https://eventos.utfpr.edu.br//sicite/sicite2020

Análise Estatística da Demanda de Energia da UTFPR-MD Usando Regressão Linear

Statistical Analysis of UTFPR-MD Energy Demand Using Linear Regression

Paulo Victor Zuffo Oyama
pvzoya@hotmail.com
Universidade Tecnológica Federal do Paraná, Medianeira, Paraná, Brasil

Diogo Marujo
diogomarujo@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Medianeira, Paraná, Brasil

Alex Lemes Guedes
alexguedes@utfpr.edu.br
Universidade Tecnológica Federal do Paraná, Medianeira, Paraná, Brasil

RESUMO

A previsão da demanda de energia elétrica desempenha um papel fundamental na gestão de diversas organizações. Desta maneira, análises de previsão quantitativa e qualitativa da demanda de energia elétrica são de grande valor. O objetivo do presente trabalho é elaborar um algoritmo na linguagem de programação Python para realizar uma análise estatística de sensibilidade da demanda do Campus de Medianeira da UTFPR e determinação de fatores influenciadores no consumo de energia elétrica. Para isto, foi utilizada a técnica de regressão linear múltipla aplicada a dados históricos de demanda, dados meteorológicos e calendário letivo. Também foram utilizadas algumas técnicas de processamento de dados, como normalização de dados, remoção de *outliers.* O modelo encontrado é testado considerando a divisão treino-teste, de acordo com métricas fornecidas pelo algoritmo de treino do modelo, e com *Mean Absolute Deviation* e *Root Mean Squared Error*. O modelo obteve resultados bastante satisfatórios e indicam que a Ocupação Proporcional e Temperatura Máxima do Ar mostraram ter mais impacto no consumo de energia elétrica.

PALAVRAS-CHAVE: Análise de regressão. Eficiência energética. Python, analise estatística, previsão de demanda.

ABSTRACT

Energy demand forecasting plays a key role in the management of several organizations. So, quantitative and qualitative forecasting analyzes of electricity demand are of great value. The present work's objective is to elaborate an algorithm in the Python programming language to carry out a statistical sensitivity analysis of the demand of the UTFPR Medianeira Campus and determine influencing factors in electricity consumption, based on multiple linear regression, historical demand data, meteorological data, and school calendar. Some data processing techniques were also used, such as data normalization, removal of outliers. The model is tested considering the training-test division, according to metrics provided by the model's training algorithm, and with Mean Absolute Deviation and Root Mean Squared Error. The results indicate that the proportional occupation and maximum air temperature showed a more significant impact on electricity consumption.

KEYWORDS: Regresion Analysis. Energy Efficiency. Python, Statistical Analysis, demand forecast.

Recebido: 19 ago. 2020.
Aprovado: 01 out. 2020.

INTRODUÇÃO

A energia elétrica é imprescindível para inúmeras atividades corriqueiras, incluindo atividades de ensino, como é o caso da Universidade Tecnológica Federal do Paraná – Campus de Medianeira. O uso consciente da energia elétrica, bem como a gestão estratégica devem estar inclusos nos planejamentos de redução de custos destas instituições.

Segundo Pellegrini (2000), as previsões de demanda desempenham um papel chave em diversas áreas na gestão de organizações, pois com a correta gestão do consumo de energia elétrica, os setores administrativos podem dispender mais recursos em outros setores do funcionamento do campus. Uma das técnicas que é amplamente utilizada é a previsão de séries temporais, embasada em diferentes algoritmos de inteligência artificial, incluindo algoritmos robustos como Redes Neurais, como em Muralitharan et. al. (2017) e Gibson e Kraft (1993).

Este trabalho tem como objetivo a obtenção de um modelo de regressão linear múltipla para a realização de uma análise estatística de sensibilidade dos fatores que influenciam na demanda elétrica. O modelo será obtido com base em dados históricos de variáveis meteorológicas obtidas da estação meteorológica do campus Medianeira, e com base nas informações dos horários de aula ao longo do ano de 2019. O ajuste do modelo é feito pelo método OLS (*Ordinary Least Squares*), que é um algoritmo de *machine learning*, implementado na linguagem de programação Python.

Vários trabalhos atingiram conclusões relevantes para as análises que serão feitas a seguir, entre eles, Engel et al. (2006), que conduziram uma pesquisa que incentivava as pessoas a mudarem seu horário de consumo de energia, e afirmam que o maior consumo de energia em um ambiente está ligado à presença de pessoas naquele local, desta maneira, pode-se supor que a presença de alunos no campus da universidade tenha influência significativa na demanda.

Becker (2014) realizou um estudo no mesmo campus do qual trata o presente trabalho e concluiu que se pode dar ao clima a responsabilidade pelo contraste de consumo de energia entre verão e inverno, e desta maneira, a temperatura pode ser um dos fatores influenciadores no consumo de energia elétrica. Também Oliveira (2006), no mesmo sentido, realizou um monitoramento detalhado do consumo de energia dos prédios da UnB em Brasília. Os resultados apresentados indicam que a curva de carga dos prédios poderia ser uniformizada, bem como poderia ser realizado um estudo de demanda ótima, ou seja, além da previsão da demanda, existem outras medidas que podem ser tomadas para tornar o consumo de energia elétrica mais eficiente.

Este artigo está organizado da seguinte forma: a próxima seção aborda o material e os métodos. Em seguida são apresentados os resultados e discussões. Por fim, são apresentadas as conclusões e as referências.

MATERIAL E MÉTODOS

Em todas etapas deste estudo, as manipulações, pré-processamento e ajuste, a programação será feita na linguagem *Python* (v. 3.7.4), pelo compilador *Spyder* (v. 4.1.3). Também serão utilizadas as bibliotecas: *Pandas* (v. 1.0.5), *Numpy* (v.

1.19.0), *Statsmodels* (v. 0.11.1), *Pyod* (v. 0.8.1) (Zhao e Nasrullah (2019)) e Sklearn (v. 0.0).

As informações utilizadas neste trabalho como entrada são: Demanda de energia, Dados meteorológicos e Ocupação proporcional (quantidade de alunos no campus).

A demanda é uma informação fornecida pela concessionária, em seu *website*, que pode ser exportada para uma planilha Excel. Pode-se escolher o período e o tempo entre cada medição. Neste caso, foram utilizadas as medições durante o ano de 2019, a cada 15 minutos.

Os dados meteorológicos são provenientes da estação meteorológica da UTFPR, campus Medianeira, que contemplam as seguintes variáveis: Precipitação, Radiação solar incidente, Temperatura média do ar, Umidade relativa, Velocidade do vento 10m, Direção do vento 10m, Temperatura máxima do ar, Temperatura mínima do ar, Umidade relativa máxima, Umidade relativa mínima, Velocidade máxima do vento 10m. Também pode ser escolhido o tempo entre cada medição.

A variável ocupação proporcional é uma variável que foi criada com o intuito de melhorar o modelo de regressão. É calculada com base na capacidade individual de alunos de cada sala, na capacidade total de alunos do campus, e no horário dos semestres letivos de 2019, de modo que os valores se aproximem de um coeficiente.

O fator Ocupação Proporcional pode ser calculado através da equação (1):

$$OcupProp = \frac{S1*C1+S2*C2+\cdots+Sn*Cn}{C1+C2+\cdots+Cn} \quad (1)$$

Na equação acima, $OcupProp$ é o fator ocupação proporcional para um dado horário de aula em um semestre letivo, Sn é cada uma das n-ésimas salas que estão ocupadas (aula) naquele horário, e Cn é a capacidade de alunos da Sala n.

Além da obtenção dos dados, algumas técnicas de processamento de dados podem ser utilizadas. Sua eficácia será avaliada conforme os resultados do modelo de regressão linear obtido. As técnicas utilizadas serão: Remoção de outliers e Normalização de dados.

Na estatística, *outliers* são valores atípicos, anômalos, popularmente chamados de fora da curva por se distanciarem excessivamente dos demais, baseado em alguma medida (AGGARWAL; YU, 2001). Quando um conjunto de dados possui *Outliers*, isso pode indicar que a amostra é enviesada e inconsistente, o que pode consequentemente implicar nas análises realizadas, tornando-as menos confiáveis do que o esperado.

O algoritmo usado para a busca de *outliers* é o *K–Nearest Neighbour (KNN),* que é um algoritmo simples que, segundo Amaral (2016), busca, em tempo de execução, quais novas instâncias são mais parecidas com os dados históricos, e, neste caso, quais não são parecidas.

A normalização dos dados, por outro lado, se trata de colocar todos os dados na mesma ordem de grandeza, para haver menos probabilidade de o modelo ser enviesado, uma vez que não há precisão ao utilizar-se de um *dataset* disforme. Também é eficiente em reduzir substancialmente a sensibilidade dos modelos de regressão.

Para este trabalho foi utilizado o módulo "*preprocessing*" da biblioteca "*Sklearn*". A função "MinMaxScaler()" se trata de mudar a escala dos dados através da equação (2):

$$X_{novo} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

onde X_{novo} são os dados normalizados, X são os dados originais, X_{min} e X_{max} são respectivamente o menor e o maior valor dentro do vetor X.

Quanto ao ajuste e treinamento do modelo de regressão linear em si, será realizado pelo método "*Ordinary Least Squares*".

Na Figura 1 pode ser observado o Resumo do modelo de regressão linear criado com a variável dependente Demanda e com a variável independente OcupProp, retornado pela função "*statisticsmodelos.ols()"*:

Figura 1 – Exemplo do resumo do modelo de regressão linear

Results: Ordinary least squares

Model:	OLS	Adj. R-squared:	0.558
Dependent Variable:	Demanda	AIC:	3678.4014
Date:	2020-07-12 15:27	BIC:	3686.2012
No. Observations:	365	Log-Likelihood:	-1837.2
Df Model:	1	F-statistic:	460.6
Df Residuals:	363	Prob (F-statistic):	1.48e-66
R-squared:	0.559	Scale:	1386.4

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
Intercept	58.1205	2.6995	21.5299	0.0000	52.8118	63.4291
OcupProp	485.1071	22.6042	21.4609	0.0000	440.6555	529.5587

Omnibus:	41.696	Durbin-Watson:	0.657
Prob(Omnibus):	0.000	Jarque-Bera (JB):	54.617
Skew:	0.826	Prob(JB):	0.000
Kurtosis:	3.928	Condition No.:	12

Fonte: Autoria própria (2020).

Este resumo contém inúmeras métricas importantes para a avaliação do modelo de regressão linear, como F-Statistic, Omnibus, Skew e Kurtosis, Durbin-Watson, Jarque-Bera, Condition No e Adj. R-Squared. Além dos testes presentes no resumo, também será utilizado o teste de Breusch-Pagan.

F-Statistic testa a relevância do modelo, Skew e Kurtosis, como em Guidolin e Timmermann (2008), testam normalidade e simetria dos dados, Omnibus e Jarque-Bera, discutidos por Doornik e Hansen (2008) e Thadewald e Büning (2007), são testes de normalidade da amostra. Já Durbin-Watson e Breusch-Pagan, apresentados em Halunga et. al. (2011) e Rutledge e Barros (2002), são testes para heterocedasticidade.

Além das métricas apresentadas no resumo do modelo, a variância e o desvio padrão da previsão em relação à amostra também são relevantes, pois permitem validar o modelo, ou seja, avaliar se o modelo possui uma precisão adequada. Também conhecidos como MAD e RMSE (*Mean Absolute Deviation* e *Root Mean Squared Error*), respectivamente, podem ser calculados através das equações (3) e (4).

$$MAD = \frac{\sum_{i=1}^{n} |y\_true_i - y\_pred_i|}{n} \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y\_true_i - y\_pred_i)^2}{n}} \quad (4)$$

Nas equações (3) e (4), *y_true* são os valores reais e *y_pred* são os valores resultantes da previsão, da mesma variável, através do modelo, enquanto *n* é o número de instâncias do *dataset*.

A validação será realizada através da técnica de *Train-Test-Split*, ou seja, com 60% dos dados sendo utilizados para o ajuste e 40% para o teste do modelo.

RESULTADOS E DISCUSSÃO

Com base nas relações de dados, e nas funções de pré-processamento construídas, foram obtidos diversos modelos de regressão linear, dentre os quais um se destacou. As variáveis foram escolhidas com base no teste de probabilidade contido no resumo do modelo e nos coeficientes de correlação entre as variáveis individualmente.

MODELO DE REGRESSÃO LINEAR

O modelo foi ajustado com as seguintes variáveis independentes: radiação incidente, velocidade do vento, temperatura máxima do ar, umidade relativa máxima e ocupação proporcional, como pode ser observado na figura 2. Os dados foram normalizados e foi realizada a remoção dos *Outliers.*

Figura 2 - Modelo: Resumo

Results: Ordinary least squares

Model:	OLS	Adj. R-squared (uncentered):	0.821
Dependent Variable:	Demanda	AIC:	-36033.3700
Date:	2020-09-03 01:59	BIC:	-35994.1298
No. Observations:	18921	Log-Likelihood:	18022.
Df Model:	5	F-statistic:	1.733e+04
Df Residuals:	18916	Prob (F-statistic):	0.00
R-squared (uncentered):	0.821	Scale:	0.0087164

	Coef.	Std.Err.	t	P>|t|	[0.025	0.975]
RadInc	0.0519	0.0037	13.8761	0.0000	0.0446	0.0593
VelocVen	-0.0506	0.0058	-8.7398	0.0000	-0.0619	-0.0392
TempMaxAr	0.2034	0.0033	61.2840	0.0000	0.1969	0.2099
UmRelMax	0.0228	0.0021	10.8320	0.0000	0.0186	0.0269
OcupProp	0.3215	0.0025	129.5907	0.0000	0.3166	0.3263

Omnibus:	5841.977	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21026.034
Skew:	1.533	Prob(JB):	0.000
Kurtosis:	7.156	Condition No.:	8

Fonte: Autoria própria (2020).

Como pode ser observado na figura 2, o modelo atingiu o coeficiente ajustado R-*Squared* de 0.821, o que significa que 82,1% da flutuação da Demanda pode ser explicada pelo modelo obtido, e este valor pode ser considerado um coeficiente mediano.

Pelo valor de "Prob(F-*statistic*)", pode-se inferir que o modelo é estatisticamente relevante, e os valores de *Omnibus* e Prob(*Omnibus*) indicam que os resíduos não são normalmente distribuídos, pois espera-se respectivamente os valores 0 e 1 quando isso acontece.

Os valores de Jarque-Bera/Prob(JB), indicam que há heterocedasticidade nos dados, e em grande quantidade. *Condition Number* também apresenta um valor um pouco elevado, indicando sensibilidade no modelo e provavelmente multicolinearidade.

A presença de heterocedasticidade na amostra também pode ser confirmada pelo teste de Breusch-Pagan, cujos resultados são apresentados na figura 3.

Figura 3 - Modelo de regressão: Teste de Breusch-Pagan

```
Lagrange multiplier statistic, p-value,f-value, f p-value
(4485.753572710967, 0.0, 1175.6295953630788, 0.0)
```

Fonte: Autoria própria (2020).

Na figura 3 pode ser observado o resultado do teste de Breusch-Pagan, que segundo Breusch e Pagan (1979), confirma a existência de heterocedasticidade no conjunto de dados de treino, devido ao baixo valor de *p-value* e alto valor de *t-value.*

A equação do modelo de regressão linear está apresentada na equação (5):

$$Demanda = 0.047 * RadInc - 0.071 * VelocVen + 0.202 * TempMaxAr + 0.083 * UmRelMax + 0.321 * OcupProp \quad (5)$$

Na equação (5) podem ser observadas as variáveis independentes escolhidas para o modelo e seus respectivos coeficientes: Radiação incidente (RadInc), Velocidade do vento (VelocVen), Temperatura máxima do ar (TempMaxAr), Umidade relativa Maxima (UmRelMax) e Ocupação proporcional (OcupProp).

Qualitativamente, com base no teste estatístico apresentado no resumo do modelo (valor t), e pelos coeficientes apresentados na equação resultante, pode-se inferir que os fatores que mais influenciam no consumo de energia elétrica são Temperatura máxima do ar e ocupação proporcional, que são justamente os fatores apontados por Becker (2014) e Engel et al. (2018).

Os demais valores presentes no modelo possuem menor relevância, porém são fatores que interferem na sensação térmica, e que possivelmente influenciam na Demanda devido ao grande número de aparelhos condicionadores de ar nas instalações.

CONCLUSÃO

O modelo obtido e apresentado obteve resultados bastante satisfatórios, haja vista que a abordagem matemática é bastante simples e o problema é de grande complexidade. Embora haja heterocedasticidade nos dados, o coeficiente de Adj. R-Squared informa que 82,1% da variação da demanda pode ser explicada pela regressão encontrada, o que é suficiente para realizar inferências acerca do consumo de energia elétrica e adotar medidas que possam melhorar o perfil de carga do campus de Medianeira da UTFPR. Quanto à temperatura, pouco há a ser feito, contudo, a maior constância da quantidade de alunos ao longo do dia e da semana pode ser decisivo para que a curva de carga seja mais uniforme.

Notou-se que a grande quantidade de dados resultou em elevados níveis de ruído, prejudicando os modelos, como pode ser observado nos resultados dos testes de Breusch-Pagan, Omnibus, Jarque-Bera, Durbin-Watson. Esse ruído foi atenuado através da normalização dos dados e remoção de outliers, o que é significativo, mas não completamente suficiente considerando análises mais completas.

Algumas funções de pré-processamento não tiveram exatamente o impacto esperado. A remoção dos outliers não teve efeito em algumas ocasiões, embora tenha melhorado o resultado do modelo apresentado. A normalização dos dados foi essencial para reduzir a medida "Condition Number" e melhorar o erro, embora tenha tido pouca influência nas outras métricas do modelo.

Menos dados, com menos ruído e menos heterocedasticidade podem render um modelo melhor do que a grande profusão de dados que há no dataset obtido. Portanto, há fortes indícios de que grandes quantidades de dados não são adequadas para algoritmos simples como regressão linear. A utilização de algoritmos mais robustos como redes neurais artificiais, que lidam melhor com este tipo de ruído, pode ser uma alternativa para este problema.

AGRADECIMENTOS

Os autores agradecem à Fundação Araucária, pelo incentivo financeiro à pesquisa, à UTFPR-MD, por fornecer a estrutura e a possibilidade de realizar este trabalho.

REFERÊNCIAS

AGGARWAL, Charu C.; YU, Philip S.. **Outlier Detection for High Dimensional Data. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA**, 01., 2001, Santa Babara California Usa. Proceedings [...] . Santa Barbara: Alta Books, 2001. v. 1, p. 37-46.

AMARAL, Fernando. **Aprenda mineração de dados: teoria e prática_**. 1. Ed. Rio de Janeiro: Editora Alta Books. 2016. 225p. ISBN 978-85-7608-988-9

BECKER, Taise Vanessa. **Otimização da demanda e consumo de energia elétrica da Universidade Tecnológica Federal do Paraná - Campus Medianeira**. 2014. 67 f. TCC (Graduação) - Curso de Engenharia de Produção, Departamento Acadêmico de Engenharia de Produção, Universidade Tecnológica Federal do Paraná, Medianeira, 2014.

BREUSCH, Trevor S.; PAGAN, Adrian R. **A simple test for heteroscedasticity and random coefficient variation**. Econometrica: Journal of the Econometric Society, p. 1287-1294, 1979.

DOORNIK, Jurgen A.; HANSEN, Henrik. **An omnibus test for univariate and multivariate normality**. Oxford Bulletin of Economics and Statistics, v. 70, p. 927-939, 2008.

ENGEL, S.; VON APPEN, J.; DÖRRE , E.; NESTLE , D.; RINGELSTEIN, J.. **Results from the operation of a Social Energy Management System**, 2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), Sarajevo, 2018, pp. 1-6, doi: 10.1109/ISGTEurope.2018.8571594.

GIBSON, Gerald L.; KRAFT, Timothy T. Electric demand prediction using artificial neural network technology. ASHRAE Journal (American Society of Heating, Refrigerating and Air-Conditioning Engineers);(United States), v. 35, n. 3, 1993.

GUIDOLIN, Massimo; TIMMERMANN, Allan**. International asset allocation under regime switching, skew, and kurtosis preferences**. The Review of Financial Studies, v. 21, n. 2, p. 889-935, 2008.

HALUNGA, Andreea G.; ORME, Chris D.; YAMAGATA, Takashi. **A heteroskedasticity robust Breusch-Pagan test for contemporaneous correlation in dynamic panel data models.** 2011.

MURALITHARAN, K. et al. **Neural network based optimization approach for energy demand prediction in smart grid**. Neurocomputing. Coimbatore, p. 199-208. 24 ago. 2017.

OLIVEIRA, Lilian Silva de. **Gestão do consumo de energia elétrica no campus da UnB**. 2006. 238 f. Dissertação (Mestrado) - Curso de Engenharia Elétrica, Departamento de Engenharia Elétrica, Universidade D Brasília, Brasília, 2006.

PELLEGRINI, Fernando Rezende. **Metodologia para implementação de sistemas de previsão de demanda**. 2000. 146 f. Dissertação (Mestrado) - Curso de Engenharia de Produção, Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2000.

RUTLEDGE, D. N.; BARROS, A. S. **Durbin–Watson statistic as a morphological estimator of information content**. Analytica Chimica Acta, v. 454, n. 2, p. 277-295, 2002.

THADEWALD, Thorsten; BÜNING, Herbert. **Jarque–Bera test and its competitors for testing normality–a power comparison**. Journal of applied statistics, v. 34, n. 1, p. 87-105, 2007.

ZHAO, Y., NASRULLAH, Z., LI, Z., 2019. **PyOD: A Python Toolbox for Scalable Outlier** Detection. Journal of machine learning research (JMLR), 20(96), pp.1-7.