Capítulo 3

Metodologia

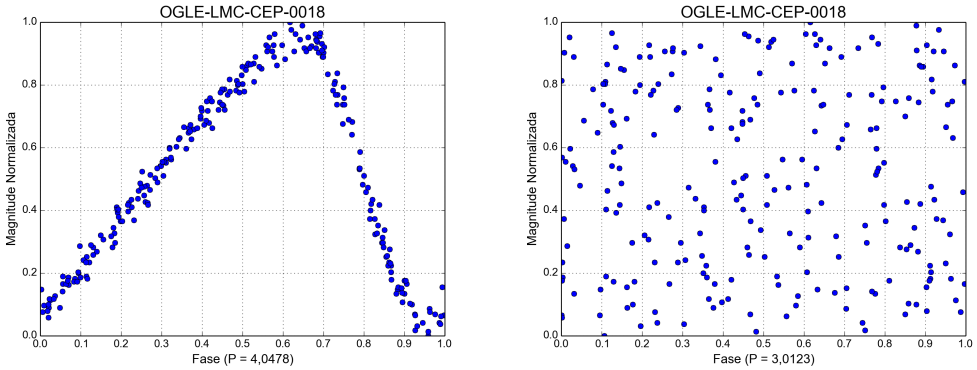
Neste capítulo será abordado o método utilizado para a determinação de períodos de estrelas variáveis pulsantes assim como a metodologia aplicada para a análise do método e criação do algoritmo, cobrindo o catálogo utilizado para obtenção dos dados e o formato dos mesmos.

3.1 Entropia de Shannon

Na teoria de informação, a entropia ou entropia de Shannon (Claude E Shannon, 1998), é a medida de incerteza de uma variável. Em outras palavras, essa grandeza mede o grau de desordem para um sinal periódico. Este sinal periódico pode ser uma curva de luz de estrela ou até mesmo observações de velocidade radial de estrelas (Cincotta, Mendez e Nunez, 1995).

A ideia deste método aplicado para a curva de luz de estrelas pulsantes se baseia na seguinte ideia: Sendo sinais periódicos as curvas de luz das variáveis pulsantes, ao fazer a transformação para o espaço de fase, a curva de luz construída com o período correto possui um certo grau de ordem, enquanto que a curva de luz construída com um período errôneo não possui ordem, gerando uma dispersão de pontos e todo o espaço de fase. Desta forma, a entropia de Shannon calculada para um sinal totalmente disperso em seu espaço de fase possui um valor maior do que essa mesma grandeza calculada para um sinal mais ordenado. Portanto, a

entropia nos informa esse grau de desordem ou incerteza da variável em estudo, que neste caso é o período, e para um conjunto de períodos que queremos analisar a entropia de Shannon deve ser mínima para o período que produz a dispersão mais ordenada no espaço de fase. Um exemplo de espaço de fase com diferentes períodos é mostrado na figura 3.1.



(A) Período correto (B) Período errado

FIGURA 3.1: Exemplos da distribuição de pontos espaço de fase para a Cefeida OGLE-LMC-CEP-0018 do catálogo OGLE. O espaço de fase da imagem na esquerda foi construído utilizando o período correto ($P = 4,0478$) e possui um valor para entropia de $H_c = 1,0762$. A imagem da direita foi utilizado um período aleatório ($P = 3,0123$) e o valor de entropia calculado é $H_c = 1,5943$.

A entropia de Shannon foi aplicada pela primeira vez em curvas de luz por Cincotta, Mendez e Nunez (1995). Eles normalizaram a magnitude das curvas de luz, transformaram para o espaço de fase e fizeram m repartições nesse espaço. Desta forma, a entropia que é definida por:

$$H = -\sum_{i}^{m} \mu_i \ln \mu_i \tag{3.1}$$

foi calculada. Nessa expressão, μ_i representa a probabilidade de ocupação da repartição i. Numericamente, a probabilidade de ocupação é calculada simplesmente contando os pontos de observação dentro da repartição e dividindo pela quantidade total de pontos. As vantagens desse método são a facilidade para lidar com sinais que possuam espaçamento variável entro os seu pontos, a simplicidade de

aplicar e possui um embasamento matemático e estatístico bem definido dentro da teoria de informação, sendo que nem todos os métodos de detecção de períodos possuem esse ultimo item bem definido (Cincotta, Mendez e Nunez, 1995).

3.1.1 Entropia de Shannon condicional

A entropia de Shannon condicional surgiu da necessidade de contornar um problema bem conhecido da analise de curvas de luz: o efeito de *Aliasing* causado pelo período $P = 1$ dia. Este efeito ocorre devido as observações serem efetuadas sempre à noite, o que ocasiona um espaçamento de um dia entre os conjuntos de observação. A figura 3.2 mostra a distribuição de pontos no espaço de fase normalizado utilizando o período $P = 1$ dia. A entropia de Shannon calculada para uma distribuição desta forma retorna um valor pequeno pois os pontos estão localizado em uma determinada parte do espaço. Para lidar com esse problema,

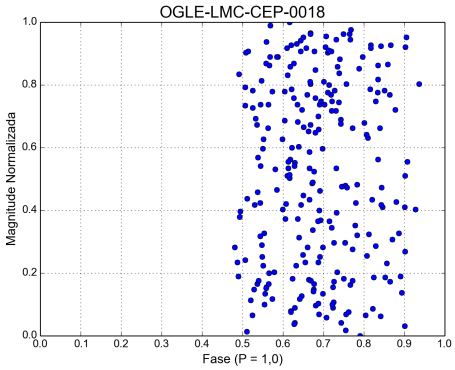


FIGURA 3.2: Efeito de *Aliasing* devido ao período de 1 dia para a Cefeida OGLE-LMC-CEP-0018 do catálogo OGLE. Os pontos se localizam em uma determinada região do espaço de fase (menos da metade) deixando o que faz com que a entropia calculada seja pequena. A entropia condicional foi proposta para lidar com este problema e para este caso o seu valor é $H_c = 1,5542$.

Graham et al. (2013b) propuseram a entropia de Shannon condicional. Nesta variação do método o espaço de fase é dividido em i repartições na magnitude e j

repartições na fase e a entropia é calculada da seguinte forma:

$$H_c = \sum_{i,j} p(m_i, \phi_j) \ln \left(\frac{p(\phi_j)}{p(m_i, \phi_j)} \right) \tag{3.2}$$

em que $p(m_i, \phi_j)$ é a probabilidade de ocupação na i-ésima repartição da magnitude e na j-ésima repartição da fase e $p(\phi_j)$ é a probabilidade de ocupação na j-ésima repartição da fase. Como estamos lidando com repartições retangulares:

$$p(\phi_j) = \sum_{i} p(m_i, \phi_j) \tag{3.3}$$

ou seja, $p(\phi_j)$ é a soma das probabilidades na i-ésima coluna.

Graham et al. (2013b) analisaram o impacto no resultado da entropia causado pela quantidade de repartições e estimaram que 5 repartições na magnitude ($\Delta m = 0,2$) e 10 repartições na fase ($\Delta \phi = 0,1$) seriam ideais pois, quanto maior a quantidade dessas repartições mais recursos computacionais são necessários e com essa escolha e entropia continua retornando bons resultados em pouco tempo.

Desta forma, a entropia de Shannon condicional será utilizada considerando 5 repartições para a magnitude e 10 para fase. Este método será aplicado para o espaço de fase normalizado para um conjunto de períodos que se quer analisar, calculando a entropia de Shannon condicional para cada espaço de fase criado para esse conjunto de períodos. O menor valor de entropia corresponde ao conjunto de pontos mais ordenado que seria o período correto da estrela (Graham et al., 2013b) . Porém, antes de entrar em detalhes no algoritmo criado é necessário entender os dados utilizado no trabalho.

3.2 Catálogo OGLE

O catálogo OGLE (*The Optical Gravitational Lensing Experiment*) consiste em 8 anos de dados observacionais cobrindo uma área de 40 graus quadrados na

TABELA 3.1: Exemplo de dados do catálogo OGLE

Tempo	Magnitude	Erro
2165,85271	15,130	0,007
2183,83450	15,326	0,008
2238,62899	15,102	0,007
⋮	⋮	⋮

direção das Nuvens de Magalhães. Esse catálogo busca por estrelas variáveis tendo monitorado mais de 200 milhões de estrelas. As observações foram feitas utilizando os filtros Cousins I e V (Cousins, 1973). Na banda I, as observações possuem um tempo de 180s de exposição tendo em média 400 medidas de observação. Por outro lado, a banda V possuem em média apenas 30 medidas de observação. Os dados da sua terceira fase (Udalski et al., 2008), chamado de OGLE-III, são públicos[1] e foram utilizados nesse trabalho, dando prioridade para as observações na banda I devido a maior quantidade de medidas em relação a banda V.

Os dados de observação disponíveis são obtidos no formato .dat e possuem três colunas que significam tempo em dias Julianos, magnitude e erro na magnitude. Um exemplo de dado pode ser visto na tabela 3.1.

Nesse trabalho foram utilizados os dados de dois tipos de estrelas variáveis pulsantes localizadas na Grande Nuvem de Magalhães, as Cefeidas Clássicas e as RR Lyraes, sendo utilizados 3056 Cefeidas classificadas entre modo fundamental (FU) e primeiro sobretom (FO) e 22651 RR Lyraes também classificas entre modo fundamental (AB) e primeiro sobretom (C), totalizando 25707 estrelas.

3.3 Algoritmo

Foi desenvolvido um algoritmo em `Python3` para calcular a entropia condicional de dados de estrelas variáveis pulsantes pertencentes ao Catálogo OGLE-III.

[1] http://ogledb.astrouw.edu.pl/~ogle/CVS/

A figura 3.3 apresenta um pseudo-código do algoritmo. O código completo é apresentado no apêndice A.

Entrada: Tempo e Magnitude
Saída: Período P que minimiza a entropia
Início
    Leitura dos dados de entrada como vetores;
    Cria um vetor com n períodos sendo P = (p_1 , p_2, $\cdots$, p_n);
    Normalização da magnitude;
    para cada p_i *em* P **faça**
        Transformar o tempo para o espaço de fase;
        Faz as repartições e contabiliza os pontos;
        Calcula a entropia de Shannon condicional;
        Armazena a entropia calculada para o período p_i
    fim
    Achar o valor mínimo de entropia: E_{min} = min(Entropia)
    Achar o período que minimiza a entropia: $P_{E_{min}}$=P[min(entropia)]
Fim
retorna $P_{E_{min}}$

FIGURA 3.3: Pseudo-código do algoritmo em português estruturado.

Para cada um dos dados das estrelas que serão analisadas, o programa faz a leitura das informações de tempo e magnitude da estrela e cria um vetor de períodos que serão analisados. Para uma Cefeida, esse vetor de períodos é criado com período inicial $p_1 = 0,1$ dias e período final $p_n = 32$ dias com um intervalo entre os períodos de $0,001$ dia. Então, para cada um dos elementos do vetor período o algoritmo faz as seguintes ações: o tempo é transformado em fase, são feitas as repartições no espaço de fase e são contabilizados a quantidade de pontos em cada repartição, a entropia de Shannon condicional é calculada e o valor armazenado em um vetor entropia. No fim, o algoritmo indica o menor valor do vetor entropia e qual período esta relacionado com este valor.

3.4 Análise Teórica

Ao utilizar o algoritmo para os dados do catálogo podemos analisar como o método responde para dados reais. Porém, se quisermos analisar qual a abrangência de atuação desta técnica, é possível calcular a entropia de Shannon para um conjunto de dados teóricos em que seja conhecido o período. Desta forma, podemos entender como o formato de um sinal influencia nos resultados finais.

De acordo com Graham et al. (2013b) e Cincotta, Mendez e Nunez (1995), um sinal periódico sintético que se assemelhe com os dados observacionais da maioria dos Surveys de estrelas variáveis pode ser construído utilizando a expressão:

$$m(t) = A_0 + \sum_{i=1}^{3} A_i \sin\left(\frac{2k\pi t}{P}\right) + B\eta \tag{3.4}$$

em que $m(t)$ é a magnitude sintética, A_0 é termo de deslocamento linear, os termos A_i são termos de escala para as funções senos, k é um parâmetro de escala para a amostragem do sinal, t é o vetor tempo, P é o período de oscilação do sinal, η é uma distribuição gaussiana com média zero e desvio unitário que tem como função introduzir ruído no sinal e B é um parâmetro de escala para esse ruído.

A amostragem, f_s, de um sinal representa a frequência de pontos de observação. Essa quantidade afeta diretamente a construção do vetor tempo pois, a amostragem é definida como:

$$f_s = \frac{1}{dt} \quad \rightarrow \quad dt = \frac{1}{f_s} \tag{3.5}$$

ou seja, o intervalo de tempo depende do valor da amostragem. Desta forma, assim que for definida a nossa amostragem, podemos variar essa grandeza para construir os vetores tempos e com isso construir o sinal sintético para calcular a entropia de Shannon condicional e analisar os resultados. A análise de todos os dados, sintéticos e reais, será discutida no capítulo 4.

Capítulo 4

Resultados e Discussão

4.1 Dados do Catálogo OGLE

Em um total foram calculados os períodos de 25707 estrelas variáveis localizadas na Grande Nuvem de Magalhães e pertencentes ao catálogo OGLE. Deste numero total, 3056 eram Cefeidas clássicas tipo FO e FU, e 22651 eram RRLyraes tipo AB e C. Os resultado obtidos foram comparados com os resultados do catálogo e o percentual de acertos pode ser visto na tabela 4.1.

De acordo com os resultados da tabela 4.1 podemos perceber que para as Cefeidas o método apresenta um resultado um pouco melhor se comparado com as RR Lyraes. Uma explicação para este resultado seria que o método de entropia de Shannon condicional funciona melhor para magnitude mais brilhantes (Graham et al., 2013a) e sendo as Cefeidas ($m \approx 15$) mais brilhantes do que as RR Lyraes ($m \approx 19$), essa afirmação é coerente com os resultados.

Tabela 4.1: Quantidade de dados analisados e resultados corretos.

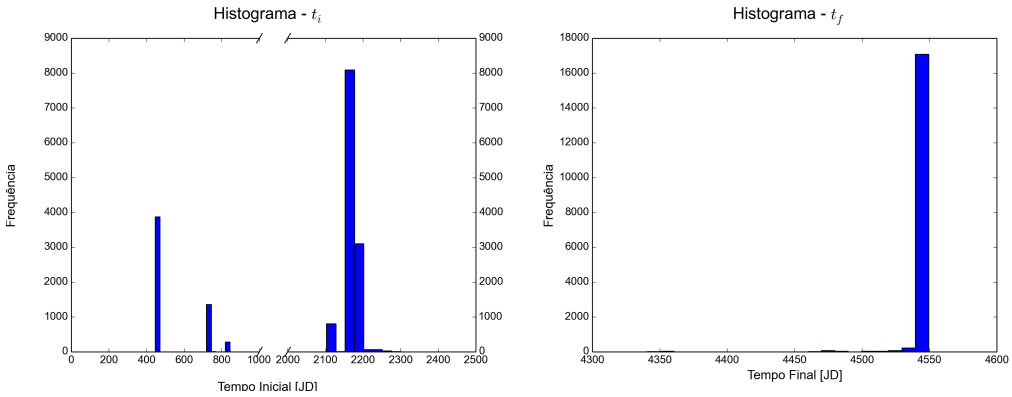
Estrelas	Quantidade	Acertos	Porcentagem
Cefeidas FU	1818	1817	$99,94\%$
Cefeidas FO	1238	1231	$99,43\%$
RRLyraes AB	17693	17540	$99,14\%$
RRLyraes C	4958	4535	$91,47\%$
Total	**25707**	**25123**	**97,73 %**

A alta taxa de acerto do método no informa que com estes resultados podemos confiar no método de entropia de Shannon condicional, porém para entender melhor o comportamento desse método será analisado os resultados para dados sintéticos.

4.2 Dados Sintéticos

Dados sintéticos foram criados a fim de explorar o método e entender até onde podemos utilizá-lo. De acordo com a tabela 4.1, as RRLyraes apresentaram uma taxa menor de acerto então elas foram utilizadas como referencia para construir os dados sintéticos. Para isto, é necessário entender os dados das RR Lyraes do catálogo OGLE para obter os parâmetro sobre o tempo e amostragem para enfim utilizar a expressão 3.4 e construir os dados.

Analisando os dados das 22651 estrelas, foram criados histogramas sobre os dados iniciais e finais do tempo e a quantidade de pontos de observação. As figuras 4.1 e 4.2 nos mostram esses histogramas.



(A) Tempo Inicial (B) Tempo Final

FIGURA 4.1: Histogramas sobre o tempo inicial e final das RR Lyraes. As imagens representam (a) tempo inicial e (b) tempo final . A partir dessa análise foram obtidos os valores $t_i = 2152,5019$ e $t_f = 4539,4593$.

A partir dessa análise foram obtidos os valores $t_i = 2152,5019$ e $t_f = 4539,4593$ e $n = 352$ como os valores de tempo inicial, final e quantidade de pontos mais

frequentes nos dados das RR Lyraes. Desta forma, utilizando esses valores é possível construir um sinal sintético que se assemelhe com os dados do catálogo. Então, a amostragem é calculada pela expressão 3.5 em que a variação do tempo e obtida da seguinte forma:

$$dt = \frac{t_f - t_i}{n} = \frac{4539,4593 - 2152,5019}{352} = 6,7888 \tag{4.1}$$

e substituindo este resultado na equação 3.5 nos obtemos:

$$f_s = 0,1473. \tag{4.2}$$

Desta forma foi determinado a partir dos dados do catálogo qual a variação média entre os pontos de observação e foi calculada a amostragem média dos dados.

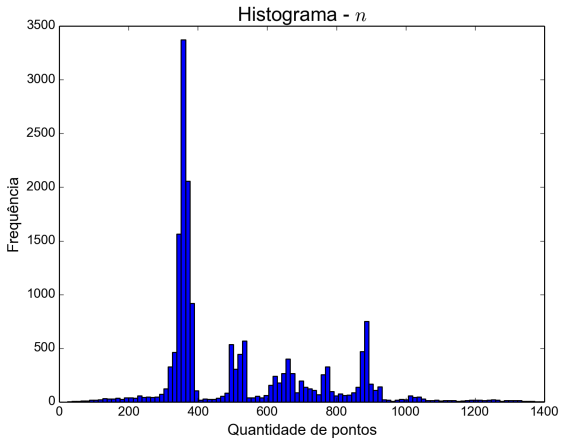
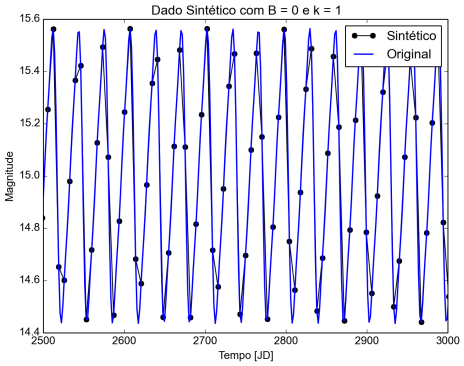


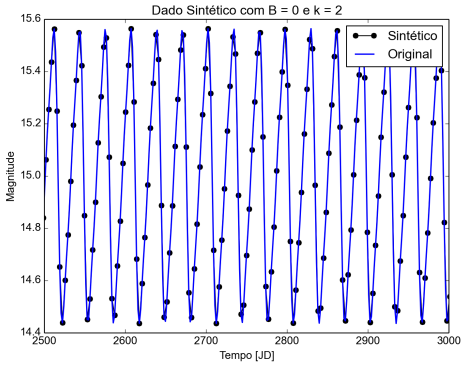
FIGURA 4.2: Histograma sobre a quantidade de pontos nos dados das RR Lyraes. A quantidade com maior frequência é $k = 352$.

Tendo obtido a amostragem, podemos construir dados sintéticos variando a frequência de pontos e o nível de ruído para estudar como o método se comporta com esses sinais. O sinal sintético é construido pela expressão 3.4 em que os termos A_i são dados por Graham et al. (2013b) e Cincotta, Mendez e Nunez (1995) como sendo $A_0 = 15$, $A_1 = -0.5$, $A_2 = 0.15$ e $A_3 = -0.05$. O período utilizado para

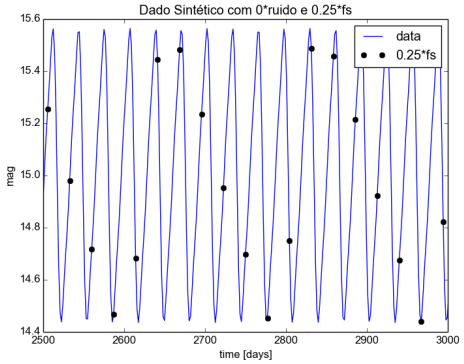
criar o sinal será $P = 0.576$ dias pois, de acordo com Soszyński et al. (2009) esse é o valor de período médio das RR Lyraes do catálogo.



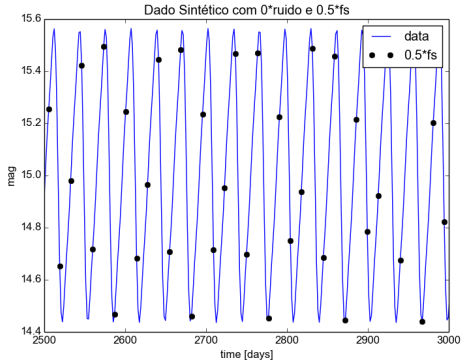
(A) Dado sem ruído e com amostragem padrão



(B) Dado sem ruído e $k = 2$



(C) Dado sem ruído e $k = 1/4$



(D) Dado sem ruído e $k = 1/2$

FIGURA 4.3: Exemplos de curvas de luz sintética. Os exemplos foram criados sem ruído, $B = 0$, e variando k. A linha azul representa o sinal original completo e os pontos e linha preta correspondem a observação com a amostragem k.

Para estudar a influência da amostragem nos dados, o vetor t será criado utilizando os valores obtidos pelos histogramas de tempo inicial ($t_i = 2152, 5019$) e final ($t_f = 4539, 4593$) e a a variação de pontos dt será construindo pela relação:

$$dt = \frac{1}{f} \tag{4.3}$$

em que $f = k \times f_s$, ou seja, a frequência de pontos f será um parâmetro de escala k vezes a amostragem f_s dos dados. Desta forma, variando o parâmetro k de $0, 25$

a $4,0$ com um intervalo de $0,25$ e variando o parâmetro de escala para o ruído B de $0,0$ até $1,0$ com intervalo de $0,05$, forma criadas 300 curvas de luz para serem analisadas. Quatro exemplos de curva de luz sintética gerada pelo método acima podem ser vistas na figura 4.3.

Na figura 4.3, os pontos e linha preta correspondem aos pontos de observação e a linha contínua azul seria o sinal original completo. Podemos perceber que quanto maior a amostragem, maior a quantidade de pontos.

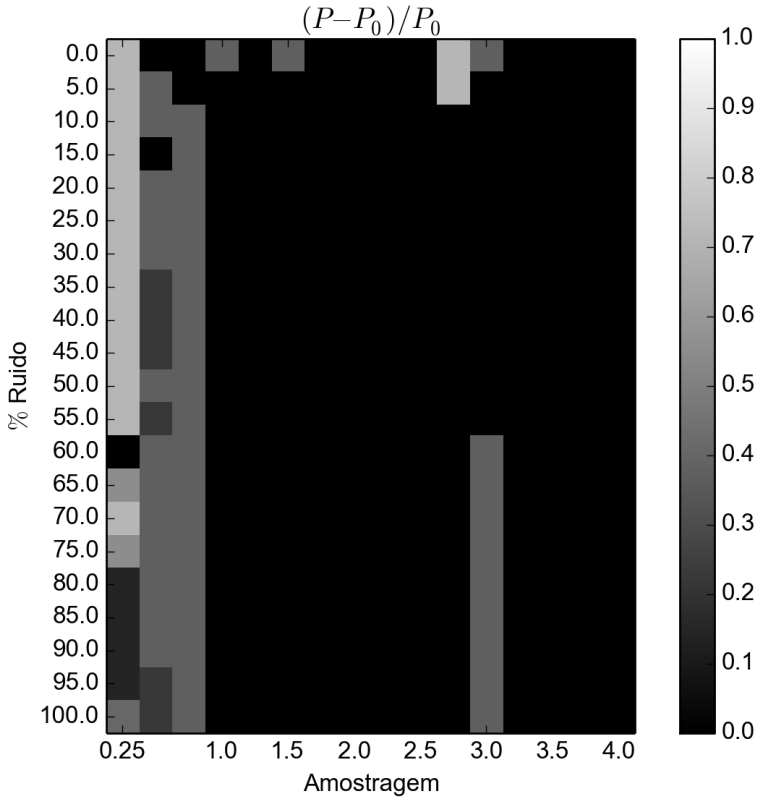


FIGURA 4.4: Resultados obtidos em escala de cinza. O eixo das abcissas representa o parâmetro de escala k da amostragem e o eixo das ordenadas representa a variação do parâmetro de escala B para o ruído. Quanto mais escura a cor do quadrado mais correto o valor calculado pela entropia de Shannon condicional.

A figura 4.4 nos mostra um mapa de cor em escala de cinza entre parâmetro de escala B do ruído e o parâmetro de escala k da amostragem. A cor representa

o valor $|(P - P_0)/P_0|$, ou seja, quanto que o período calculado está variando em relação ao período original. A cor mais escura representa o valor 0 (período calculado = período real) e quanto mais clara a cor, maior o desvio do período.

Podemos observar que a partir do parâmetro de escala $k = 1,0$ para a amostragem, todos os resultado calculados foram corretos não importando o nível de ruído, com exceção da faixa entre os ruídos 60% e 100% para $k = 3$ que apresentam um resultado $\approx 0,4$. Essa exceção significa que o resultado obtido é aproximadamente a metade do período real.

4.3 Conclusão

A observação e detecção de períodos de estrelas variáveis é fundamental para descrição desses objetos astronômicos e para a determinação de distâncias. Embora existam diversos métodos para o calculo de período, o desenvolvimento de técnicas que sejam confiáveis e possam ser aplicadas para dados com espaçamento variável entre os pontos de observação é de grande importância em uma realidade em que há dificuldades para observação dos telescópios, sendo essas dificuldades devido ao tempo disponível de observação e as condições climáticas. O método apresentado neste trabalho. a entropia de Shannon condicional, é uma técnica simples de ser entendida e aplicada, possuindo um embasamento matemático dentro da teoria da informação o que faz com que a sua análise estatística seja conhecida, fato que não é verdade para alguns métodos de detecção de períodos. Além disso, o método apresenta um desempenho mais do que satisfatório com uma taxa de acerto maior do que 97% para as 25707 estrelas pulsantes do catálogo OGLE-III. Além disso, a análise dos dados sintéticos afirma que o método é confiável para qualquer nível de ruído desde que a frequência de pontos dos dados seja maior do que $f_s = 0,1473$. Por fim, com a figura 4.4 foi possível construir uma ferramenta

que nos indica como os dados influenciam no resultado do método, ou ainda, partindo do resultado que se espera obter, é possível determinar como a observação nos telescópios devem ser conduzidas. Parte dos resultados obtidos nesse trabalho foram apresentados em Ramos, Bellinger e Kanbur (2014) e Ramos, Ferrari e Kanbur (2015).