



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

PROJETO LOMB-SCARGLE

Leonardo Sattler Cassará

Relatório apresentado à Prof.^a
Margarete O. Domingues como ati-
vidade do curso Análise Wavelet I.

Repositório deste projeto:
[github/leosattler/projeto-lomb-scargle](https://github.com/leosattler/projeto-lomb-scargle)

INPE
São José dos Campos
5 de outubro de 2020

RESUMO

Neste trabalho, dados de fluxo solar na faixa de 10.7 cm são manipulados de modo a simular amostragem não uniforme dos mesmos. O objetivo é investigar o efeito da amostragem não uniforme e introduzir o periodograma de Lomb-Scargle, conceito pertinente à disciplina Análise Wavelet I. Tal ferramenta é discutida e implementada a partir do pacote `astropy` (da linguagem `Python`) com a classe `LombScargle`. Os diferentes periodogramas gerados são analisados e a performance da ferramenta investigada à luz do conhecido ciclo de atividade solar, cujo período é de onze anos.

Palavras-chave: Fluxo solar. Análise de sinal. Periodograma de Lomb-Scargle. Método dos mínimos quadrados. Séries temporais.

LISTA DE FIGURAS

	<u>Pág.</u>
2.1 Efeitos da amostragem finita.	2
2.2 Efeitos do sampling rate.	3
2.3 Efeitos do sampling não uniforme.	4
3.1 Espectro de potência dos dados originais (média diária).	5
3.2 Análise das médias diárias com 4 intervalos.	6
3.3 Análise das médias diárias com 5 intervalos.	7
3.4 Análise das médias diárias com 6 intervalos.	7
3.5 Análise das médias diárias com 7 intervalos.	8
3.6 Análise das médias diárias com 8 intervalos.	8
3.7 Análise com exclusão aleatória até 1% dos dados.	9
3.8 Análise com exclusão aleatória até 0.5% dos dados.	10
3.9 Análise com exclusão aleatória até 0.1% dos dados.	10
3.10 Análise com exclusão aleatória até 0.05% dos dados.	11
3.11 Análise com exclusão aleatória até 0.04% dos dados.	11

SUMÁRIO

	<u>Pág.</u>
1 INTRODUÇÃO	1
2 FERRAMENTA LOMB-SCARGLE	1
2.1 Efeitos da amostragem	2
2.2 Periodograma de Lomb-Scargle	4
3 RESULTADOS E DISCUSSÃO	5
3.1 Cenário 1 - diferentes intervalos de observação	6
3.2 Cenário 2 - exclusão aleatória de dados	9
4 CONSIDERAÇÕES FINAIS	12
REFERÊNCIAS BIBLIOGRÁFICAS	13

1 INTRODUÇÃO

O fluxo solar na faixa de 10.7 cm (doravante chamado F10.7) é uma medida da intensidade da emissão do sol na faixa do rádio, mais precisamente em 10.7 cm (ou 2800 MHz). Este índice é um indicador da atividade magnética do Sol, fornecendo informações da atividade solar no ultravioleta e raio-X. Por isso, esse índice é muito relevante em ramos como astrofísica, meteorologia e geofísica. Com aplicações em modelagem climática, seu monitoramento é importante para a manutenção dos sistemas de comunicação via satélite ([HUANG et al., 2009](#)).

Uma das ferramentas mais usadas para trabalhar com séries temporais deste tipo é a análise espectral, que objetiva representar um sinal como a combinação linear de funções periódicas. Para dados obtidos com um *sampling rate* uniforme, i.e., sob a mesma taxa de registro durante toda a observação, o espectro de potência via FFT (do inglês, Fast Fourier Transform) é o método padrão utilizado. Porém, nem sempre o sinal disponível foi adquirido sob intervalo uniforme. Por exemplo, o registro da variação do brilho de estrelas via telescópios terrestres está sujeito a diversas interrupções, umas de natureza periódica (rotação e translação terrestre) e outras de natureza não-periódica (mal tempo, problemas do equipamento, etc.).

O espectro de potência não é apropriado para dados não uniformes, e uma nova ferramenta se faz necessária para esses casos. O periodograma de Lomb-Scargle ([LOMB, 1976; SCARGLE, 1982](#)) é um algoritmo para detectar e caracterizar a periodicidade de séries temporais com sampling rates não uniformes. Ele utiliza o método de mínimos quadrados para ajustar funções senoidais aos dados ([VanderPlas, 2017](#)).

O presente trabalho é um follow-up de [Cassara \(2020\)](#). Os dados F10.7 são manipulados com o fim de simular aquisição não uniforme. Experimentos são efetuados com a simulação de diferentes cenários de sampling rates não uniformes, com a geração do periodograma de Lomb-Scargle utilizando a biblioteca `astropy`. O presente manuscrito está assim organizado: na Seção 2 a metodologia empregada é introduzida; na Seção 3 os resultados são apresentados com uma breve discussão; na Seção 4 são oferecidas as considerações finais do autor.

2 FERRAMENTA LOMB-SCARGLE

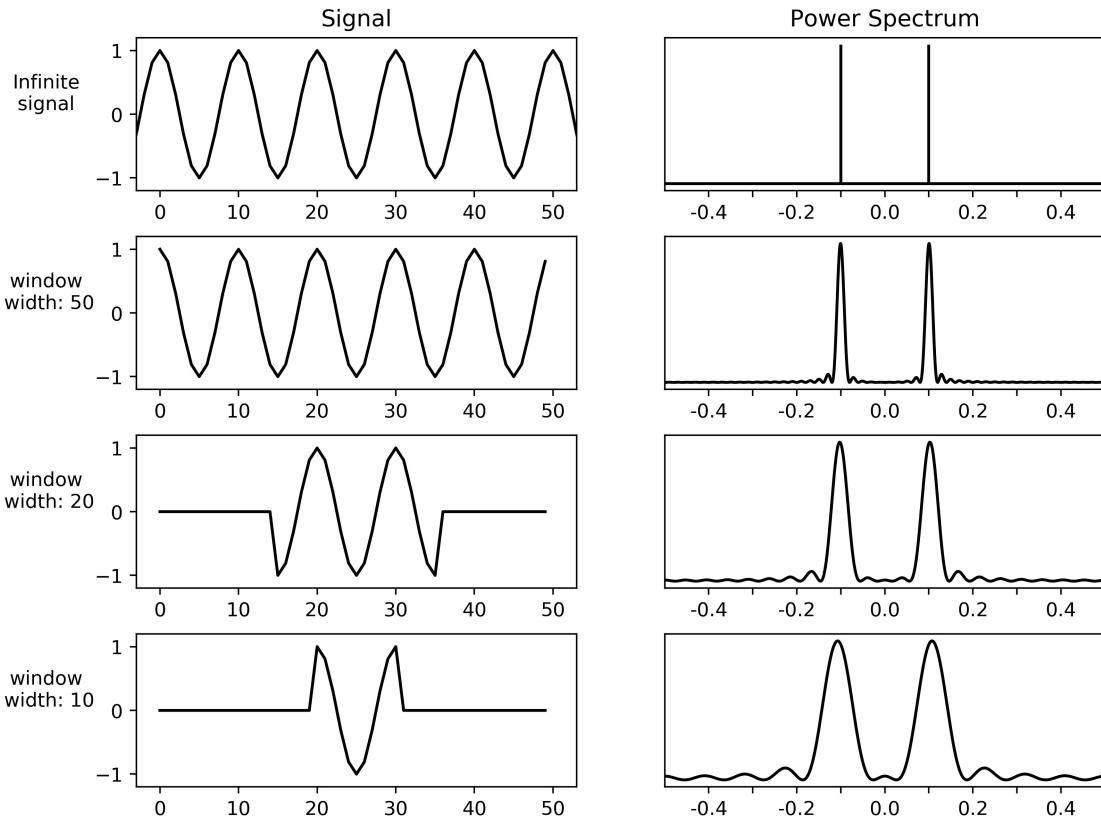
O periodograma de Lomb-Scargle é a ferramenta padrão para dados com sampling rates não uniformes. Dito isto, é muito importante a compreensão dos diferentes

artefatos presentes na análise espectral e a diferenciação de suas origens.

2.1 Efeitos da amostragem

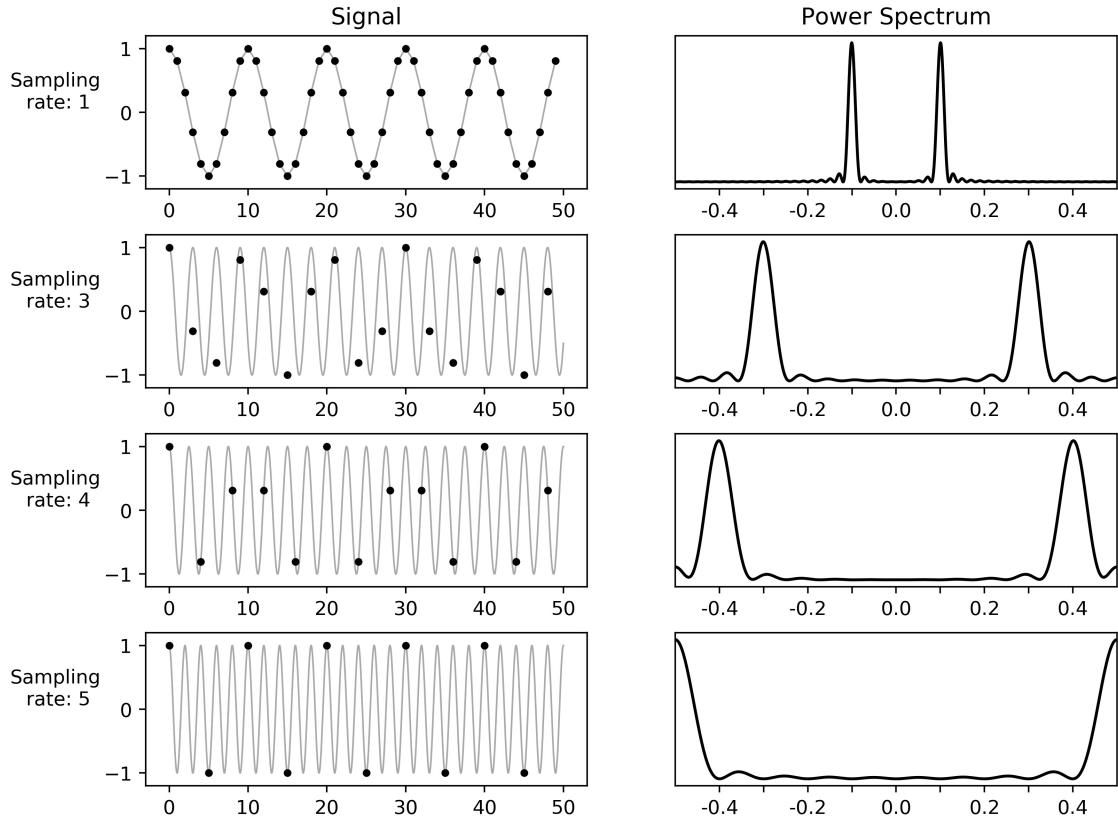
Os principais artefatos da análise espectral são o *aliasing* e o *spectral leakage*. Leakage espectral é o “vazamento” da energia devida a uma frequência existente no sinal para outras, por exemplo os lóbulos laterais presentes em muitos espectros. Aliasing é um tipo de leakage, que é o efeito do espectro apresentar assinaturas falsas de sinais (alias vem do inglês e significa pseudônimo). As Figuras 2.1 e 2.2 exemplificam tais fenômenos e ilustram suas causas.

Figura 2.1 - Efeitos da amostragem finita.



Efeitos do tamanho da janela de observação sobre o espectro de potência de um sinal. No topo à esquerda o sinal está representado por uma função analítica que se estende infinitamente, cuja transformada de Fourier (topo à direita) é a função delta sobre a frequência do sinal (no caso, 0,1). Abaixo estão sinais com janelas de observação diferentes e seus respectivos espectros. Quanto menor a janela de observação, maior o efeito dos lóbulos laterais sobre a função delta original, o chamado leakage espectral.

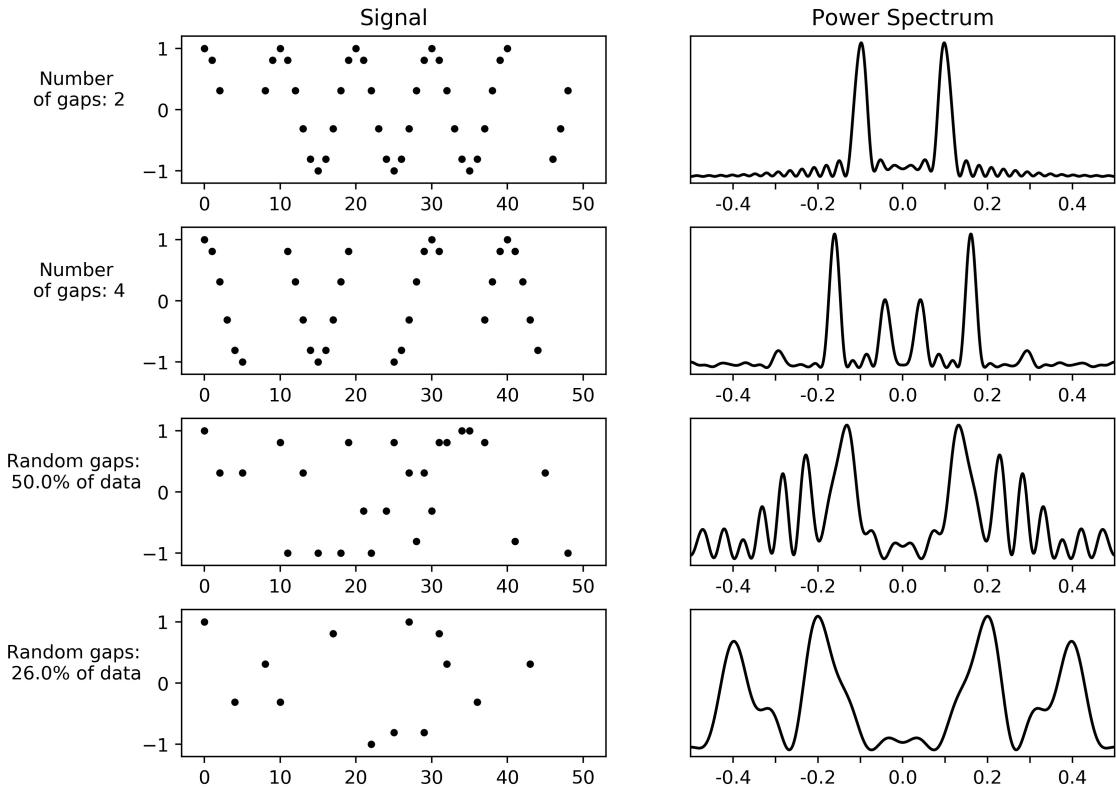
Figura 2.2 - Efeitos do sampling rate.



Efeitos do sampling rate sobre o espectro de potência de um sinal. No topo, um sinal com um sampling rate igual a um (que corresponde ao sinal com janela de tamanho 50 na Figura 2.1). Abaixo, o mesmo sinal sob diferentes sampling rates e seus respectivos espectros. A linha em cinza claro ilustra o sinal falsamente identificado, tanto pelo nosso cérebro quanto pela FFT, conforme indicado nos seus espectros de potência à direita. Fica evidente que para diferentes taxas, diferentes aliases do sinal original são gerados, de modo que a transformada inversa retornaria um sinal totalmente diferente do original.

Os lóbulos laterais (leakage spectral) são um artefato devido ao intervalo de observação ser finito. As falsas frequências (aliasing) é uma artefato que surge da natureza do sampling. Sabemos que nossos dados de fluxo solar F10.7 são finitos no tempo e apresentam um sampling rate uniforme e satisfatório para gerar bons espectros via FFT, conforme explicitado em Cassara (2020). Mas qual seria o efeito de sampling não uniforme sobre o sinal das figuras anteriores? A Figura 3 ilustra dois cenários de ausência de dados e seus respectivos espectros. Fica evidente a incrível irregularidade do espectro de potência resultante da FFT devido a um espaçamento desigual da amostragem do sinal.

Figura 2.3 - Efeitos do sampling não uniforme.



Efeitos do sampling não uniforme sobre o espectro de potência de um sinal. No topo, o sinal do topo da Figura 2.1 se apresenta com dois gaps (intervalos) aleatórios sem dados. Abaixo deste, o mesmo sinal com quatro gaps aleatoriamente posicionados. Os dois últimos sinais representam um cenário com remoção aleatória de dados, um permanecendo com 50% dos dados e o outro com 26% apenas. Somente o espectro de potência do primeiro sinal (topo à direita) possui um pico consistente (posicionado na frequência esperada de 0.1).

2.2 Periodograma de Lomb-Scargle

O periodograma de Lomb-Scargle é a principal ferramenta para análise de séries temporais com amostragem desigual. Ele pertence a um grupo de ferramentas de análise espectral que explora o método de mínimos quadrados, estimando frequências do sinal a partir de testes sobre frequências pré-determinadas com o fim de ajustar funções senoidais aos dados. O periodograma de Lomb-Scargle é dos métodos de análise espectral por mínimos quadrados desenvolvido por [Lomb \(1976\)](#) com posterior contribuição de [Scargle \(1982\)](#). Ele está disponível no pacote `astropy` (com complexidade $O[N \log N]$) através da classe `LombScargle`, e pode ser facilmente implementado:

```
from astropy.timeseries import LombScargle
```

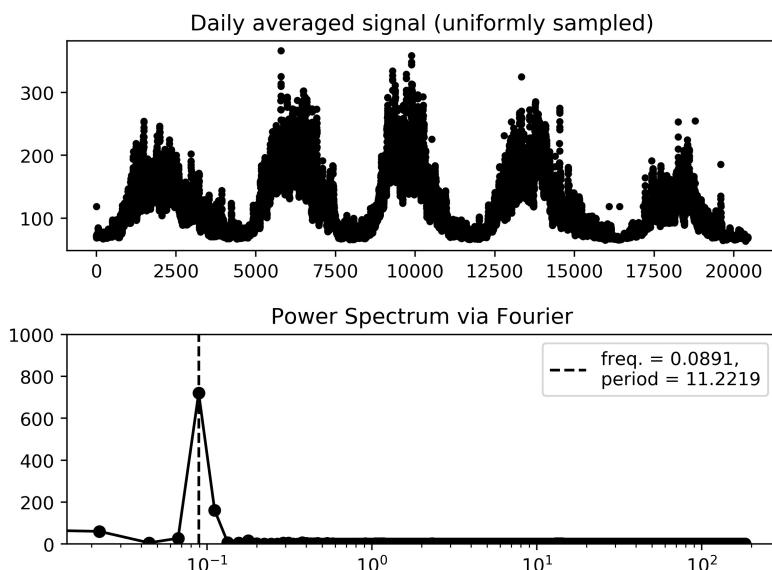
```
frequency, power = LombScargle(t, f).autopower()
```

No exemplo acima, o periodograma de Lomb-Scargle foi aplicado a um sinal `f` amostrado em tempos irregulares conforme o array `t`, gerando um array com as frequências testadas (`frequency`) e o periodograma resultante (`power`). O método `autopower()` aplica uma heurística para selecionar frequências adequadas ao teste de mínimos quadrados. Pode-se ajustar essa mesma heurística para testar frequências mais altas e com maior resolução através da palavra-chave `nyquist_factor`. O valor de dois foi empregado nas análises deste trabalho. Além disso, é possível passar como um terceiro input a incerteza dos dados, pois a classe `LombScargle` é capaz de considerar incertezas em seus cálculos (assume-se incerteza gaussiana). O sinusóide de melhor ajuste pode ser computado a partir do método `model()`. O output da classe `LombScargle` é adimensional e por padrão normalizado (seus valores estão entre 0 e 1). Alguns desses recursos são explorados na próxima seção.

3 RESULTADOS E DISCUSSÃO

A presente seção apresenta e discute o resultado da aplicação da classe `LombScargle` do pacote `astropy` sobre os dados do fluxo solar F10.7. Mas antes, a Figura 3.1 exibe o espectro de potência dos dados de média diárias do índice F10.7. Ele foi obtido via FFT conforme explicitado em [Cassara \(2020\)](#).

Figura 3.1 - Espectro de potência dos dados originais (média diária).



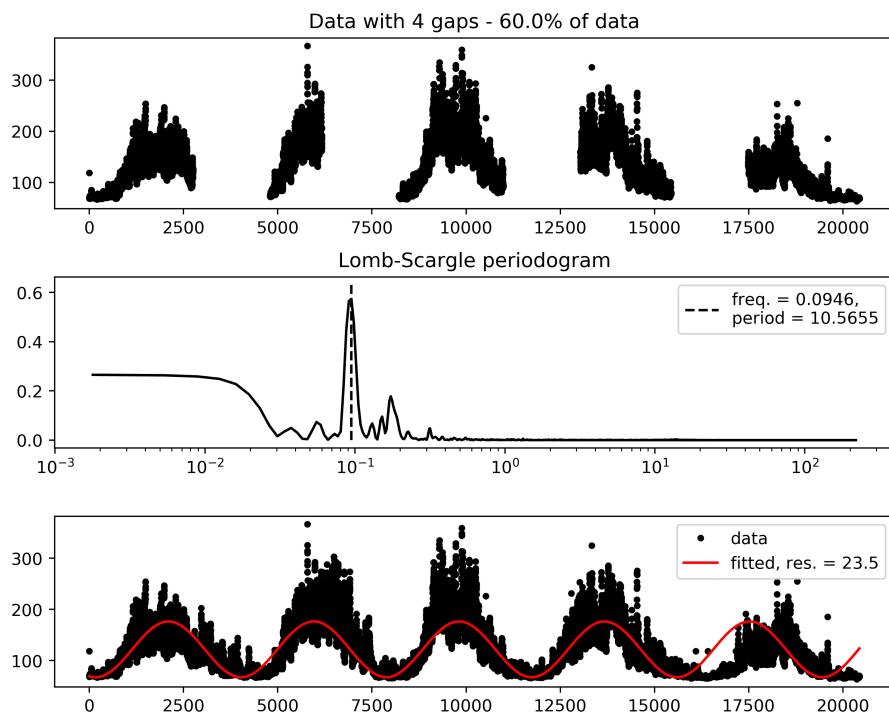
Espectro de potência via FFT das médias diárias do índice solar F10.7. Com amostragem uniforme, longa janela de observação e sampling rate satisfatório, a técnica do espectro de Fourier via algoritmos de FFT é um método robusto e amplamente empregado.

Nas seções a seguir, os dados de médias diárias do fluxo F10.7 são investigados sob diferentes cenários de amostragem aleatória. O primeiro cenário se baseia em gaps (intervalos) de interrupção na aquisição dos dados. São testados cinco números de gaps diferentes, posicionados aleatoriamente e com o mesmo tamanho. O segundo cenário simula a ausência aleatória dos dados, considerando cinco porcentagens diferentes do tamanho inicial da amostra para exclusão. Em todos os casos o resultado e a performance da classe `LombScargle` do pacote `astropy` são discutidos. A heurística da ferramenta foi configurada com `nyquist_factor = 2` durante os testes.

3.1 Cenário 1 - diferentes intervalos de observação

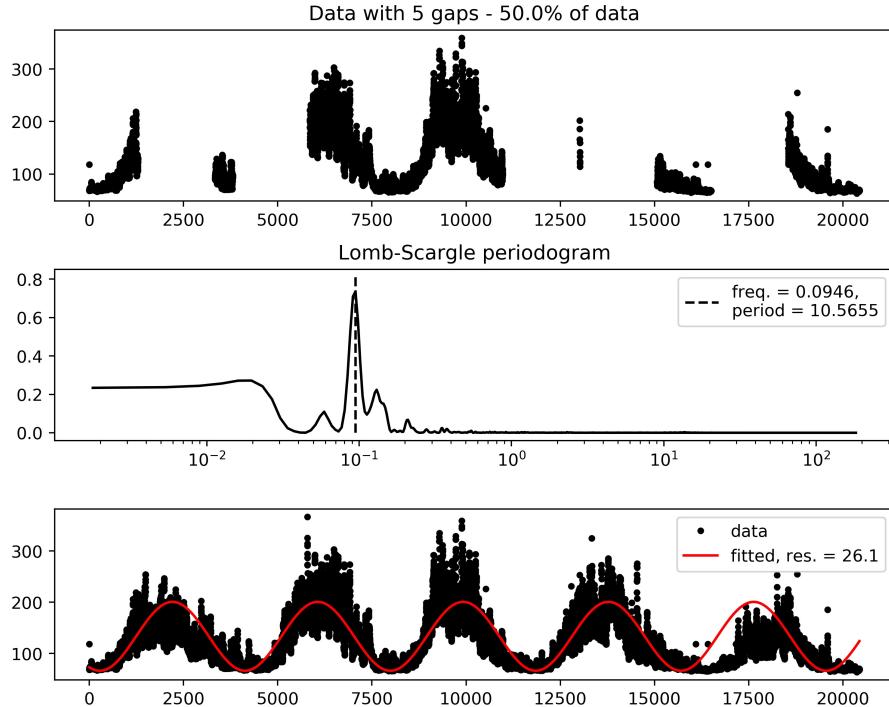
Aqui são apresentados os resultados do periodograma de Lomb-Scargle para o cenário de interrupção de observação com intervalos aleatoriamente posicionados nos dados. Em todos os testes o intervalo tem tamanho fixo e igual a 10% do tamanho da série total. A seguir são exibidos as figuras referentes às médias diárias do índice F10.7.

Figura 3.2 - Análise das médias diárias com 4 intervalos.



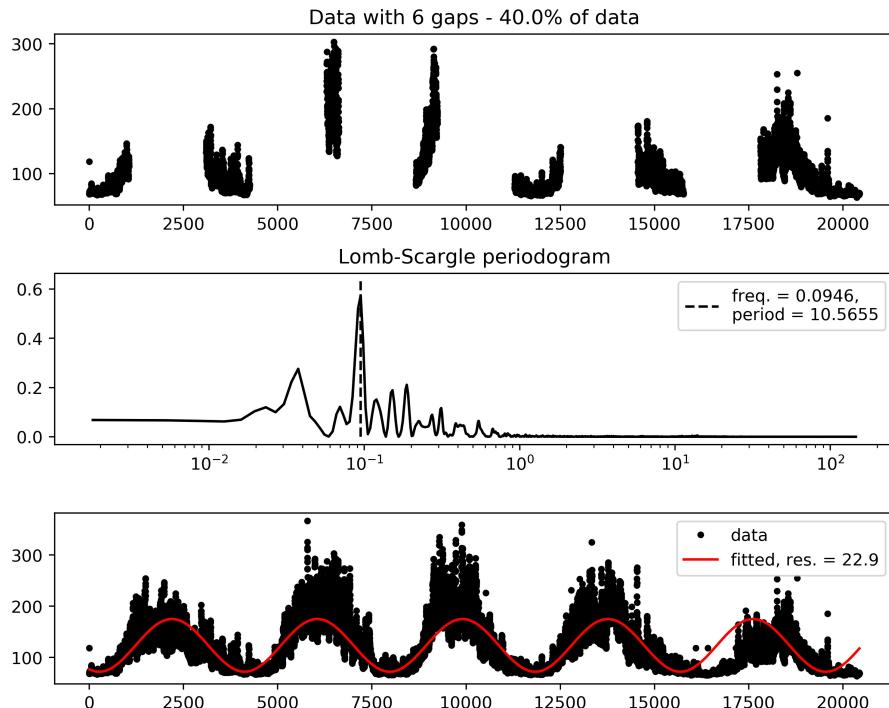
Topo: série de médias diárias do fluxo 10.7 com a presença de 4 gaps na obtenção de dados, restando assim 60% dos dados originais. Meio: periodograma de Lomb-Scargle com indicação da frequência predominante determinada pela localização do pico. Abaixo: em preto a série original, em vermelho a senóide determinada a partir do método `model()`. O resíduo médio é indicado, quantificando o quanto a função ajustada pelo pacote `LombScargle` se aproxima da série original.

Figura 3.3 - Análise das médias diárias com 5 intervalos.



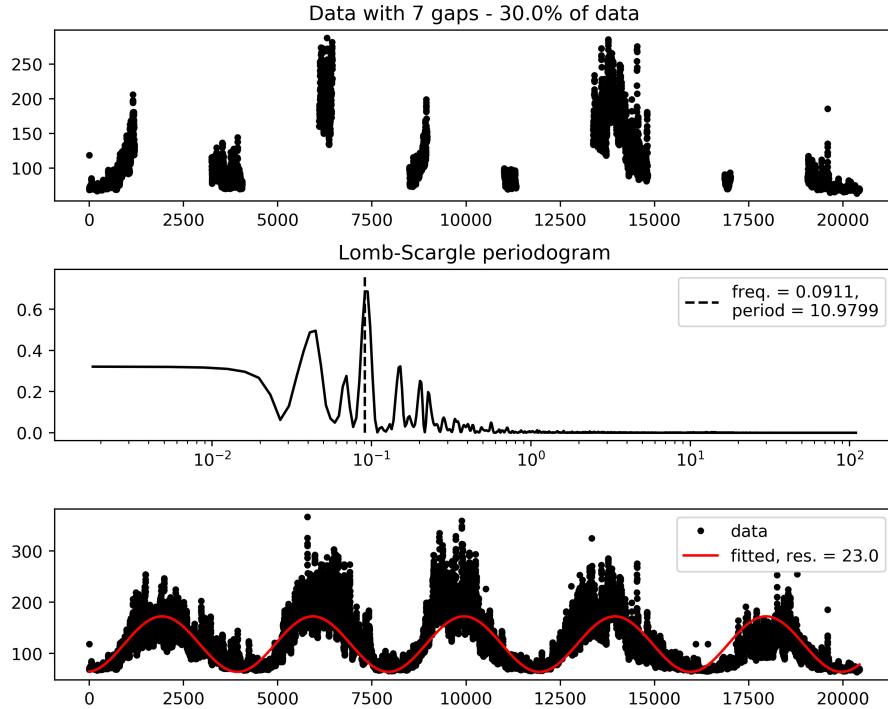
Resultado para 5 intervalos. Novamente a ferramenta utilizada corretamente identificou a periodicidade do sinal, e a função ajustada também corresponde bem à série original.

Figura 3.4 - Análise das médias diárias com 6 intervalos.



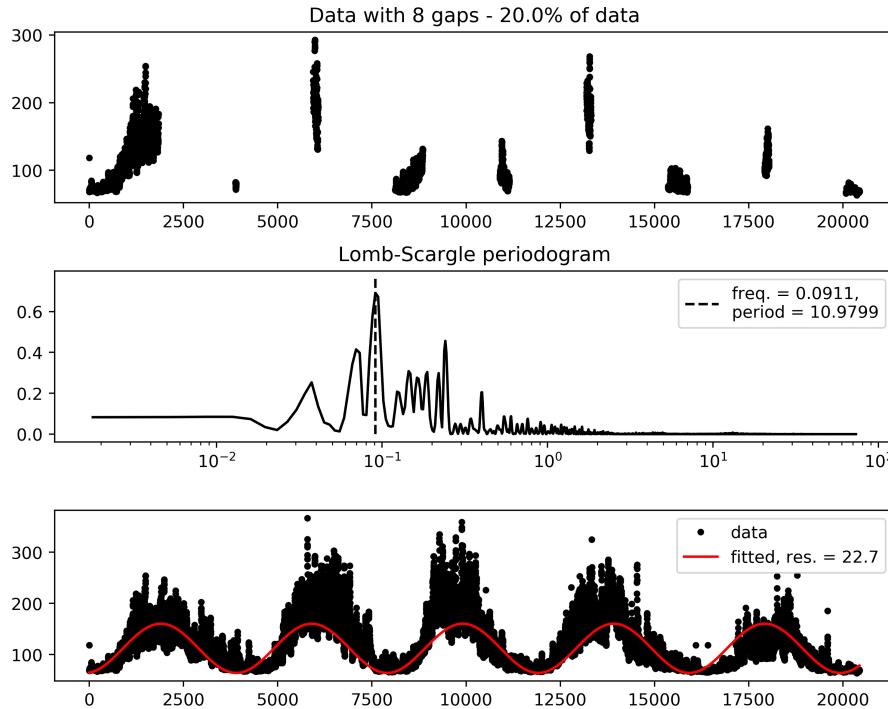
Resultado para 6 intervalos. Aqui mais frequências espúrias começam a surgir no periodograma, mas a principal assinatura ainda é bem identificada.

Figura 3.5 - Análise das médias diárias com 7 intervalos.



Resultado para 7 intervalos. Mesmo com somente 30% dos dados, o resultado do periodograma de Lomb-Scargle se mostra consistente.

Figura 3.6 - Análise das médias diárias com 8 intervalos.



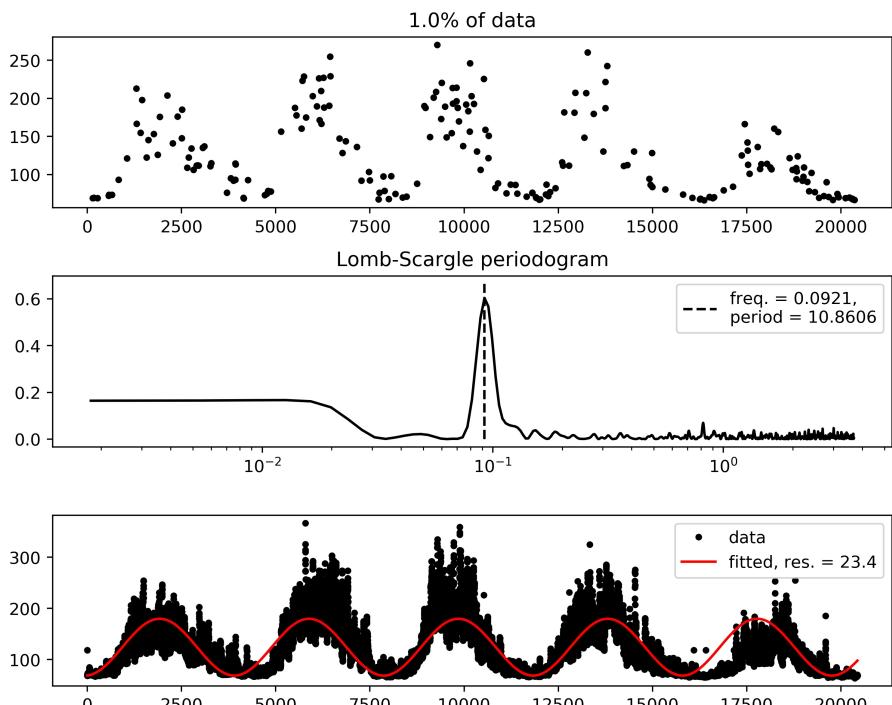
Resultado para 8 intervalos. A quantidade de picos espúrios próximo ao principal é maior que nos testes anteriores.

Os resultados das Figuras 3.2 a 3.6 indicam que o periodograma de Lomb-Scargle é satisfatório em diversas situações de ausência de dados. A princípio, a Figura 3.2 indica que a presença de poucos intervalos, tomando $\sim 40\%$ dos dados, não causa tantas anomalias ao periodograma. Os picos próximos ao principal (em 0.0946) são frequência que também se ajustaram bem aos dados. O resultado da melhor frequência foi aproximadamente o mesmo em todos os testes. Ao mesmo tempo, quanto maior o número de intervalos, maior foi a presença de picos espúrios. Neste sentido, as Figuras 3.5 e 3.6 ilustram uma característica (aleatória) do experimento: não só a quantidade de gaps, mas também a distribuição destes afetou a performance da ferramenta `LombScargle`, ainda que em menor grau. Ou seja, dependendo da posição dos intervalos durante um teste, os resultados com 6, 7 e 8 intervalos podiam ser igualmente bons, ruins, ou diferir substancialmente, mas sempre apresentando mais picos espúrios que os resultados com 4 ou 5 intervalos.

3.2 Cenário 2 - exclusão aleatória de dados

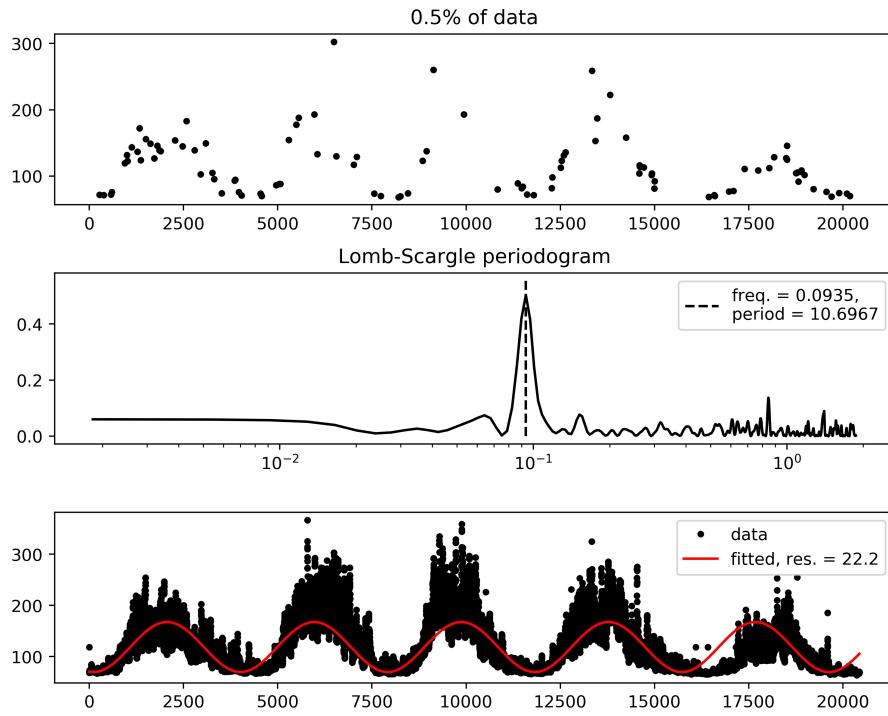
O cenário testado a seguir se baseia na exclusão aleatória das amostras até que se chegue a um limite estabelecido de porcentagem do total inicial. Esse limite foi variado cinco vezes, de modo que restasse entre 1% e 0.04% dos dados de média diária do fluxo 10.7.

Figura 3.7 - Análise com exclusão aleatória até 1% dos dados.



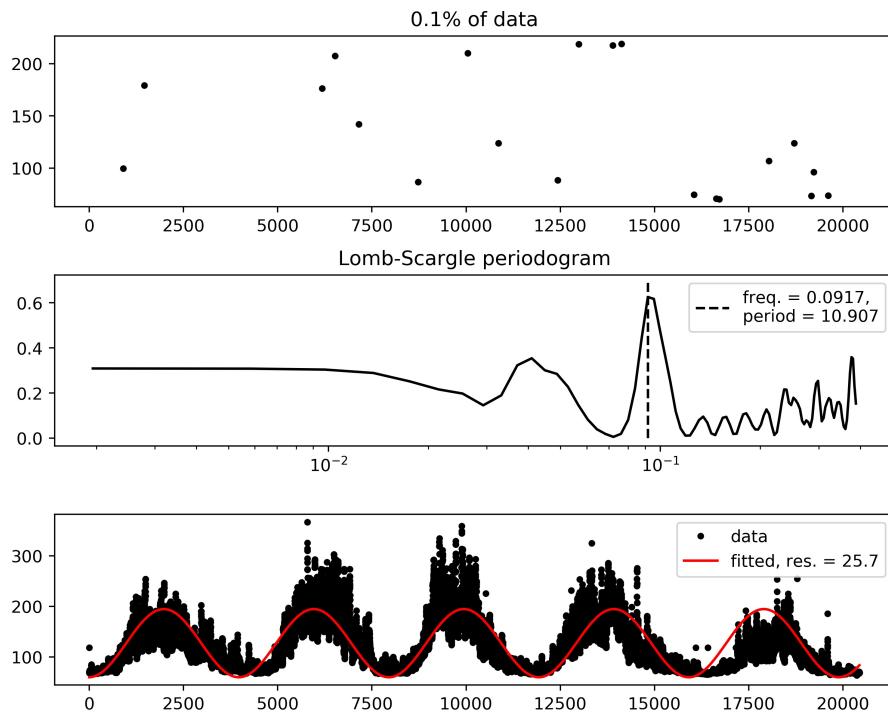
A série original continha ~ 20000 amostras, de modo que 1% destes dados, ainda que aleatoriamente distribuídos, é capaz não só de indicar o formato original da série (graças aos nossos olhos e nossa capacidade de identificar padrões) mas também de ser analisada com excelência via periodograma de Lomb-Scargle (graças à classe `LombScargle`).

Figura 3.8 - Análise com exclusão aleatória até 0.5% dos dados.



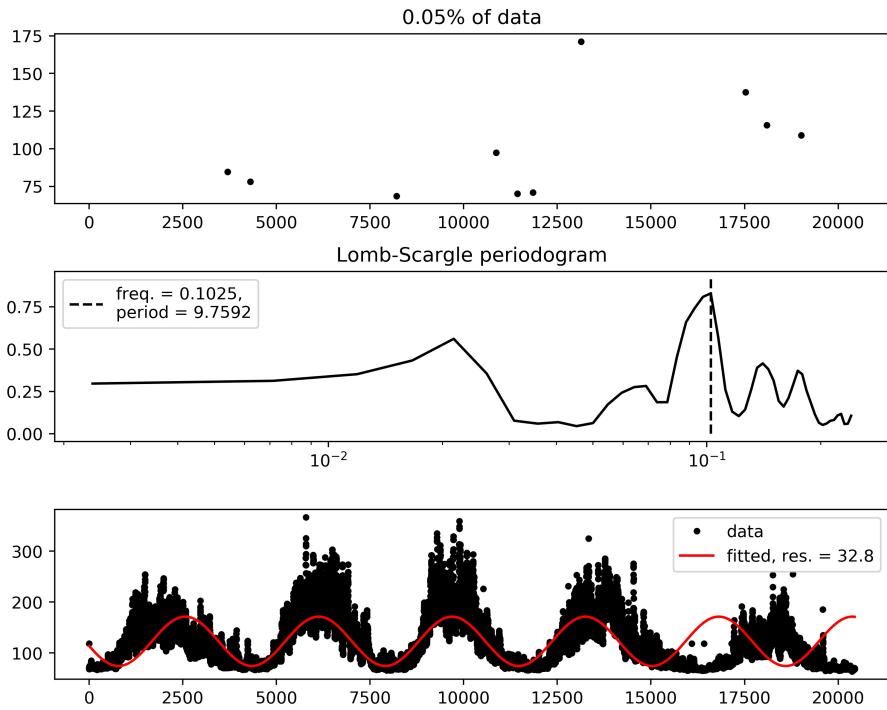
Com 0.5% dos dados a série se apresenta mais descaracterizada, mas com pontos suficientes para se assemelhar à a série original. O periodograma de Lomb-Scargle foi aplicado com sucesso.

Figura 3.9 - Análise com exclusão aleatória até 0.1% dos dados.



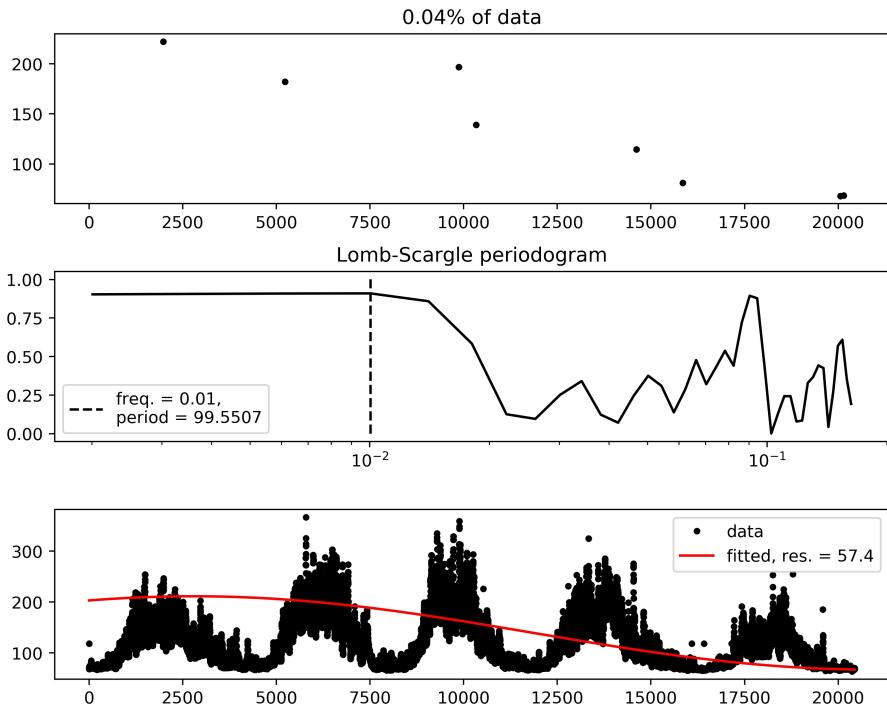
Com somente 0.1% dos dados, o perfil da série não é mais aparente e no periodograma surgem frequências espúrias. O período de ~ 11 anos continua sendo corretamente determinado pela ferramenta.

Figura 3.10 - Análise com exclusão aleatória até 0.05% dos dados.



Aqui o periodograma de Lomb-Scargle apresenta seus picos com baixa resolução (picos espalhados). Ainda assim, a frequência determinada é satisfatória e a senóide resultante se ajusta bem à série, conforme ilustrado pelo plot de baixo.

Figura 3.11 - Análise com exclusão aleatória até 0.04% dos dados.



Neste teste, com 0.04% dos dados, o periodograma de Lomb-Scargle falhou. A assinatura da série é completamente perdida, e a função ajustada é completamente diferente do esperado.

Para os fins desta discussão, considera-se um bom resultado do `LombScargle` um periodograma suave, próximo de zero em todas as frequências a não ser pela presença de um pico na frequência conhecida de 0.089 ou próximo desta, e com pouca ou nenhuma frequência espúria. A Figura 3.7 (experimento com 1% dos dados), se contrastada com as Figuras 3.2 a 3.6, indica que a baixa quantidade de dados não necessariamente afeta a performance da ferramenta. Em outras palavras, o resultado sob as condições da Figura 3.7 foi tão bom ou melhor que sob as condições de todos os testes do cenário anterior. O teste com 0.5% (Figura 3.8) obteve um resultado tão bom quanto o teste anterior. O teste com 0.1% dos dados (Figura 3.9) foi particular: mesmo quando o perfil oscilatório (visual) dos dados está totalmente perdido, o uso do `LombScargle` permitiu recuperar com êxito a característica original do sinal. Abaixo de 0.1%, assim como durante os testes do cenário anterior, a performance da ferramenta caiu demasiadamente e se tornou igualmente inconsistente para todos os valores. Ou seja, a depender dos dados aleatoriamente excluídos, o resultado da ferramenta era o mesmo com 0.05% (Figura 3.10) ou 0.04% (Figura 3.11) dos dados, e estes eram muito piores que os resultados com 0.1% dos dados ou mais.

4 CONSIDERAÇÕES FINAIS

As atividades realizadas no presente trabalho tiveram como objetivo introdução à ferramenta conhecida como periodograma de Lomb-Scargle. Os dados de fluxo solar na faixa de 10.7 cm foram manipulados a fim de simular as condições em que o periodograma de Lomb-Scargle é aplicado: na análise de séries temporais com amostragem aleatória. Os experimentos foram realizados em dois cenários: (1) com N intervalos aleatoriamente distribuídos sobre os dados e de tamanho igual a 10% do mesmo, e (2) com exclusão aleatória de amostras até $p\%$ de amostras restantes. Cinco valores de N e cinco valores de p foram aplicados sobre os dados do fluxo solar, e em seguida seu periodograma produzido com a classe `LombScargle` do pacote `astropy` (da linguagem Python).

Os resultados obtidos com a ferramenta empregada foram consistentes na maioria dos testes, indicando robustez da mesma e domínio de seu uso durante a análise. O parâmetro `nyquist_factor` empregado foi variado de modo a testar um valor ideal, uma vez que a heurística padrão não permitia enxergar adequadamente a frequência principal na maioria dos testes. Com o valor de 3, o periodograma tinha performance muito ruim na maioria dos testes. A partir deste valor, a ferramenta passou a retornar um pico para a frequência de valor 1, ou seja, um alias persistia nas

análises. Para o valor de 2, a ferramenta se tornou consistente e com performance vastamente superior à heurística padrão.

Em resumo, os diferentes efeitos da amostragem foram explorados. Em particular, o efeito de amostragem não uniforme. Sob tal condição, o espectro de potência via FFT não é mais aplicável e o periodograma de Lomb-Scargle se torna a ferramenta ideal. Num pipeline de análise, sua implementação através da classe `LombScargle` do pacote `astropy` requer cuidados com a escolha da heurística. Durante os testes aqui realizados, essa ferramenta foi capaz de corretamente identificar o período de ~ 11 anos do ciclo solar através dos dados de média diária do fluxo F10.7. O ajuste de mínimos quadrados para análise espectral se mostrou robusto na maioria dos cenários testados, e a senóide recuperada se ajustou bem aos dados originais.

REFERÊNCIAS BIBLIOGRÁFICAS

CASSARA, L. **Projeto Fourier**. [S.l.]: INPE, 2020.

<https://github.com/charlespwd/project-title>. 1, 3, 5

HUANG, C.; LIU, D.-D.; WANG, J.-S. Forecast daily indices of solar activity, f10.7, using support vector regression method. **Research in Astronomy and Astrophysics**, IOP Publishing, v. 9, n. 6, p. 694, 2009. 1

LOMB, N. R. Least-squares frequency analysis of unequally spaced data. **Astrophysics and space science**, Springer, v. 39, n. 2, p. 447–462, 1976. 1, 4

SCARGLE, J. D. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. **The Astrophysical Journal**, v. 263, p. 835–853, 1982. 1, 4

VanderPlas, J. T. Understanding the Lomb-Scargle Periodogram. **ArXiv e-prints**, mar. 2017. 1