Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 8, n. 1, 2021.

Trabalho apresentado no XL CNMAC, Evento Virtual - Co-organizado pela Universidade do Mato Grosso do Sul (UFMS).

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Técnicas de imputação de dados aplicadas ao ambiente das smart grids: uma revisão

Jonas F. Schreiber[1]
DCEEng/UNIJUI, Ijuí, RS
Paulo S. Sausen[2]
DCEEng/UNIJUI, Ijuí, RS
Maurício de Campos[3]
DCEEng/UNIJUI, Ijuí, RS
Airam T. Z. R. Sausen[4]
DCEEng/UNIJUI, Ijuí, RS

Resumo. O mundo está cada dia mais conectado, qualquer ação é passível de registro, e o setor elétrico pode ser caracterizado como um dos serviços que mais tem agregado novas tecnologias de forma muito intensa nos últimos anos. As chamadas *Smart Grids*, ou Redes Elétricas Inteligentes, se caracterizam pela adoção de tecnologias de monitoramento e comunicação em praticamente todo o sistema. Com a aplicação e uso destas novas tecnologias tem-se como efeito colateral, o aumento significativo na quantidade de dados dificultando de sobremaneira a sua análise. Um dos grandes desafios atuais passa a ser a definição de métodos, estratégias e procedimentos para tratar esse volume de dados, também conhecido como *Big Data* e *Big Data Analytics* que surge com a necessidade de trabalhar e manipular uma grande quantidade de dados que também possuí várias lacunas e falhas de continuidade. Neste contexto, neste artigo é proposto uma revisão geral das possibilidades e uso de técnicas de imputação de dados que podem ser aplicadas nas *Smart Grids* objetivando viabilizar a transformação destas bases de dados em conhecimentos.

Palavras-chave. Imputação de Dados, Smart Grid, Big Data Analytis

1 Introdução

Uma concessionária de energia elétrica, que é parte do SEP (Sistema Elétrico de Potência), é responsável pelo suprimento de energia elétrica aos consumidores, abrangendo as fases de geração, transmissão, distribuição e carga [18]. O processamento tradicional de informação a partir da geração e distribuição de energia não se traduz apenas em fazer uma operação simples nos dados obtidos de uma subestação de energia elétrica, i.e. consulta, estatística, modificação entre outras, pois relações e regras mais aprofundadas são necessárias para tomadas de decisões por parte de profissionais das concessionárias de energia elétrica, ou seja, transformar dados em informações, fazendo com que possamos denominá-la de Redes Inteligentes. As *Smart Grids* são definidas pelo *Electric Power Research Institute* (EPRI) [17] como a superposição de um sistema unificado de comunicação e controle ao sistema existente de distribuição de energia, sensores são instalados nos equipamentos da concessionária de energia para medir seu funcionamento, bem como o estado do seu ambiente de funcionamento.

[1]jonasfs@gmail.com
[2]sausen@unijui.edu.br
[3]campos@unijui.edu.br
[4]airam@unijui.edu.br

A capacidade de integrar os dados adquiridos por sensores para analisar, monitorar, processar e responder perto do tempo real é necessária para resolver os problemas e se beneficiar das oportunidades criadas. Portanto, as redes inteligentes e os medidores inteligentes abrem oportunidades, sem precedentes, em termos de negócios para empresas de serviços públicos, mas também apresentam enormes desafios no manuseio e gerenciamento de grandes volumes de dados, também conhecido como *Big Data* [1]. As técnicas tradicionais de análise não conseguem processar esta quantidade gigantesca de dados gerados de forma efetiva, sendo necessária a utilização de soluções derivadas do conceito de *Big Data Analytics* [15], pois trabalha com a aplicação de técnicas avançadas específicas para grandes conjuntos de dados.

A ampliação do uso de sistema de monitoramento em conjunto com a crescente utilização de diferentes meios de transmissão de dados, que conectam muitas fontes diferentes e transmitem uma enorme quantidade de dados distribuídos, podem, ocasionalmente, duplicar um dado a partir do registro proveniente de sensores em transformadores que já poderiam estar disponíveis em outros dispositivos de monitoramento. Além disso, pode haver corrompimento de alguns dados, tanto por erros na aquisição ou mesmo na transmissão, estarem incompletos, devido a falhas de um sistema de proteção ou ainda estarem ausentes em decorrência de um erro de transmissão.

Fica claro que no cenário das *Smart Grids* existe uma necessidade crescente de aplicar algum procedimento ou técnica que possibilite, detectar e corrigir a ausência de dados. Usualmente, essas técnicas são denominadas de técnicas de imputação de dados e são aplicadas em uma diversidade de áreas a vários anos, porém no segmento das redes inteligentes não existe estudos que mostram quais e como essas técnicas podem ser aplicadas, nem sequer se é possível aplica-las diretamente ao problema ou mesmo se devem ser adaptadas. É justamente neste contexto, que o presente artigo propõe realizar um estudo no formato de *survey* das técnicas existentes verificando se as mesmas podem ser aplicadas ao segmento elétrico, mais especificamente ao ambiente das *Smart Grids*.

O restante deste artigo está organizado da seguinte forma: Na Seção 2 é feita uma breve descrição do conceito de *Big Data* e suas técnicas de tratamento de informações. E, na Seção 3, são apresentadas e avaliadas um conjunto de técnicas clássicas de tratamento de dados faltantes, também é apresentado uma tabela resumo com as vantagem e desvantagens destas técnicas quando aplicadas no ambiente das Smart Grids. E finalmente na Seção 4 são apresentadas as considerações finais e apontado os trabalhos futuros.

2 Big data analytics

Com o aumento expressivo de dados gerados a partir das *Smart Grids*, as concessionárias de energia elétrica estão recebendo e armazenando volumes cada vez maiores de dados, de uma crescente variedade de fontes, i.e. medidores inteligentes, Unidade de Medição Fasorial, Sistema de Supervisão Controle e Aquisição de Dados (SCADA), entre outros [3]. Junto com o aumento significativo de dados surgiu o conceito de *Big Data*, que possuí muitos significados e interpretações, mas cinco recursos principais estão sempre presentes, são eles: volume, velocidade, variância, veracidade e valor [12]. Os dados coletados de subestações de distribuição de energia elétrica satisfazem estes três recursos, já que, por exemplo, sensores enviam dados quase que de forma ininterrupta, gerando milhares de informações por mês, multiplicado pelo número de transformadores e subestações que a concessionária possui, acaba se transformando em um grande base de dados para ser gerenciada. Estes dados, coletados em períodos frequentes, podem auxiliar e virar uma grande vantagem para a concessionária de energia se utilizados de forma correta.

Com o aumento expressivo no volume de dados no segmentos das Smart Grids, técnicas tradicionais de análise não conseguem realizar satisfatoriamente o processamento, havendo a necessidade de utilizar técnicas oriundas de *Big Data Analytics*, que propiciam a aplicação de técnicas avançadas de análise para grandes conjuntos de dados. As informações gerada e armazenadas, de acordo com

DOI: 10.5540/03.2021.008.01.0487

Dola [6], podem ser convertidas em conhecimento sobre padrões históricos e tendências futuras para resolver problemas ou mesmo tomar decisões de negócios.

3 Imputação de dados

Com o rápido desenvolvimento da Tecnologia da Informação (TI), especialmente os grandes avanços da Internet das Coisas (IoT), redes sociais, comércio eletrônico e redes inteligentes, a quantidade de dados está crescendo e sendo acumulados a uma taxa sem precedentes. No entanto, o surgimento de registros incompletos crescem no mesmo ritmo, degradando a qualidade e usabilidade destas bases, tornando o campo de pesquisa em imputação de dados, em especial no setor elétrico, consolida-se como uma área profícua para estudos e pesquisas. Uma revisão do estado da arte destas técnicas e métodos será apresentada da Seção 3.3.

No setor elétrico é usual que a maioria dos sistemas de monitoramento usem os dados apenas para gerar alarmes de determinadas situações e eventos. No entanto, com a ampliação da aplicação de novas tecnologias e o aumento do número de sensores tornando as redes elétricas ambientes inteligentes os dados passaram a ser armazenados e principalmente analisados. Como já mencionado, essas bases passaram rapidamente a serem consideradas Big Data exigindo que sejam aplicadas técnicas de análise adequadas e entre as questões de análise de Big Data está a complexidade computacional dos algoritmos e dimensionamento dos algoritmos adequados para manipularem grandes quantidades de dados, adiciona-se a esta problemática o problema dos valores ausentes nos dados. Às vezes, os dados não são registrados devido à interrupção/falha no sensor, enquanto que, em outras ocasiões, o valor que é reportado/armazenado está longe do intervalo esperado ou não é relevante, tornando-se um valor inválido. Nestes cenários, o valor relatado não é o real, sendo considerado inválido e rotulado como um valor ausente.

Os estudos sobre dados incompletos pode ser encontrados desde o ano de 1962, onde Sebestyen [16] propôs uma solução baseada na probabilidade de hipóteses; já na década de 1970, Rubin [13] apresenta o método de múltiplas imputações, muito utilizado para resolver o problema de dados faltantes. A proporção de dados faltantes em relação ao conjunto de dados é definida na equação (1) por Chang [5] como:

$$p = \frac{P}{T} \tag{1}$$

onde p é a relação entre P (número de dados faltantes) e T (número total de dados da amostra).

3.1 Mecanismos de não-resposta

A ausência de dados, também conhecida como Variáveis Ausentes (VA), não deve ser considerada apenas como um problema, mas uma questão de interpretação dos resultados. A presença de VA acontece frequentemente em diversas áreas do conhecimento, e com base no sistema de classificação criado por Rubin [14], consideremos uma matriz D de dados coletados. Esta matriz possui R linhas representando os registros e A colunas representando os atributos, sendo $d_i = (d_{i1}, ..., d_{iA})$, onde d_{ij} é o valor do atributo j para o registro i. Com estas informações, conseguiremos dividir a matriz D em dois conjuntos:

$$D = \{D_{existente}, D_{inexistente}\} \tag{2}$$

onde $D_{existente}$ são os dados não faltantes e $D_{inexistente}$ são os dados faltantes. Correspondendo a cada matriz D existe uma matriz F de mesma dimensão, um identificador de dados faltantes, onde $f_{ij} = 1$ se d_{ij} existe, e $f_{ij} = 0$ caso contrário.

O mecanismo de dados faltantes, ou de não-resposta, é considerado através da distribuição condicional de F dado D, $P(F|D)$, podendo ser:

DOI: 10.5540/03.2021.008.01.0487

1. **MCAR** - Dado Faltante completamente aleatório (*Missing Completely at Random*) Onde os dados são faltantes completamente ao acaso e não relacionadas à quaisquer variáveis, ou seja, a probabilidade de uma observação não ser registrada não depende de qualquer outra observação da matriz D:
$$P(F|D) = P(F) \tag{3}$$
o que implica que a probabilidade de ocorrência do dado ausente é a mesma para todos os casos, que a causa que levou aos dados falantes é um evento aleatório;

2. **MAR** - Dado faltante aleatório (*Missing at Random*) Estes tipos de dados faltantes possuem um padrão de perda previsível a partir de outras variáveis, ou seja, os dados faltantes dependem apenas das informações registradas, sendo correlacionadas com a variável que possui dados faltantes:
$$P(F|D) = P(F|D_{existente}) \tag{4}$$
se os dados faltantes não dependem dos valores de $D_{inexistente}$ e apenas dos valores de $D_{existente}$. Neste caso, os dados faltantes são causados por alguma variável observada, disponível para análise e correlacionada com a variável que possui dados faltantes;

3. **NMAR** - Dado faltante não-aleatório (*Missing Not at Random*) Tipo de dado mais difícil de ser tratado em uma análise, onde são relacionados à valores não observados, mais altos ou mais baixos do que o padrão da amostra, ou seja, a probabilidade de dados faltantes varia de razões desconhecidas. Quando F depende dos dados faltantes que constam na matriz D ($D_{inexistente}$), também podendo depender dos dados existentes $D_{existente}$:
$$P(F|D) \neq P(F|D_{existente}) \tag{5}$$

3.2 Métodos de imputação

O processo de fornecer a melhor estimativa para um valor ausente é chamado de imputação de dados. É chamado de método de imputação simples, ou única, a substituição de dados faltantes quando são substituídos uma única vez, por algum dos métodos analisados. Uma desvantagem da imputação simples é a superestimação e subestimação dos erros padrão das estimativas obtidas pelas técnicas, já que os valores são preenchidos apenas uma vez com o mesmo valor para todas as observações faltantes, desconsiderando que outros valores poderiam ser incluídos na estimativa. O método de imputação múltipla está em grande expansão dentro do estudo de imputação de dados e tem como ideia central a substituição de cada valor ausente por dois ou mais valores imputados, seguindo basicamente três etapas:

1. Para cada valor ausente, são gerados X valores, sendo que X deve ser maior ou igual a duas amostras.

2. Estes valores são organizados de forma que o primeiro valor imputado para cada dado ausente produza o primeiro conjunto de dados completado, o segundo valor produz o segundo conjunto e assim por diante, até completar a quantidade de X. Cada conjunto é analisado usando métodos para dados completos.

3. Os resultados das X análises são combinados permitindo que a incerteza da imputação seja considerada.

A imputação múltipla pode ser compreendida basicamente através da Figura 1, onde A representa o banco com os dados faltantes. É feita a etapa de imputação de valores plausíveis para cada valor faltante, o que acarreta a criação de novos bancos de dados completos (B_x), sendo

DOI: 10.5540/03.2021.008.01.0487

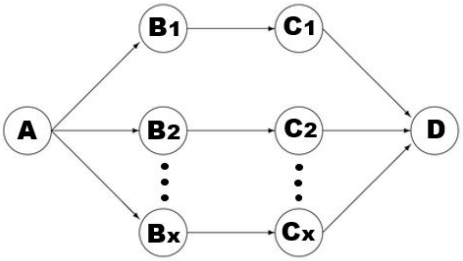

Figura 1: Imputação múltipla.

repetido x vezes. A partir desta etapa que serão feitas as validações das etapas posteriores, sendo que o modelo utilizado para a imputação deve ser o mesmo utilizado para a análise dos dados. Depois, é efetuada a etapa da análise, estimando parâmetros de interesse de cada banco de dados B_x separadamente, gerando x estimações para os C_x. Após a conclusão das etapas anteriores, é feito o agrupamento dos resultados em D.

3.3 Métodos de tratamento para dados faltantes

Existem diversas técnicas que são utilizadas para o tratamento de dados faltantes em banco de dados. Técnicas mais simples, onde é realizada a supressão de registros, e técnicas onde registros são, através de métodos, inseridos onde existe falha de dados. A simples eliminação do registro onde está ocorrendo a falta de dado é uma prática rápida e simples, porém acarreta em perda de dados, já que a informações que existem ali são simplesmente ignoradas. O mais interessante, nos casos de dados faltantes, é eleger uma técnica de imputação de dados, onde os registros incompletos são preenchidos, evitando perda de informação e quantidade de dados a ser analisados.

Como exemplos de técnicas de eliminação de registros afetados por dados faltantes, temos a Deleção de casos incompletos, onde todo o registro que contém uma variável faltante é removido. E também a opção de Deleção *Pairwise*, onde apenas são excluídas as amostras com dados omissos nas variáveis que serão necessárias para a análise. Já no caso de substituição dos valores ausentes, temos diversas técnicas, onde escolhemos algumas para analisar. A Imputação por Média permite que se substitua um valor omisso pela média dos valores presentes na variável de interesse, e possui uma variação que é a imputação pela Mediana, uma das medidas de tendência central, que para ser usada necessita primeiramente ordenar os dados, e em seguida escolher a amostra que divide este conjunto de dados no meio, ou seja, em partes iguais. Para casos específicos, podemos utilizar a imputação por Zero, onde os dados faltantes são substituídos por uma constante, no caso o zero.

Outro opção de imputação existente é a substituição *Hot Deck*, onde é feita a substituição de um valor faltante a partir de um caso similar no conjunto de dados atual, ou seja, para cada caso faltante, este irá ser preenchido por outro valor semelhante presente na própria variável. Uma variação do *Hot Deck* é a substituição *Cold Deck*, que se difere por utilizar um valor de outro conjunto de dados semelhante (dados de amostras anteriores).

Existem também métodos mais elaborados, como a imputação por Regressão que substitui dados faltantes por valores preditos a partir de um modelo de regressão, ou seja, imputa os dados omissos baseado em outras variáveis no conjunto de dados. Método KNN (*K-Nearest Neighbor*), onde k vizinhos são escolhidos com base em alguma medida de distância e sua média é usada como uma estimativa de imputação. Outra técnica existente é a *Expectation Maximization* (EM), que procura estimar os parâmetros da distribuição conjunta dos dados, tais como o vetor de média e matriz de covariância, resultando em estimativas pontuais destes vetores.

DOI: 10.5540/03.2021.008.01.0487

Na Tabela 1 é apresentado uma análise comparativa das técnicas de imputação de dados analisadas nesta Seção destacando os pontos positivos e negativo de cada uma delas

Tabela 1: Análise comparativa métodos imputação de dados.

Tipo	Vantagem	Desvantagem
Deleção de casos incompletos [4]	Implementação simples	Potencial perda de informações.
Deleção Pairwise [2]	Implementação simples.	Causa perda clara de informação disponível nos dados eliminados.
Imputação pela Média [7]	Insere resultados razoáveis e rápidos.	Não leva em consideração a relação entre os atributos, que é útil no processo de tratamento dos valores faltantes.
Imputação pela Mediana [10]	Frequentemente tem um bom desempenho como uma medida de tendência central, quando a distribuição desvia muito da distribuição normal padrão.	Reduz a variação no conjunto de dados.
Imputação por zero [10]	Muito utilizado onde a variável é binária ou o valor zero é plausível.	Abordagem arriscada, depende muito do conhecimento do analista.
Substituição *Hot Deck* [9]	Todos os valores imputados são valores realmente observados.	Dependendo do número de variáveis classificadas pode tornar-se intratável em grandes pesquisas.
Substituição *Cold Deck* [9]	Utilização de outros conjuntos semelhantes de dados.	Dados são de origem externa ao conjunto analisado, que pode ter um viés na análise, não sendo recomendada.
Imputação por Regressão [14]	Como os valores substituídos foram preditos a partir de outras variáveis, eles tendem a se encaixar bem e, portanto, o erro padrão é esvaziado.	Pode-se assumir que existe uma relação linear entre as variáveis usadas na equação de regressão quando pode não haver.
K-Nearest Neighbor [8]	Uma das características mais atraentes do algoritmo KNN é que é simples de entender e fácil de implementar.	O algoritmo KNN é que ele consome tempo ao analisar grandes conjuntos de dados porque ele procura instâncias semelhantes em todo o conjunto de dados.
Expectation Maximization [11]	Tem a vantagem de assegurar a obtenção de uma convergência, de exigir pouco espaço de memória, baixo custo por iteração e facilidade de ser programado.	Lento para convergir em algumas situações práticas, principalmente em dados completos de alta dimensão e sensível diante da presença de dados discrepantes (outliers).

4 Conclusões

Este artigo apresentou uma revisão de um conjunto de técnicas utilizadas no processamento de dados faltantes, um dos grandes desafios do conceito de *Big Data*. Tecnologias utilizadas para analisar este grande volume de dados em diversas áreas têm de lidar, não apenas com uma quantidade imensa de registros, mas também com um problema corriqueiro de dados faltantes. A análise de uma base de dados com muitos registros faltantes afeta a tomada de decisões, pois pode acarretar em informações errôneas devido à supressão excessiva de registros falhos. Para isso, algumas técnicas são utilizadas, tanto de exclusão quanto de imputação de dados. No caso das *Smart Grids*, técnicas de deleção devem ser evitadas, pois acarretam em perdas consideráveis de informações, e como existe a necessidade de altas frequências de aquisição, isso passa a ser um problema. Os métodos tradicionais de imputação por média e mediana, apesar de ser de simples aplicação, tende a reduzir a variação no conjunto de dados, pois preenchem as lacunas sem levar em conta possíveis alterações de valores. Dentre as técnicas relatadas neste trabalho pode-se sugerir a adoção das técnicas de imputação por Regressão, KNN e EM, por conseguirem analisar não apenas a variável faltante, mas todo o histórico dos dados e demais variáveis correspondentes. Em trabalhos futuros pretende-se avaliar estas técnicas a partir de dados reais de um sistema de monitoramento de Subestações Subterrâneas de Energia de uma concessionária do sul do Brasil.

Referências

[1] Alahakoon, D. and Yu, X. Advanced Analytics for Harnessing the Power of Smart Meter Big Data. In *Intelligent Energy Systems (IWIES), IEEE International Workshop*, pages 40-45, 2013.

[2] Allison, P. D. *Missing data*. Sage publications Inc, 2001.

[3] Bhuiyan, S. M. A., Khan, J. F. and Murphy, G. V. Big Data Analysis of the Electric Power PMU Data from Smart Grid. In *SoutheastCon*, 2017.

[4] Brand, J. et al. Multiple imputation as a missing data machine. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, American Medical Informatics Association, 1994.

[5] Chang, G. and Ge, T. Comparison of Missing Data Imputation Methods for Traffic flow. In *2011 International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE)*, pages 639-642, 2011.

[6] Dola, H. M. and Chowdhury, B. H. Data Mining for Distribution System Fault Classification. In *Proceedings of the 37th Annual North American*, pages 457-462, 2005.

[7] Fichman, M. and Cummings, J.M. Multiple Imputation for Missing Data: Making the Most of What you Know. *Organizational Research Methods*, volume 6, n. 3, pages 282-308, 2003.

[8] Keerin, P., Kurutach, W. and Boongoen, T. Cluster-based KNN missing value imputation for DNA microarray data. *2012 IEEE International Conference on Systems, Man, and Cybernetics(SMC)*, pages 445–450, 2012.

[9] Lakshminarayan, K., Harp, S. A. and Samad, T. Imputation of Missing Data in Industrial Databases. *Applied Intelligence*, volume 11, n. 3, pages 259–275, 1999.

[10] Mcknight, P. et al. *Missing data: A gentle introduction*. Guilford Press, 2007.

[11] Mclachlan, G. and Krishnan, T. *The EM Algorithm and Extensions*. John Wiley & Sons, 2nd Edition, 2008.

[12] Narayanan, U., Paul, V. and Joseph, S. Different Analytical Techniques for Big Data Analysis: A Review. In *International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 2017.

[13] Rubin, D. B. Multiple imputations in sample surveys: a phenomenological bayesian approach to non response. In *Proceedings of the Survey Research methods Section*, 1978.

[14] Rubin, D. B. *Multiple imputation for nonresponse in surveys*. John Wiley and Sons, 1987.

[15] Russom, P. *Big Data Analytics. TDWI Best Practices Report, Fourth Quarter, v. 19, n. 4, pp. 1-34*, 2011. Disponível em: https://vivomente.com/wp-content/uploads/2016/04/big-data-analytics-white-paper.pdf. Acesso em: 25 abr. 2020.

[16] Sebestyen, G. S. *Decision-making processes in pattern recognition*. Macmillan Publishing Co., Inc, New York, 1962.

[17] The Green Grid - Energy Savings and Carbon Emissions Reductions Enabled by a Smart Grid, EPRI, 2007. Disponível em: https://tinyurl.com/rbhrf3q. Acesso em: 25 abr. 2020.

[18] Tleis, N. *Power Systems Modelling and Fault Analysis: Theory and Practice*. Elsevier, 2007.

