Estimação de consumo de energia elétrica para novos consumidores baseado no consumo da vizinhança.

Carolina L.S. Cipriano[1], Weldson A. Corrêa[1], Arthur G. S. Fernandes [1], Mayara G. Silva [1], Stelmo Netto [1], Eliana Monteiro [1], Marcia Izabel A. da Silva [1], Aristófanes C. Silva [1]

[1] Núcleo de Computação Aplicada – Universidade Federal do Maranhão (UFMA)
São Luís – MA – Brasil

{carol, weldson.amaral, arthurgsf, mayara, stelmo.netto, ari}@nca.ufma.br
{eliana.monteiro,marcia.silva}@cemar-ma.com.br

Abstract. *Electricity consumption forecasting is currently an area of major interest for most power companies. Despite being a trend in this area, forecasting can be very challenging and even impractical, especially for consumers with little or nonexistent consumption history. We propose in this work an alternative electricity consumption prediction model for consumers without consumption history. The proposed model is based on the x-means algorithm, which uses the k-nearest neighbours' consumptions to determine consumer groups, and stochastic gradient descent regressor to create a consumption estimate. The proposed method achieved promising results, in which we highlight the mean absolute percentage error of 38.76% and Theil Inequality Coefficient of 29.56%.*

Resumo. *A tarefa de predição do consumo de energia elétrica dos consumidores é atualmente uma tendência nas companhias de fornecimento de energia elétrica. Essa predição torna-se difícil ou impraticável em consumidores sem nenhum ou com curto histórico de consumo. Dessa forma, esse trabalho trata de uma alternativa à predição do consumo de energia, para consumidores sem histórico de consumo, baseado no consumo dos k-vizinhos mais próximos. A abordagem proposta utiliza o agrupador x-means na definição dos grupos de consumidores e o regressor gradiente descendente estocástico para estimar o consumo. O método proposto alcançou resultados promissores, obtendo erro médio absoluto percentual de 38,76% e Coeficiente de Desigualdade de Theil de 29,56%.*

1. Introdução

É uma tendência atual das companhias de fornecimento de energia elétrica investirem em inteligência artificial e aprendizagem de máquina, com o objetivo de prever o comportamento mensal do consumo de energia dos seus consumidores. A previsão é benéfica tanto para as companhias de energia, quanto para os consumidores. Esse benefício mútuo advém da redução dos gastos da companhia de energia durante a distribuição de energia e aumento do seu faturamento, por reduzir perdas financeiras ocasionadas por medições erradas ou furtos de energia, que então pode repassar aos seus consumidores uma taxa de faturamento menor do consumo de energia. Assim, essa tarefa de prever o comportamento do consumo de energia elétrica é relativamente simples quando existe o histórico de consumo dos seus consumidores, mas torna-se difícil ou impraticável em consumidores

sem nenhum histórico de consumo, isto é, aos casos de novas instalações consumidoras, ou que possuem curto histórico de consumo de energia. Portanto, com o objetivo de sanar o problema de predição do consumo de energia de consumidores sem histórico de consumo, este trabalho propõe um método de estimação de consumo de energia, que usa informações de consumo da vizinhança e do perfil de consumo desses vizinhos.

A predição do consumo de energia elétrica dos consumidores é uma etapa essencial na verificação de inconsistências na medição do consumo de energia. A verificação de inconsistências evita, tanto um faturamento incorreto para um consumidor, quanto pode indicar que o mesmo, por possuir um consumo anormal de energia, pode estar se utilizando de arranjos técnicos para diminuir seu consumo de energia e, portanto, não ser registrada adequadamente. Por essa razão, as companhias de fornecimento de energia elétrica têm investido em métodos de reconhecimento de padrões (RP) para prever o consumo de energia individual dos seus clientes e, assim, melhorar a etapa de verificação das inconsistência de medição do consumo de energia. Nesse sentindo, esse trabalho visa substituir a predição do consumo de energia de novos consumidores, que usualmente é realizada pelo registro do consumo de energia igual a zero.

Na etapa de predição do consumo de energia, é comum o agrupamento de consumidores com histórico similar de consumo, que em nosso trabalho, cada grupo criado é denominado de perfil de consumo. Esse agrupamento geralmente é realizado por métodos RP como o *k-means* e os grupos de consumidores encontrados são utilizados para a geração de modelos de predição de consumo de energia específicos para cada grupo [McLoughlin et al. 2015]. Portanto, conforme será detalhado nos resultados, a informação dos perfis de consumo criados aparecem como uma importante informação na predição do consumo de energia, quando consumidores não possuem histórico de consumo.

Para melhor compreensão do presente trabalho, este está dividido em: Seção 2 que apresenta os trabalhos relacionados, a Seção 3, onde é apresentado a metodologia proposta de predição do consumo individual de energia dos consumidores. Já a apresentação dos resultados é dada na Seção 4. Por último, uma conclusão sobre os resultados obtidos será dada na Seção 5.

2. Trabalhos Relacionados

A predição do consumo de energia diário em apartamentos da República da Coreia do Sul aparece como um problema no trabalho de Wahid e Kim (2016). Nesse trabalho, o método *Nearest Neighbors* (KNN) é utilizado como preditor do consumo de energia sobre os dados horários de consumo, provenientes de 520 apartamentos. Do histórico horário de consumo, foram extraídos quatro características, a média, variância, assimetria e curtose, que juntas, conseguem prever com acurácia de até 95,96% o consumo dos apartamentos.

Lora et al. (2002) compararam o desempenho de predição de séries temporais do preço da energia de dois modelos, um gerado por rede neural recorrente perceptron de múltiplas camadas e o outro criado por uma combinação do KNN e algoritmo genético (GA). Nesse trabalho, o GA é utilizado para ajustar os pesos para a distância euclidiana. Assim, o desempenho dos dois modelos foram comparados em um pequeno conjunto de dados dos preços de energia no período de janeiro a agosto de 2001, apresentando um erro médio absoluto de 0,3464, no período de março a maio, e de 0,428, no período de

junho a agosto.

Poloczek et al. (2014) e Kim et al. (2017) utilizam o KNN para prever os valores perdidos no processo de aquisição dos dados de sensores, em razão da inoperatividade dos mesmos. Ambos os trabalhos mostraram que o método KNN pode gerar resultados muito próximos dos valores reais, usando tanto a proximidade dos valores dos dados, quanto a informação espacial desses sensores. Assim, este resultado nos motivou a utilizar tanto a informação espacial das instalações consumidoras de energia, quanto ao KNN, pela sua simplicidade em gerar dados.

Os trabalhos acima citados realizam a predição de dados utilizando apenas o KNN e informações espaciais, em razão da sua excelente performance no processo de regressão dos dados. De maneira diferente, nosso trabalho utiliza o KNN, informações espaciais e de perfil de consumo de energia, para estimar o consumo, usando o regressor gradiente descendente estocástico.

3. Materiais e métodos

Esta seção apresenta os materiais e os métodos necessários para a execução do método proposto. Os mesmos são apresentados na sequência em que foram empregados, como mostrado na Figura 1. Primeiro, é descrita a aquisição da base de dados. Segundo, os dados passam por um pré-processamento. Terceiro, é feita uma análise de perfil, com objetivo de agrupar clientes com mesmo perfil de consumo. Na quarta etapa, realiza-se a estimação de consumo, utilizando duas abordagens que usam o KNN: (1) obtendo a mediana do consumo estimado para a vizinhança e (2) criando uma nova série, baseada no consumo da vizinhança e em seguida, estima-se o consumo para a série criada, usando o *Stochastic Gradient Descent* (SGD) [Ruder 2016]. Por fim, há uma etapa de validação do método proposto.

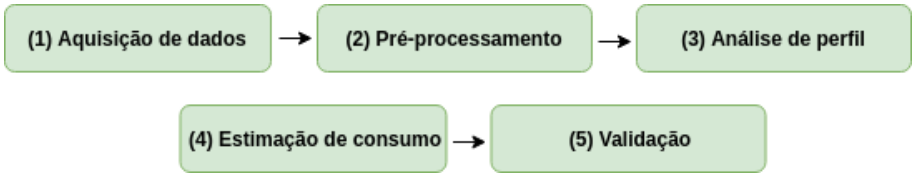

Figura 1. Etapas do método proposto.

3.1. Aquisição de Dados

A base de dados é composta por dados de consumo de energia de 91.262 consumidores residenciais ativos do município X do Brasil. Assim, esses dados de consumo obtidos em um intervalo de tempo formam a série temporal de consumo de cada consumidor. Dessa forma, a base foi formada a partir do consumo registrado, mensalmente, no período de janeiro de 2017 à abril de 2019, formando 27 registros de consumo por cliente. Os testes foram realizados apenas em clientes que possuem série histórica de 27 registros de consumo.

3.2. Pré-processamento

Antes da etapa de estimação de consumo, os dados são pré-processados. Verificou-se que a base contém séries onde não há o consumo registrado mensal ou onde todos os

registros mensais do cliente são iguais a zero. Nesses casos, optou-se por removê-los do conjunto de estimação. No primeiro caso, há uma quebra da sequência a ser estimada. E no segundo, o cliente que contém todos os registros mensais iguais a zero, não teria como ser validado pelo modelo proposto. Após o pré-processamento restaram 53.226 clientes.

3.3. Análise do perfil

O objetivo desta etapa é agrupar os clientes com padrão de consumo semelhante. Haja vista que existe variação na quantidade de dias de consumo faturado, utiliza-se o consumo médio diário, que é a razão do consumo registrado no mês pela quantidade de dias de consumo de cada cliente.

Em seguida, foi realizada uma análise dos dados para identificar possíveis anomalias no consumo registrado, para que estes valores não influenciem a definição do perfil. Para tanto, foi feita a identificação de *outliers*, por meio da análise do *box plot* [Williamson et al. 1989], resultando num total de 49.372 clientes removidos do *dataset*.

Após a remoção de *outliers* e clientes com histórico de consumo igual a 0, o conjunto de dados resultante foi utilizado na identificação da quantidade ótima de perfis, por meio da aplicação do algoritmo *x-means* [Pelleg et al. 2000]. O *x-means* é uma variante do *k-means*, que determina automaticamente a quantidade de grupos (k), em um intervalo informado, através do Bayesian Information Criterion (BIC).

3.4. Estimação de consumo

No estudo realizado, foram testadas duas abordagens de estimação do consumo para um cliente novo, isto é, que não possui histórico de consumo. A Abordagem I avaliou a utilização do consumo estimado para a vizinhança em duas etapas: (1) encontram-se os k vizinhos mais próximos do novo cliente, todos pertencentes a mesma região de leitura, que é definida pela companhia. (2) Após a normalização entre 0 e 1, de cada série da vizinhança, realiza-se a estimação de consumo de todos os k vizinhos, individualmente. Por fim, avaliou-se a utilização da média e da mediana dos consumos estimados dos vizinhos, para ser o consumo estimado do novo cliente. A Figura 2 ilustra a estratégia testada na Abordagem I.

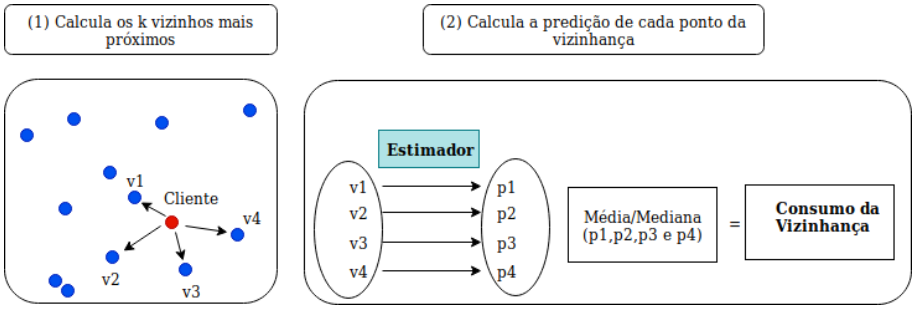

Figura 2. Abordagem I: estimação do consumo baseada na média ou mediana dos consumos dos k vizinhos.

Já na Abordagem II, avaliou-se a geração de uma nova série, também, em duas etapas: (1) encontram-se os k vizinhos mais próximos do novo cliente, todos pertencentes a mesma região de leitura, que é definida pela companhia e (2) gera-se uma série simulada para o novo cliente, baseada no histórico de consumo desses vizinhos. Cada mês da série

simulada é obtida pela média ou mediana, do referido mês, de seus vizinhos. Por fim, a série é normalizada entre 0 e 1, para ser utilizada no treinamento do regressor SGD, que estima o próximo mês de consumo do cliente. A Figura 3 ilustra a estratégia testada na Abordagem II.

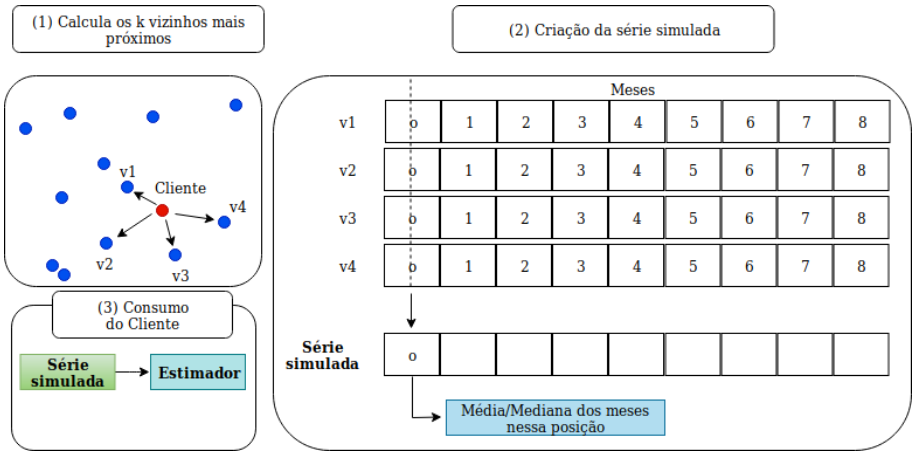

Figura 3. Abordagem II: estimação do consumo baseada na média ou mediana dos consumos dos k vizinhos.

3.5. Validação dos Resultados

O método proposto foi avaliado utilizando o erro médio absoluto percentual (MAPE) e o Coeficiente de Desigualdade de *Theil* (TIC). Estas métricas são comumente utilizadas para avaliar técnicas de estimação de valores. Quanto menores forem as métricas de erro, melhor.

$$MAPE = \frac{1}{n}\sum_{i}^{n}\left|\frac{valReal_i - valPred_i}{valReal_i}\right| \tag{1}$$

O MAPE é um erro relativo percentual, que expressa quanto o erro absoluto entre o valor real ($valReal_i$) e o valor predito ($valPred_i$) é superior ao valor real, para um ponto i da série temporal. Segundo Yorucu (2003), uma previsão com percentual de MAPE inferior a 10% é interpretada como altamente precisa; previsão superior a 10% e inferior a 20% é interpretada como boa; previsão superior a 20% e inferior a 50% é razoável; e previsão superior a 50% é considerada imprecisa.

O Coeficiente de Desigualdade de *Theil* (TIC) [Leuthold 1975] é a razão entre a diferença e a soma da predição e do *label*. O TIC retorna sempre resultado entre 0 e 1, onde zero significa previsão perfeita. Esta métrica é definida pela Equação 2, onde y_t é o valor real e $\hat{y}$ é o valor da predição, no ponto t, dos N pontos disponíveis.

$$TIC = \frac{\sqrt{\frac{1}{N}\sum_{t=1}^{N}(\hat{y} - y_t)}}{\sqrt{\frac{1}{N}\sum_{t=1}^{N} y_t^2} + \sqrt{\frac{1}{N}\sum_{t=1}^{N} \hat{y}_t^2}} \tag{2}$$

4. Resultados

Os experimentos foram realizados no conjunto de 53.226 clientes, que permaneceram após o pré-processamento, dos quais 10% do conjunto foram selecionados aleatoriamente

para a predição do último mês disponível. Logo, apenas as séries com 27 meses permaneceram, sendo utilizados 26 meses para treino e 1 mês para teste. Utilizou-se a distância euclidiana como métrica de proximidade dos pontos e foram definidos, experimentalmente, dez vizinhos mais próximos do novo cliente.

As duas abordagens (Seção 3.4) foram avaliadas em função das métricas MAPE e TIC. Em ambas abordagens, as métricas foram analisadas, também, após a exclusão dos consumidores que possuem anomalias de consumo. Essa análise foi denominada de corte do MAPE. Assim, a média das métricas de avaliação foram recalculadas para todos os consumidores que tiveram MAPE inferior a 125%. Então, uma pequena parcela de consumidores fica de fora da média das métricas

A Tabela 1 apresenta os resultados obtidos, sem considerar a seleção de clientes por perfil de consumo. Observa-se um MAPE superior a 50%, considerado impreciso [Yorucu 2003], em ambas as abordagens. Após o corte, obteve-se MAPE razoável, de 38%, em ambas abordagens, mantendo cerca de 90% (coluna Porc. da Tabela 1) dos consumidores. Analisando o TIC, percebe-se uma pequena melhora e um resultado percentual que pode ser considerado bom, por está próximo a zero. A abordagem II produziu o menor TIC, igual a 29,56%.

Tabela 1. Resultado da estimação de consumo sem o perfil.

	Abordagem I			Abordagem II		
Corte	Porc. (%)	MAPE (%)	TIC (%)	Porc. (%)	MAPE (%)	TIC (%)
-	100,00	165,70	33,39	100,00	171,05	32,97
$< 125\%$	90,39	38,97	30,21	89,97	38,76	29,56

A Tabela 2 apresenta os resultados obtidos considerando o perfil de consumo dos vizinhos. Assim, como na avaliação anterior, o MAPE obtido foi superior a 50% em ambas as abordagens. Já após o corte, obteve-se MAPE razoável, de cerca de 40% em ambas abordagens, preservando cerca de 90% dos consumidores. Analisando o TIC, assim como no teste anterior, houve uma pequena melhora e um resultado percentual que pode ser considerado bom. A abordagem II produziu o menor TIC, igual a 31,65%.

Tabela 2. Resultado da estimação de consumo baseada no perfil.

	Abordagem I			Abordagem II		
Corte	Porc. (%)	MAPE (%)	TIC (%)	Porc. (%)	MAPE (%)	TIC (%)
-	100,00	163,30	35,07	100,00	169,06	34,48
$< 125\%$	90,97	41,00	32,22	90,15	40,13	31,65

5. Conclusão

No presente trabalho, foi exposto um método de predição do consumo de energia elétrica para consumidores sem histórico de consumo. O método fez uso de técnicas de aprendizagem de máquina, como os k vizinhos mais próximos, para geração de uma série simulada, e do SGD, que é utilizado na predição do consumo de energia dos consumidores que tiveram suas instalações recentemente energizadas.

Foram realizados testes utilizando o consumo médio e a mediana dos k vizinhos em duas abordagens: (1) criando uma séria simulada e (2) utilizando o consumo estimado. Das abordagens avaliadas, a abordagem II sem perfil, após o corte do MAPE, se mostrou superior a abordagem I. A utilização do perfil não mostrou melhora significativa.

Pelo exposto, conclui-se que o método de estimação de consumo de energia elétrica, baseado na vizinhança, é um método promissor para novos consumidores, ainda sem histórico de consumo. Apesar de promissor, o método pode ser melhorado, utilizando outras técnicas de agrupamento de clientes, como o *mean shift* e agrupamento hierárquico, outros regressores para estimar o consumo, como o *Gradient Boosting regression* e *Random Forest*; e técnicas de otimização de parâmetros, como o *Particle Swarm Optimization* e o *Randomized Search*.

Agradecimentos

Os autores agradecem a CEMAR (contrato Nº 604/2018) e CELPA (contrato Nº 695/2018) pelo suporte financeiro disponibilizado através do Programa de Pesquisa e Desenvolvimento (PD) da Agência Nacional de Energia Elétrica (ANEEL).

Referências

Kim, M., Park, S., Lee, J., Joo, Y., and Choi, J. K. (2017). Learning-based adaptive imputation methodwith knn algorithm for missing power data. *Energies*, 10(10).

Leuthold, R. M. (1975). On the use of theil's inequality coefficients. *American Journal of Agricultural Economics*, 57(2):344–346.

Lora, A. T., Santos, J. R., Santos, J. R., Ramos, J. L. M., and Exposito, A. G. (2002). Electricity market price forecasting: Neural networks versus weighted-distance k nearest neighbours. In Hameurlain, A., Cicchetti, R., and Traunmüller, R., editors, *Database and Expert Systems Applications*, pages 321–330, Berlin, Heidelberg. Springer Berlin Heidelberg.

McLoughlin, F., Duffy, A., and Conlon, M. (2015). A clustering approach to domestic electricity load profile characterisation using smart metering data. *Applied energy*, 141:190–199.

Pelleg, D., Moore, A. W., et al. (2000). X-means: extending k-means with efficient estimation of the number of clusters. In *Icml*, volume 1, pages 727–734.

Poloczek, J., Treiber, N. A., and Kramer, O. (2014). Knn regression as geo-imputation method for spatio-temporal wind data. In de la Puerta, J. G., Ferreira, I. G., Bringas, P. G., Klett, F., Abraham, A., de Carvalho, A. C., Herrero, Á., Baruque, B., Quintián, H., and Corchado, E., editors, *International Joint Conference SOCO'14-CISIS'14-ICEUTE'14*, pages 185–193, Cham. Springer International Publishing.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Wahid, F. and Kim, D. (2016). A prediction approach for demand analysis of energy consumption using k-nearest neighbor in residential buildings. *International Journal of Smart Home*, 10:97–108.

Williamson, D. F., Parker, R. A., and Kendrick, J. S. (1989). The box plot: a simple visual method to interpret data. *Annals of internal medicine*, 110(11):916–921.

Yorucu, V. (2003). The analysis of forecasting performance by using time series data for two mediterranean islands. *Review of Social, Economic & Business Studies*, 2:175–196.