entendida da seguinte maneira: no topo da árvore encontra-se o nó raiz, que é o primeiro nó da árvore, já os nós internos são chamados de nós de decisão. Cada nó possui um teste sobre alguma variável independente e os seus respectivos resultados formam os ramos da árvore, bem como os nós folhas nas extremidades da árvore representam as classes.

As classes são os próprios métodos de preenchimento de falhas. Para os nós de decisão, foram utilizadas as características das falhas nas séries de dados:

- tipo da falha: falhas em sequência ou em pontos aleatórios;
- quantidade: porcentagem de falhas em relação à quantidade de registros na série de dados;
- tipo da variável meteorológica: radiação, temperatura/umidade e outras;
- falhas relacionadas: se houve ou não falhas de outras variáveis que poderiam ajudar na estimativa do dado que está ausente.

4. Resultados

A árvore de decisão foi desenvolvida em Python, com a ajuda de bibliotecas como a Scikit-learn e Pandas. Ela foi construída com um total de 3.102 exemplos sendo separada em dois conjuntos para treinamento e teste, destes 2481 utilizados para treinamento e 621 para teste. Foi realizado um pós-processamento para simplificar a árvore removendo nós desnecessários. A Figura 1 ilustra a árvore de decisão final.

Os nós de decisão contém as informações sobre a característica que está sendo analisada, o coeficiente de Gini e os exemplos dividido por classes, na ordem: RLM, Média, SLR e SVM. O coeficiente de Gini é um índice de dispersão estatística que calcula a pureza do nó. Assim, uma pontuação de Gini igual a zero significa que o nó é puro, ou seja, que possui inteiramente a uma classe. Essa medida é frequentemente usada em prática e é mais sensível do que a classificação incorreta de erro para alterações na probabilidade do nó [Moisen 2008].

Segundo a árvore de decisão, a característica mais importante para a classificação é o tipo de falha, podendo ser aleatório ou sequencial. Essa é uma característica que realmente impacta na precisão da estimativa de preenchimento de falha. Ter dados próximos para ser utilizado na estimativa é algo útil. Assim, falhas sequenciais aumentam a dificuldade em realizar o preenchimento.

Em seguida, é verificado se há também falhas em variáveis meteorológicas que possuem relação com o dado ausente. Como os dados meteorológicos têm relação entre si, é comum a utilização de outros dados coletados no mesmo momento que ocorreu a falha, para auxílio na estimativa do valor ausente (dado de umidade ser utilizado para estimar temperatura, por exemplo). A presença ou não de dados de variáveis relacionadas também influencia na escolha do método a ser utilizado.

A próxima característica é o tipo da variável que está tentando ser tratada. É verificado se a variável é de radiação, de temperatura ou umidade, ou outro tipo, como velocidade do vento ou precipitação. Por fim, se depois das análises anteriores ainda não chegou-se a um resultado, é avaliado a quantidade de falhas presentes.

De acordo com a árvore de decisão gerada, o método de média foi selecionado na maioria dos cenários. Se as falhas são aleatórias, a única situação em que não é indicada a média é quando há variáveis relacionadas para auxiliar a estimativa e quando a variável alvo é do tipo de radiação. Mesmo quando há falhas sequenciais, em alguns cenários é indicado utilizar a média para diminuir o erro das estimativas.

O SVM foi o segundo mais indicado para preencher falhas. Este método aparece nas decisões de variáveis de temperatura e umidade, além de ser escolhido quando há uma grande quantidade de falhas na série de dados, no caso de haver falhas nas variáveis relacionadas.

O método de RLM aparece em dois cenários. Em ambos quando a variável a ser tratada é de radiação e quando existe a possibilidade de utilizar dados de variáveis relacionadas ao dado ausente.

Por fim, o método de regressão linear simples não foi selecionado em nenhum cenário. Dese modo, em qualquer situação, a média dos dados, a regressão linear múltipla ou métodos como o SVM, terão um desempenho superior ao de regressão linear simples.

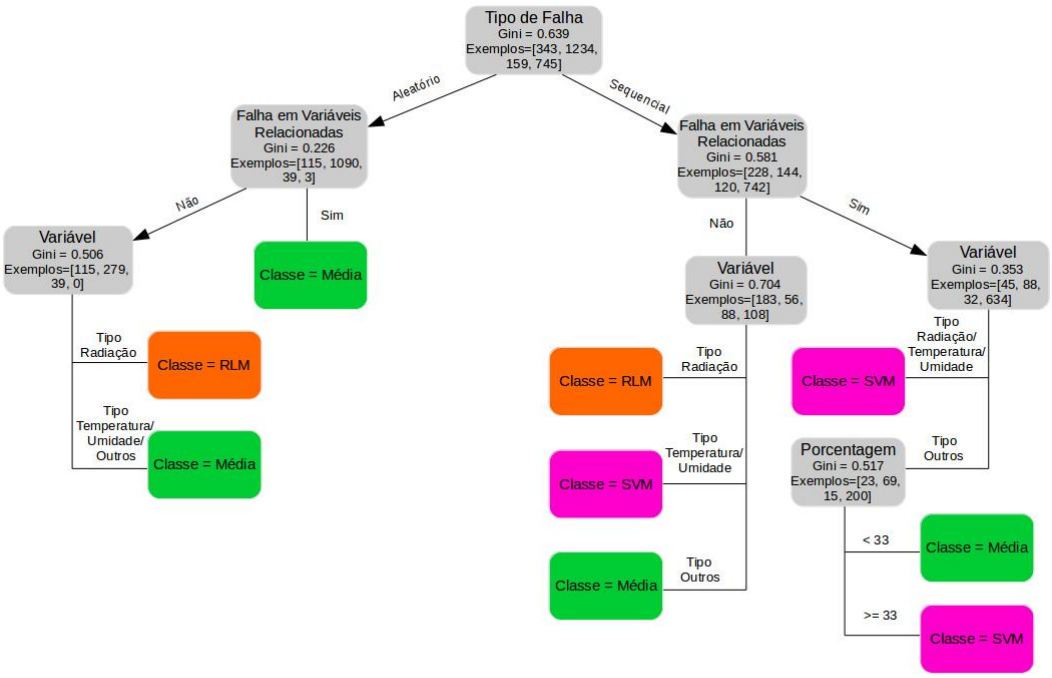

Figura 1. Árvore de decisão para determinar o melhor método de preenchimento de falhas de acordo com as características da série de dados.

5. Considerações Finais

Na escolha de um método de preenchimento de falhas para tratar dados meteorológicos é necessário tanto conhecimento de diversos métodos quanto uma análise detalhada da série que está sendo tratada. Desta forma, para um tratamento com precisão, é necessário tempo e um especialista na área, recursos que normalmente não estão disponíveis.

Por meio do desenvolvimento deste trabalho foi possível propor uma metodologia de escolha de métodos de preenchimento de falhas. Foi utilizada uma árvore de decisão, na qual a geração de diversos cenários possibilitaram o treinamento da mesma. Com esta árvore de decisão, é possível tratar grandes séries de dados com maior qualidade, utilizando os melhores métodos de preenchimento de falhas de acordo com as características das lacunas encontradas.

De posse desta árvore de decisão, é possível desenvolver algoritmos para tratar uma série de dados completa de maneira automática, sempre escolhendo os melhores métodos de preenchimento de falhas para cada variável ou mesmo para cada faixa de falhas. Além disso, mais testes podem ser realizados com outros métodos de preenchimento de falhas. Por fim, outras ações de tratamento de dados, como detecção de outliers, podem ser processadas para encontrar as respectivas árvores de decisão para escolha dos melhores métodos em cada situação.

6. Agradecimentos

Os autores agradecem a Fundação de Amparo a Pesquisa do Estado do Mato Grosso (FAPEMAT) pela bolsa de iniciação científica.

Referências

Correia, T. P., Dohler, R. E., Dambroz, C. S., and Binoti, D. H. B. (2016). Aplicação de redes neurais artificiais no preenchimento de falhas de precipitação mensal na região serrana do Espírito Santo. *UNESP, Geociências*, 35(4):560–567.

Curral, J. (1994). Statistics packages: A general overview. *Universidade de Glasgow*.

da Silva Júnior, A. A., de Souza Rosa Gomes, R., Ventura, T. M., Rodrigues, T. R., de Souza Nogueira, J., de Oliveira, A. G., and de Figueiredo, J. M. (2019). Visão geral sobre o tratamento de dados meterológicos no Brasil. *Natural Resources*, 9(2):59–66.

de Oliveira, A. G. (2009). A importância dos dados das variáveis climáticas nas pesquisas em geografia: Um estudo de caso empregando a precipitação pluviométrica. *Caminhos de Geografia - revista online*, pages 9–21.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). Advances in knowledge discovery and data mining. *American Association for Artificial Intelligence*, pages 1–34.

Giroto, D. B., Guldoni, B., and Tommaselli, J. T. G. (2015). A escola na estação meteorológica: a importância da meteorologia no cotidiano humano. *8º Congresso de extensão universitária da UNESP*, pages 1–11.

Han, J. and Kamber, M. (2001). *Data mining: concepts and techniques*. Morgan Kaufmann Publishers, 1st edition.

Júnior, L. C. G. V., Ventura, T. M., Gomes, R. S. R., de S. Nogueira, J., de A. Lobo, F., Vourlitis, G. L., and Rodrigues, T. R. (2020). Comparative assessment of modelled and empirical reference evapotranspirations methods for a brazilian savanna. volume 232.

Mello, Y. R. d., Kohls, W., and Oliveira, T. M. N. d. (2017). Uso de diferentes métodos para o preenchimento de falhas em estações pluviométricas. *Boletim de Geografia*, 35(1):112–121.

Mentzer, J. T. and Bienstock, C. C. (1998). Sales forecasting management. *California: Sage*.

Moisen, G. G. (2008). Classification and regression trees. In Sven Erik Jorgensen, B. D. F., editor, *Encyclopedia of Ecology*, volume 1, pages 582–588. Elsevier.

Nascimento, R. F. F., de Âlcantara, E. H., Kampel, M., Stech, J. L., de Moraes Novo, E. M. L., and Fonseca, L. M. G. (2009). O algoritmo *Support Vector Machines* (SVM): avaliação da separação ótima de classes em imagens CCD-CBERS-2. *Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil*, pages 2079–2086.

Rangel, R. H. O., de Oliveira-Júnior, J. F., Júnior, A. R. T., Pimentel, L. C. G., and de Gois, G. (2018). Série e transformada de fourier aplicadas no preenchimento de falhas de séries temporais de intensidade do vento na central nuclear almirante Álvaro Alberto, Rio de Janeiro - Brasil. *Anuário do Instituto de Geociências - UFRJ*, 41(2):74–84.

Ventura, T. M., de Oliveira, A. G., Marques, H. O., Oliveira, R. S., Martins, C. A., de Figueiredo, J. M., and Bonfante, A. G. (2013). Uma abordagem computacional para preenchimento de falhas em dados micro meteorológicos. *Revista Brasileira de Ciências Ambientais*, (27).

Ventura, T. M., Martins, C. A., de Figueiredo, J. M., de Oliveira, A. G., and Montanher, J. R. P. (2019). Mannga: A robust method for gap filling meteorological data. *Revista Brasileira de Meteorologia*, 34(2):315–323.

Ventura, T. M., Santana, L. L. R., Martins, C. A., and de Figueiredo, J. M. (2016). Análise da aplicabilidade de métodos estatísticos para preenchimento de falhas em dados meteorológicos. *Revista Brasileira de Climatologia*, 12:168–177.

Wanderley, H. S., de Amorim, R. F. C., and de Carvalho, F. O. (2012). Variabilidade espacial e preenchimento de falhas de dados pluviométricos para o estado de Alagoas. *Revista Brasileira de Meteorologia*, 27(3):347–354.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann Publishers and Elsevier, 2nd edition.