Recuperação das propriedades da trajetória em 2D por inversão Bayesiana semi-empírica de séries temporais irregularmente amostradas

George Caminha-Maciel e Marcia Ernesto, Departamento de Geofísica, IAG/USP
e-mail: caminha@iag.usp.br , marcia@iag.usp.br

Resumo

Métodos para investigação de oscilações coerentes em séries temporais existem em grande abundância na literatura. No entanto, a grande maioria das ferramentas conhecidas pode ser aplicada apenas a dados regularmente espaçados e estatisticamente homogêneos. Esta limitação se deve principalmente ao tratamento estatístico de 'ensemble', ou seja, as incertezas nos parâmetros do modelo são definidas através de transformações entre conjuntos homogêneos de variáveis aleatórias. Neste trabalho propomos métodos 'de pequenas amostras', desenvolvidos para situações onde os tempos de amostragem não estão à nossa disposição ou os dados disponíveis não são conjuntos assintoticamente grandes, possibilitando validar a aplicação *a priori* do Teorema Central do Limite.

Apresentamos os resultados da inversão Bayesiana com funções de estado de informação semi-empíricas baseadas nas propriedades espectrais de séries temporais aleatoriamente amostradas, ou seja, sem intervalo de amostragem característico.

Introdução

A obtenção das propriedades de uma trajetória, a partir de múltiplas séries temporais incompletas, ou irregularmente amostradas, é um problema que aparece em diversas aplicações práticas. Usualmente o tratamento se dá por redução aos casos regulares, considerando-se as diferenças como pequenos desvios de uma grade regular, com tais desvios sujeitos a alguma lei supostamente conhecida (ou assumida). Nesse tipo de aproximação a análise das incertezas no modelo final se torna difícil de estabelecer, tendo em vista a heterogeneidade inerente aos dados irregularmente amostrados. Ainda há casos em que, além da amostragem em tempos irregulares, os dados possuem incertezas diferentes.

Em todos estes casos seria desejável um tratamento 'de pequenas amostras', que não se inicie da suposição de que as amostras seriam pertencentes a um *ensemble* de variáveis independentes e identicamente distribuídas (*i.i.d.*).

Um dos melhores exemplos são as seqüências magnetoestratigráficas que permitem obter registros da dinâmica de longo período do campo geomagnético principal e exibem diversos eventos notáveis de longa duração (milhares ou milhões de anos), como excursões geomagnéticas e inversões de polaridade. Essas seqüência permitem conhecer com certo detalhe, as trajetórias descritas pelo pólo de referência, através de suas estimativas a cada instante, os pólos geomagnéticos virtuais (PGVs). Em seqüências de lava, por ser um processo intermitente, o registro magnetoestratigráfico pode ser muito incompleto e descrever pobremente a trajetória dos PGVs.

Quando múltiplas seqüências magnetoestratigráficas da mesma formação geológica estão disponíveis, é possível buscar as propriedades da trajetória dos PGVs que são recorrentes nas seqüências amostrais. Entre outras opções, escolhemos propriedades espectrais, ou de correlação com funções seno e cosseno, pois o fenômeno em questão aparenta comportamento oscilatório e, para esta base de funções, a obtenção dos parâmetros é mais intuitiva.

Problema Investigado

No caso em questão, não se dispõe de marcadores de fase, pois as incertezas nas datações radiométricas são bem maiores que as separações entre os pontos. Sem um intervalo característico que possamos definir como 'taxa de amostragem', pois a média dos intervalos pode ser pouco representativa, buscamos um funcional que nos dê uma medida da chance de uma determinada freqüência ser uma das componentes de Fourier da série original, presente em cada série amostral. As opções são: funcionais dependentes do inverso do resíduo do ajuste linear ('misfit functions'), dependentes da média das correlações e dependentes da média das correlações ponderada pelas auto-correlações das funções da base, entre outros.

Metodologia

Em qualquer destes casos, o periodograma fornece um ponto de partida conveniente, sendo o maior desafio suavizá-lo. Escolhemos um algoritmo semi-empírico baseado em periodogramas de Lomb-Scargle (Lomb,

1976; Scargle, 1982), onde o critério de suavização, expresso pela densidade (e intervalo) de amostragem no domínio da freqüência, depende de uma escolha do usuário dentro de limites pré-estabelecidos. Assim, a largura da banda pesquisada e o intervalo de discretização são os parâmetros mais importantes. Estes sofrerão modificações dentro do limite correspondente ao período de uma onda, com comprimento igual ao da série e até a freqüência limite de Nyquist, até atingir-se uma curva suave onde apareçam as feições de interesse.

As funções de estado de informação obtidas devem ser tais que, para as séries pouco informativas, a curva deve se aproximar de uma reta constante em toda a banda pesquisada. A partir daí, é preciso definir um critério de normalização para as funções de estados de informação amostrais. Podemos normalizar, por exemplo, pela entropia, pelo conteúdo de informação ou mesmo pela soma da probabilidade sobre todos os estados. Daí então só se precisa combinar os vários estados de informação amostrais.

A idéia principal em que se baseia o procedimento é a de que a informação de cada série é utilizada para *falsear* (Tarantola, 2006) possíveis soluções em um subconjunto do espaço paramétrico. Isto é, combinar a informação geral que dispomos, para toda a trajetória, em relação a determinado parâmetro – neste caso, a probabilidade de que determinada freqüência seja uma verdadeira componente de Fourier da trajetória original. Combinando estas funções amostrais obteremos dois invólucros no espaço paramétrico, um maior, inclusivo e outro menor, contido no anterior, exclusivo. Estes representam a união e intersecção, respectivamente, dos diversos estados de informação e será o resultado do nosso procedimento de inversão, descrevendo a incerteza que temos sobre o valor do parâmetro com grande generalidade. Também será através das propriedades geométricas destas funções de estado de informação que descreveremos a incerteza para qualquer escolha de 'melhor' modelo que se queira fazer.

Escrevemos um programa de Matlab para realizar as operações descritas anteriormente, além de um simulador simples, adaptado a gerar cinco diferentes séries de pontos. Estes foram escolhidos uniformemente sobre fragmentos de uma trajetória, cuja composição em termos de componentes de Fourier é conhecida. Adicionamos um ruído proporcional à amplitude de cada ponto e comparamos a amostra (série) com uma amostra de igual densidade, porém uniformemente distribuída sobre o intervalo.

Apresentamos alguns exemplos de simulações de séries amostrais e as inversões conseguintes (Figs. 1, 2 e 3). Além de enfatizar as freqüências (ou, comprimentos de onda) impostas e recuperadas, cabe notar o importante efeito da amostragem insuficiente que, na maioria dos casos, conserva pouco da aparência do fenômeno original.

Em cada uma das figuras, exibimos as cinco amostras da série original em (a) e os resultados da inversão em (b). Em cada quadro (b), mostramos na janela superior as funções de estado de informação para cada série separadamente, e nas janelas seguintes o resultado das operações de combinação 'OU' e 'E' equivalentes à união e intersecção, respectivamente, das funções de estado de informação amostrais (Tarantola e Valette, 1982).

Devemos esclarecer que estas séries representam apenas uma entre muitas possíveis escolhas de 'aleatoriedade', ou de distribuição randômica, que teríamos para modelar séries 'irregularmente espaçadas' e que são similares às séries de PGVs amostrais.

Resultados

Devemos relembrar que, através deste procedimento, ao invés de obtermos um valor que elegeremos como 'sinal' ('signal detection problem'), obteremos uma medida de probabilidade sobre um subconjunto do espaço das freqüências, que nos indicará quais modelos se ajustam melhores que outros, segundo um critério especificado. Assim, a interpretação dos fenômenos de 'falseamento' e vazamento de potência se dá naturalmente, como efeitos inevitáveis numa análise do conjunto de pontos finito através de correlação com bases funcionais infinitas.

Assim, estamos interessados não somente nos valores máximos da ordenada dos 'espectros' obtidos. Estamos também procurando máximos locais de correlação, sendo muito importante a continuidade, ou o comportamento 'local' da distribuição.

Também a obtenção de espaço solução muito restrito não significa necessariamente que conhecemos o valor do parâmetro com precisão absoluta. Devemos interpretar com cautela os valores obtidos, com base na incerteza estimada nos dados amostrais e na quantidade (e qualidade) das séries. Neste sentido, cada série representa um experimento de medida espectral, ou da 'rugosidade', da série original. O problema é análogo ao do experimentador que medindo duas ou três vezes uma determinada grandeza física e obtendo coincidentemente valores muito próximos, acredita conhecer o valor da tal grandeza com grande precisão, pois sua estimativa da variância da distribuição a partir do desvio-padrão amostral é pequena. Ou seja, a distribuição da variável combinada (estimativa) é concentrada somente porque as distribuições combinantes, apesar de serem pouco informativas, têm pequena intersecção ou poucos modelos comuns.

Apesar das inúmeras dificuldades encontradas, inclusive na interpretação dos resultados obtidos, o procedimento mostra excelente resultado, haja vista a restrição de possíveis soluções no espaço paramétrico, da estrutura aparente encontrada nas soluções, além da recuperação da informação original, na maioria dos casos. As periodicidades (freqüências) impostas à série original foram satisfatoriamente recuperadas, encontrando-se em cada caso uma região de alta correlação próxima às freqüências verdadeiras no espectro 'E' ou 'OU' e, consistentemente, na maioria das funções de estado de informação individuais de cada série, em cada exemplo. Além destas, é claro, surgiram algumas outras regiões de

alta correlação (máximos do espectro), oriundas de fenômenos de natureza espectral, como vazamento de potência para outras freqüências, além das regiões onde não há resolução nos modelos a partir das funções de estado de informação amostrais utilizadas.

Discussão e Conclusões

Duas idéias principais suportam os procedimentos realizados: i) Se a série original contém um número finito (e pequeno) de componentes sabidamente periódicas, as séries amostrais irão também apresentar alta correlação com as funções base, nos comprimentos de onda próximos ao das periodicidades verdadeiras. ii) Então é possível definir um estimador (embora não seja não-viciado) que nos dê, para baixos valores, uma probabilidade pequena de que uma determinada freqüência seja uma verdadeira componente de Fourier da série original, embora valores altos do estimador não signifiquem necessariamente que o valor é de uma freqüência verdadeira. Este é o significado de *falsear* possíveis soluções num subconjunto do espaço paramétrico. Significa que, para cada função de estado de informação amostral, mais do que procurar uma 'melhor solução', devemos procurar classes de não-admissibilidade dentro de uma região do espaço paramétrico. Daí então, combinamos estas informações e achamos os mencionados 'invólucros' que englobam tais classes de não-admissibilidade no espaço paramétrico, comuns e não-comuns às séries amostrais. O complementar desses invólucros aponta as regiões de alta correlação comuns a todas as séries e onde deve estar certamente o 'sinal' original, os 'efeitos' espectrais ligados ao vazamento de potência e falseamento, bem como todos os pontos do espaço paramétrico onde não há resolução. Portanto a informação final se aproximaria da distribuição *a priori*.

Estritamente falando, o maior perigo deste tipo de procedimento está na discretização do espaço amostral, que no nosso caso está intimamente ligada ao procedimento semi-empírico de suavização dos modelos. Devemos proceder com cautela, pois uma grade escolhida muito densa adicionará muito ruído ao modelo final, enquanto uma grade escolhida com poucos pontos deixará de 'ver' as feições desejadas. O 'Teorema da amostragem ou de Nyquist' nos ensina que não devemos tentar obter informações de freqüências acima da freqüência (comprimento de onda) limite de Nyquist, o dobro da taxa de amostragem da série, que no caso de séries uniformes coincide com o dobro da *menor* distância entre dois pontos. Este comprimento de onda limite também é uma boa sugestão inicial para o incremento entre os pontos no domínio da freqüência. O perigo é que sempre poderemos incorrer no erro de deixar de considerar um 'fenômeno' que ocorre entre dois pontos da grade (como uma singularidade). Portanto, este ajuste deverá ser sempre subjetivo, considerando-se as incertezas estimadas nos dados das séries amostrais, a classe de modelos considerada (neste caso, funções seno e cosseno de Fourier) e a resposta esperada do fenômeno estudado (modelo *a priori*).

Com isto em mente, o procedimento acima descrito tem se mostrado muito eficiente, não para a busca da 'melhor solução', mas para descartar uma ampla classe de modelos como não aceitáveis a todas as séries amostrais em comum. Como complementar, obtemos um conjunto-solução no qual a resposta certamente deve estar contida. Além disso, a incerteza nos resultados é visualizada diretamente nas distribuições obtidas.

Agradecimentos

Os autores agradecem o apoio da FAPESP, através de bolsa de doutoramento a G.C.M..

Referências

Lomb, N.R., 1976. Least-squares frequency analysis of unequally spaced data. Ap.Space Sci. 39:447-462.

Scargle, J.D., 1982. Studies in astronomical time series analysis. II. Statistical aspects of spectral analysis of unevenly spaced data. The Astrophys. J. 263:835-853.

Tarantola, A., 2006. Popper, Bayes and the inverse problem. Nature, 2, 492-494.

Tarantola, A., Valette, B., 1982. Inverse problems = Quest for information. J. Geophysics, 50: 159-170.

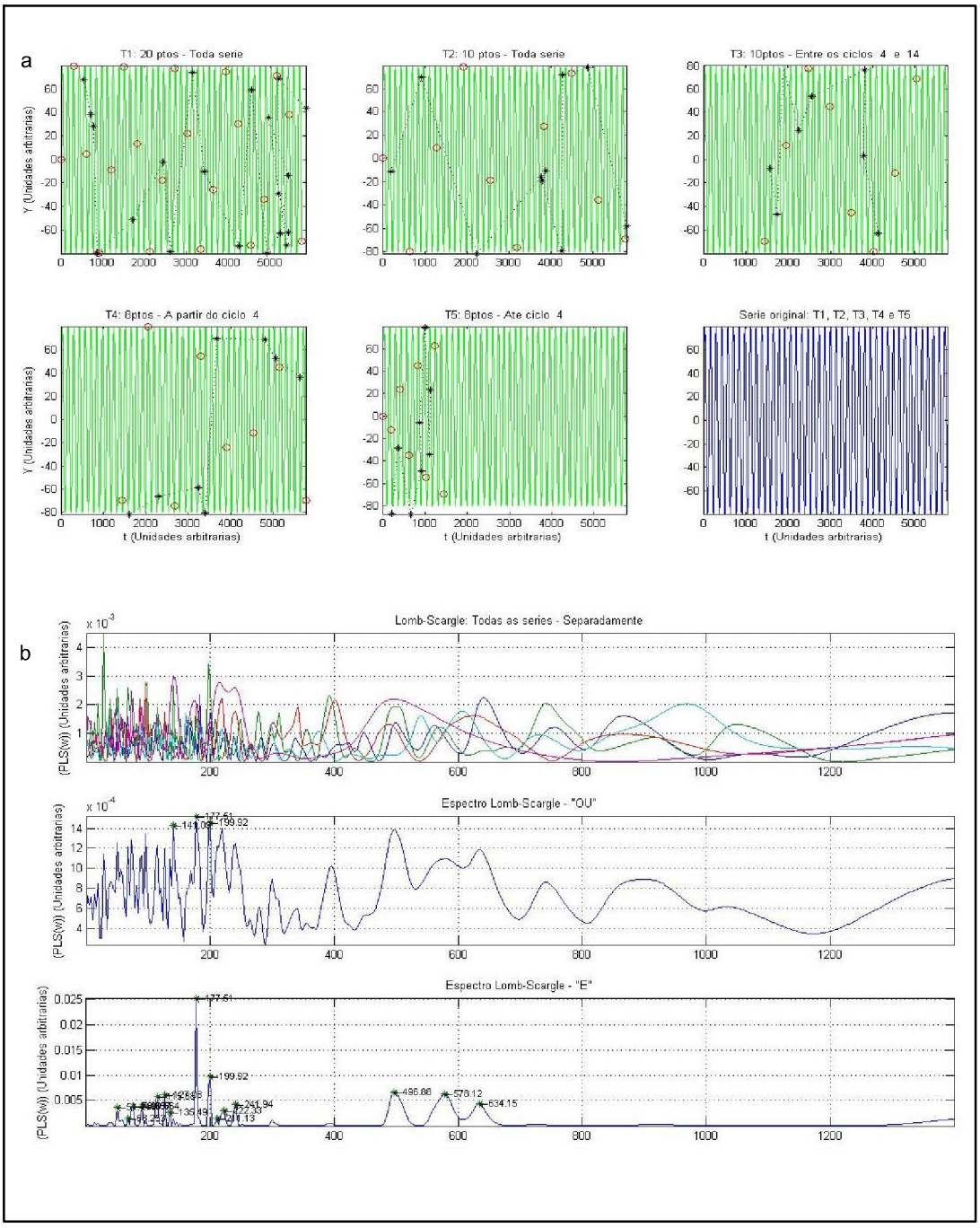

Figura 1. Superior: Série original (T=135) = curva em azul; séries amostrais com ruído gaussiano de 5% (*); (o) representa amostra de igual densidade no intervalo, sem ruído (taxa de amostragem constante). Inferior: Espectros de Lomb-Scargle para cada série individual e o resultado dos operadores OU e E.

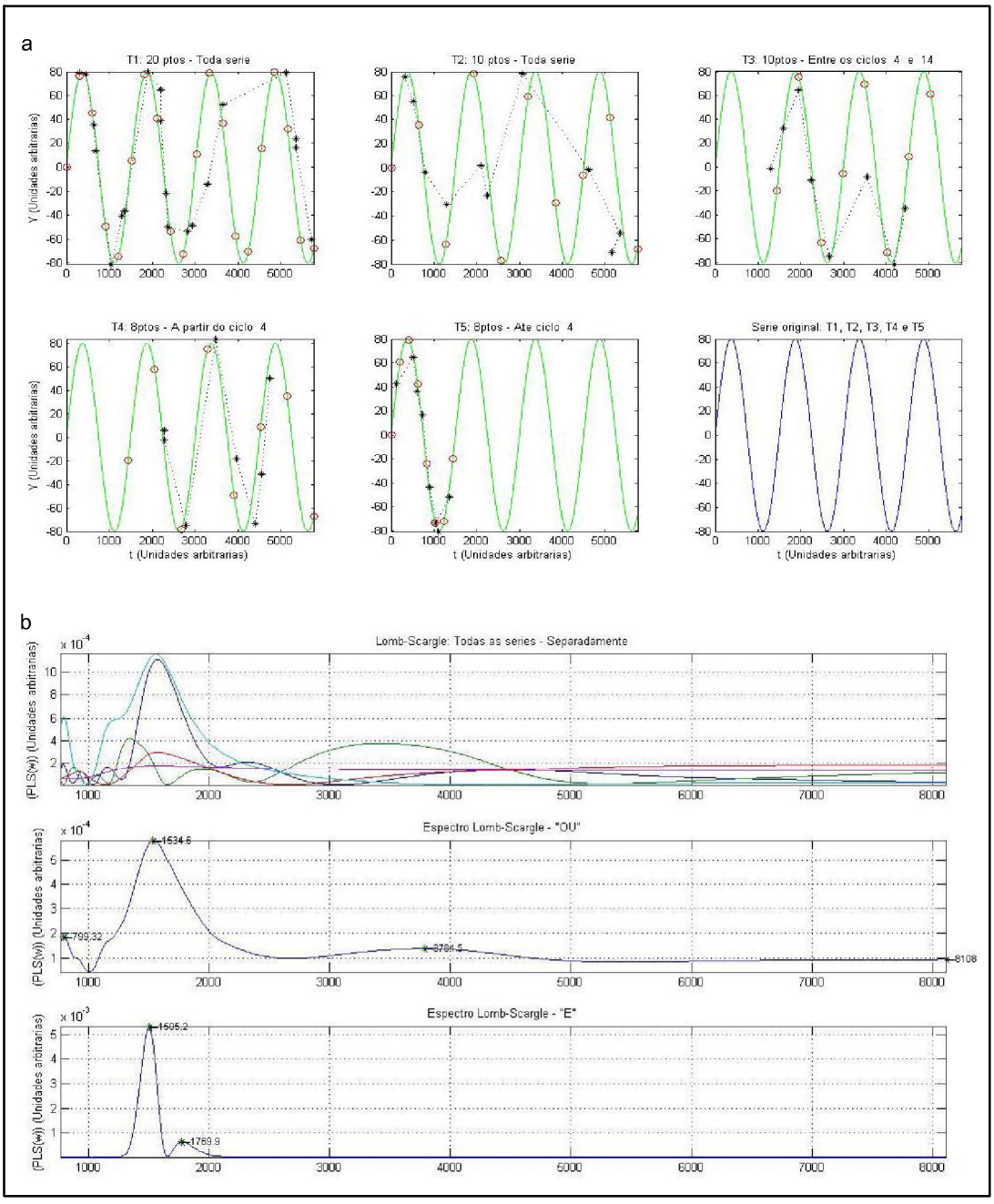

Figura 2. Superior: Série original (T=1500) = curva em azul; séries amostrais com ruído gaussiano de 5% (*); (o) representa amostra de igual densidade no intervalo, sem ruído, com taxa de amostragem constante. Inferior: Espectros de Lomb-Scargle para cada série individualmente e o resultado dos operadores OU e E.

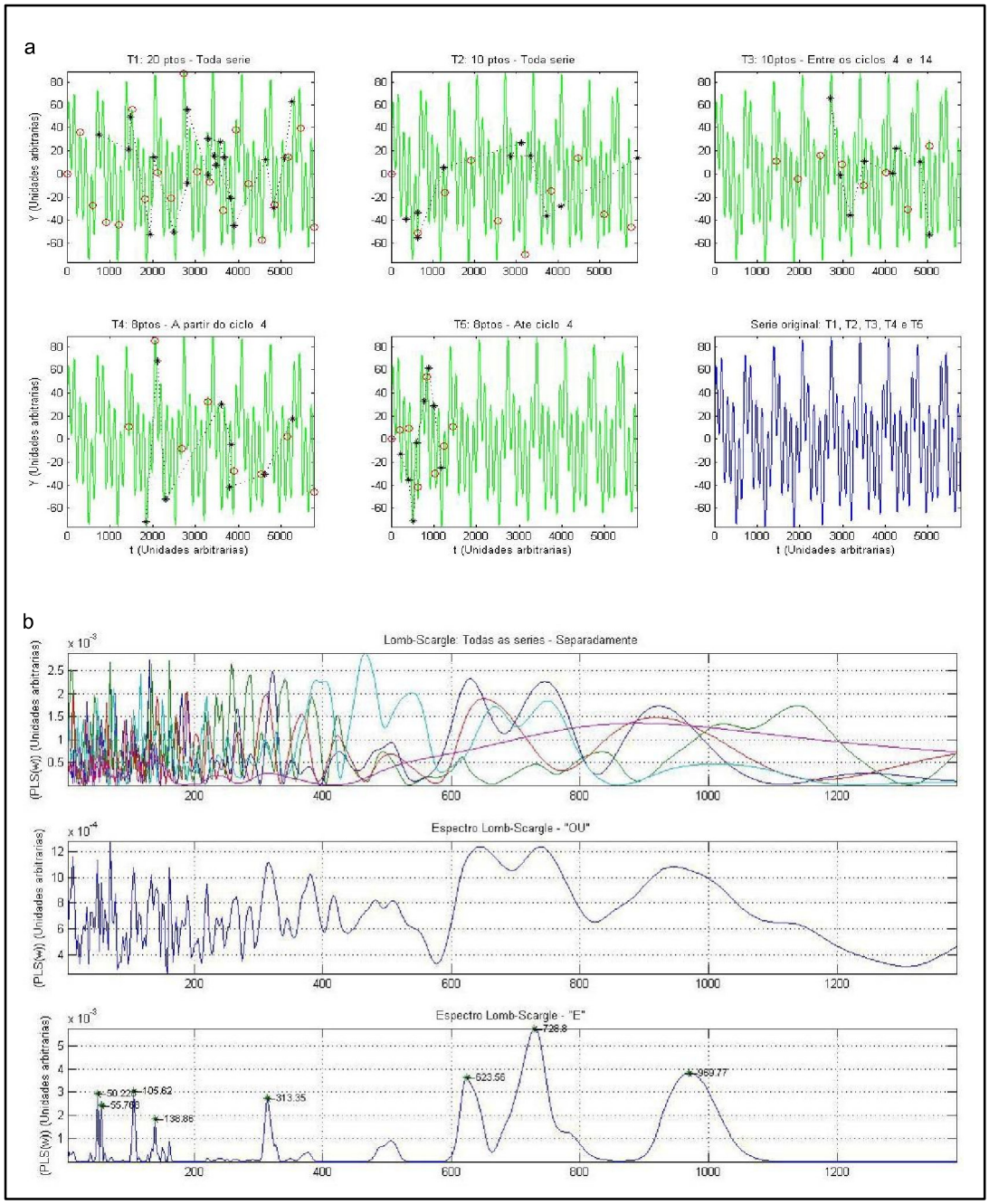

Figura 3. Superior: Série original (T=135 +335+650) = curva em azul; séries amostrais com ruído gaussiano de 5% (*); (○) representa amostra de igual densidade no intervalo, sem ruído, com taxa de amostragem constante. Inferior: Espectros de Lomb-Scargle para cada série individual e o resultado dos operadores OU e E.