Regularização e detecção de periodicidade em séries temporais com *XGBoost*

Samuel Bueno Soltau [1,2] Luiz Claudio Lima Botti[1,3,4,5]

[1] CAGE/UPM. Programa de Pós-Graduação em Ciências e Aplicações Geoespaciais da Universidade Presbiteriana Mackenzie. [2] UNIFAL-MG. Universidade Federal de Alfenas.

[3] CRAAM, Centro de Radioastronomia e Astrofísica Mackenzie. Escola de Engenharia da Universidade Presbiteriana Mackenzie. [4] DAS/CEA/INPE/MCTIC, Instituto Nacional de Pesquisas Espaciais. [5] ROI/INPE, Rádio Observatório do Itapetinga

Introdução

Dados de radiofontes extragalácticas têm amostragem irregular devido a fatores como as condições meteorológicas, a manutenção de receptores, entre outras, que influenciam na aquisição de dados por estações terrestres. Tais dificuldades produzem séries temporais desigualmente espaçadas que impõem limitações ao uso de métodos de análise convencionais. Estudos com múltiplas frequências exploraram aspectos distintos de radiofontes compactas, em particular variações de densidade de fluxo para determinar periodicidades em curvas de luz (Botti, 1990; Gastaldi, 2017). Métodos comumente usados para determinar periodicidades em radiofontes incluem Transformada de Fourier, Periodograma Lomb-Scargle, Transformada Wavelet e *Cross Entropy*, entre outros (Cincotta & Nunez, 1995).

Avanços em Inteligência Artificial forneceram algoritmos de *machine learning*, como Redes Neurais, *Ensemble* e *Deep Learning*, que ao ingressarem em estudos astrofísicos, forneceram abordagens computacionais diferentes dos métodos consagrados, incluindo aplicações potenciais para a análises de radiofontes (Witten, Frank, Hall & Pal, 2016). Assim, apresenta-se o *XGBoost* (*eXtreme Gradient Boosting*), uma biblioteca de *software* para *machine learning* (Chen & Guestrin, 2016), como um recurso aplicável tanto na geração de uma série temporal regularmente espaçada, quanto na detecção de periodicidades.

XGBoost: eXtreme Gradient Boosting

O *XGBoost* é um conjunto de métodos de *machine learning* baseados em árvores de decisão, reunidos em uma biblioteca projetada e otimizada para para extrair o máximo de *performance* das arquiteturas computacionais disponíveis e criar um modelo mais geral (Chen & Guestrin, 2016). O mecanismo utilizado pelo *XGBoost* como modelo de escolha é chamado *ensemble* de árvores de decisão. Trata-se de um conjunto de árvores de classificação e regressão denominado *CART* (*Classification and Regression Trees*) (Breiman, 1996; Chen & Guestrin, 2016). Um exemplo simples de *CART* (ver Figs. 1 e 2) é o que classifica se uma pessoa gostará ou não de jogos de computador (*games*).

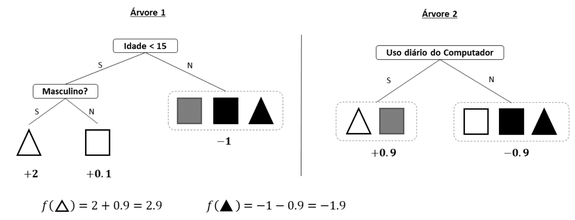

Figura 1: O modelo de ensemble: duas árvores para completeza do modelo e somatório dos scores.

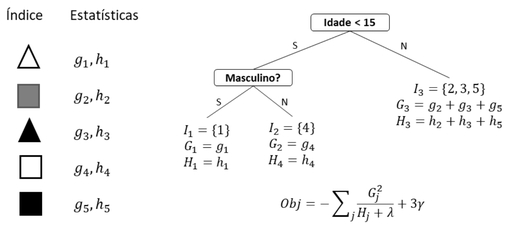

Figura 2: *Score* da estrutura

Materiais e Métodos

O quasar PKS 1921-293 (OV 236) é uma das mais compactas radiofontes conhecidas (Tornikoski, 1996) que emite em toda faixa de energia, apresenta grande variabilidade e pode ser classificada como um objeto *BL Lacertae* (Impey, Brand, Wolstencroft & Williams,1982) devido à sua polarização de cerca de 14% e à variabilidade de $2,2\mu m$ (137 THz, no infravermelho médio) e *redshift* $z = 0.3525$ (Wills & Wills, 1981).

Neste estudo utiliza-se dados do Observatório de Radioastronomia da Universidade de Michigan (*University of Michigan Radio Astronomy Observatory*, *UMRAO*) (Fig. 3) e aplica-se o *XGBoost* (Fig. 4) para efetuar a regularização da série temporal (Fig. 5) e analisar a evolução da curva de luz, nas frequências de $4,8$ GHz, entre 1980 e 2011, $8,0$ GHz de 1975 a 2011 e de $14,5$ GHz de 1976 a 2011. Os dados PKS 1921-293 constituem uma série temporal de amostragem irregular.

Resultados

A fase de análise, que inicialmente forneceu dados de treinamento para o algoritmo criado com a biblioteca *XGBoost*, identificou segmentos de dados (Fig. 6) que permitiram estimar a periodicidade da ocorrência de eventos entre 14, 36, 62 meses na série temporal de 1975 a 2011, em 4.8 GHz; 15, 31, 57 meses a 8.0 GHz e 15, 53, 76 anos a 14.5 GHz.

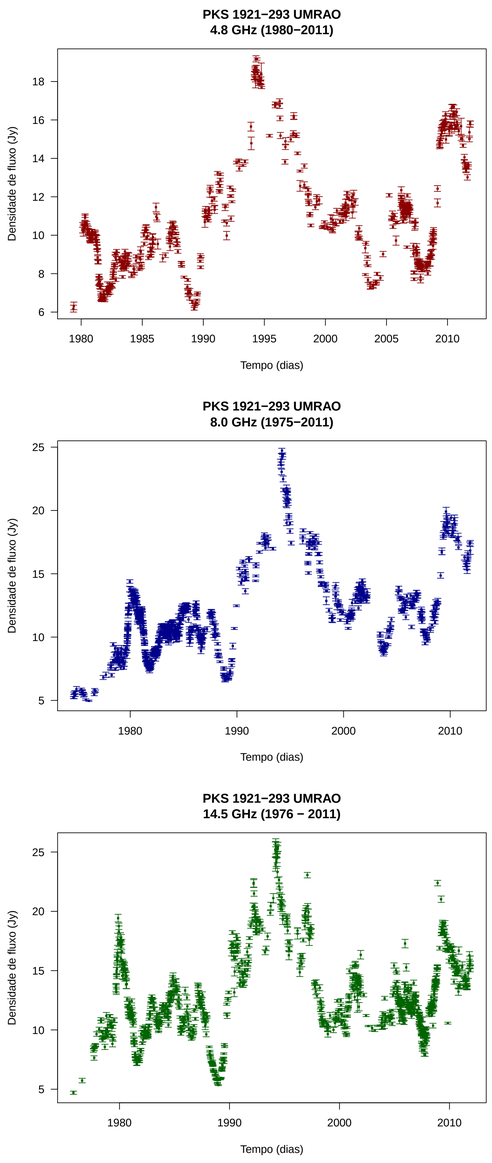

Figura 3: Curvas de luz PKS 1921-293. Faixas 4.8 GHz, 8.0 GHz e 14.5 GHz UMRAO. Dados sem processamento.

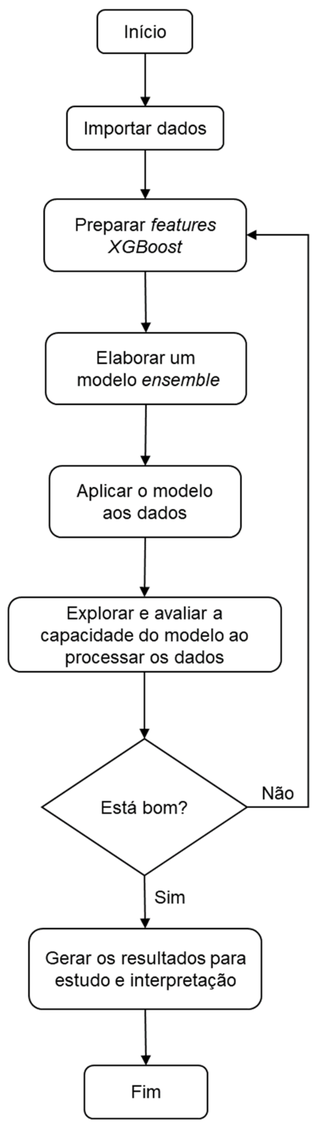

Figura 4: Fases e triagem computacional

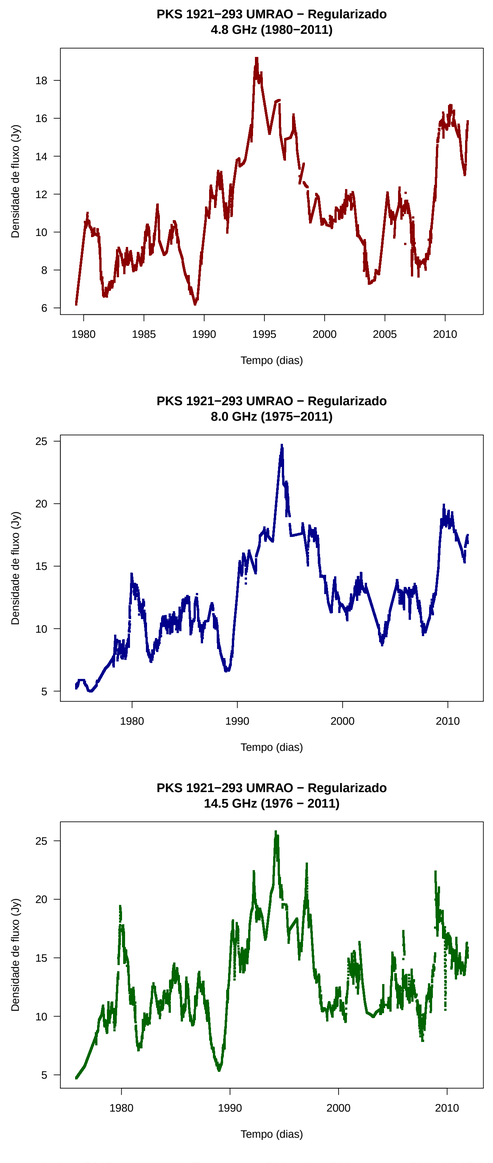

Figura 5: Curvas de luz PKS 1921-293. Faixas 4.8 GHz, 8.0 GHz e 14.5 GHz UMRAO, após o procedimento computacional para regularização da série temporal.

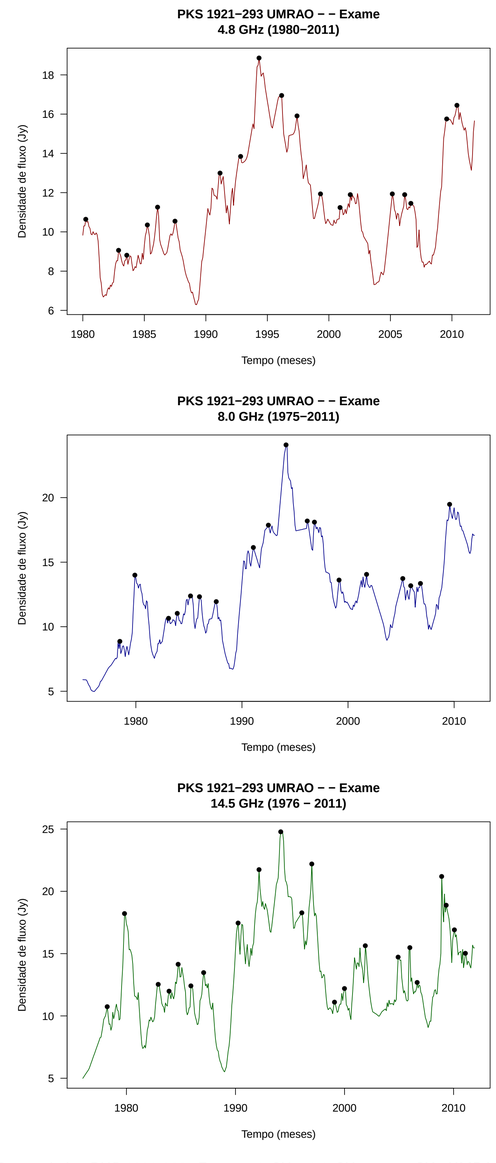

Figura 6: Curvas de luz PKS 1921-293. Faixas 4.8 GHz, 8.0 GHz e 14.5 GHz UMRAO. Detecção de *outburst*.

Conclusão

O *XGBoost* foi capaz de capturar o comportamento da curva de luz contornando distorções e suavizações indesejáveis que ocorrem tipicamente em métodos tradicionais de tratamento de séries temporais. A identificação de faixas de densidade de fluxo acentuadas, típicas de *outbursts*, foi utilizada para estimar a periodicidade desses eventos nos objetos estudados. A robustez do *XGBoost* proporciona a criação de métodos inovadores que associam poder computacional com a *expertise* do pesquisador, viabiliza o refino de técnicas de *data mining* aplicáveis inclusive à investigação de grandes volumes de dados astrofísicos, a exemplo do que fizeram (Tamayo et al., 2016), seguindo tendências de trabalhos desenvolvidos em outras áreas como, por exemplo, Chen & He (2015), Bethapudi & Desai (2018).

Agradecimentos

Ao meu orientador, Prof. Botti, pela paciência e orientação. Ao CAGE / UPM - Programa de Pós-Graduação em Ciências e Aplicações Geoespaciais da Universidade Presbiteriana Mackenzie, ao CRAAM - Centro de Radioastronomia e Astrofísica da Universidade Presbiteriana Mackenzie e à Universidade Presbiteriana Mackenzie pelo financiamento da pesquisa. Ao Observatório de Rádio Astronomia da Universidade de Michigan pelos dados sobre os quais esta pesquisa se desenvolveu. À Naíma Ferrão pela companhia, paciência, dedicação e pelo *design* das figuras.

Referências

Bethapudi, S.; Desai, S. 2018. Separation of pulsar signals from noise using supervised machine learning algorithms. Astronomy and Computing, v. 23, p. 15–26, apr.

Botti, L. C. L.1990, Spectrum variability study of radio sources in the 22 to 43 GHz range (Unpublished doctoral thesis). Instituto de Pesquisas Espaciais, S ao José dos Campos (Brazil).

Breiman, L. 1996. Bagging predictors. Machine Learning, Kluwer Academic Publishers, Boston, v. 24, p. 123140, 1996.

Chen, T., He, T. 2015. Higgs Boson Discovery with Boosted Trees. JMLR: Workshop and Conference Proceedings 42:69–80.

Chen, T., Guestrin, C. 2016. in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16

Cincotta, P. M., Mendez, M., & Nunez, J. A. 1995, Astrophysical Journal, Supplement, 449 , 231.

D'Onofrio, M. and Marziani, P. & Sulentic, J. W, 2012, Fifty Years of Quasars: From Early Observations and Ideas to Future Research, (New York: Springer).

Gastaldi, M. R. 2017, Uso do Periodograma de Lomb e da transformada Wavelet para detecção de periodicidades em radiofontes extragalácticas (Unpublished doctoral thesis). CAGE/UPM, S ao Paulo (Brazil)

Impey, C. D., Brand, P. W. J. L., Wolstencroft, R. D., & Williams, P. M. 1982, Monthly Notices of the Royal Astronomical Society, 200, 19.

Tamayo, D., et al. 2016. A machine learns to predict the stability of tightly packed planetary systems. The Astrophysical Journal Letters, v. 832, p. L22, dec.

Tornikoski, M., et al. 1996, Astronomy and Astrophysics, Supplement, 116, 157

University of Michigan Radio Astronomy Observatory 2018, UMRAO Database Interface. Retrieved from `https://dept.astro.lsa.umich.edu/datasets/umrao`.

Véron-Cetty, M.-P., & Véron, P. 2010, Astronomy and Astrophysics, 518 , A10.

Wills, D., & Wills, B. J. 1981, Nature, 289, 384.

Witten, I. H. and Frank, E. and Hall, M. A. and Pal, C. J. 2016, Data Mining: Practical Machine Learning Tools and Techniques, (Cambridge, MA: Elsevier Science).