Imputação de dados faltantes no monitoramento de consumo energético residencial em *Smart Grids*

Matheus T. M. Barbosa[1], Alexandre Lima[1], Bruno T. Kuehne[2], Bruno G. Batista[2], Dionisio M. L. Filho[3], Maycon L. M. Peixoto[1]

[1] Universidade Federal da Bahia (UFBA)
Salvador – BA – Brasil

[2]Universidade Federal de Itajubá (UNIFEI)
Itajubá – MG – Brasil

[3]Universidade Federal do Mato Grosso do Sul (UFMS)
Ponta Porã – MS – Brasil

{matheus.thiago, maycon.leone}@ufba.br, alexandrecurylima@hotmail.com,

{brunoguazzelli, brunotardiole, dionisio.mlf}@gmail.com

Abstract. *Smart Grids are networks that are responsable for the energy distribuition on a safe way and promote a fair measure of consumption, for having a big quantity of sensors to monitor and record diferente quantity of data throughout the day, they may fail to collect informations, producing missing data or invalid, affecting the quality of the service. Therefore, this article presents the proposal of an adaptive algorithm, built from the performance evaluation of two algorithms used for the imputation of missing data, Spline and Singular Spectrum Analysis-SSA. Performance evaluation shows significant improvements in the imputation of missing data using the built algorithm, allowing a more accurate measurement of consumption even with missing data.*

Resumo. *As Smart Grids são redes responsáveis por distribuir a energia de forma segura e promover uma medição justa do consumo. Por terem uma grande quantidade de sensores para monitorar e registrar diferentes quantidades de dados ao longo do dia, podem deixar de coletar informações, produzindo dados ausentes ou inválidos, afetando a qualidade do serviço. Esse artigo apresenta a proposta de um algoritmo adaptativo, construído a partir da avaliação de desempenho de dois algoritmos utilizados para a imputação de dados faltantes, o Spline e o Singular Spectrum Analysis(SSA). A avaliação de desempenho mostra melhorias significativas na imputação de dados faltantes com o algoritmo construído, permitindo uma medição mais precisa do consumo mesmo com dados faltantes.*

1. Introdução

O atual sistema de energia elétrica permanece inalterado desde o século passado e nestes sistemas, a energia é transportada de alguns geradores centrais de forma radial para um grande número de centros de carga onde estão os usuários [Siddiqui et al. 2008]. As leituras do consumo desses usuários são feitas periodicamente pelos funcionários das companhias elétricas, demandando tempo e custo deste trabalho. Além disso, esses sistemas são ineficientes quanto a entrega de energia (pois uma grande parte desta é perdida no processo de transmissão); são

fisicamente e virtualmente inseguros; bem como propensos a possíveis falhas [Lo and Ansari 2012].

Devido a isso, nos últimos anos, houve um aumento nos esforços pelas empresas de serviços públicos de geração de eletricidade, governos e pesquisadores na construção de sistemas de leitura automática de medidores- *Automatic Meter Reading* (AMR) [Khalifa et al. 2011], que, como o nome já diz, é um hardware de coleta automática de dados de medidores de energia que são transferidos para um sistema centralizado para o processamento subsequente. Acompanhando essa crescente, as *Smart Grids* que tem como espinha-dorsal tais sistemas, que nesse caso são chamados de *Smart Meters*, vem crescendo e implantando uma nova forma de entrega e produção de energia elétrica [Zheng et al. 2013].

As *Smart Grids* são responsáveis por distribuir a energia de forma segura e com tolerância às possíveis falhas, promovendo uma medição justa do consumo. Além disso, tais redes permitem o fluxo de informação e de energia de forma bidirecional, logo, além de consumir, o usuário pode também produzir energia, como por exemplo, a eólica e a solar, impulsionando a implantação de fontes de energia renováveis, aumentando a eficiência da geração, transmissão e uso. Sendo assim, as *Smart Grids* são consideradas uma evolução das redes elétricas existentes, pois a integração de tecnologias avançadas de computação e comunicação oferece ganhos de desempenho e confiabilidade para os sistemas de energia [Zheng et al. 2013] [Yaacoub and Abu-Dayya 2014].

Mesmo no cenário de *Smart Grids*, existe a possibilidade de ocorrerem problemas de perda ou leitura incorreta dos dados nos *Smart Meters* devido a fraudes ou falhas no processo de transmissão e medição, de maneira que dados corrompidos ou faltantes são medidos e transmitidos à central gerando inconsistência [Cemgil et al. 2017] [Chen et al. 2010]. Tais problemas não são enfrentados apenas nessa área, podendo ser também um empecilho na geofísica, como relatado por [Li et al. 2017], onde a falta de dados é um fenômeno comum nas medidas do ambiente espacial. Em sistemas inteligentes de transporte, de acordo com [Qu et al. 2009], os problemas de ausência de dados são inevitáveis. No ano de 2009, em Pequim (China), esse sistema possuía uma taxa de informação perdida de 10%, podendo variar em até 25% em algumas situações. Dessa forma, a maioria dos modelos e teorias disponíveis precisam de dados completos, especialmente quando se tratam de dados que são analisados em tempo real. As perdas trazem sérios desafios para o processo de modelagem e análise subsequente. Estimar os valores para substituir os dados ausentes mantendo as principais características dos valores originais é uma tarefa complexa.

O modelo mais simples de preenchimento de lacunas é o **LOCF** - (*Last Observation Carried Forward*), que substitui valores inválidos pelo último valor válido encontrado, enquanto outros métodos se baseiam em dados históricos para tal. Além desses, algumas técnicas utilizam o método **Spline**, que faz a imputação por regressão, e há também, quem utilize o SSA, que emprega correlações espaço-temporal para fazer o preenchimento das lacunas.

Desse modo, a partir das análises dos resultados dos algoritmos clássicos da literatura de imputação de dados: Spline e SSA, é proposto neste artigo um algoritmo chamado de AdaptS. O AdaptS foi concebido seguindo rigorosamente a metodologia de Planejamento de Experimentos descrita em [Jain 1991]. Essa metodologia permitiu identificar para cada cenário experimentado do ambiente de *Smart Grids*, qual a abordagem mais indicada de imputação de dados: considerando a relação combinatória entre os fatores, níveis e variáveis de reposta.

Os resultados apresentados neste artigo mostram que é possível reconstruir os dados ausentes por meio do algoritmo AdaptS com mais acurácia do que os métodos tradicionais da literatura utilizados de maneira isolada.

Este trabalho foi divido em seções, de maneira que na Seção 2 é apresentada uma revisão da literatura com algumas das abordagens existentes acerca do tema, enquanto na Seção 3 são apresentados os métodos para preenchimento de dados ausentes encontrados também na literatura. Para tornar possível a reprodução dos experimentos, a Seção 4 apresenta as variáveis e informações utilizadas nos ensaios, além dos resultados dos métodos Spline e SSA, que foram utilizados como base para a criação do algoritmo AdaptS. A Seção 5, faz uso do planejamento de experimentos utilizados na seção anterior para mostrar os resultados obtidos pelo projeto fatorial 2^k com os algoritmos SSA e AdaptS, podendo ser verificado o grau de influência que os fatores exercem sobre a variável de resposta Acurácia. Na Seção 6 todas as considerações finais são apresentadas.

2. Trabalhos Relacionados

Nos últimos anos, há uma crescente nos esforços na industria de energia elétrica, principalmente quando se diz em redes inteligentes de energia, chamadas de *Smart Grids*. Elas trazem o conceito do uso intensivo de tecnologia de informação e comunicação na rede elétrica, com o intuito de conectar diversos dispositivos e criar estratégias de controle e otimização. O conceito trazido por [Siddiqui et al. 2008], mostra que *Smart Grids*, são responsáveis por distribuir a energia de forma segura e com tolerância às possíveis falhas, promovendo uma medição justa do consumo.

Neste contexto, o fluxo de energia elétrica e de informações se dá de forma bidirecional. Assim, a energia tradicionalmente gerada, transmitida e distribuída de forma radial a partir de instalações das concessionárias, poderá também, ser gerada e integrada às redes elétricas a partir de unidades consumidoras. Logo o consumidor passa a ser produtor e consumidor, pois produz e fornece energia à rede. Dessa forma, as *Smart Grids* dão ao usuário a capacidade de gerenciar e tomar decisões relacionadas ao seu consumo, que hoje é feito pelas empresas de geração de energia. Assim, esse conceito tenta mudar um sistema que permanece praticamente inalterado desde o século XIX, com estações geradoras e um sistema de entrega de energia eletromecânico, operado por centrais de controle.

Para criar estratégias de otimização nas *Smart Grids*, são utilizados os *Smart Meters*. Eles são dispositivos inteligentes agregados à rede que podem tanto coletar dados do consumo de energia, como coletar informações sobre o uso desta pelos clientes, quase que em tempo real [Fang et al. 2012]. Este conhecimento antecipado de carga, torna possível ser feita a análise, o tratamento e o armazenamento de dados para uso futuro, porém, devido à grande quantidade dos mesmos, é fundamental que eles estejam corretos e precisos para que seu uso seja efetivo.

Em seu trabalho [Li et al. 2017] mostra, por exemplo, que dados faltantes são bastantes comuns na geofísica, principalmente em dados dinâmicos. As causas principais para essas perdas se dão às falhas dos equipamentos de medição, saturação do sensor, condições de medição, valores anormais, entre outros. Para resolver esse problema, foi proposto um algoritmo baseado no Singular Spectrum Analysis (SSA) tendo resultados significativos quanto ao uso neste tipo de pesquisa.

[Lecomte et al. 2017] traz também, o uso do SSA com o intuito de substituir valores faltantes no sistema de vigilância no ambiente das Smart Cities, onde precisão é fundamental.

Devido às grandes áreas distribuídas e à quantidade de dispositivos de vigilância, podem correr falhas com os monitores de carga de trabalho por motivos que vão desde o funcionamento inadequado dos elementos de monitoração de dados, perda de pacotes durante a transmissão dos mesmos e até pontos cegos do sistema de vigilância. Esses eventuais erros podem causar problemas no rastreamento de objetos e na qualidade do serviço, afetando a experiência do usuário. Fazendo um comparativo entre o Spline e o SSA, o segundo apresentou valores mais precisos para os dados ausentes, com acurácia média de 96,59% em relação aos dados originais, sendo 31,79% melhor que o Spline. Os resultados comprovaram a importância da utilização das técnicas de substituição de dados ausentes no ambiente de Smart Cities, principalmente no setor de vigilância.

Utilizando outra abordagem algorítmica, [Genes et al. 2016] analisa a performance na recuperação de dados faltantes feitas por um algoritmo de *matrix completion* utilizando *singular value thresholding* (SVT), em comparação com o algoritmo de estimativa de erro quadrático médio mínimo (MMSE). Ele utilizou dados reais da *Electricity North West Limited*, para a avaliação dos dados tanto para amostragem aleatória, quanto para amostras de codificação linear ótima. Os resultados numéricos mostraram que o algoritmo SVT, superou o estimador MMSE quanto à sua recuperação quando o número de observações disponíveis é baixo e as estatísticas de dados não são perfeitamente conhecidas.

[Chen et al. 2010] aborda um problema enfrentado pelos *AMR's* na medição em sistemas de energia, a limpeza de carga. Tal problema se dá pelo motivo que os dados registrados podem vir com falhas, seja no processo de medição, como no processo de transmissão, assim, as curvas de dados podem apresentar dados corrompidos e faltantes. A solução é proposta pela modelagem da estrutura subjacente dos dados da curva de carga usando técnicas de regressão não paramétrica, o B-Spline smoothing e o Kernel smoothing. Utilizando dados reais das curvas de carga da *British Columbia Transmission Corporation* (BCTC), o experimento e a avaliação foram realizados, mostrando a eficácia da solução apresentada.

[Cemgil et al. 2017] assim como em [Chen et al. 2010] trata do problema de dados faltantes e atípicos em um *AMR*, e mostra que os dados de consumo de energia elétrica carregam boas indicações de fraude no sistema, caso estas estejam presentes. Os autores criaram dois algoritmos para poder interpolar dados faltantes e, dessa forma, poder detectar fraudes, denominados *Auto-Regressive* (AR) e *Non-negative Matrix Factorization* (NMF). Tais algoritmos foram comparados com o *Alternating Least Square* (AML), e os resultados da simulação mostraram que a detecção de Outliers ajuda o algoritmo NMF a melhorar sua interpolação, e comparando com o algoritmo ALS o NMF funciona melhor e mais rápido.

Os trabalhos citados acima tem como foco mostrar como a falta de dados pode ser resolvida de diversas formas. O presente artigo, considera a perspectiva do ambiente de perda de dados em *AMR* de baixo custo em *Smart Grids*.

3. Métodos Para o Preenchimento de Dados Ausentes

A revisão bibliográfica apontou que as técnicas de aprendizado de máquinas podem ser utilizadas para a substituição de dados ausentes. Dentre os resultados encontrados, as abordagens utilizadas neste trabalho são: **Spline** e **SSA**.

3.1. Spline

A interpolação cúbica Spline é um método geralmente utilizado como uma alternativa aos métodos de aprendizado de máquina [Richardson et al. 2015] e no processamento de sinais

temporais [Hussain et al. 2015]. Essa técnica consiste em analisar um conjunto de dados $A = \{a_1, a_2, a_3, \ldots, a_{n-1}, a_n\}$ que tem seus pontos gerados por uma regra $g(\cdot)$ desconhecida. O algoritmo tenta, então, estimar uma função $g'(\cdot)$ para a qual $g(a) = g'(a), \forall a \in A$.

Desta forma, para se definir $g'(\cdot)$ é utilizado um polinômio de grau 3 para cada intervalo entre observações. Ou seja, uma spline $S(x)$ é definida por:

$$S3(x) = \begin{cases} C_0(x), & x_0 \leq x \leq x_1 \\ C_i(x), & x_{i-1} \leq x \leq x_i \\ C_n(x), & x_{n-1} \leq x \leq x_n \end{cases} \tag{1}$$

Onde cada função $C_i(\cdot)$ é definida sobre a forma de $C_i(x) = a + bx + cx^2 + dx^3$ e deve passar pelas observações de forma que $C_i = a_i$ tal que $1 \leq i \leq n$.

3.2. SSA

O SSA é um método não paramétrico usado na análise de séries temporais e que quase não exige o conhecimento comportamental prévio [de Miranda Esquivel et al. 2013]. Ao contrário do Spline, ele identifica e utiliza os padrões geradores da série temporal como ruído, sazonalidades e tendência do conjunto de dados observados para fazer a recuperação das lacunas. Esse método faz a sua investigação a partir do comportamento histórico através da decomposição e reconstrução dos seus componentes que constituem a série. Cada estágio é composto por quatro passos da técnica: incorporação (*embedding*), decomposição do valor singular (SVD), agrupamento (*grouping*) e média diagonal (*diagonal averaging*) [Hassani 2007].

A reconstrução da série temporal inicia-se através da decomposição de dados observados em uma soma de poucas subséries, sendo estas identificadas e interpretadas como componentes constitutivos. Na decomposição dos dados, vem o primeiro passo que é a incorporação, onde uma matriz de trajetórias é produzida transformando o conjunto de dados unidimensional em uma série de dimensões L, onde L é dito o comprimento da janela, sendo o único parâmetro que representa a quantidade de componentes em que a série é decomposta. Este valor deve ser inteiro, entre $2 \leq L \leq N$, e o tamanho de L deve ser suficientemente grande, mas não superior a $\frac{N}{2}$ [Hassani 2007]. Considerando X, um conjunto de dados tal que $|X| = n$ e um $K = n - L + 1$, onde K é o número de vetores deslocados no tempo, temos:

$$\begin{aligned} \mathbf{X} &= [X_1 : \ldots : X_K], \\ &= (x_{ij})_{i,j=1}^{L,K}, \\ &= \begin{pmatrix} x_1 & x_2 & x_3 & \ldots & x_K \\ x_2 & x_3 & x_4 & \ldots & x_{K+1} \\ x_3 & x_4 & x_5 & \ldots & x_{K+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_L & x_{L+1} & x_{L+2} & \ldots & x_N \end{pmatrix} \end{aligned} \tag{2}$$

No terceiro passo, a decomposição de valor singular (SVD) é usada para decompor a matriz de trajetórias X em uma soma de matrizes elementares, de modo que $X = E_1 + E_2 + \ldots + E_d$, onde d é igual ao número de autovalores não nulos da matriz $M = XX^T$.

Após o processamento da SVD, é começada a fase de reconstrução pelo passo de agrupamento, onde é feita a junção das matrizes elementares E_i em vários grupos. Os grupos $I_1, \ldots, I_m$ representam subconjuntos distintos dos índices das matrizes elementares, representados por $1, \ldots, d$.

Então, obtém-se uma nova representação da matriz de trajetórias, de modo que $X = E_{Ii}, i \in \{1, \ldots, m\}$. Dessa forma, como resultado deste passo, é dada a representação da matriz de trajetória como uma soma de matrizes resultantes $E_{Ii}, i \in \{1, \ldots, m\}$, onde pelo quarto e último passo, obtém-se para cada uma das matrizes resultantes, uma aproximação da série original [de Miranda Esquivel et al. 2013].

4. Ambiente de Experimentos

Durante o projeto deste artigo foi realizado um estudo e uma série de experimentos avaliando a capacidade de imputação de dados faltantes utilizando os algoritmos Spline e SSA. Para tal, utilizou-se o planejamento de experimentos fatorial 2^k, o qual foi apresentado por [Jain 1991], onde são definidos dois níveis para cada fator. Foi elaborada a Tabela 1 que apresenta os fatores e seus respectivos níveis utilizados.

Tabela 1. Fatores e níveis

Fator	**Níveis**	
Abordagem (A)	Spline	SSA
Quantidade de dados inválidos (B)	12960	20760
Tipo de Lacuna (C)	Contínua	Individuais(Randômicas)
Fonte de dados (D)	ÉdaSuaConta	IoTaWatt

Este modelo utiliza um arranjo de 2^4, que é o número de níveis elevado ao número de fatores, sendo alcançado a partir da Equação 3.

$$\begin{aligned} y &= q_0 + q_A x_A + q_B x_B + q_C x_C + q_D x_D + q_{AB} x_{AB} + q_{AC} x_{AC} + q_{AD} x_{AD} \\ &+ q_{BC} x_{BC} + q_{BD} x_{BD} + q_{CD} x_{CD} + q_{ABC} x_{ABC} + q_{ABD} x_{ABD} q_{ACD} x_{ACD} \\ &+ q_{BCD} x_{BCD} + q_{ABCD} x_{ABCD} \end{aligned} \tag{3}$$

Substituindo os valores dos experimentos, obtêm-se os valores de $q_A, q_B, q_C, q_D, q_{AB}, q_{AC}, q_{AD}, q_{BC}, q_{BD}, q_{CD}, q_{ABC}, q_{ABD}, q_{ACD}, q_{BCD}, q_{ABCD}$ como mostra a Equação 4, onde é calculado o valor de q_0.

$$\begin{aligned} q_0 = 1/16 * (&y_1 + y_2 + y_3 + y_4 + y_5 + y_6 + y_7 + y_8 + y_9 \\ &+ y_{10} + y_{11} + y_{12} + y_{13} + y_{14} + y_{15} + y_{16}) \end{aligned} \tag{4}$$

A partir dos valores obtidos pode-se determinar a soma dos quadrados. A variação total ou Soma Total dos Quadrados (SST), é dada pela Equação $SST = \sum_{i,j} (y_{ij} - \bar{y})$. Nesta equação, $\bar{y}$ representa a média das respostas de todas as repetições de todos os experimentos. Na simulação realizada o SST é dado por: $SST = 2^4(q_A^2 + q_B^2 + q_C^2 + q_D^2 + ... + q_{ABCD}^2)$.

Por meio da utilização do modelo de regressão, a SST fornecerá a variação total das variáveis de resposta e a influência de cada fator e suas interações. Para obter a influência de um determinado fator, por exemplo o fator A, é necessário utilizar *y = SSA/SST*, onde *SSA* = $2^4 * q_A^2$..

Dessa forma, apesar de ter um grande número de experimentos e ter um alto custo para a avaliação, este modelo foi utilizado pelo motivo de trazer vantagens como a possível avaliação de todos os fatores, podendo-se assim, determinar a influência de qualquer fator. Além disso, também pode ser verificada as interações entre os fatores.

A abordagem foi definida pelos algoritmos acima citados, que são os objetos principais de tal estudo. As fontes de dados diferentes ÉdaSuaConta e IoTaWatt, foram escolhidas para que fosse possível validar os métodos de imputação em dois ambientes diferentes. Tais dados, que compõem as séries temporais, foram obtidos a partir do banco de dados da leitura do consumo em *Watts* de duas residências que utilizam *AMR*'s *open hardware* e *open source* de baixo custo, o IoTaWatt e o ÉdaSuaConta. Ambos foram feitos a partir da colaboração do projeto *OpenEnergyMonitor (OEM)*[1].

O IoTaWatt[2] é baseado na plataforma ESP8266, usa um adaptador MCP3208 para ler as amostras de tensão e transformadores de correntes (CT's) não invasivos SCT-013 para fazer a leitura das mostras da corrente. Já o ÉdaSuaConta, que é baseado nas plataformas Raspberry Pi e Arduino, utiliza também os transformadores de corrente não invasivos SCT-013 para as leituras das amostras de corrente, e para as leituras de tensão, utiliza apenas a tensão de referência do local.

As amostras de tensão e corrente foram lidas a cada segundo, sendo somadas e salvas no banco de dados a cada 10 segundos. Dessa forma, os dados da 00h do dia 1º de Agosto de 2017 até às 23:59 do dia 30 deste mesmo mês, foram utilizados como a série temporal desse trabalho, com um total de 259200 dados de leituras.

Após a definição da fonte de dados, e a série temporal total em estudo, foram definidos os valores de T, que representa a quantidade total de dados faltantes, Q, que representa a quantidade de lacunas e L, que é dado por $L = \frac{T}{Q}$, representando, a quantidade de valores inválidos (NA) em cada lacuna. A quantidade de lacunas contínuas (L) foi definida como 30, sendo uma lacuna contínua por dia. Na quantidade total de dados faltantes T, foram escolhidos dois valores, 12960 e 20736, que são respectivamente 5 e 8% da quantidade total de dados (259200). Como 20736 não é divisor inteiro de 30, os valores de L, ficaram em 692 e 431, representando, desta forma, a perda diária por falhas no processo de medição. Nas lacunas aleatórias, T valores inválidos (NA) são distribuídos uniformemente em toda a série, representando as perdas de dados inválidos, ou seja, quando o valor lido é muito distante do intervalo esperado ou quando o valor não chega ao *AMR* por perda de pacotes, por exemplo.

A variável de resposta adotada foi a acurácia, que mede quão próximo o valor obtido do experimento está do valor original. Para isso, utilizou-se o cálculo da discrepância relativa que é a diferença entre dois valores medidos de uma mesma grandeza e a acurácia é tanto maior quanto menor, a depender da discrepância relativa. Uma medida x, valor do experimento, pode ser avaliada pela discrepância relativa, $\Delta = \left|\frac{x - x_{ref}}{x_{ref}}\right|$, onde x_{ref} é o valor original. Assim, a acurácia é dada por $\Theta = (1 - \Delta) * 100$. A escolha dessa variável de resposta possibilita observar a influência dos fatores em torno da proximidade dos valores

[1]Disponível em `https://openenergymonitor.org`

[2]https://github.com/boblemaire/IoTaWatt/wiki

imputados em relação aos dados originais.

4.1. Resultados sobre o SSA e Spline

Com o intuito de explicar o processo de obtenção da acurácia por meio do preenchimento de dados ausentes, considere a figura 1. Essa série temporal contém 86400 observações. Neste exemplo, existem 8640 valores aleatórios inválidos distribuídos uniformemente em toda a série. E

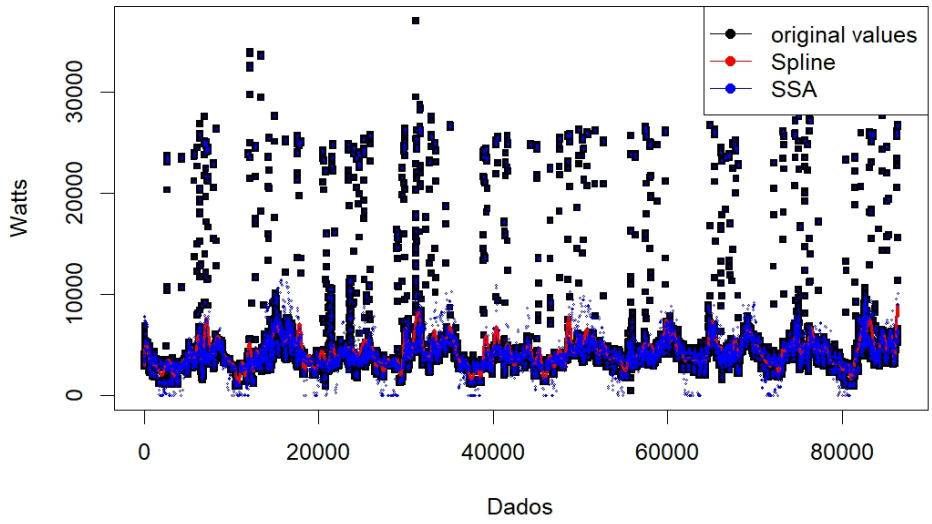
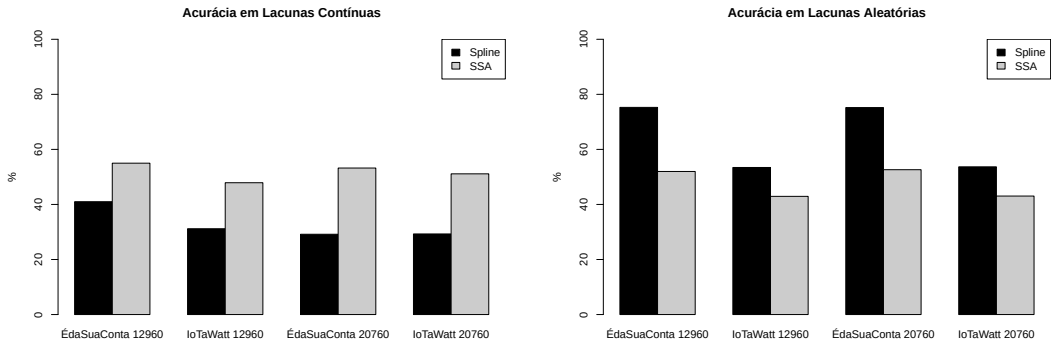


Figura 1. Série temporal com 86.4000 após a análise pelos métodos Spline e SSA

Os resultados obtidos mostraram que o desempenho quanto a acurácia dos algoritmos Spline e SSA se comportam de formas diferentes de acordo com o tipo de lacuna, mesmo aumentando a quantidade de dados inválidos, de 12960 para 20760. O SSA por exemplo, como pode ser visto nas figuras 2(a) 2(b) tem um desempenho melhor em lacunas contínuas, ao contrário do Spline, que tem uma acurácia melhor em lacunas aleatórias. Além disso, a partir da análise do planejamento de experimentos foi possível verificar um limiar de até 6 perdas contínuas para que o Spline funcione com melhor acurácia que o SSA.



(a) Lacunas Contínuas (b) Lacunas Aleatórias

Figura 2. Comparação da acurácia do Spline e SSA em lacunas de 12960 e 20760.

4.2. O Algoritmo AdaptS

O algoritmo de imputação de dados proposto neste trabalho verifica a quantidade de dados faltantes por série, para fazer a melhor escolha algorítmica a partir dos resultados obtidos pelo estudo feito na na seção 4.1.

O Algoritmo 1, chamado de **AdaptS**, é chamado quando no relatório se percebe que houve perda de um ou mais dados. Por ser um algoritmo em tempo de projeto, pode-se ter mais lacunas de diferentes tamanhos, dessa forma, é feita uma verificação do tamanho da lacuna, para a partir daí, escolher qual é o melhor algoritmo a ser usado. Utilizamos uma função $f(tamanholacuna)$ para identificação do tipo de lacuna, caso a função retorne o – $Tipo\ (1)$ – (lacuna contínua) é escolhido o algoritmo SSA, por outro lado, caso o – $Tipo\ (2)$ – (lacuna aleatória) é escolhido o algoritmo Spline. Ao receber a série temporal total e os vetores de pisos e tetos (que são os intervalos de perda de dados, onde se inicia e termina cada lacuna), o algoritmo começa a interação pegando a primeira fatia de dados que vai do início da série temporal ao primeiro teto (o fim da primeira lacuna). A partir daí, é feita a chamada da função $f(tamanholacuna)$ para escolha do algoritmo Spline ou do algoritmo SSA. Após fazer a imputação desses dados, o método troca é chamado, sendo responsável por fazer a imputação de dados originais, que vai da primeira posição da série até o teto da interação, pela fatia da série dos dados imputados.

Algoritmo 1: Algoritmo AdaptS

```
Data: serieTemporal, vetorPisos, vetorTetos
Result: serieTemporalImputada
for i ← 1 to i ≤ tamanho(vetorPisos) do
    piso ← vetorPisos [i];
    teto ← vetorTetos [i];
    tamanhoLacuna← teto−piso;
    for j ← 1 to j ≤ teto do
        temporaria ← serieTemporal [j];
    end
    switch f(tamanhoLacuna) do
        case 1 do
            serieTemporalImputada ← SSA(temporaria);
        end
        case 2 do
            serieTemporalImputada ← Spline(temporaria);
        end
    end
    troca(serieTemporal, serieTemporalImputada);
end
return(serieTemporalImputada);
```

5. Modelo de Avaliação: Planejamento de Experimentos

Para a avaliação do algoritmo proposto, utilizou-se o planejamento de experimentos fatorial 2^k, da mesma forma que explicado na seção 4, modificando apenas a abordagem, com os algoritmos AdaptS e o SSA, este sendo utilizando por ter uma acurácia média em relação ao Spline. A tabela 2 apresenta os fatores e seus respectivos níveis utilizados no experimento.

Para validar os dados observados, a figura 3 representa a distribuição residual obser-

Tabela 2. Fatores e níveis

Fator	Níveis	
Abordagem	AdaptS	SSA
Quantidade de dados inválidos	12960	20760
Tipo de Lacuna	Contínua	Individuais(Randômicas)
Fonte de dados	ÉdaSuaConta	IoTaWatt

vada nos resultados. Ela apresenta a observação da normalidade na execução dos experimentos. O esperado é que os pontos do gráfico, relacionados aos experimentos, residam sobre ou próximos à linha normal, como é observado em tal.

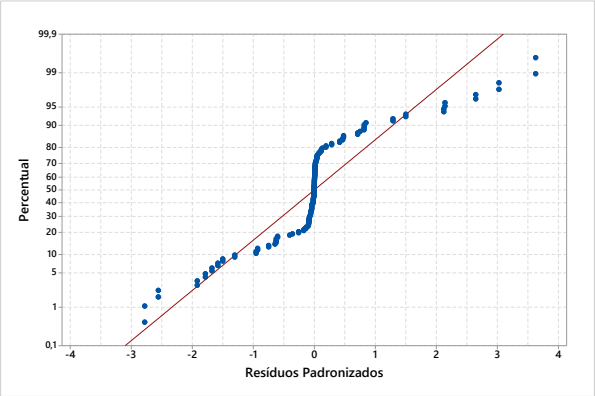


Figura 3. Distribuição residual das observações

A figura 4 mostra o gráfico pareto dos efeitos para o projeto fatorial 2^k com a combinação dos algoritmos SSA e AdaptS. Com esse gráfico é possível verificar o grau de influência que os fatores exercem sobre a variável de resposta Acurácia. Os efeitos mais significativos são Fonte de Dados (A), Abordagem (D), a interação entre os fatores B e D, e A e B. além do fator Tipo de Lacuna (B). Isso significa que a mudança na fonte de dados, de algoritmo e dos tipos de lacuna alteram significativamente a acurácia.

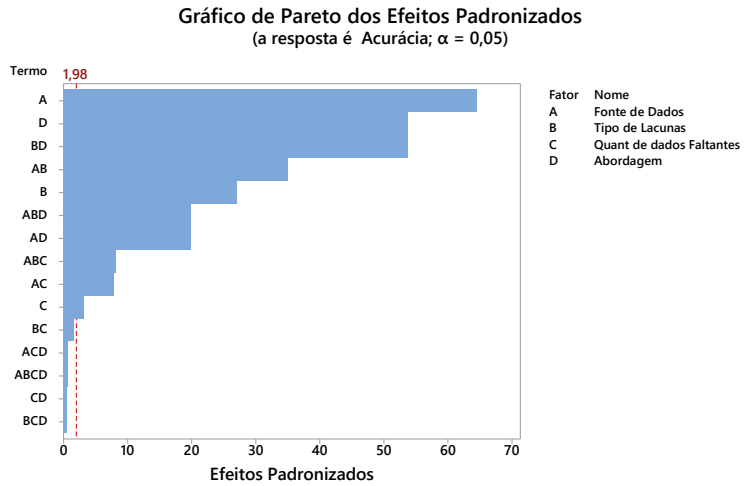


Figura 4. Influência dos Fatores

Na figura 5 há um aprofundamento nas informações em relação aos efeitos de in-

fluências dos fatores observados. À medida que o fator B, dirige-se à direita da linha central vermelha normalizada, ocorre o acréscimo no valor obtido da variável de resposta. No entanto à medida que os fatores A, D e BD, se encontram à esquerda da linha normalizada, sugere-se uma diminuição na acurácia.

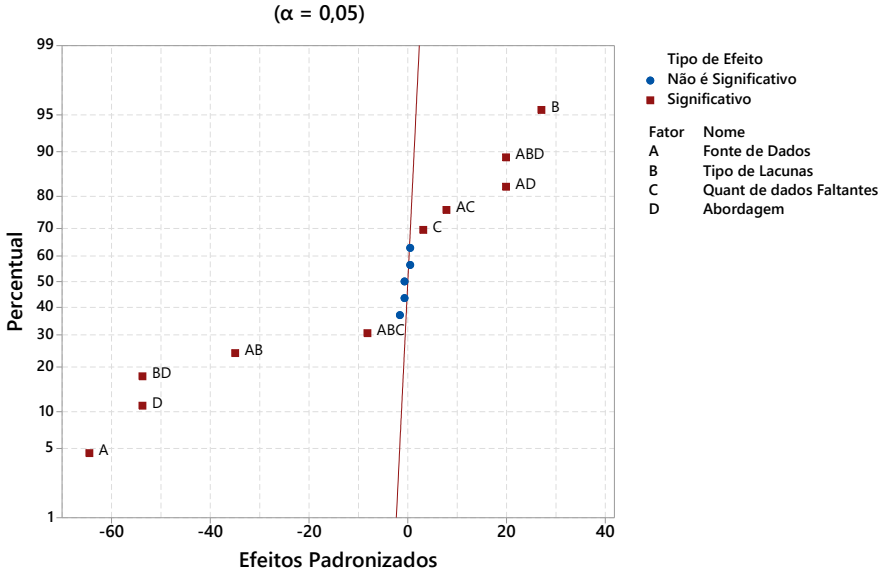


Figura 5. Influência dos Fatores

Sob a análise da influência das interações entre os fatores, a figura 6 mostra que o algoritmo proposto possui uma acurácia maior que o método SSA. À medida em que as lacunas diminuem de tamanho, é possível verificar a tendência de aumento da acurácia do algoritmo proposto em relação ao SSA, como pode ser visto na interação entre a Abordagem e o Tipo de Lacunas.

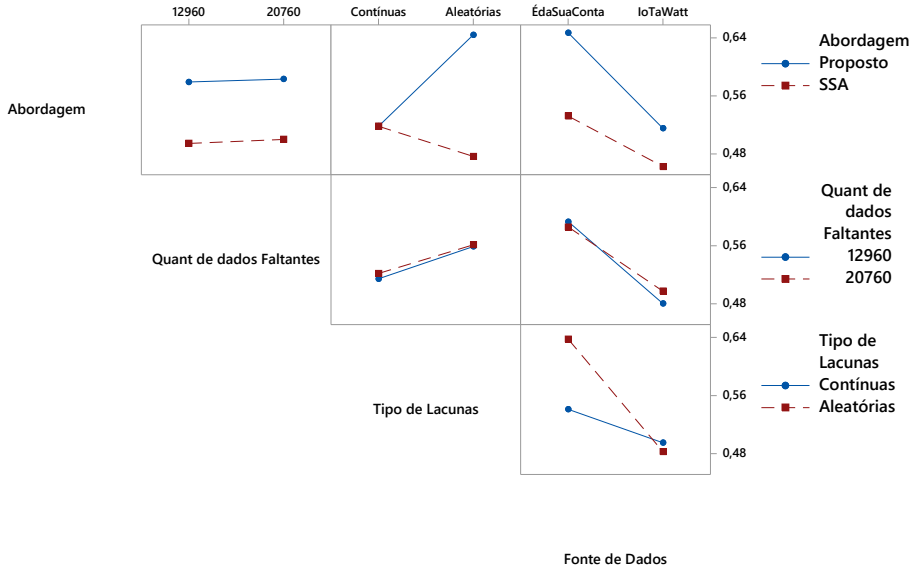


Figura 6. Gráfico de Interação entre os fatores.

Em comparação ao SSA, o algoritmo AdaptS proposto nesse trabalho, conseguiu maximizar a média da acurácia, mesmo nos cenários em que as lacunas aumentam e a fonte de

dados muda. De acordo com os cenários experimentados, o AdaptS mostrou ser capaz de imputar de melhor forma que o SSA e o Spline, tendo acurácia média de 58.10 contra 49.73 e 48.52%, respectivamente, como pode ser visto na figura 7.

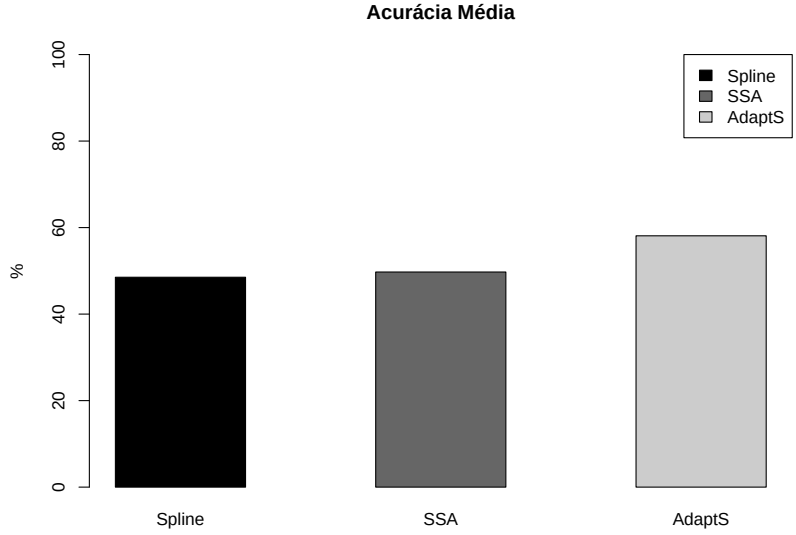


Figura 7. Média da acurácia nos algoritmos Spline, SSA e AdaptS.

6. Conclusão

No contexto de uma *Smart Grid*, para promover uma medição justa do consumo de energia são necessários mecanismos de tolerância à falhas. Essas falhas podem ocorrer por fraudes ou funcionamento inadequado nos aparelhos de medição, chamados de AMR ou *Smart Meters*, ou até perda de pacotes na transmissão de dados destes aparelhos para a central. Os erros de dados inadequados ou dados ausentes nesses aparelhos de medição podem causar perdas na qualidade do serviço se não forem tratados apropriadamente.

Dessa forma, para resolver os problemas de dados ausentes na série temporal de estudo deste artigo, que são dados reais de dois AMR's; o IoTaWatt e o ÉdaSuaConta, nos trinta primeiros dias do mês de agosto de 2017, foi proposto um algoritmo adaptativo chamado AdaptS, que utiliza os algoritmos Spline, que é uma técnica indicada para cenários com lacunas menores, uma vez que esse método consegue bons resultados por meio da interpolação de dados, e o SSA que é caracterizado dentro da área de séries temporais por realizar uma estimativa espectral não-paramétrica com as correlações espaço-temporal.

Os resultados comprovaram que o AdaptS produziu valores mais precisos para os dados ausentes, maximizando assim a sua acurácia média em relação aos dados originais, sendo 58,10% contra 49,73% e 48,52%, do SSA e Spline. Além disso, a aplicação da técnica estatística de planejamento de experimento permitiu comparar minuciosamente os algoritmos Spline e SSA, e a partir daí propor um novo algoritmo.

Ressalta-se a contribuição o algoritmo AdaptS construído para a substituição de dados ausentes, estimando valores que não foram corretamente monitorados, aumentando a acurácia no monitoramento de energia elétrica no ambiente *Smart Grid*. A acurácia quando analisada separadamente do contexto, pode-se passar a impressão de que 60% é baixo, mas isso se

deve a rigorosidade na construção da métrica, onde não analisamos a precisão e sim acurácia, incluindo a impossibilidade de obter melhores valores que é dada condição aleatória dos GAP's analisados pela carga de trabalho real. Mesmo assim, a acurácia produzida no nosso trabalho é em média 10% maior que as dos algoritmos clássicos que hoje são referência na literatura para o problema de Gap Filling.

Além disso, esse trabalho permitiu demonstrar que para o cenário de *Smart Grid* o Spline é uma técnica indicada para cenários com dados ausentes aleatórios em relação à amostra total (lacunas menores) e o SSA apresenta bons resultados independente do tamanho da lacuna, porém com acurácia significativamente menor do que o AdaptS.

7. Agradecimentos

Os autores agradecem a FAPEMIG, FAPESB, CAPES, CNPq. Em especial ao MCTI-UFBA, pelo apoio financeiro por meio do Edital PROPCI/PROPG – PROPESQ/UFBA 004/2016.

Referências

Cemgil, T., Kurutmaz, B., Cezayirli, A., Bingol, E., and Sener, S. (2017). Interpolation and fraud detection on data collected by automatic meter reading. In *2017 5th International Istanbul Smart Grid and Cities Congress and Fair (ICSG)*, pages 51–55.

Chen, J., Li, W., Lau, A., Cao, J., and Wang, K. (2010). Automated load curve data cleansing in power systems. *IEEE Transactions on Smart Grid*, 1(2):213–221.

de Miranda Esquivel, R., de Senna, V., and Soares da Silva Gomes, G. (2013). Análise espectral singular: Comparação de previsões em séries temporais. *Revista ADM. MADE*, 16(2):87–101.

Fang, X., Misra, S., Xue, G., and Yang, D. (2012). Smart grid x2014; the new and improved power grid: A survey. *IEEE Communications Surveys Tutorials*, 14(4):944–980.

Genes, C., Esnaola, I., Perlaza, S. M., Ochoa, L. F., and Coca, D. (2016). Recovering Missing Data via Matrix Completion in Electricity Distribution Systems. In *17th IEEE International workshop on Signal Processing advances in Wireless Communications*, Edinburgh, United Kingdom.

Hassani, H. (2007). Singular spectrum analysis: methodology and comparison. *Journal of Data Science*. p.239-257.

Hussain, M. Z., Irshad, M., Sarfraz, M., and Zafar, N. (2015). Interpolation of discrete time signals using cubic spline function. In *Information Visualisation (iV), 2015 19th International Conference on*, pages 454–459. IEEE.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. Wiley Professional Computing. John Wiley & Sons Chichester.

Khalifa, T., Naik, K., and Nayak, A. (2011). A survey of communication protocols for automatic meter reading applications. *IEEE Communications Surveys Tutorials*, 13(2):168–182.

Lecomte, G., Hipolito, V., Batista, B. G., Kuehne, B. T., Filho, D. M. L., Martins, J. A. C., and Peixoto, M. L. M. (2017). Gap filling of missing streaming data in a network of intelligent surveillance cameras. In *WebMedia*.

Li, X., Liu, S., Li, Z., and Gong, J. (2017). Improved gap filling method based on singular spectrum analysis and its application in space environment. *Proc.SPIE*, 10605:10605 – 10605 – 13.

Lo, C. H. and Ansari, N. (2012). The progressive smart grid system from both power and communications aspects. *IEEE Communications Surveys Tutorials*, 14(3):799–821.

Qu, L., Li, L., Zhang, Y., and Hu, J. (2009). Ppca-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Transactions on Intelligent Transportation Systems*, 10(3):512–522.

Richardson, J., Reiner, P., and Wilamowski, B. M. (2015). Cubic spline as an alternative to methods of machine learning. In *Industrial Informatics (INDIN), 2015 IEEE 13th International Conference on*, pages 110–115. IEEE.

Siddiqui, O., Parmenter, K., Hurado, P., LLC., G. E. P., and Institute, E. P. R. (2008). *The Green Grid: Energy Savings and Carbon Emission Reductions Enabled by a Smart Grid*. Electric Power Research Institute.

Yaacoub, E. and Abu-Dayya, A. (2014). Automatic meter reading in the smart grid using contention based random access over the free cellular spectrum. *Computer Networks*, 59(Supplement C):171 – 183.

Zheng, J., Gao, D. W., and Lin, L. (2013). Smart meters in smart grid: An overview. In *2013 IEEE Green Technologies Conference (GreenTech)*, pages 57–64.