António Jorge Teixeira Falcão

Licenciado em Engenharia Informática

Detecção de Correlação e Causalidade em Séries Temporais não Categóricas

Dissertação para obtenção do Grau de Mestre em
Engenharia Informática

Orientador: Joaquim Francisco Ferreira da Silva, Professor Auxiliar, Faculdade de Ciências e Tecnologia, UNL

Co-orientador: Ivan Dorotovič, Ph.D., Observatório Astronómico Central Hurbanovo, Eslováquia

Júri:

Presidente: Prof. Doutor Adriano Martins Lopes
Arguente(s): Prof. Doutora Rita Almeida Ribeiro
Vogal(ais): Prof. Doutor Joaquim Francisco Ferreira da Silva

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Junho 2012

Universidade Nova de Lisboa
Faculdade de Ciências e Tecnologia
Departamento de Informática

Dissertação de Mestrado em Engenharia Informática
1º Semestre, 2011/2012

Detecção de Correlação e Causalidade em Séries Temporais não Categóricas

Nº 30007 – António Jorge Teixeira Falcão

Orientador
Prof. Doutor Joaquim Francisco Ferreira da Silva

Co-Orientador
Ivan Dorotovič, Ph.D.

Junho de 2012

Nº do aluno: 30007
Nome: António Jorge Teixeira Falcão

Título da dissertação:
Detecção de Correlação e Causalidade em Séries Temporais não Categóricas

Palavras-Chave:

- Séries temporais
- Correlação
- Auto-correlação
- Detecção de periodicidades
- Causalidade

Keywords:

- Time-series
- Correlation
- Auto-correlation
- Periodicity detection
- Causality

Agradecimentos

Começo por agradecer o meu orientador, o Prof. Doutor Joaquim Silva, por todo o seu apoio e em especial a sua paciência e disponibilidade ao longo deste "percurso" do mestrado. Os meus sinceros agradecimentos.

Em relação ao caso de estudo e à validação dos resultados, contei com a valiosa colaboração do Dr. Ivan Dorotovič, astrofísico solar, com doutoramento na área do ciclo da actividade da corona solar, e autor de múltiplas publicações na área. Dakujem, Ivan.

À Prof.ª Doutora Rita A. Ribeiro, um especial obrigado. Agradeço a oportunidade de fazer parte de um grupo de investigação ligado a uma área de aplicação que me fascina, e que tanta experiência me tem proporcionado ao longo destes anos.

Calorosos cumprimentos a todos membros actuais e passados do grupo CA3, que animaram sempre essa experiência.

Refiro que durante os trabalhos preliminares da tese foram utilizados dados do observatório de Lomnický štít, suportado pelo IEP SAS em Košice pela agência Eslovaca APVV.

Aos meus pais, que fomentaram a minha curiosidade pela ciência em criança.

À minha «Chica Adorada», por ser tão especial.

Resumo

As séries temporais estão presentes em múltiplos domínios do nosso quotidiano – áreas tão distintas como a astronomia, geofísica, economia, medicina, entre outras. As tecnologias de informação actuais têm a capacidade de gerar grandes quantidades de dados, representando séries temporais. Para extrair informação, e consequentemente gerar conhecimento, a partir de uma quantidade tão vasta de dados, torna-se necessário recorrer a técnicas para automatizar a análise destes dados de uma forma exequível e eficiente.

Com esta tese pretende-se contribuir especificamente para a análise de séries temporais não categóricas, mais concretamente de valores numéricos reais, com um conjunto de ferramentas que auxiliem na detecção de correlações entre múltiplas séries temporais e na detecção de possíveis periodicidades existentes. Para além dos métodos conhecidos de correlação, desenvolveu-se uma variante aplicada à detecção de picos nas séries de modo a lidar com determinados tipos de parâmetros, com resultados muito positivos.

No âmbito da tese, foi também desenvolvida uma metodologia de modo a determinar relações de causalidade entre variáveis. Esta permite detectar situações de causa-efeito a partir de séries temporais não categóricas. Esta dissertação fica assim a focar duas partes; uma onde se aborda o tema da correlação entre séries temporais, e outra onde se trata da questão da causalidade existente entre elas.

Como caso de estudo, utilizou-se o domínio da astrofísica solar, analisando séries temporais provenientes de parâmetros solares. Não obstante, manteve-se o objectivo de os métodos e ferramentas resultantes poderem ser aplicados a qualquer domínio expresso em séries temporais, pelo que não foram introduzidos nos algoritmos factores relativos a domínios específicos.

Palavras-chave: Séries temporais, Correlação, Auto-correlação, Detecção de periodicidades, Causalidade

Abstract

Time series are present in many areas of our daily lives - areas as diverse as astronomy, geophysics, economics and medicine, among others. Information technologies currently have the ability to generate large amounts of data, in part represented as time-series. Analysing the huge amount of generated data is a task that is exceeding human capabilities. To extract information and therefore generate knowledge from such a vast amount of data, it is necessary to use techniques to automate the analysis of these data efficiently.

This thesis aims to contribute specifically to the analysis of non-categorical time-series (i.e., numeric values), with a set of tools that aid in the detection of correlations among multiple time series, and the detection of periodicities associated to them. Besides using known correlation methods, a variant applied to peak detection was developed in order to handle certain types of parameters, with positive results.

Within the scope of the thesis, a methodology for determining causality between parameters was developed. It allows detecting cause-effect relationships from non-categorical time-series. As such, this dissertation focuses on two parts: the topic of correlation detection between time-series, and the matter of causality relationships between them.

Solar astrophysics shall be the main focus as case study, analysing time series from solar parameters. Nevertheless, it is intended that the methods and tools resulting from this approach can be applied to any domain expressed in time-series, so no artefacts related to specific domains were introduced into the algorithms.

Keywords: Time-series, Correlation, Auto-correlation, Periodicity detection, Causality

Índice

Índice de Figuras

Índice de Tabelas

1. Introdução

Foi estimado que em 2005 a nível mundial foram gerados 150 exabytes[1] de dados. Para o ano de 2010 estimou-se um valor próximo dos 1200 exabytes. No caso particular de um telescópio, o VST (VLT Survey Telescope do European Southern Observatory), estima-se que a câmara, constituída por 32 CCD's totalizando uma imagem de 268 megapixéis, irá produzir 30 terabytes de dados anualmente.

Analisar esta enorme quantidade de dados gerada é uma tarefa que começa a ultrapassar as capacidades humanas. São necessários mecanismos para a detecção automática de padrões e relações.

Com essa realidade em mente, esta tese propõe contribuir com um conjunto de mecanismos para auxiliar a tarefa, focando as séries temporais numéricas não categóricas.

Os objectivos resumem-se à:

- detecção de correlações positivas e negativas, em conjuntos de parâmetros em número superior a 1;
- detecção, por interface simples, de correlações tendo em conta desvios temporais;
- detecção, por interface simples, de periodicidades em séries temporais;
- determinação de causalidade entre parâmetros/séries temporais;

Quanto à detecção de periodicidades, existem várias técnicas e são largamente utilizadas, tais como as transformadas de Fourier e as *wavelets*. Não obstante, nesta tese pretendeu-se apresentar uma abordagem alternativa, por um lado baseada na definição de correlação de Pearson (apresentado na secção 2.1.1.3), computacionalmente mais simples, e por outro, disponibilizando uma visualização muito mais intuitiva do que as abordagens atrás referidas.

[1] Exabytes = 10^{18} bytes (equivalente a 1 000 000 terabytes)

A determinação de causalidade entre parâmetros surgiu no decorrer da tese, não fazendo parte dos objectivos iniciais. Detectar a causalidade a partir de dados categóricos é uma tarefa relativamente fácil a partir de técnicas conhecidas; o mesmo não se pode dizer em relação a dados não categóricos em séries temporais. Propõe-se no âmbito desta tese um processo simples e eficiente na detecção de causalidades, não sendo baseado em métodos de regressão linear, e não dependendo por isso, da qualidade dessas regressões. O método desenvolvido revelou-se como sendo consistente nos resultados produzidos, sendo validado em vários casos de domínios diferentes (astronomia, meteorologia, entre outros).

Como caso de estudo, utilizou-se o domínio da astrofísica solar, analisando séries temporais provenientes de parâmetros solares. Algumas periodicidades e correlações neste domínio são já conhecidas [1], [2], [3], o que constituiu uma boa base de validação da abordagem. Com ela, espera-se proporcionar a oportunidade de novas descobertas. Tive a felicidade da estreita colaboração de um perito na área da astrofísica, que me deu a validação necessária e orientação relativamente ao conjunto de dados a explorar para os testes inicias. Não obstante este domínio ter sido tomado como caso de estudo, pretende-se que os métodos e ferramentas resultantes desta abordagem possam ser aplicados a qualquer domínio expresso em séries temporais, pelo que não foram introduzidos nos algoritmos factores relativos a domínios específicos.

O protótipo desenvolvido no âmbito desta tese foi apresentado em conferência internacional [4] - JENAM2010 – Joint European and National Astronomy Meeting, recebendo comentários positivos por parte da comunidade.
Nesta dissertação, o conteúdo relativo à parte da detecção da causalidade em séries temporais foi submetido e aceite para publicação, prevista para Abril de 2012, pela Springer em *book-chapter* no livro Behaviour Computing.

1.1 Motivação

A motivação por detrás desta tese prende-se com a grande quantidade de diferentes parâmetros e volume de dados com as quais os cientistas são hoje confrontados. Isto leva a um trabalho moroso e dificultado pela inexistência de ferramentas de fácil utilização e adequadas ao tratamento dos dados. A detecção de correlações, periodicidades e outros

padrões em séries temporais não categóricas é na maior parte dos casos feita de uma forma manual, por análises gráficas das respectivas séries.

Um dos elementos motivadores desta tese consistiu no desenvolvimento de um conjunto de ferramentas que permitisse de forma automática e expedita analisar uma grande quantidade de parâmetros, assinalando potenciais correlações e padrões em séries temporais, cuja utilidade esperava-se viesse a ser apreciada pela comunidade.

Tendo em conta o caso de estudo, tornava-se desde logo importante que como resultado da tese fosse produzido algo facilmente aplicável ao ambiente de trabalho dos astrónomos, sem necessidade de conhecimento aprofundado dos algoritmos / métodos envolvidos, e que fosse computacionalmente eficiente, produzindo bons resultados em pouco tempo.

1.2 Contexto

Uma série temporal é uma colecção de valores de um parâmetro, feita sequencialmente ao longo do tempo, normalmente em intervalos equitativamente espaçados [5]. Nesta tese, o âmbito será em séries temporais não categóricas, mais concretamente em séries temporais onde os parâmetros apresentam valores numéricos reais.

O caso de estudo para efeitos de aplicação e validação da tese é na área da astrofísica solar, como já foi referido. A Figura 1.1 mostra um exemplo de dois parâmetros do domínio do caso de estudo. Trata-se de medições de fluxo de electrões provenientes do sol e medidos por dois satélites (GOES11 e GOES12) ao longo do tempo, indicado no eixo das abcissas em "número de amostra" (com taxa de amostragem de 5 minutos). Os dados foram obtidos a partir do SWENET – Space Weather European Network [6] em formato de texto; o respectivo gráfico foi gerado para uma melhor leitura e interpretação dos dados.

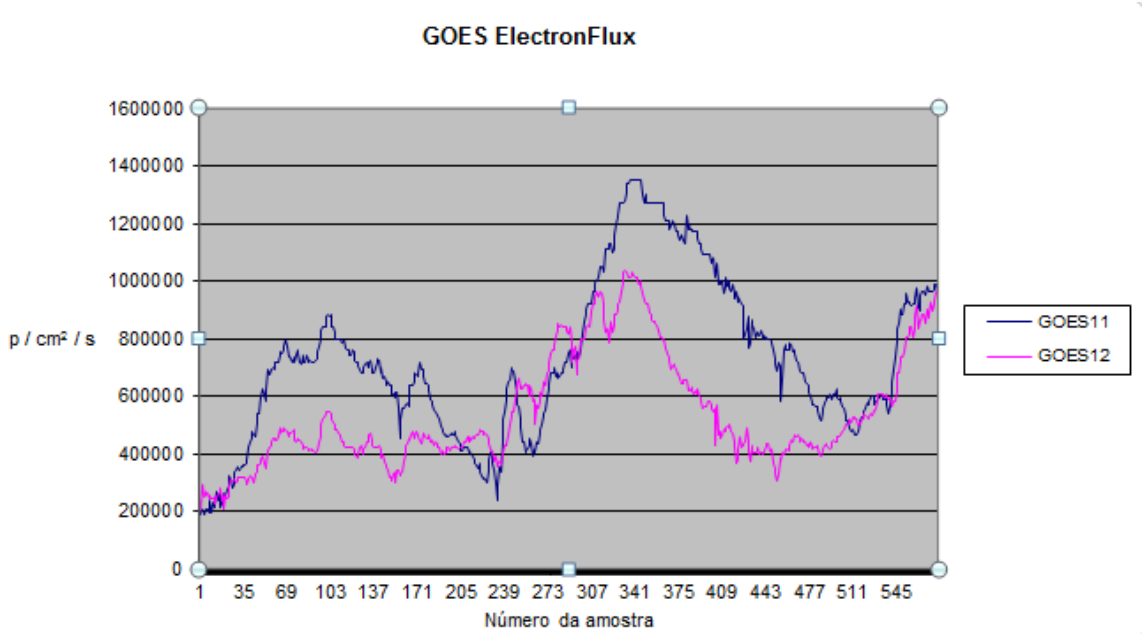


Figura 1.1 Parâmetros correlacionados

(O eixo das abcissas indica o número da amostra, e o eixo das ordenadas indica o fluxo de electrões em partículas / cm^2 / s)

A Figura 1.2 ilustra uma situação de onde se podem observar parâmetros correlacionados, mas com um desvio temporal associado. No protótipo desenvolvido foi possível encontrar, de forma automática, o valor do desvio entre estes dois parâmetros.

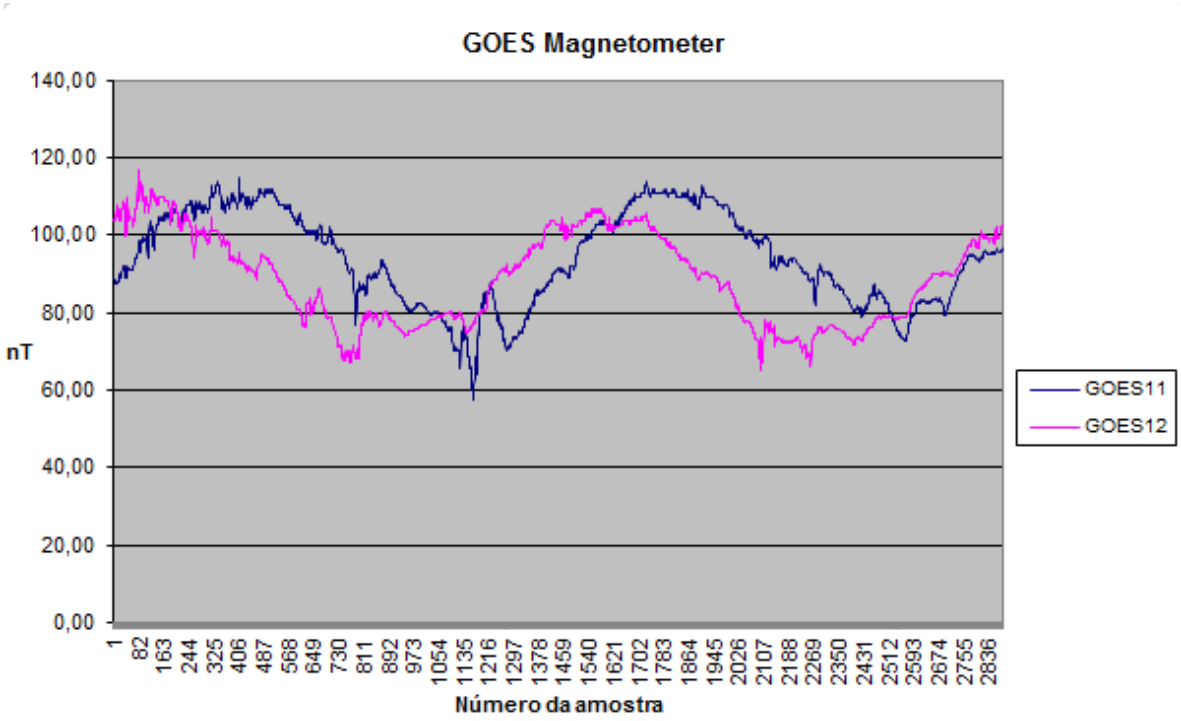


Figura 1.2 Parâmetros correlacionados com desvio temporal associado

(O eixo das abcissas indica o número da amostra, e o eixo das ordenadas indica o valor medido pelo magnetómetro em nanoTesla – nT)

Na Figura 1.3, podemos observar outra situação de interesse no caso de estudo: o caso de um parâmetro com uma periodicidade associada. Neste caso trata-se do número de manchas solares (*solar sunspot number*) observadas no disco solar. Dados retirados do SIDC – Solar Influences Data Analysis Center [7], do Observatório Real da Bélgica, para a geração do gráfico.

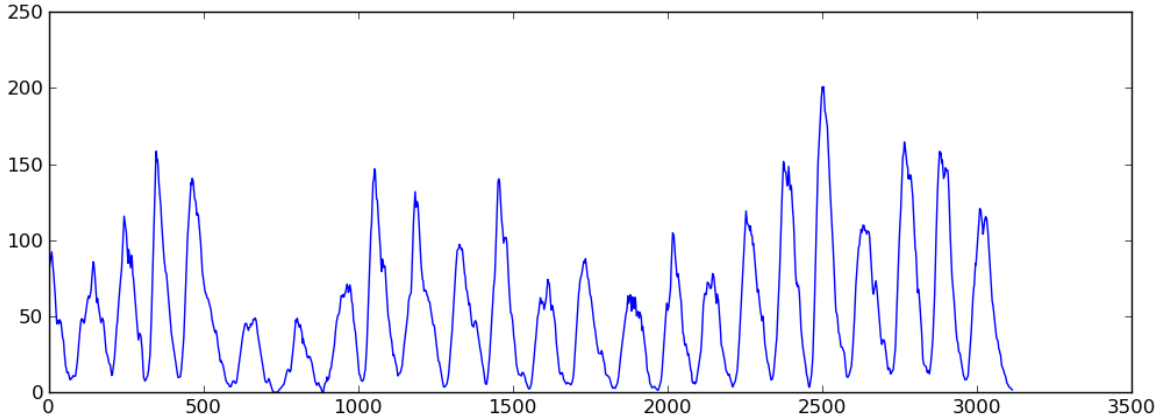


Figura 1.3 Parâmetro com periodicidades

(O eixo das abcissas indica o número da amostra, e o eixo das ordenadas corresponde ao valor da média mensal de manchas solares)

Cada amostra no eixo das abcissas é tirada mensalmente, iniciando-se no ano de 1749, data a partir da qual se considera que os dados são fiáveis e normalizados.

De seguida, podemos ver uma imagem da Agência Espacial Europeia (European Space Agency - ESA) onde figuram várias perturbações causadas pela actividade solar, desde efeitos a nível do espaço extraterrestre, efeitos a nível da atmosfera, ou até efeitos que se observam no solo. Estes efeitos podem influenciar não só satélites que se encontrem em órbita e as respectivas transmissões, mas também equipamento que se encontra no solo, como por exemplo linhas de alta-tensão. Em situações extremas, pode-se dar a incidência excessiva de radiação em passageiros que se encontrem em voos internacionais.

Dada a sua importância, o estudo do *Space Weather* é agora foco de variados programas, incluindo o "Space Situational Awareness Programme" da ESA [8].

O termo *Space Weather* ("meteorologia espacial") refere-se a condições no Sol e no ambiente do espaço que podem influenciar o desempenho e a fiabilidade de engenhos espaciais ou sistemas tecnológicos em terra, e que pode colocar em perigo a saúde e a vida humana [9].

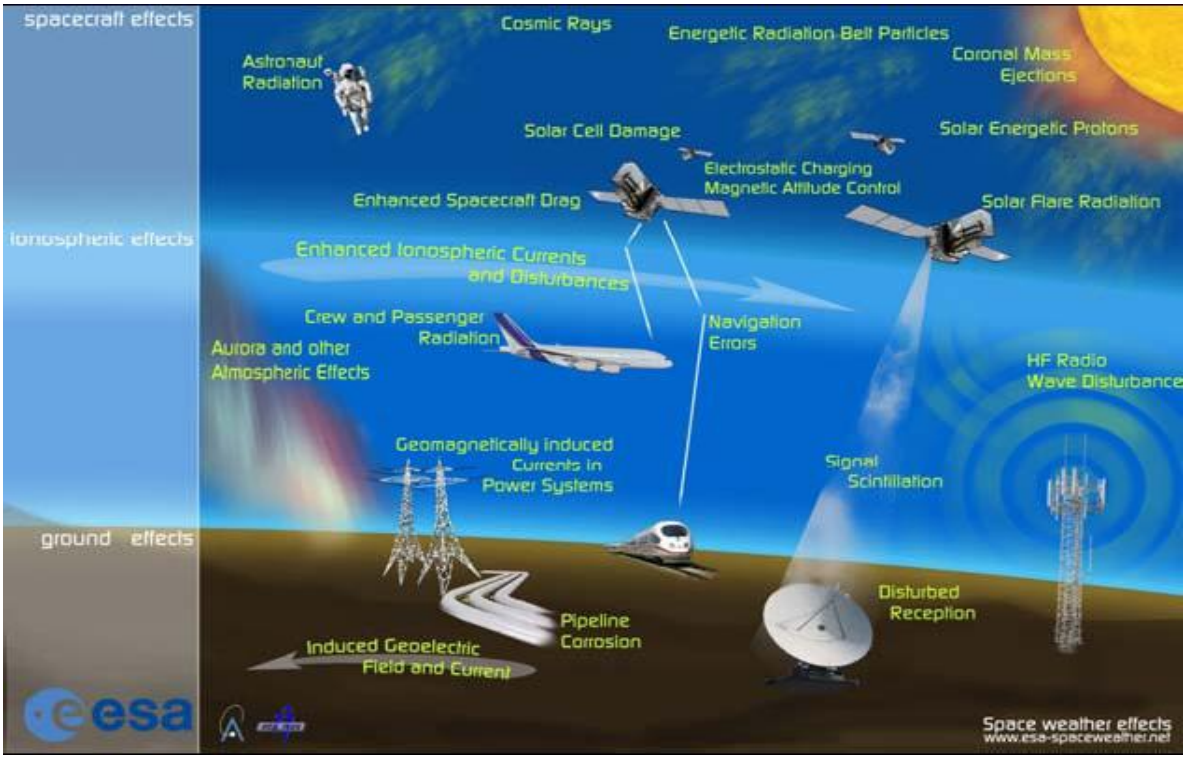


Figura 1.4 Efeitos nocivos da actividade solar

Muitos fenómenos no Sol, no espaço interplanetário e na proximidade da Terra, desempenham um papel importante no sistema sol-terra, exibindo um grande conjunto de relações complexas. O forte impacto da actividade solar no espaço interplanetário e no meio ambiente da Terra (atmosfera, biosfera) é geralmente bem conhecido. O sol produz radiação em todos os comprimentos de onda através de uma variedade de mecanismos e de uma grande multiplicidade de contextos físicos. A energia da radiação solar, juntamente com o plasma e as partículas carregadas do vento solar, regulam os processos na atmosfera e na superfície terrestre. Assim, o Sol é o principal impulsionador do clima espacial.

Um dos temas de investigação do co-orientador da tese é no estudo da radiação cósmica. Radiação cósmica é um fluxo de partículas energéticas vindas do espaço, e é constituída por radiação cósmica primária e secundária. Partículas de radiação cósmica primária não chegam à superfície da Terra - transformam-se em partículas de radiação cósmica secundária ao colidir com núcleos de átomos na atmosfera, sendo estas partículas secundárias registadas por detectores de neutrões no solo. O Sol é também uma fonte de radiação de partículas, mas muito variável, tanto no fluxo como nos níveis energéticos. É geralmente sabido que existe uma correlação negativa forte entre o nível de actividade solar e a radiação cósmica. São observadas variações mais ou menos regulares da

radiação cósmica, relacionadas com o ciclo de actividade solar (11 anos), a duração da rotação solar (27 dias) ou até relacionado com a rotação da Terra (24 horas). Estas variações da radiação cósmica são por vezes interrompidas por ocorrências de diminuições súbitas no sinal, com uma recuperação lenta, a chamada *Forbush Decrease* (explorado na secção 2.2.3).

Para resolver estas questões fundamentais da física solar e ciência espacial, existem em curso várias missões espaciais e adicionalmente vários missões programadas, como por exemplo, o Solar Probe Plus (NASA) e o Solar Orbiter (ESA). Todos estes instrumentos proporcionam / proporcionarão uma grande quantidade de dados e/ou imagens para estudos no âmbito de *Space weather*, que leva à necessidade de ferramentas eficientes para estudos estatísticos. Uma melhor compreensão dos processos de actividade solar, tais como erupções solares, manchas (*sunspots*), filamentos e ejecções de massa coronal (*coronal mass ejections*), contribuirão para melhorar o conhecimento sobre a evolução dos ciclos solares, os modelos de previsão do *Space weather* e, espera-se, proporcionar melhores sistemas de alerta antecipada.

1.3 Contribuições

No âmbito do trabalho desta tese, tendo em conta o caso de estudo, procurou-se em primeiro lugar responder às necessidades dos físicos solares. Como primeiro passo, foi necessário estudar a forma de trabalho actual destes cientistas. É importante compreender a forma como lidam com os dados, os métodos e as ferramentas utilizadas.

Para procurar responder às suas necessidades, foram propostas ferramentas para lidar com o tipo de problema estudado. Foi proposto a aplicação de uma medida que permite a identificação de parâmetros de séries temporais correlacionados, num domínio de uma grande quantidade de parâmetros e dados. Utilizando essa mesma métrica, propôs-se um método alternativo, prático e intuitivo de detecção de periodicidades.

Um ponto importante consistiu em tornar este conjunto de ferramentas / métodos de utilização intuitiva, sem que fosse necessário adquirir conhecimentos sobre a sua implementação para tirar melhor partido dele. Adicionalmente, teve-se a preocupação de as manter independentes do domínio, sem introduzir nenhuma especificidade do tema, e assim possibilitando a sua aplicação directa noutras áreas. Houve também a implementação de um protótipo simples disponibilizando uma interface amigável para a

realização das suas tarefas, incluindo métodos de visualização das séries temporais e resultados da aplicação dos métodos desenvolvidos.

Em síntese, esta tese contribui para os seguintes tópicos:

- **Detecção automática de correlações em séries temporais**: contribuição com um conjunto de métodos e ferramentas que permitam a detecção automática e quantificação de correlações em conjuntos de séries temporais não categóricas. Esta detecção de correlações inclui tanto correlações positivas como negativas. Foi também possível a detecção de correlações em séries temporais com um desvio temporal associado.
- **Detecção de periodicidades em séries temporais**: desenvolvimento de um método baseado no Coeficiente de Correlação de Pearson para detecção de periodicidades, proporcionando uma visualização de leitura simples e intuitiva das suas componentes, relativamente à frequência e à amplitude.
- **Determinação de causalidade entre dois parâmetros:** propõe-se um mecanismo para obter o sentido de causa-efeito entre dois parâmetros de séries temporais não categóricas de valores reais.
- **Desenvolvimento de um protótipo intuitivo e útil**: contribuição com um protótipo funcional de uma ferramenta orientada para um utilizador poder usufruir dos métodos desenvolvidos no âmbito desta tese.
- **Contribuição no domínio da física solar**: com a disponibilização das ferramentas já referidas, espera-se contribuir futuramente para a descoberta de novos conhecimentos no domínio da astrofísica solar.

1.4 Estrutura do Documento

Esta tese lida com duas temáticas: correlação e causalidade. Estas são abordadas separadamente nos capítulos 2 e 3, respectivamente. Cada um dos dois capítulos abordam o trabalho relacionado, o trabalho realizado e incluem uma discussão dos respectivos resultados obtidos. O capítulo 4 descreve o protótipo desenvolvido, e demonstra os conceitos apresentados na tese. O capítulo 5 encerra a dissertação com conclusões e propostas de trabalho futuro.

2. Correlação

2.1 Trabalho Relacionado

O tipo de dados em estudo nesta tese assenta em variáveis contínuas. Pretende-se acima de tudo medir correlações/dependências entre parâmetros/variáveis. Exemplos de parâmetros, tendo em conta o nosso caso de estudo são: *Solar Rádio Flux*, *Sunspot Number*, *Neutron Counts*, etc. Uma vez que a correlação e a independência são conceitos estreitamente relacionados, numa primeira pesquisa consideraram-se as abordagens baseadas em probabilidades condicionais, tipicamente usadas para testar dependências. Uma destas abordagens largamente conhecida e usada é o Teorema de Bayes:

$$P(A\,|B) = \frac{P(B|A\,)P(A)}{P(B)} \qquad (1)$$

Este teorema permite calcular a probabilidade de um evento *A* dada a observação B, em função das probabilidades *a priori* P(*A*) e P(*B*), da probabilidade *a posteriori* de *B* condicionada pelo evento *A*.

Se for conhecida a distribuição probabilística dos dados, o Teorema de Bayes tem um desempenho eficiente [10].

Tal como outras abordagens baseadas em probabilidades condicionais, o teorema permite avaliar a dependência entre variáveis categóricas, por exemplo, "Estado do Tempo", "Ida à praia", etc. Para tal, estas variáveis são instanciadas com domínios limitados de valores: "Chuvoso, "Nublado" e "Solarengo", para a primeira variável; e "Sim" e "Não" para a segunda variável (Ida à praia). No entanto, não faria sentido tentar medir a probabilidade do valor de "Solar Rádio Flux" ser exactamente igual a "1016.2546" num certo instante, já que, se este valor ocorrer, provavelmente não se repetirá. Por outro lado, a hipótese de tentar formar pequenos domínios de valores para transformação das variáveis contínuas

em categóricas, seria inviável nos casos em que se desconhece a natureza dos dados das séries temporais, não conhecendo por isso quais os limites naturais (ou que fariam sentido) para cada grupo de valores.
Dada a não aplicabilidade deste tipo de abordagens ao domínio de trabalho nesta tese, foi necessário encontrar uma alternativa adequada.

2.1.1 Medidas de Correlação

Considerou-se a utilização de métricas de correlação existentes nos métodos estatísticos mais ou menos clássicos, dado que são adequados ao tratamento de valores contínuos.
A correlação é uma das possíveis relações estatísticas entre variáveis aleatórias. A correlação entre duas variáveis num contexto de séries temporais mede o grau de concordância relativamente ao sentido da evolução dos valores assumidos por cada variável ao longo do tempo. Assim sendo, foram analisadas as seguintes métricas: correlação de Kendall, correlação de Spearman, e o coeficiente de correlação de Pearson.

2.1.1.1 Correlação de Kendall (*Kendall Rank Correlation Coefficient*)

Também conhecida como o coeficiente τ de Kendall, originalmente proposto em [11], é uma medida da dependência estatística entre duas variáveis usando para tal o *rank* dos valores assumidos por cada variável.
O coeficiente é dado por:

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n\,(n-1)} \qquad (2)$$

Onde n_c corresponde ao número de pares concordantes, n_d ao número de pares discordantes e n o número de elementos assumidos pela variável. O denominador representa o número total de pares dado o conjunto de n elementos.
O par de observações $\{x_1, y_1\}$ e $\{x_2, y_2\}$, sendo x_1, x_2 instâncias de uma variável X e y_1, y_2 os valores temporalmente correspondentes a x_1, x_2 de uma outra variável Y, considera-se par concordante se verificar a seguinte condição:

$$\text{sgn}\,(x_2 - x_1) = \text{sgn}\,(y_2 - y_1) \qquad (3)$$

Caso contrário, o par será discordante.

O valor do coeficiente de Kendall assume valores reais entre -1 e 1, sendo os valores próximos de 1 correspondentes a correlações fortes; os valores negativos indicam, obviamente, correlações negativas. Valores próximos de 0 reflectem correlações fracas.
Um pequeno exemplo:

Tabela 2-1 Exemplo usando o coeficiente de Kendall

X	*Y*	*Rank (X)*	*Rank (Y)*
100	80	4	4
20	30	2	2
30	10	3	1
2	50	1	3

Neste exemplo, n_c = 2, e n_d = 2, logo, τ = 0, pelo que as duas séries não estão correlacionadas de acordo com o coeficiente de Kendall.

Este coeficiente apresenta duas desvantagens tendo em conta o propósito da tese. Em primeiro lugar, tem em conta apenas a posição relativa (*rank*) dos valores, portanto não quantificando com rigor as variações dos valores assumidos pelas variáveis.
Por outro lado, a necessidade de computar as concordâncias / discordâncias de todos os possíveis pares de valores, torna este método computacionalmente ineficiente para grandes conjuntos de dados. De notar que o número de possíveis pares é o número de combinações de *n* instâncias de cada variável, 2 a 2.

2.1.1.2 Correlação de Spearman (*Spearman's Rank Correlation Coefficient*)

O coeficiente de Correlação de Spearman, inicialmente apresentado em [12] mede o grau de dependência entre duas variáveis baseando-se no *rank* dos valores assumidos por elas. Esta métrica valoriza a proximidade entre os valores dos *ranks* de x_i e de y_i para cada par de valores $\{x_i, y_i\}$ das séries temporalmente alinhadas.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \qquad (4)$$

Onde d_i corresponde à distância dos *ranks* das duas variáveis. A correlação de Spearman assume valores reais entre -1 e 1.

Usando o exemplo anterior, temos a seguinte tabela com os valores das distâncias calculadas:

Tabela 2-2 Exemplo usando coeficiente de Spearman

X	***Y***	***Rank* (*X*)**	***Rank* (*Y*)**	d_i	d_i^2
100	80	4	4	0	0
20	30	2	2	0	0
30	10	3	1	2	4
2	50	1	3	-2	4

Aplicando a fórmula para o cálculo do coeficiente de Spearman, teremos um valor de ρ = 0.2. Este valor aponta para uma correlação relativamente fraca entre as duas séries temporais.

Igualmente ao de Kendall, este coeficiente tem a desvantagem de não ter em conta os desvios quantitativos dos valores assumidos pelas séries temporais (apenas dos seus *ranks*). Apesar de já contabilizar a distância entre os *ranks*, esta medida não reflecte a realidade dos desvios que interessa quantificar no âmbito desta tese.

2.1.1.3 Correlação de Pearson (*Pearson Product-Moment Correlation*)

Uma forma comum de medir a correlação é utilizando o Coeficiente de Pearson. Estudos recentes utilizam o coeficiente de Pearson para calcular a correlação da localização de moléculas fluorescentes [13], melhorar o alinhamento de imagens de ressonância magnética [14], e na investigação de correlações entre erupções solares e oscilações de alta frequência na atmosfera solar [15].

O coeficiente de Pearson, historicamente apresentado em [16] e [17], é vulgarmente representado pela letra grega ρ:

$$\rho_{(X,Y)} = \frac{cov\ (X,Y)}{\sqrt{cov\ (X,X)}\ .\sqrt{cov\ (Y,Y)}} \qquad (5)$$

Também na forma:

$$\rho_{(X,Y)} = \frac{cov\ (X,Y)}{\sigma_X\ .\ \sigma_Y} \tag{6}$$

Sendo:

$$cov\ (X,Y) = \frac{1}{N}\sum_{i=1}^{i=N}(x_i - \bar{x})(y_i - \bar{y}) \tag{7}$$

Onde, *N* é igual ao número de elementos da série temporal, x_i é o i-ésimo elemento da série *X*, e $\bar{x}$ o seu valor médio (analogamente para y_i e $\bar{y}$).

Ou seja, o coeficiente de correlação de Pearson calcula-se pelo quociente da co-variância das duas variáveis pelo produto dos seus desvios padrão.
A Correlação de Pearson indica o grau de relação entre duas variáveis. Tal como nas correlações anteriores, o seu valor situa-se no intervalo entre -1 e 1, sendo que o valor de 1, indica uma correlação *perfeita* positiva entre os parâmetros.

Este coeficiente revela ser uma óptima medida de correlação para aplicação ao tema em estudo. Lida com variáveis de valor real, produzindo um resultado entre -1 e 1, que indica tanto correlações negativas, como positivas. Uma das vantagens sobre os coeficientes de Kendall e Spearman, é que trata os valores directos da série, não dos seus respectivos *ranks*. Isto reflectirá uma relação mais estreita entre duas variáveis, caso exista, e é computacionalmente mais leve, uma vez que não é necessário uma ordenação prévia dos valores para obter os seus *ranks*.

2.1.2 Detecção de Periodicidades

2.1.2.1 Transformadas de Fourier

Tendo em conta o domínio contínuo, as transformadas de Fourier surgiram como uma hipótese a considerar, tendo em conta as suas diversas aplicações [18], [19] e também [20].
A Transformada de Fourier permite obter as frequências e amplitudes presentes num sinal de variável real, pelo que pode ser denominada como a *representação no domínio frequência* da função original. Esta transformada decompõe um sinal como um somatório

(série) de sinusóides, cada uma com a sua amplitude e frequência. A série de Fourier de uma função *f* define-se como:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos(n\,x) + b_n \sin(n\,x)) \qquad (8)$$

No entanto, esta abordagem mostra-se incompleta face ao que necessitamos. Com efeito, as transformadas de Fourier conseguem detectar as componentes no domínio da frequência, mas não informam quando é que cada componente ocorre no sinal.

Além do mais, seria sempre necessário encontrar uma outra ferramenta para detectar correlações entre parâmetros, sendo este um problema de outra natureza.

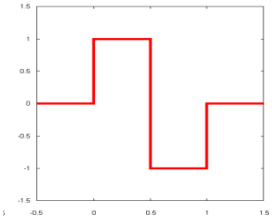
2.1.2.2 Wavelets

A incapacidade das transformadas de Fourier de fornecerem informação no domínio do tempo, levou inicialmente os investigadores a desenvolverem a *Short Time Fourier Tranformation* que, basicamente, aplica as transformadas de Fourier a janelas temporais. O desenvolvimento desta ferramenta levou à criação das *wavelets*, que são adequadas para a análise de sinais não estacionários [21], [22].

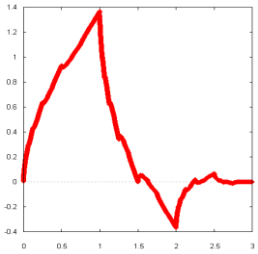
De certo modo, as *wavelets* são construídas à medida das necessidades específicas para análise de cada sinal. São conhecidas várias famílias de *wavelets*. As w*avelets* podem ser combinadas usando técnicas de soma, desvio e multiplicação (convoluções) a partir de vários sinais de modo a extrair a informação que se pretende do sinal em análise. Por outras palavras, uma *wavelet* é uma função matemática utilizada para dividir uma função ou sinal contínuo nas suas componentes de frequência e tempo.

Exemplos de algumas famílias de wavelets

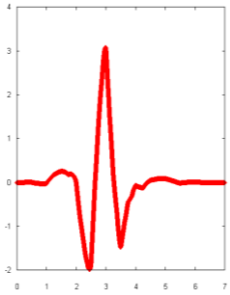
Haar Wavelet (caso particular da família de wavelets Daubechies - D1)



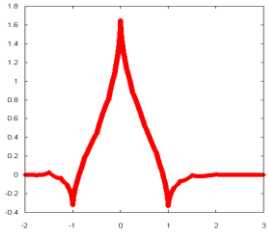
Daubechies D2



Symmlet



Coiflet K1



Apesar das potencialidades desta ferramenta, ela não se mostra de fácil utilização, no sentido em que requer do utilizador quer um conhecimento técnico da metodologia em causa, quer do problema que pretende analisar, no que toca à escolha adequada da família da *wavelet* a utilizar; esta exigência pode ser desencorajadora.

As *wavelets* tendem a ser por um lado mais informativas em certos contextos, mas por outro lado têm um preço a pagar elevado em termos de usabilidade.

Na verdade é necessário conhecer quais as famílias de *wavelets* mais adequadas para resolver os problemas específicos com que estamos a lidar. Isto é muitas vezes um motivo de desencorajamento na utilização desta técnica.

Não obstante, em certas aplicações uma utilização desta ferramenta por peritos pode revelar detalhes importantes tais como em [23]. Neste trabalho, o autor usa as *wavelets* para medir a auto-correlação entre séries temporais onde o principal objectivo se centra no estudo do ciclo solar dos 11 anos. Esta aplicação permite observar com mais detalhe outras possíveis correlações a diferentes escalas na frequência, embora com menos relevância de acordo com o autor.

Em [24] os autores apresentam uma técnica para seleccionar uma função de *wavelet* que tenha boas características para identificação perturbações de qualidade em sinais eléctricos. O próprio artigo demonstra a dificuldade existente na aplicação correcta de *wavelets* aos dados que se estão a analisar.

Estamos portanto perante uma ferramenta que para os propósitos desta tese é, por um lado, excessiva tendo em conta a mais-valia que proporciona, e por outro, de maior sobrecarga (*overhead*) por não ser de fácil exploração, quando comparada com a correlação e auto-correlação que é apresentada nesta tese.

2.1.2.3 Auto-correlação

Tendo em conta a desvantagem associada à complexidade na utilização das *wavelets*, e como consequência da utilização de medidas de correlação de dois parâmetros, verificou-se no decorrer desta tese que fazendo sucessivos cálculos de correlação de uma variável com ela própria, associando um desvio temporal, passamos a dispor de uma forma muito prática de detectar periodicidades.

Não sendo inovadora esta técnica (auto-correlação), [23], a forma de o implementar é no entanto muito mais simplificada, não sendo necessário recorrer à integração dos intervalos.

O resultado desta aplicação torna a existência de periodicidades bastante fácil de observar (ver Figura 2.10 Análise de periodicidade no S*unspotNumber* utilizando auto-correlação), permitindo analisar a frequência ao longo do tempo (não possível com transformadas de Fourier) e sem necessidade de saber detalhes das famílias de *wavelets* e respectiva aplicação.

Mais detalhes sobre a implementação deste tópico podem ser vistos na secção de trabalho realizado e na secção 4 de elaboração do protótipo.

2.2 Trabalho Realizado

2.2.1 Detecção de correlações entre dois parâmetros

O objectivo inicial desta tese foi o de encontrar de forma automática a correlação entre dois parâmetros. Como foi referido, optou-se pelo coeficiente de correlação de Pearson (secção 2.1.1.3).

A título de exemplo de aplicação, consideremos as duas figuras abaixo, representando respectivamente a medida de *NeutronMonitor* e a de *SolarRadioFlux*:

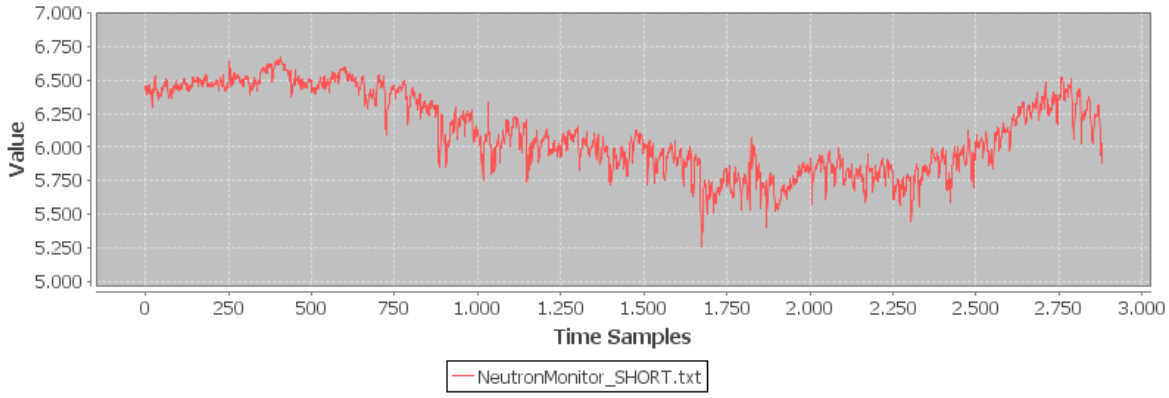


Figura 2.1 *NeutronMonitor*

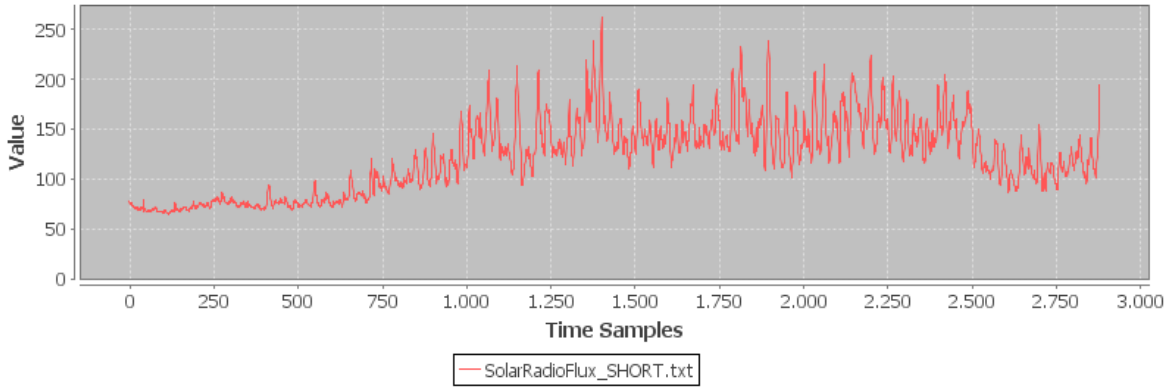


Figura 2.2 *SolarRadioFlux*

Aplicando a correlação de Pearson, obteve-se um valor de correlação de -0.79, um valor claramente significativo, e coerente com o conhecimento que os cientistas desta área têm. Muitos outros parâmetros do domínio (e não só) foram testados (e.g., valores de magnetómetros de satélites GOES, radiómetros ACE), com resultados igualmente coerentes.

Como foi mencionado, torna-se útil poder dispor da possibilidade de detectar correlações entre um grande número de pares de parâmetros. Em consequência disso, foi desenvolvida uma funcionalidade que permite visualizar em formato matriz, a correlação entre todos os possíveis pares de parâmetros, dando realce (com recurso à intensidade de cor), às correlações mais significativas. A figura seguinte mostra um exemplo:

<>	SolarRadioFlux_SHORT.txt	goes12_mag_java.txt	goes11_mag_java.txt	NeutronMonitor_SHORT.txt
SolarRadioFlux_SHORT.txt	1	-0.22930451244620687	-0.2395524131823749	-0.7931512238867362
goes12_mag_java.txt	-0.22930451244620687	1	0.46490235062139296	0.19735179246918202
goes11_mag_java.txt	-0.2395524131823749	0.46490235062139296	1	0.11373526834015732
NeutronMonitor_SHORT.txt	-0.7931512238867362	0.19735179246918202	0.11373526834015732	1

Figura 2.3 Matriz de correlações entre múltiplos parâmetros

(escala de cores variável de acordo com o valor da correlação; tons verdes indicando correlações positivas e tons azuis correlações negativas, sendo o tom mais forte quanto maior a correlação).

2.2.1.1 Detecção de correlações com desvios temporais

Por vezes, um par de parâmetros aparentemente não apresenta uma correlação significativa. No entanto, a correlação pode existir se for considerado um desvio temporal entre os dois parâmetros. Na verdade, este desvio pode dever-se a dois motivos: ou porque a influência de um parâmetro sobre o outro manifesta-se ao fim de algum tempo (relação de causa-efeito – discutido no capítulo 3 da Causalidade), ou porque se trata do mesmo fenómeno medido em locais diferentes.

As duas figuras seguintes apresentam a segunda situação. Neste caso, o valor de fluxo magnético solar é medido minuto a minuto por dois satélites distintos, que se encontram em órbitas desfasadas, dando origem ao ligeiro desvio que se observa.

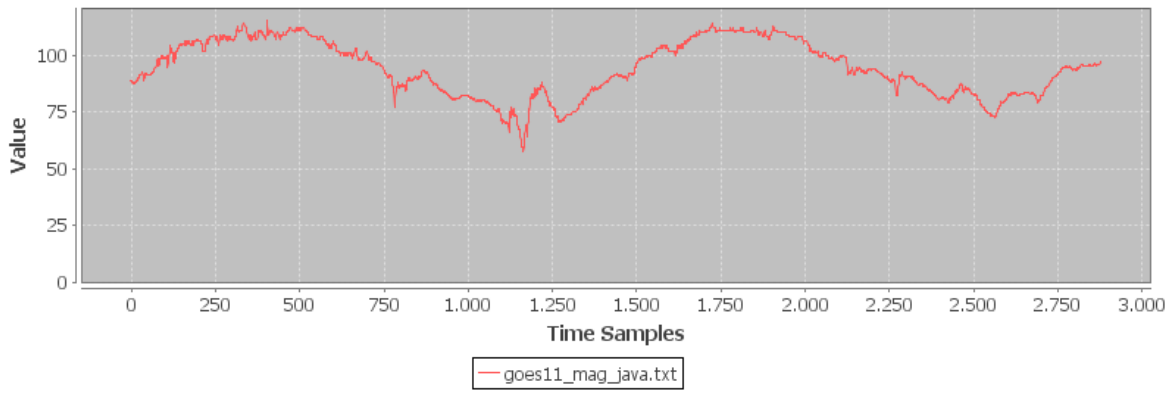


Figura 2.4 Fluxo magnético medido pelo satélite GOES11

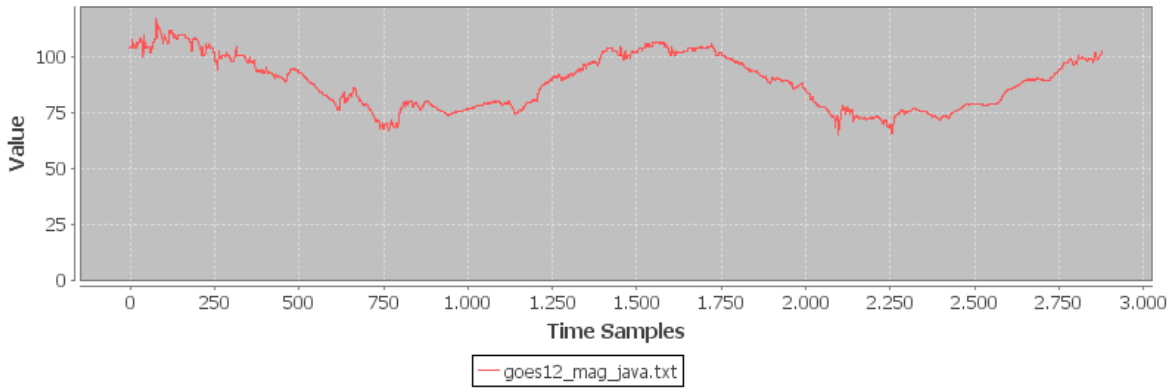


Figura 2.5 Fluxo magnético medido pelo satélite GOES12

Para abordar este problema, implementou-se uma funcionalidade que considera vários desvios temporais, apresentado num gráfico os valores de correlação para os diferentes desvios. A Figura 2.6 mostra o resultado obtido relativamente aos parâmetros representados nas figuras anteriores.

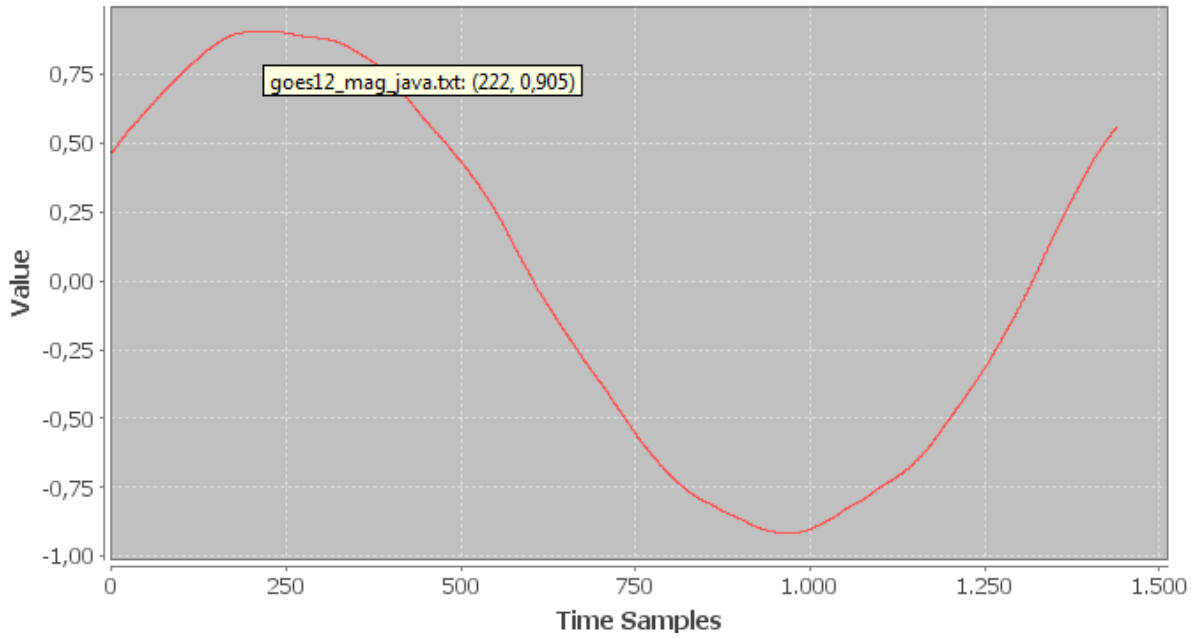


Figura 2.6 Valores de correlação com desvios temporais

Como se pode observar, o valor da correlação varia com o desvio. O valor obtido sem desvio foi de 0.46, enquanto que com o desvio 222 (desvio correspondente a 222 × 1 min = 222 min), obtém-se uma correlação de 0.905.

2.2.2 Detecção de periodicidades usando auto-correlação

No âmbito da fase de preparação da tese, e tendo sido analisada a bibliografia, decidiu-se que seria importante realizar uma série de experiências para comparar as várias técnicas possíveis para detecção de periodicidades. Como tal, no domínio do caso de estudo, optou-se por um tema habitual da análise da actividade solar – o ciclo solar de 11 anos.

O ciclo solar de 11 anos é bem conhecido. Foi descoberto por Schwabe no século XIX [25]. Este ciclo relaciona-se com a actividade solar. Uma das formas de observar esta actividade é recorrendo à contagem de manchas solares. Já apresentado no Capítulo 1, o gráfico seguinte é uma representação do número de manchas solares desde 1749 (gráfico feito na biblioteca Matplotlib na linguagem de programação Python – o eixo das ordenadas corresponde ao valor da média mensal do número de manchas solares).

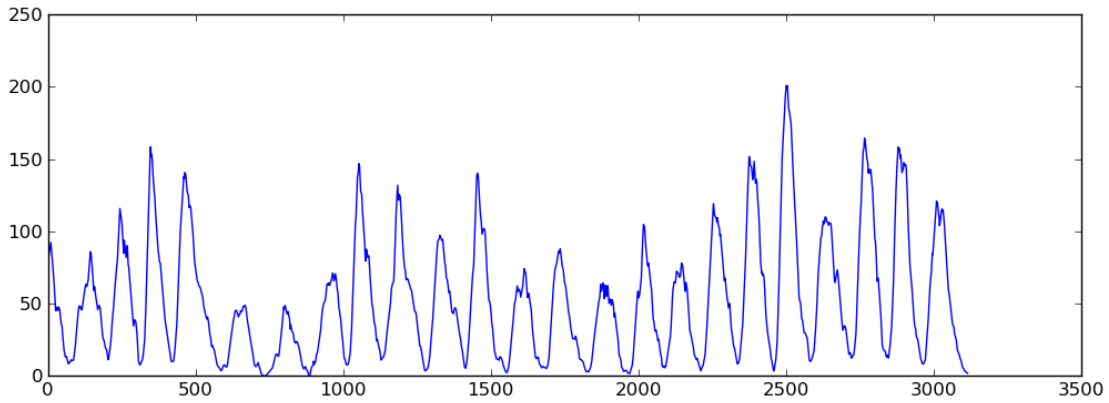


Figura 2.7 Número de manchas solares (Solar Sunspot Number)

(O eixo das abcissas indica o número da amostra, e o eixo das ordenadas corresponde ao valor da média mensal de manchas solares)

Como se pode observar, é facilmente visível a periodicidade no sinal.

Tendo feito uma análise da série temporal aplicando uma transformada de Fourier resultou no seguinte gráfico (produzido em Matlab®):

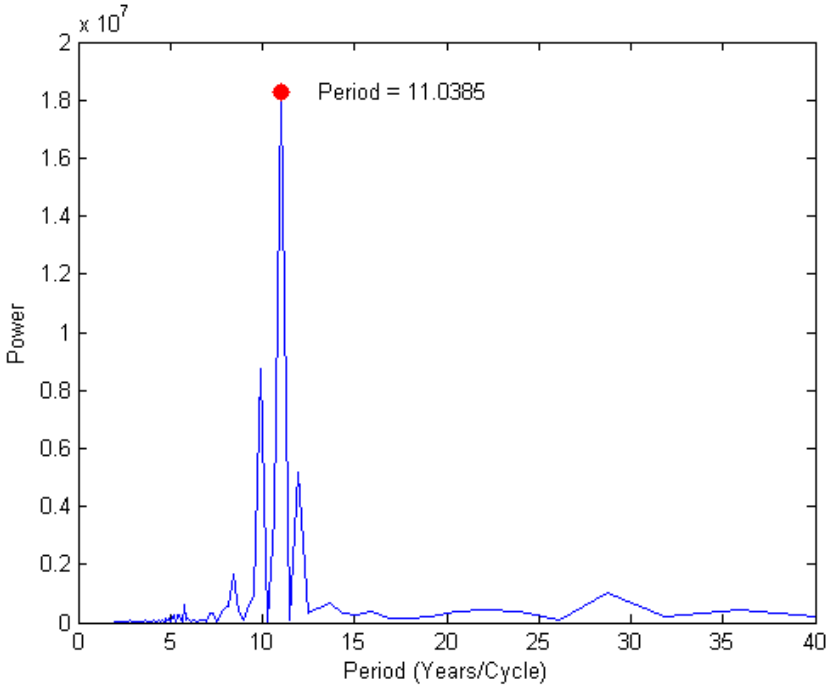


Figura 2.8 Análise do sinal periódico das manchas solares usando FFT

Pode-se observar que a frequência predominante indica um período de cerca de 11 anos tal como esperado. Apesar da facilidade de leitura do gráfico, que rapidamente indica o período detectado, não é possível localizar temporalmente esse sinal, como já foi referido.

Por outro lado, através da utilização das *wavelets*, depois de escolhida a família mais adequada, o que exige conhecimento específico, obtém-se o seguinte gráfico para o mesmo sinal (experiência realizada também em Matlab®). Como se pode observar, já é possível localizar temporalmente as variações. As periodicidades no sinal observam-se como padrões repetitivos no gráfico.

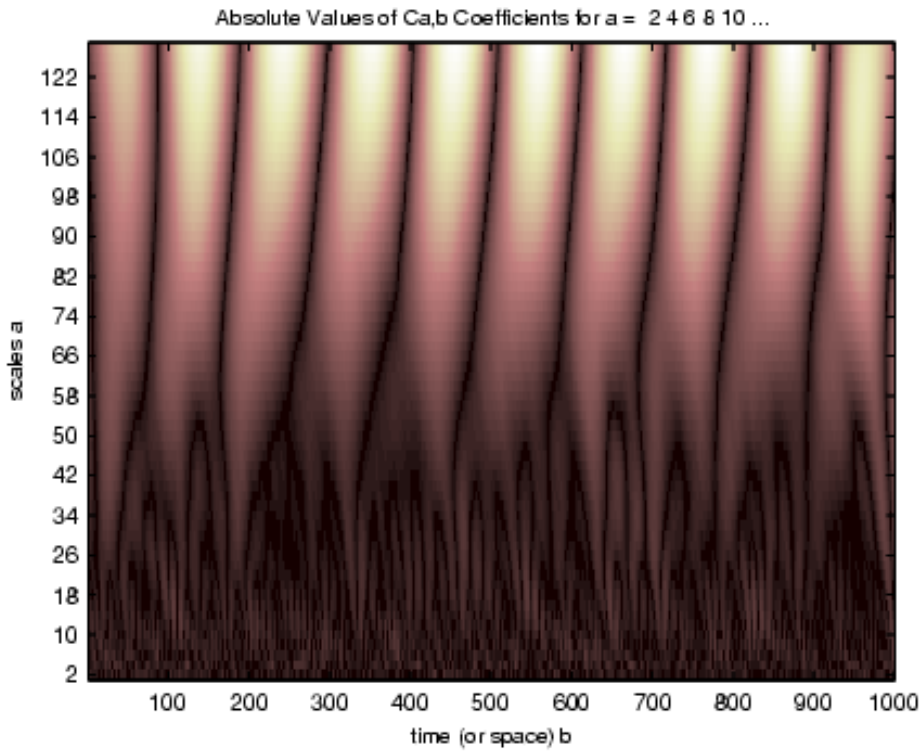


Figura 2.9 Análise de um sinal periodico (manchas solares) usando *wavelets*

Por outro lado, aplicando uma auto-correlação com base no Coeficiente de Correlação de Pearson, com sucessivos desvios temporais (à semelhança do que foi feito relativamente à detecção de correlações entre parâmetros com desvio temporal) é possível detectar periodicidades existentes no sinal. Aplicando esta técnica à série das manchas solares, obtemos o seguinte gráfico (recorrendo a uma implementação em Python, usando Matplotlib para geração do gráfico):

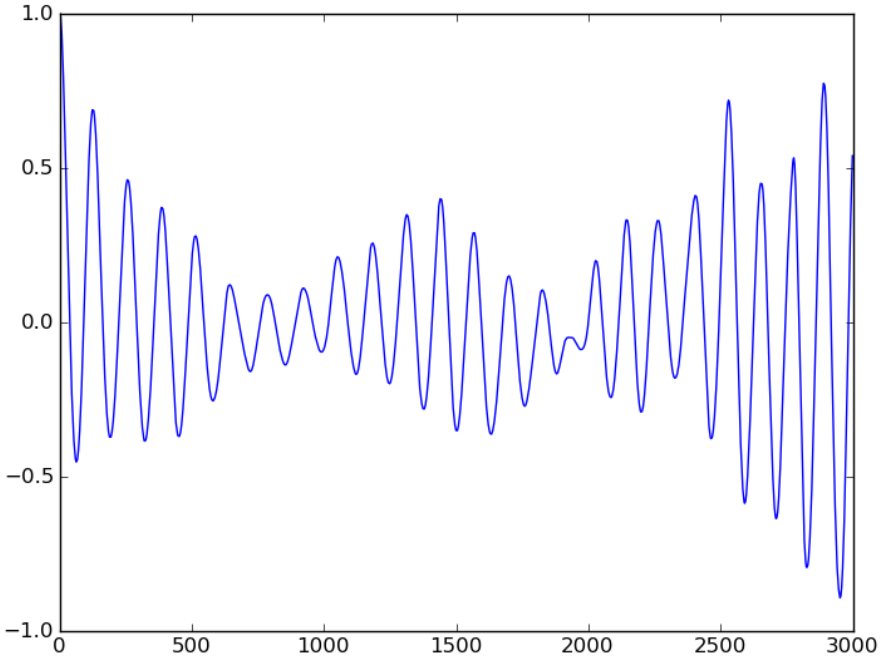


Figura 2.10 Análise de periodicidade no S*unspotNumber* utilizando auto-correlação

(O eixo das abcissas indica o tempo em meses, e o eixo das ordenadas o valor da auto-correlação após sucessivos desvios temporais)

Também neste gráfico, o eixo das abcissas indica o tempo em meses e o das ordenadas o valor da correlação (auto-correlação), fazendo sucessivos desvios temporais, cuja variação ao longo do tempo nos permite uma leitura sobre as periodicidades do sinal. O gráfico mostra duas frequências sobrepostas: uma maior que é possível identificar através duma ampliação deste gráfico de forma a medir com rigor a distância em meses entre dois máximos ou dois mínimos consecutivos e que corresponde sensivelmente a 132 meses, isto é 11 anos; quanto à frequência menor, ao confrontar o co-orientador desta tese com estes dados, verificou-se que se trata de uma periodicidade conhecida – o ciclo de Wolf-Gleissberg [3], [26]. Apesar de não se tratar de uma descoberta, esta verificação serve como validação positiva relativamente ao método proposto.

Em comparação com a abordagem das *wavelets*, o método que aqui se propõe faculta uma leitura gráfica mais simples e intuitiva, adequada ao tipo de informação que se pretende obter. Não é necessário conhecimento prévio sobre o sinal nem sobre qual a *wavelet* que melhor se ajustaria à análise do sinal. No entanto, tendo em conta as potencialidades apresentadas pelas *wavelets*, é possível que a abordagem proposta nesta tese venha a ser futuramente complementada com funcionalidades dessa abordagem, que a valorizem, sem a tornar mais complexa quanto ao seu uso por parte do utilizador.

2.2.3 Correlações provenientes de sinais com picos significativos

Foi desenvolvida nesta tese uma técnica para lidar com a detecção de correlações em sinais contendo picos significativos. Um dos desafios propostos pelo co-orientador desta tese relacionou-se com uma situação observada pelos astrofísicos envolvendo duas variáveis com um comportamento particular. Existe uma situação onde se observa uma correlação no comportamento de dois parâmetros físicos: fluxo de neutrões – *NeutronFlux* – (uma medição da radiação cósmica que atinge a Terra, feita a partir do observatório de Lomnický štít na Eslováquia) e o fluxo de iões *H1* provenientes do sol (medição feita pelo satélite ACE).

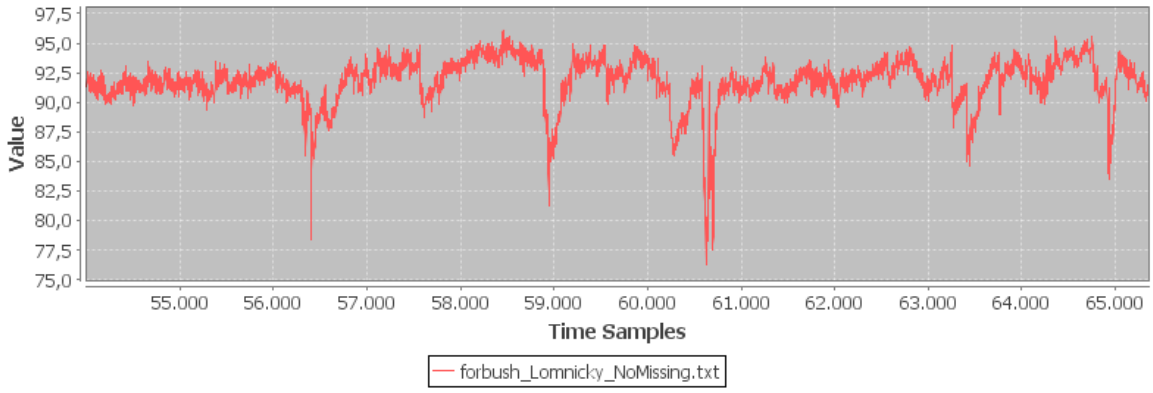


Figura 2.11 *NeutronFlux* observado em Lomnicky stit, Eslováquia

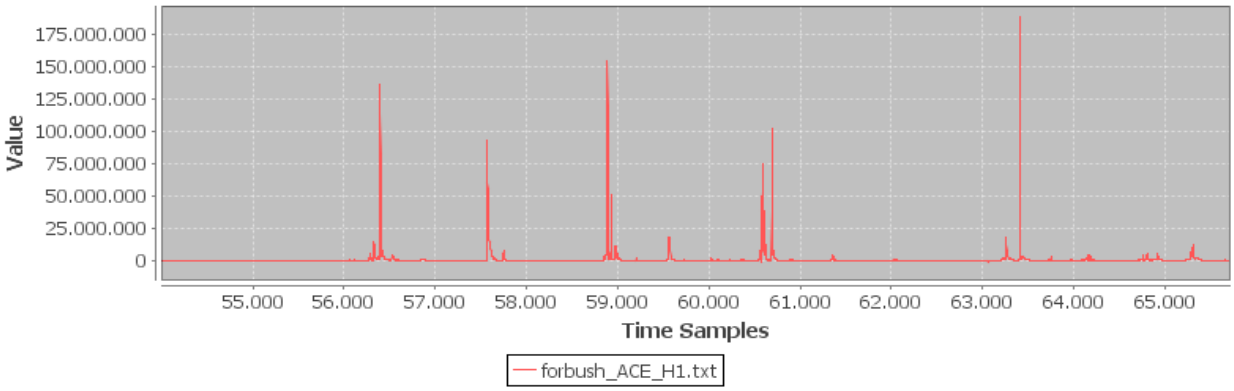


Figura 2.12 *H1* observado pelo satélite ACE

Como se pode observar pelos gráficos dos dois parâmetros (temporalmente alinhados), onde se regista um pico no sinal do *H1*, observa-se uma queda nos valores do *NeutronFlux*. Isto é o chamado "*Forbush Decrease*". Este fenómeno é observado repetidamente, e tem sido objecto de estudo por parte dos físicos solares [2], [27]. Existe

uma quantidade elevada de outros potenciais parâmetros onde este fenómeno poderá ser observado, conhecendo-se já várias outras situações semelhantes. Com o objectivo de detectar correlações neste tipo de situações, desenvolveu-se um método que se resume aos seguintes passos:

Cálculo da correlação das duas séries. Calcula-se a correlação de Pearson (ρ) entre as duas séries temporais.

Detecção dos picos. A detecção de picos no sinal faz-se por uma análise de máximos (picos) ou mínimos (depressões) locais no sinal.

Extracção das vizinhanças. Tendo a lista de picos (ou depressões) do sinal seleccionado, "extraem-se" 60 pontos vizinhos desses picos para uma série temporal auxiliar que contém os valores correspondentes às vizinhanças. A opção por este número assenta na necessidade de obter robustez estatística na amostra extraída. Do segundo sinal, extraem-se os conjuntos de valores temporalmente correspondentes, para uma segunda série auxiliar. Assim, têm-se duas séries reduzidas aos intervalos de valores correspondentes às vizinhanças dos picos.

Cálculo da correlação. De seguida, calcula-se a correlação entre estas duas séries temporais auxiliares (ρ').

Comparação entre ρ e ρ'. Por fim, comparam-se os valores das correlações das duas séries temporais originais com as séries reduzidas à união das vizinhanças. Cabe à sensibilidade do utilizador decidir se a diferença entre correlações é significativa ou não.

Numa situação onde se observa o fenómeno acima descrito, obtém-se um aumento (em módulo) no valor da correlação ρ'. Isto deve-se ao facto de nas vizinhanças dos picos / depressões, se observar sempre o mesmo comportamento na segunda série temporal, denotando uma relação mais forte nessas situações.

É também possível que o fenómeno apenas se observe com um determinado atraso entre as duas séries temporais, i.e., é preciso que decorra algum tempo para que o efeito seja observado num dos sinais. Para contemplar esta possibilidade, adicionou-se uma funcionalidade a este processo que permite observar os valores de correlação que se obtêm ao efectuar sucessivos desvios temporais às séries. Assim, é possível aferir um aumento da correlação (em módulo) no caso deste desvio temporal existir.

Note-se que tanto a dimensão das vizinhanças como a gama de desvios temporais aplicados no método são parâmetros ajustáveis pelo utilizador, permitindo-o(a) flexibilizar a pesquisa conforme necessário.

2.2.4 Correlação entre múltiplos parâmetros (> 2)

Como já foi dito, no contexto desta tese surgiu o objectivo de desenvolver e aplicar uma métrica para encontrar correlações entre parâmetros/variáveis. Esta necessidade prende-se com dois factores: por um lado, a correlação de maior interesse, por ser mais utilizada, é a correlação entre apenas dois parâmetros; por outro, tanto quanto se conseguiu apurar, não existe métrica para detectar a correlação entre um número de parâmetros superior a 2. Porém, dado que se pretende disponibilizar a capacidade de detectar as correlações significativas entre todos os possíveis parâmetros, há que ter em conta que o número total de pares possíveis cresce rapidamente com o número de parâmetros. Por exemplo, entre 8 parâmetros existem (8!/(6! . 2!)) = 28 possíveis pares; entre 16 parâmetros existem 120 pares. Assim, com o propósito de realizar alguma economia computacional, evitando medir a correlação entre todos os possíveis pares, no âmbito desta tese tem vindo a ser desenvolvida uma métrica para detectar a correlação simultânea entre *n* parâmetros, sendo *n* um qualquer número superior a 1.

Conforme o plano inicial de trabalho desta tese, houve a intenção de desenvolver uma métrica que avaliasse a correlação entre mais do que dois parâmetros simultaneamente. Mais do que uma utilidade prática, o propósito desta métrica seria o de obter uma eventual economia na detecção das correlações entre todos os possíveis pares num conjunto de parâmetros.

Desta forma, consideremos os parâmetros $X_1, \ldots, X_P$, correspondentes às séries temporais com o mesmo nome, com *n* linhas. A correlação ente estes parâmetros pode obter-se por:

$$corr(X_1, \ldots, X_p) = \frac{\frac{1}{n}\sum_{r=1}^{n}\left(\prod_{c=1}^{p}\left(x_{r,c} - \bar{x}_{.,c}\right)\right)}{\prod_{c=1}^{p}\left(\left[\frac{1}{n}\sum_{r=1}^{n}\left(x_{r,c} - \bar{x}_{.,c}\right)^{p}\right]^{\frac{1}{p}}\right)} \qquad (9)$$

Sendo $x_{r,c}$ o valor que o parâmetro X_c toma na linha *r*, e sendo $\bar{x}_{.,c} = \frac{1}{n}\sum_{r=1}^{n} x_{r,c}$ a média de valores do parâmetro X_c.

O leitor pode verificar que a correlação de Pearson é um caso particular da fórmula apresentada, para $p = 2$.

O valor da correlação varia também entre [-1, +1].

Sempre que todos os parâmetros do conjunto estejam correlacionados, obtém-se uma correlação diferente de zero. Por outro lado, basta que um dos parâmetros não esteja

correlacionado com qualquer dos outros, para que a correlação dê um valor igual ou próximo de zero.

Esta métrica apresentou resultados coerentes quando o número de parâmetros era par. No entanto, para um número ímpar, observaram-se limitações no comportamento da métrica.

Tendo em conta a complexidade prevista para ultrapassar esta limitação, e dado que foi entretanto verificado que o cálculo das correlações entre pares de parâmetros se mostrou muito eficiente, optou-se por não prosseguir neste desenvolvimento, decidindo antes aplicar o tempo disponível para solucionar as questões mais relevantes no âmbito desta tese.

2.3 Discussão de Resultados

Relativamente à detecção de correlações entre pares de parâmetros, os exemplos mostrados na secção 2.2.1, correspondentes à Figura 2.1, Figura 2.2 e Figura 2.3 são ilustrativos da adequabilidade do método utilizado para calcular estas correlações. No entanto, põe-se a questão de saber qual o valor de correlação que nos deve levar a considerar a existência de uma correlação significativa. Por apreciação empírica, apercebemo-nos de que mesmo quando obtemos valores de correlação muito baixos, esses valores não devem ser negligenciados, pois podem representar correlações reais, apesar de fracas. Com efeito, quando duas séries temporais são de facto independentes, como é o caso em que umas das séries contém valores aleatórios, obtêm-se valores de correlação próximos de zero. A figura seguinte mostra uma série gerada com valores pseudo-aleatórios, que foi usada no cálculo da correlação entre ela e a série de *NeutronMonitor* (Figura 2.1):

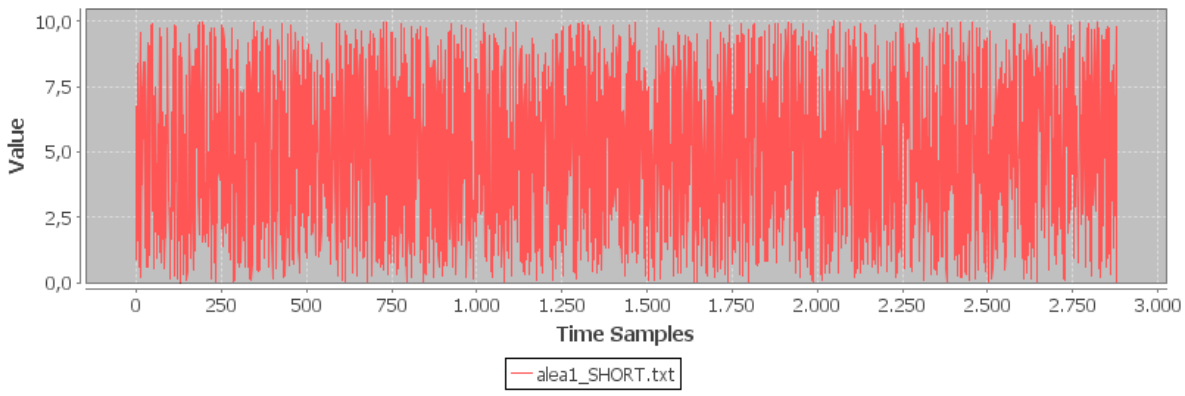


Figura 2.13 Série de valores aleatórios

Deste cálculo, obtemos uma correlação de 0.0091, um valor ilustrativo da independência entre duas séries.

No que respeita à detecção de periodicidades usando a auto-correlação e desvio temporal, para além do exemplo mostrado na secção 2.2.2, foram testados outros parâmetros como mostra o exemplo seguinte:

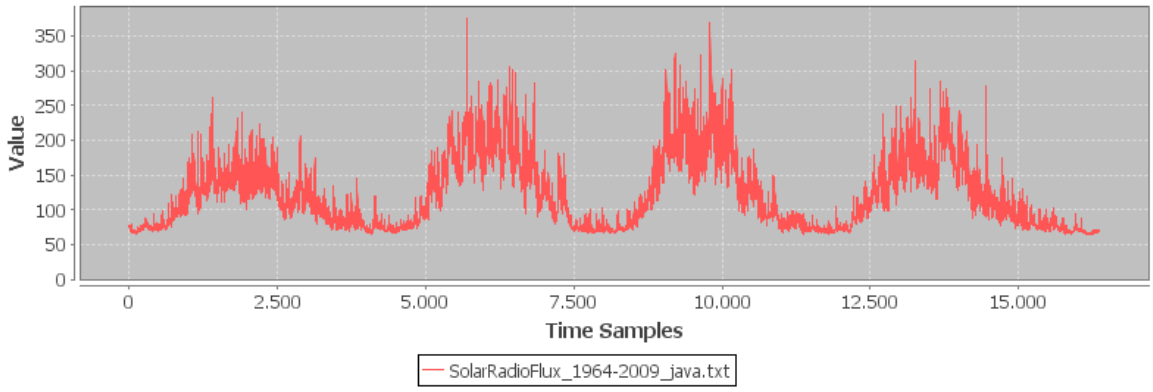


Figura 2.14 *Solar Radio Flux* (medido de 1964 – 2009)

(O eixo das abcissas está marcado em dias, e o eixo das ordenadas indica o fluxo solar de rádio)

Aplicando a funcionalidade de detecção de periodicidades, obtemos o seguinte:

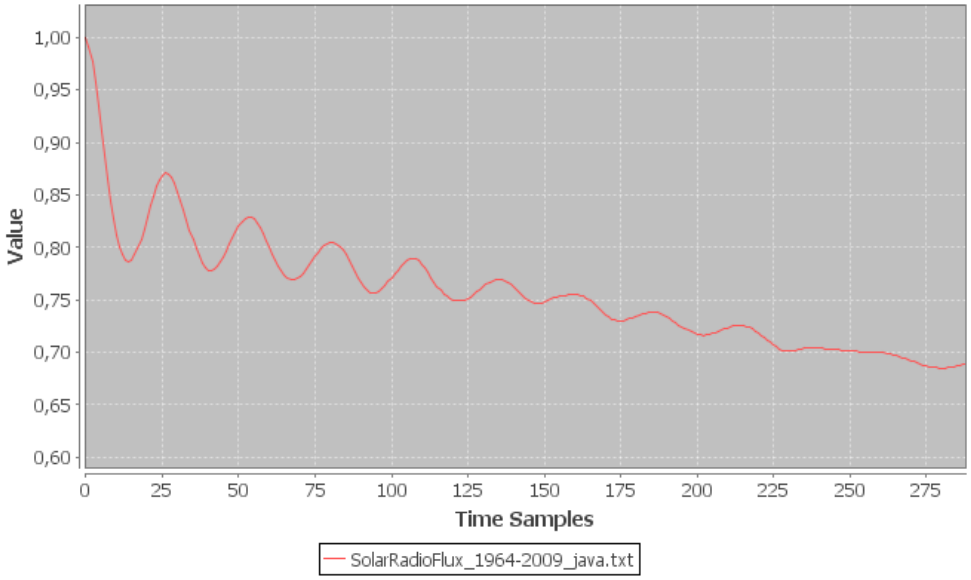


Figura 2.15 Resultado da auto-correlação no parâmetro *SolarRadioFlux*

Pela Figura 2.15, podemos reconhecer uma periodicidade aproximada de 27 dias. Foi indicado pelo co-orientador da tese que esta reflecte o período de rotação do Sol.

Para melhor validar a adequabilidade deste método, o mesmo foi testado com uma série de valores aleatórios (presente na Figura 2.13). O resultado pode-se ver de seguida:

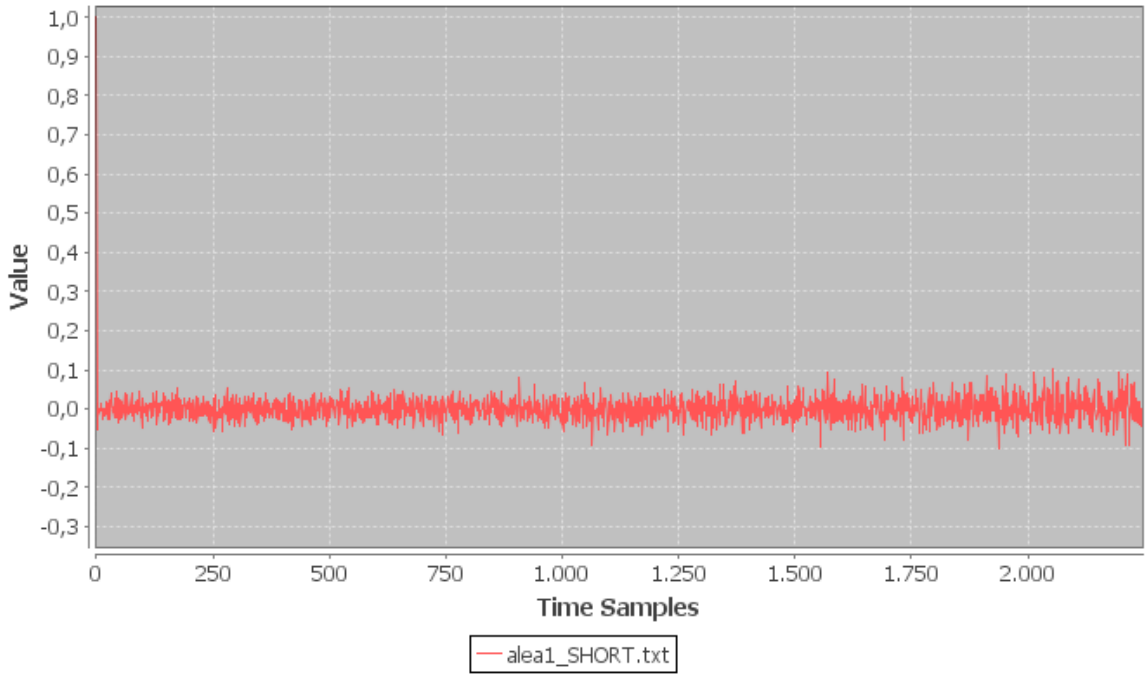


Figura 2.16 Análise de periodicidade em série de valores aleatórios

Como se vê pela figura, não se reconhecem periodicidades no sinal (aleatório), o que consideramos credibilizar o método.

Como se pode constatar pelas figuras de análise de periodicidades (em particular Figura 2.10 e Figura 2.15), o método aplicado oferece uma leitura intuitiva, sendo fácil identificar as periodicidades dos sinais.

Relativamente à detecção de correlações em sinais com picos significativos, consideremos as figuras seguintes (cujos parâmetros foram descritos na secção 2.2.3):

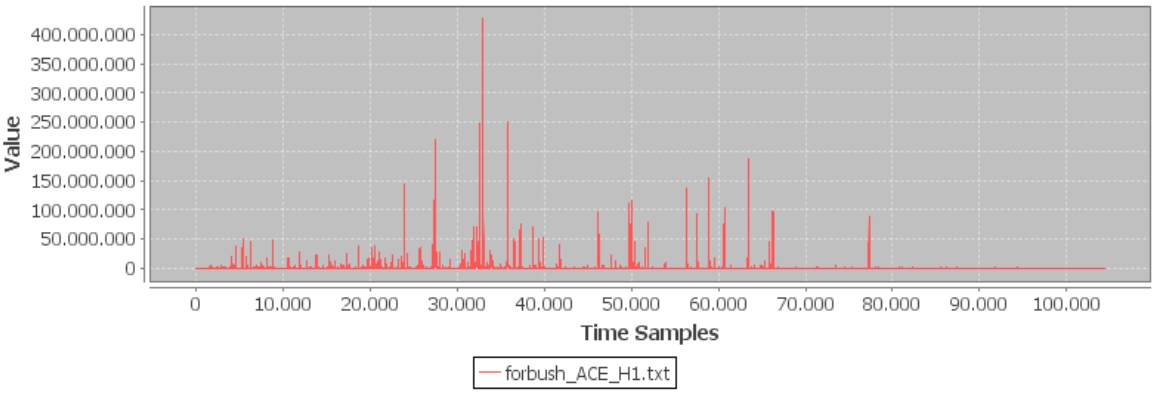


Figura 2.17 Sinal completo de *H1* observado pelo satélite ACE

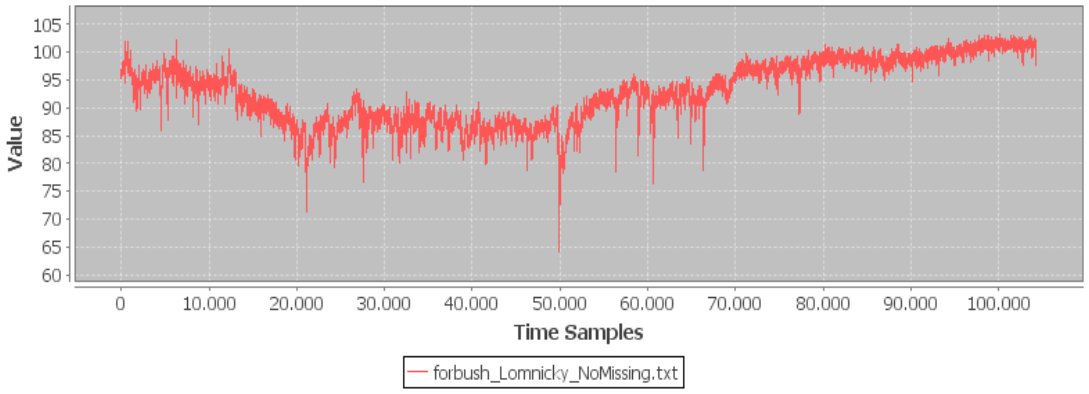


Figura 2.18 Sinal de *NeutronFlux* observado em Lomnicky stit

Aplicando o método, pelo cálculo da correlação para o sinal completo, obteve-se um valor de -0.11. De seguida, tomando as séries formadas pelas uniões das vizinhanças dos sinais (conforme descrito em 2.2.3), a correlação resultante foi de -0.14. Neste caso o aumento não é significativo.

Porém, ao considerar a detecção de picos associado a desvios temporais, obtêm-se correlações que atingem cerca de 0.25, como mostra a figura seguinte:

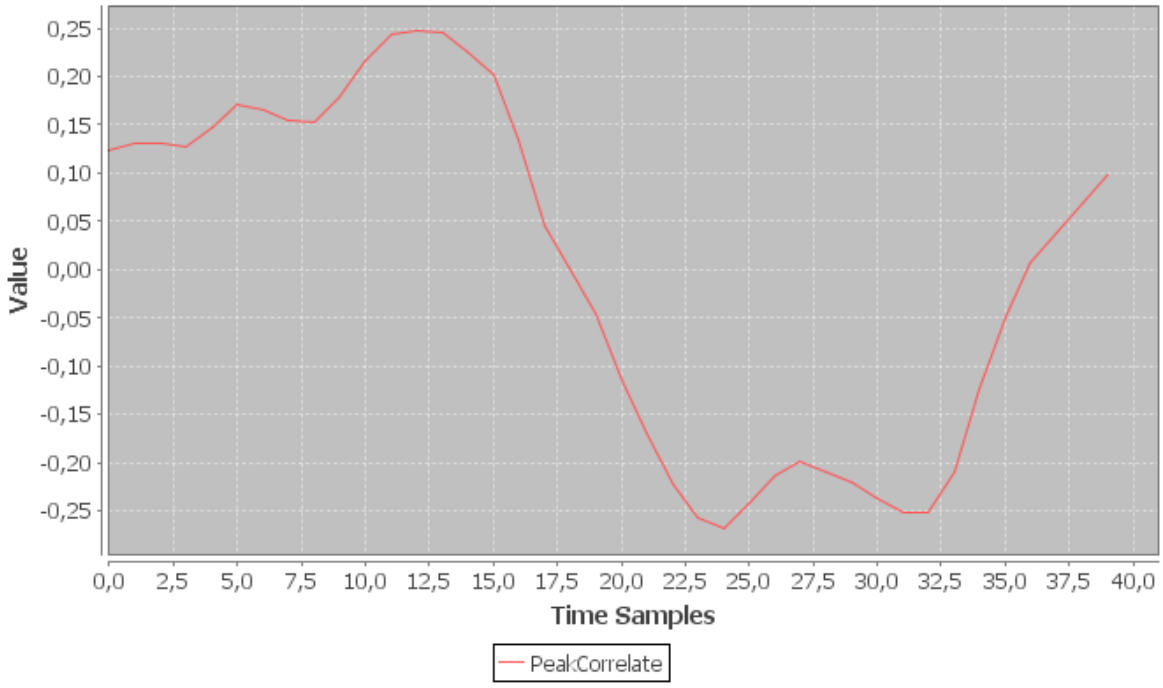


Figura 2.19 Correlações obtidas por desvios temporais associados à detecção de picos

Isto mostra que, também no contexto da detecção de correlações provenientes de sinais com picos, deve ser contemplada a hipótese de os sinais estarem desviados temporalmente.

3. Causalidade

3.1 Causalidade no Contexto da Tese

As séries temporais têm uma grande utilização a nível científico, e o caso da astronomia não é excepção. Um elevado número de séries temporais são geradas no domínio, tanto por observações a partir de satélites como por sensores baseados na terra. Cada uma destas séries temporais pode ser vista como um parâmetro ou variável. As medidas de correlação podem ser usadas para avaliar a relação entre estes parâmetros, no entanto não demonstram o sentido da causalidade, i.e., qual o efeito e qual a causa.

No decorrer da tese, surgiu esta problemática, e detectou-se a importância que a questão da causalidade teria para os astrónomos (e não só). Isso motivou-nos a aprofundar o assunto e a desenvolver um método que permitisse disponibilizar uma quantificação do sentido da causa-efeito entre dois conjuntos de dados de valores numéricos reais, e novamente sem implementar nos algoritmos conhecimentos específicos a um domínio.

Tendo em conta que se está a lidar com conjuntos de dados numéricos não categóricos, de valor real, revelou-se ser um desafio interessante, uma vez que os métodos de detecção de causalidade disponíveis tratam valores categóricos, partições de dados reais ou revelam limitações na sua aplicação em dados não categóricos. Cada um destes pontos são discutidos abaixo.

3.2 Trabalho Relacionado

3.2.1 Redes Bayesianas

Uma das formas de detecção de causalidade é a utilização de redes bayesianas. Em [28], os autores fazem uso de redes bayesianas para inferir relações de causa-efeito na expressão génica em células cancerígenas e normais. Esta aproximação implica o cálculo

das probabilidades condicionadas usando variáveis categóricas tais como $Y = 0$ e $Y = 1$ representando células normais e células cancerígenas respectivamente. No entanto, tendo em conta que se pretende detectar e medir relações de causa-efeito entre séries temporais que representam, por exemplo o fluxo solar e incidências de neutrões, estamos a lidar com dados não categóricos, como se pode ver na tabela abaixo.

Tabela 3-1 Extracto de valores

Amostra	*Solar Radio Flux*	*Neutron Monitor*
1	75,5	6405
2	76,1	6468
3	76,2	6453
…	…	…

A aplicação de redes bayesianas faz-se com dados discretizados. É possível aplicar técnicas de discretização dos dados contínuos (e.g. particionamento por frequência), no entanto isso influencia negativamente o desempenho da rede, podendo levar à perda de informação tal como à interacção e dependência de variáveis.

No caso em estudo, o facto de se tratar de dados não categóricos, ou que sejam passíveis de categorizar (sem o impacto negativo mencionado), faz com que a utilização de redes bayesianas não seja adequada.

Uma importante contribuição na componente da causalidade pode ser vista em [29], [30]. O autor define os conceitos de *efeito total* e *efeito directo*:

> *O efeito total de X sobre Y é dado por P(y*|do*(x)), nomeadamente, a distribuição de Y enquanto X é constante de valor x, e as restantes variáveis alteram naturalmente.*

Do mesmo modo, o conceito de *efeito directo* também se define com recurso às probabilidades condicionadas. De notar que estas probabilidades são calculadas para variáveis categóricas. Nitidamente, estes conceitos não são adequados para a detecção de relações de causa-efeito entre variáveis de dados numéricos não categóricos.

3.2.2 Regras de Associação

As regras de associação são geradas a partir de bases de dados relacionais de grandes dimensões, contendo tanto atributos quantitativos como categóricos [31], e em [32] restrições determinadas pelo utilizador são incluídas nessas regras de associação. Um exemplo de uma das referidas regras de associação:

```
<Idade: 30..39> AND <Casado: Sim> => <Num.Carros: 2> com 100% confiança
```

Significa que todas as pessoas casadas com idades compreendidas entre os 30 e 39 têm pelo menos dois carros, dando-nos um sentido de causa-efeito. Este tipo de abordagens lidam com atributos quantitativos tais como *Idade, Número de Carros, Salário,* etc. por um particionamento fino dos valores do atributo e subsequentemente combinando partições adjacentes conforme necessário. Deste modo, os atributos quantitativos podem ser tratados como se fossem atributos categóricos. Assim sendo, supondo que a regra acima tivesse sido obtida pela análise de uma base de dados, poder-se-ia dizer que "se uma pessoa entre os 30 e 39 é casada, *implica/causa* que tem pelo menos dois carros".

Para determinados tipos de aplicações esta aproximação pode ser usada na detecção de relações causa-efeito. No entanto, não é possível a sua aplicação quando o domínio dos dados é desconhecido ou não permite um particionamento dos valores por forma a gerar um conjunto de categorias relevantes, algo que acontece com muitas séries temporais.

3.2.3 Modelos de Regressão

Em [33], [34] encontram-se abordagens ao problema da causalidade recorrendo a modelos de regressão. Estes modelos têm um grande inconveniente uma vez que podem gerar funções de regressão inaceitáveis. Tomemos por exemplo duas séries de dados, (a) e (b), compostas pelos seguintes pares de (*x*, *y*):

Tabela 3-2 Séries (*a*) e (*b*)

Série (a)	***Série (b)***
(10; 8,04)	(8; 6,58)
(8; 6,95)	(8; 5,76)
(13; 7,58)	(8; 7,71)
(9; 8,81)	(8; 8,84)
(11; 8,33)	(8; 8,47)
(14; 9,96)	(8; 7,04)
(6; 7,24)	(8; 5,25)
(4; 4,26)	(8; 12,50)
(12; 10,84)	(8; 5,56)
(7; 4,82)	(8; 7,91)
(5; 5,68)	(8; 6,89)

Cada uma das séries origina o mesmo resultado quando se aplica um programa típico de regressão. Neste caso:

Número de observações (n) = 11;
Média dos x's ($\overline{x}$) = 9.0;
Média dos y's ($\overline{y}$) = 5.5;
Coeficiente de regressão ($b1$) de y em x = 0.5;
Equação da linha de regressão: $y = 3 + 0.5x$;
Somatório dos quadrados $x - \overline{x}$ = 110.0;
Regressão do somatório dos quadrados = 27.5 (1 d.f.[2]);
Soma residual dos quadrados de y = 13.75 (9 d.f);
Desvio padrão estimado $b1$ = 0.118;
Coeficiente de correlação múltiplo R^2 = 0.667.

Tendo em conta estes valores, a mesma função de regressão seria aceite para ambas as séries. Nas figuras abaixo podemos ver os conjuntos de pontos, juntamente com a função de regressão obtida para ambos. Como se pode ver, para conjuntos completamente distintos de valores, obtém-se a mesma função de regressão.

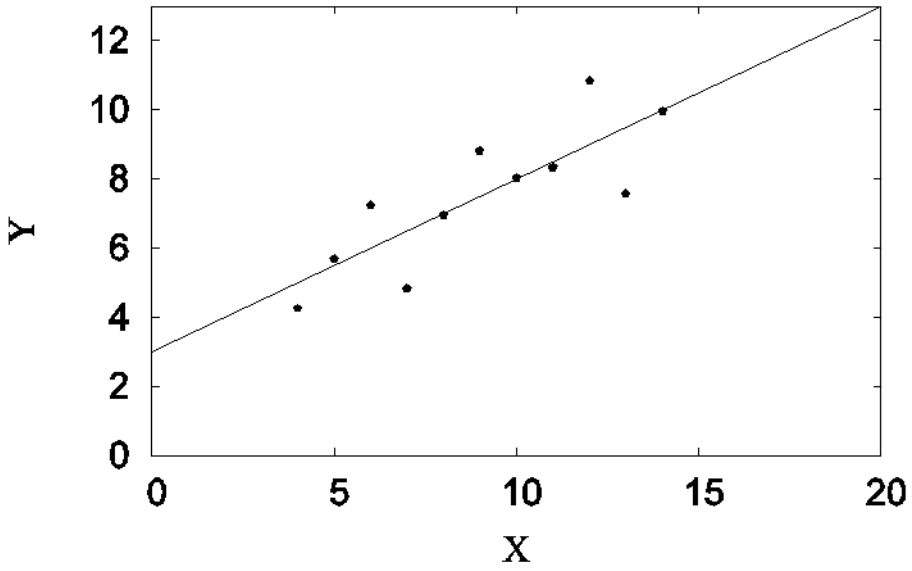


Figura 3.1 Série de dados (*a*) e respectiva linha de regressão

[2] d.f. - "degrees of freedom"

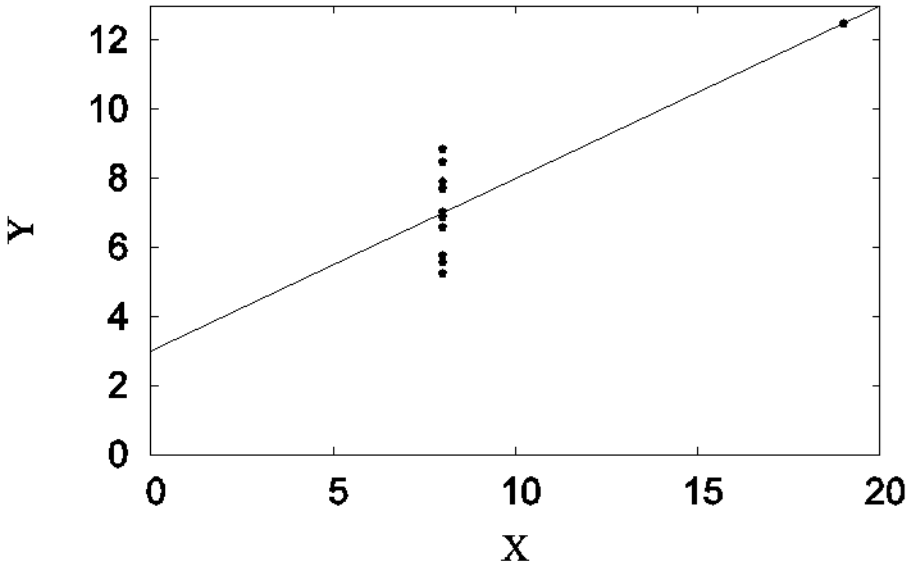


Figura 3.2 Série de dados (*b*) e linha de regressão igual à anterior

Enquanto na série de dados (*a*) obtemos uma linha de regressão que aproxima os pontos da série, na série de dados (*b*) o mesmo não se verifica. Na realidade, a série (*b*) tem apenas dois valores distintos para o *x*, e o que determina a inclinação da linha de regressão resume-se apenas ao ponto $x = 19$. Caso esse ponto fosse removido, nem seria possível calcular a inclinação. Assim, a linha de regressão calculada para a série (*a*) seria uma boa base para a detecção da relação de causa-efeito, mas o mesmo não se pode dizer para a série (*b*).

3.3 Trabalho Realizado – Detectar Relações de Causa-Efeito

Nesta secção propõe-se uma abordagem alternativa e simples, que não depende da qualidade de funções de regressão, nem de regressões em si. Aqui será explicado o método proposto para detectar relações de causa-efeito e a medição da respectiva força da relação.

3.3.1 Abordagem Proposta

Considerando duas variáveis *X* e *Y* contendo séries de dados numéricos não-categóricos, assume-se que *X* determina *Y* se para todos os possíveis intervalos de dimensão reduzida do domínio de *X* correspondem pequenas dispersões no domínio de *Y*. Ou seja, se *X* causa

Y então se $Y = y$ quando $X = x$, teremos que $Y = y_1$ quando $X = x_1$ sendo que x_1 e y_1 serão valores muito próximos de *x* e *y* respectivamente.

Assim sendo, de modo a determinar se *X* causa *Y*, propõe-se um método com três passos:

1. Particionamento de *Y* em pequenos conjuntos determinados em *X*
2. Medição da dispersão de cada partição em *Y*
3. Teste da relação de causa-efeito a partir das dispersões.

3.3.1.1 Partição dos Dados

Sejam *X* e *Y* dois conjuntos de dados tais que $X = \{x_1, x_2, \dots, x_n\}$ e $Y = \{y_1, y_2, \dots, y_n\}$. Seja *XY* uma composição de *X e Y* tal que $XY = \{(xi, yi): xi = X(i), yi = Y(i)\}$ onde $X(i)$ e $Y(i)$ representam os i-ésimos elementos de *X* e *Y* respectivamente.

O seguinte algoritmo é usado de forma a particionar *Y*:

Algoritmo 1

```
1:    partitioning(XY, p,OutputList)
2:    X ← {x|(x, y) ∈ XY }
3:    Y ← {y|(x, y) ∈ XY }
4:    if max(X) > min(X).(1 + p) ∧ ‖X‖ > MinSize then
5:          XY_left ← {(x, y)|(x, y) ∈ XY ∧ x ≤ median(X)}
6:          XY_right ← {(x, y)|(x, y) ∈ XY ∧ x > median(X)}
7:          partitioning(XY_left , p, OutputList)
8:          partitioning(XY_right , p, OutputList)
9:          return OutputList
10:   else
11:         append Y to OutputList
12:         return OutputList
```

Temos como parâmetros de entrada:

- *XY* (composto pelos conjuntos *X* e *Y)*
- *p* controla a gama de valores nos subconjuntos de *X*
- *OutputList* é a lista resultante de partições de *Y* (inicialmente vazia).

Este algoritmo retorna uma lista contendo um subconjunto de valores de *Y* tal que para cada um destes subconjuntos, os subconjuntos de valores correspondentes de *X* têm apenas uma pequena dispersão dos seus valores. O intervalo de valores dos subconjuntos de *X* é controlado pelo parâmetro *p*, que deverá ser suficientemente pequeno para assegurar desvios pequenos, mas ao mesmo tempo manter pelo menos *MinSize* elementos no conjunto. (*MinSize* por omissão é 2 enquanto que *p* é definido como parâmetro pelo utilizador).

Assim, o procedimento *partitioning* do algoritmo inicia-se obtendo um conjunto com os valores de *X*, e outro com os valores de *Y* a partir do conjunto *XY*. Se o intervalo de valores de *X* exceder o valor especificado por *p* (linha 4), então *XY* é dividido em dois conjuntos:

- XY_{left} contendo valores de *XY* cujos valores em *X* são inferiores ou iguais ao valor do central de *X* considerando que *X* está ordenado segundo os seus valores. Este valor central foi obtido através da mediana (linha 5);
- XY_{right} conterá obviamente os valores de *XY* cujos valores em *X* são superiores ao valor central de *X* (linha *6*).

Este processo é depois executado recursivamente sobre os conjuntos XY_{left} e XY_{right} (linhas 7 e 8) até que o intervalo de valores do subconjunto seja suficientemente pequeno (linha 4), e nesse caso o conjunto *Y* é adicionado ao *OutputList*.

Para exemplificar, tomemos o seguinte caso:

XY = {(537.21, 7.70), (537.06, 7.73), (536.75, 7.75), (535.81, 7.78), (534.80, 7.80), (534.56, 7.81), (328.01, 21.03), (327.45, 21.94), (326.30, 22.03), (325.04, 22.15), (207.03, 29.03), (206.44, 29.47), (205.91, 29.61), (205.39, 30.51), (205.18, 30.54), (204.72, 30.60)};

Consideremos agora particionar *Y* de acordo com intervalos de *X* tais que $p = 1\%$. Pelo algoritmo, obteríamos a seguinte lista de subconjuntos de *Y*: {30.60, 30.54, 30.51, 29.61}, {29.47, 29.03}, {22.15, 22.03}, {21.94, 21.03}, {7.78, 7.75, 7.73, 7.70}, {7.81, 7.80}.

Tomando a primeira partição como exemplo, {30.60, 30.54, 30.51, 29.61}, pode-se confirmar que os valores correspondentes de *X* são: {204.72, 205.18, 205.39, 205.91}, respeitando a restrição imposta pelo *p*, uma vez que $205{,}91 \leq 204{,}72 \times (1 + p)$, para $p = 1\%$ (linha 4 do algoritmo).

3.3.1.2 Medição da Dispersão

O segundo passo envolve medir a dispersão dos valores de cada partição resultante da aplicação do algoritmo. Inicialmente, considerou-se a hipótese de efectuar a medição da gama de valores em cada partição de *Y*, da mesma forma como é feito para o conjunto *X*, comparando os valores extremos da partição. No entanto, apesar da simplicidade associada, concluiu-se que essa comparação não seria uma boa métrica tendo em conta a sua demasiada sensibilidade a valores de *outliers*[3], o que poderia mascarar a dispersão

[3] Uma possível tradução para português seria "valor discrepante", mas optou-se pelo termo em inglês "outlier"

média associada a esses valores. Exemplificando, uma partição contendo 10 elementos, poderá ser constituída por 9 elementos muito próximos enquanto que um se desvia consideravelmente. Este *outlier* poderá surgir porque, na realidade, os valores de *Y* são muitas vezes determinados não somente pelos valores de *X*, mas também por outros parâmetros. No entanto, neste trabalho apenas queremos determinar relações causa-efeito entre *X* e *Y* (e.g. a precipitação é determinada não apenas pela *temperatura do ar*, mas também pela *pressão atmosférica*, entre outros factores).

Assim sendo, comparando valores extremos na partição não seria possível *minimizar* o efeito do *outlier* e determinar a provável baixa dispersão da partição. Assim, por forma a medir a dispersão, considerou-se inicialmente o *desvio padrão da amostragem.*

$$\sigma(P) = \sqrt{\frac{1}{\|P\|}\sum_{pi \in P}(p_i - \mu(P))^2}\,, \quad \mu(P) = \frac{1}{\|P\|}\sum_{pi \in P} p_i \qquad (10)$$

Sendo p_1, p_2, ... , p_n os elementos da amostragem. $\|P\|$ representa o número de elementos de *P*. No entanto, visto que esta métrica é sensível à escala, não se mostrou ser uma boa ferramenta para avaliar a dispersão. Por exemplo, sejam P_1 e P_2 duas partições tais que:
P_1 = {2, 2.01, 2.02, 2.03, 2.04} e P_2 = {200, 201, 202, 203, 204}. Os desvios padrão são:
σ (P_1) = 0.0141 e σ (P_2) = 1.41.

Assim sendo, de acordo com esta métrica, a dispersão de P_2 é 100 vezes superior do que P_1. Esta diferença não capta a *dispersão relativa* semelhante que está presente em ambas as partições. Na realidade, não contemplando a questão de escala, ambos os conjuntos apresentam o mesmo desvio relativo nos seus valores.

O próximo passo consistiu em aplicar o *coeficiente de variação da amostra,* uma medida normalizada de dispersão, definida pelo rácio entre o desvio padrão da amostra σ(*P*) e a sua média μ(*P*).

$$Cv(P) = \sigma(P) / \mu(P) \qquad (11)$$

Desta forma C*v* (P_1) = C*v* (P_2) = 0.00700. Esta mostrou ser uma boa métrica para os casos estudados. No entanto esta medida não se encontra definida quando a média é zero. Mais, a média poderá ser muito próxima de zero, não pelos elementos do conjunto serem valores positivos muito pequenos, mas porque o somatório dos elementos positivos poderá ser muito próximo do valor absoluto do somatório dos elementos negativos. Nestes casos, C*v*(.) não funciona como pretendido. E.g.:

Sejam P_3 = {-0.0100, -0.0110, 0.0100, 0.01101} e P_4 = {-0.0100, -0.0110, 0.0100, 0.01095}; assim C*v*(P_3) = 4205,805 e C*v*(P_4) = −839.906, sendo valores que não reflectem a magnitude de dispersão relativa esperada para estas partições.

De forma a superar este problema, propõe-se uma métrica definida pelo rácio do desvio padrão da amostra σ(*P*) e a média dos valores absolutos dos elementos da amostra:

$$Rd(P) = \frac{\sigma(P)}{\frac{1}{\|P\|}\sum_{pi \in P} |p_i|} \quad (12)$$

Assim, para o mesmo exemplo, *Rd*(P_3) = 1.00108 e *Rd*(P_4) = 1.00114, sendo valores que reflectem a magnitude da dispersão relativa esperada para estas partições.

Note-se que *Rd*(.) e C*v*(.) retornam os mesmos valores para partições contendo apenas valores positivos. E, ao contrário de C*v*(.), *Rd*(.) está definido quando a média é zero, desde que pelo menos um dos elementos da partição seja diferente de zero.

3.3.1.3 Teste de Causa-Efeito

Chegado aqui ao terceiro passo, é necessário decidir se existe uma relação de causa-efeito do conjunto *X* para o conjunto *Y*, utilizando a dispersão relativa dada pela métrica *Rd*(.) para auxiliar na decisão. Assim, supondo que partições em *Y* são obtidas de acordo com a gama de valores em *X*, se a dispersão relativa dos valores de uma partição genérica *P* de *Y* é significativamente inferior à dispersão relativa da série temporal completa *Y*, dizemos que *X* causa *Y*. Por outro lado, caso a dispersão relativa seja similar, dizemos que *X* não causa *Y*.

Assim, estabelecemos a seguinte hipótese nula:

H_0: considerando que P é uma partição genérica do conjunto de dados de Y, obtida de acordo com o algoritmo enunciado em 3.3.1.1, *então o valor de Rd*(*P*) *é próximo de Rd*(*Y*).

Assim, a verificar-se a hipótese nula podemos considerar que *X* não causa *Y*. Utilizamos a medida estatística chi-quadrado de Pearson para testar H_0:

$$X^2 = \sum_{i=1}^{i=k} \frac{(O_i - E_i)^2}{E_i} \quad (13)$$

Na equação acima, temos que *k* = 2, uma vez que existem dois casos diferentes a tratar: o primeiro caso, associado ao número de partições com uma dispersão relativa menor ou

igual à dispersão relativa do conjunto *Y*, $Rd(P) \leq Rd(Y)$, sendo *P* uma partição genérica; o segundo caso, associado ao número de partições tal que $Rd(P) > Rd(Y)$.

Assim, O_1 representa a frequência de observações associadas ao caso 1, e E_1 representa a frequência esperada de observações associadas ao caso 1, considerando H_0. O_2 e E_2 estão associados ao caso 2. Tomemos por exemplo um caso onde o algoritmo retorne 8000 partições a partir do conjunto *Y*, das quais 7500 obedecem à condição $Rd(P) \leq Rd(Y)$. Considerando H_0, esperam-se 8000 / 2 partições para as quais $Rd(P) \leq Rd(Y)$, e o mesmo número para os casos em que $Rd(P) > Rd(Y)$. Assim, $O_1 = 7500$, $O_2 = 500$, $E_1 = 4000$, $E_2 = 4000$ e $X^2 = (7500 - 4000)^2 / 4000 + (500 - 4000)^2 / 4000 = 6125$.

Recorrendo a uma tabela de distribuição cumulativa do chi-quadrado, para um nível de significância α, rejeita-se H_0 *sse*:

$$X^2 > \chi^2_{df}(\alpha) \qquad (14)$$

sendo *df* o número de graus de liberdade (dado por $df = k - 1 = 1$).

Utilizou-se um nível de significância de $\alpha = 0.05$. Para estes valores, o valor crítico de X^2 é de $\chi^2_1(0.05) = 3.84$. Note-se que este teste apenas poderá ser feito com um número suficientemente elevado de observações. De acordo com o autor em [35], este valor deverá ser de pelo menos 4 ou 5 vezes o número de *células*. Dado que no caso do exemplo, o número de células é dois ($k = 2$), e o número de observações (partições) é de pelo menos várias dezenas (e por vezes centenas ou milhares), o teste regido por (14) pode ser tomado como válido.

Verificou-se que sempre que H_0 era rejeitado, existia de facto uma relação de causa-efeito de *X* para *Y*.

Contudo, em teoria, H_0 pode ser rejeitada por O_2 ser muito maior do que E_2; não por O_1 ser muito maior do que E_1. Este caso, embora improvável, a verificar-se implicaria a rejeição de H_0 o que constituiria uma decisão errada. Assim, de forma a impedir este tipo de decisões, estabelecemos que:

Existe uma relação de causa-efeito de X para Y sse H_0 é rejeitado e $AvgRd(\wp) < Rd(Y)$.

Sendo $\wp$ o conjunto de partições de *Y* retornadas pelo Algoritmo 1 e $AvgRd(\wp)$ representa a média dos valores $Rd(.)$ de cada partição de *Y*, dada por:

$$AvgRd(\wp) = \frac{1}{||\wp||} \sum_{P_i \in \wp} Rd(P_i) \qquad (15)$$

3.3.2 Sentido Principal da Causa-Efeito

Durante a fase de experimentação, notou-se que por vezes acontece que havendo uma relação de causa-efeito do conjunto *X* para o conjunto *Y*, existe também uma relação de *Y* para *X*. De facto, ambos os conjuntos poderão influenciar-se um ao outro. Pode-se observar isso com uma série temporal de temperaturas do ar e outra de pressão atmosférica, ambas medidas simultaneamente e para o mesmo local. Apesar de, na maioria das vezes, existir um sentido *principal* na influência, quando comparada com o sentido oposto, ao qual poderemos chamar o sentido *secundário*. Podemos pois determinar qual dos sentidos é o principal e quão dominante é.

Assim, seja $\wp_{\text{right}}$ o conjunto de partições de *Y* obtido de acordo com pequenas variações de valores de *X*, de acordo com o Algoritmo 1 em 3.3.1.1. E $\wp_{\text{left}}$ o conjunto de partições de *X* obtido de acordo com pequenas variações de valores de *Y*, obtido pelo mesmo algoritmo;

seja ainda $Left = AvgRd(\wp_{\text{left}})$ e $Right = AvgRd(\wp_{\text{right}})$. A direcção principal é dada por:

$$Dir = \frac{Left - Right}{\max(Left, Right)} \tag{16}$$

Os valores de *Dir* variam entre -1 e +1. Como exemplo, se *Dir* = 0.99, significa que o sentido principal é do conjunto *X* para o conjunto *Y*, e este sentido é o dominante. Se pelo contrário, *Dir* = -0.99, significa que o sentido principal é dominante de *Y* para *X*. Caso *Dir* seja próximo de zero, significa que não existe sentido principal.

3.4 Discussão de Resultados

De modo a validar o método, recorreu-se a pares de séries temporais, alinhadas pelo tempo. Para cada par foram verificadas relações de causa-efeito, bem como o sentido principal da relação. Manteve-se o parâmetro *p* do algoritmo o mais pequeno possível, com a condição de que as partições continham pelo menos dois elementos. Em relação ao α, a escolha do valor deste parâmetro afecta a fronteira de rejeição / aceitação de H_0: um valor muito pequeno (e.g. α = 0.001) pode levar a uma perda de sensibilidade para detectar relações de causa-efeito (χ^2_1 (0.001) = 10.83). Por outro lado, um valor elevado

(e.g. $\alpha = 0.1$) pode levar a *falsos positivos* (χ^2_1 (0.1) = 2.71). Optou-se por utilizar o valor de $\alpha = 0.05$, uma escolha comum em aplicações estatísticas (χ^2_1 (0.05) = 3.84).

A tabela seguinte mostra o teste a duas séries, uma do parâmetro *NeutronMonitor* e a outra de *SolarRadioFlux*, de amostras diárias temporalmente alinhadas. Para este par de séries, colocou-se $p = 1\%$.

Tabela 3-3 Excerto de séries temporais *Neutron Monitor* e *Solar Radio Flux*

NeutronMonitor	*SolarRadioFlux*
6449	77.5
6433	75.5
6430	76.9
6404	76.9
6450	76.1
6383	75.7
6388	75.6
6403	73.6
6411	75.1
6406	72.9
6450	74.0
6453	72.8
6468	73.1
6405	71.7
⋮	⋮

Foi detectada a relação causa-efeito em ambos os sentidos, i.e., a relação de *NeutronMonitor* para *SolarRadioFlux* e o inverso, de *SolarRadioFlux* para *NeutronMonitor*. No entanto, identificou-se a segunda como sendo a relação causa-efeito principal e claramente dominante, dada pelo valor de *Dir* = -0.993. Este resultado é consistente com o conhecimento de que a actividade solar (medido pelo *SolarRadioFlux*), influencia os valores de medição de radiação cósmica (dado pelos valores de *NeutronMonitor*).

A próxima tabela mostra um excerto de valores vindos do "ACE Science Center", para parâmetros de *HydrogenH-S1* e *NeutronFlux*. Para este par, manteve-se $p = 1\%$.

Tabela 3-4 Excerto de séries temporais *HydrogenH_S1* e *NeutronFlux*

HydrogenH_S1	*NeutronFlux*
4.18E+03	93.625
8.42E+02	93.458
2.56E+03	93.674
1.67E+03	94.039
3.39E+03	94.167
2.52E+03	94.204
5.09E+03	94.165
2.56E+03	94.054
1.68E+03	93.699
8.42E+02	93.678
2.52E+03	93.610
8.48E+02	93.207
2.52E+03	93.131
8.48E+02	93.229
⋮	⋮

Para este caso, foram detectadas relações causa-efeito em ambos os sentidos. Neste caso, a relação de *HydrogenH-S1* para *NeutronFlux* foi identificada como a relação causa-efeito principal, novamente dominante, com um valor de *Dir* = 0.997. Este é novamente um resultado coerente.

Consideremos agora um outro par de séries temporais, fora do âmbito da astronomia, composto por uma série temporal contendo valores de *Minimum Daily Air Temperature* em graus Celcius, e outra alinhada no tempo contendo valores de *Precipitation* em milímetros. Estes dados foram obtidos a partir da "Daily Temperature and Precipitation Data for 223 Former-USSR Stations", ficheiro f.20674.dat, acessível em http://cdiac.ornl.gov/ftp/ndp040/. A tabela seguinte mostra parte do par construído, contendo dados de 1936 a 2001.

Tabela 3-5 Excerto de séries temporais *MinimumDailyAirTemperature* e *Precipitation*

MinimDailyAirTemperature	*Precipitation*
6.0	0.1
6.0	3.2
7.5	3.2
7.5	0.0
4.9	0.0
6.2	3.6
4.2	10.1
1.7	0.7
3.0	1.8
3.9	7.0
2.5	0.5
1.4	0.5
−0.5	2.5
0.9	1.7
⋮	⋮

Para este par, foi necessário usar p = 2%. Também aqui, foram detectadas relações causa-efeito em ambos os sentidos. Contudo, a relação causa-efeito de *MinimumDailyAirTemperature* para *Precipitation* foi identificada como sendo a principal, já que *Dir* = 0.917. Este é um outro resultado coerente, já que a precipitação é afectada pela temperatura mínima do ar.

Como despiste, foram testadas duas séries temporais, cada uma com 20000 valores pseuso-aleatórios, gerados entre 1 e 10. A Tabela 3-6 mostra um excerto desse par.

Tabela 3-6 Duas séries temporais de valores aleatórios

Random 1	*Random* 2
1.70990730576	3.27963949203
6.75568659158	7.03206162118
2.36628948303	2.49647126814
1.59258546664	1.76795665048
6.70111556089	0.233526916995
8.36871360842	4.39915014498
8.21712564928	1.27329866438
4.44562604687	4.80382818154
5.86858487717	5.20883064475
0.587672998956	6.4485330581
1.34406834937	7.78643170293
6.22154597205	5.25530285293
9.56123292807	5.37890969541
6.10342111996	6.48056520001
⋮	⋮

Nenhuma relação de causa-efeito foi detectada entre estas duas séries, o que denota um resultado consistente. Para este par, foi usado p = 1%. Nenhuma relação principal foi detectada para qualquer um dos sentidos já que Dir = 0.0312. Este é também um resultado coerente uma vez que o carácter aleatório dos valores das séries torna-as independentes.

4. Protótipo Desenvolvido

Um dos objectivos iniciais desta tese foi o de desenvolver um protótipo operacional onde fosse possível testar e comprovar os conceitos desenvolvidos no âmbito da tese, e consequentemente disponibilizar uma ferramenta útil e de fácil utilização.

Este capítulo descreve a ferramenta resultante, e disponibiliza uma sinopse das funcionalidades disponíveis.

Com foi dito na introdução, uma versão deste protótipo foi apresentado em conferência internacional ligada ao domínio da astronomia, tendo recebido comentários positivos por parte da comunidade astronómica presente.

4.1 Plataforma de Desenvolvimento

Na fase de preparação da dissertação, utilizou-se a linguagem de programação Python para prototipagem rápida e validação de conceitos. Apesar do aspecto bastante aceitável dos gráficos gerados (com recurso à biblioteca Matplotlib), o facto de ser necessário interagir com o programa a partir de uma consola de texto demonstrou ser pouco prático. A implementação de um ambiente gráfico para interacção com o utilizador tornou-se prioritário, mas a disponibilidade de componentes para criar interfaces gráficas em Python revelava-se muito limitativa.

Assim, optou-se pela plataforma Java como base de desenvolvimento. Por um lado pela necessidade de utilizar uma linguagem utilizável em múltiplos sistemas operativos, mas por outro, a disponibilidade de bons ambientes integrados de desenvolvimento com respectivos editores para o desenvolvimento de interfaces gráficas.

O protótipo desenvolvido permite ao cientista analisar os seus dados, de uma forma visual e com recurso a um conjunto de ferramentas elaboradas durante a tese. Teve-se o cuidado

de criar uma ferramenta útil, fácil de utilizar e sem a necessidade de ter conhecimentos específicos sobre os métodos utilizados, ou obter formação prévia na ferramenta.

4.2 Funcionalidades do Protótipo

A interface gráfica desenvolvida faz uso de áreas de visualização redimensionáveis de acordo com as preferências do utilizador. A figura seguinte mostra o aspecto geral da aplicação:

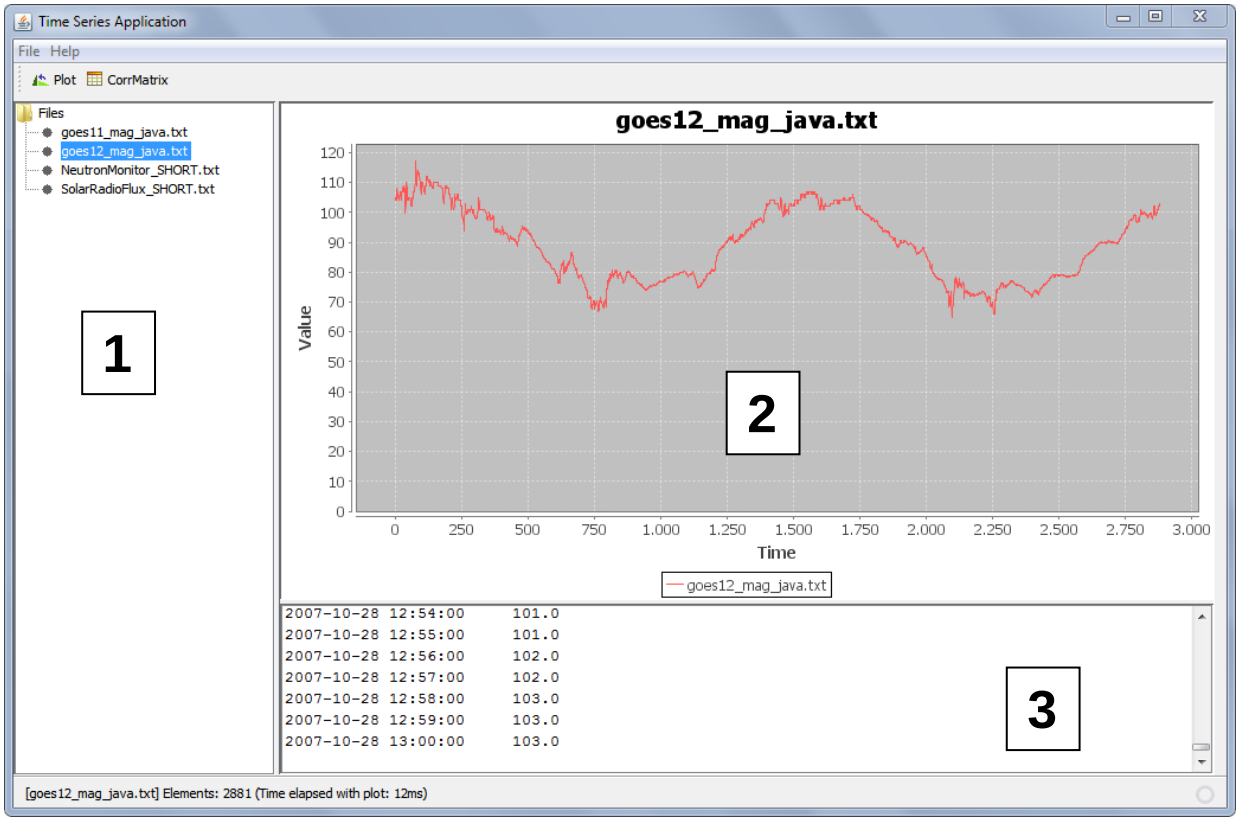


Figura 4.1 Ecrã principal da ferramenta

A aplicação está dividida em três áreas principais:

1 – Árvore de listagem de ficheiros actualmente abertos na ferramenta;

2 – Área que mostra o gráfico do parâmetro do ficheiro seleccionado;

3 – Consola com o conteúdo lido a partir do ficheiro seleccionado.

4.2.1 Carregamento de Ficheiros

Um primeiro passo na utilização da ferramenta será o carregamento de dados. Na opção File – Open (também disponível a partir do atalho "*Ctrl-O*"), a aplicação apresenta uma janela de escolha de ficheiros geral do respectivo sistema operativo. Esta permite a abertura de múltiplos ficheiros simultaneamente.

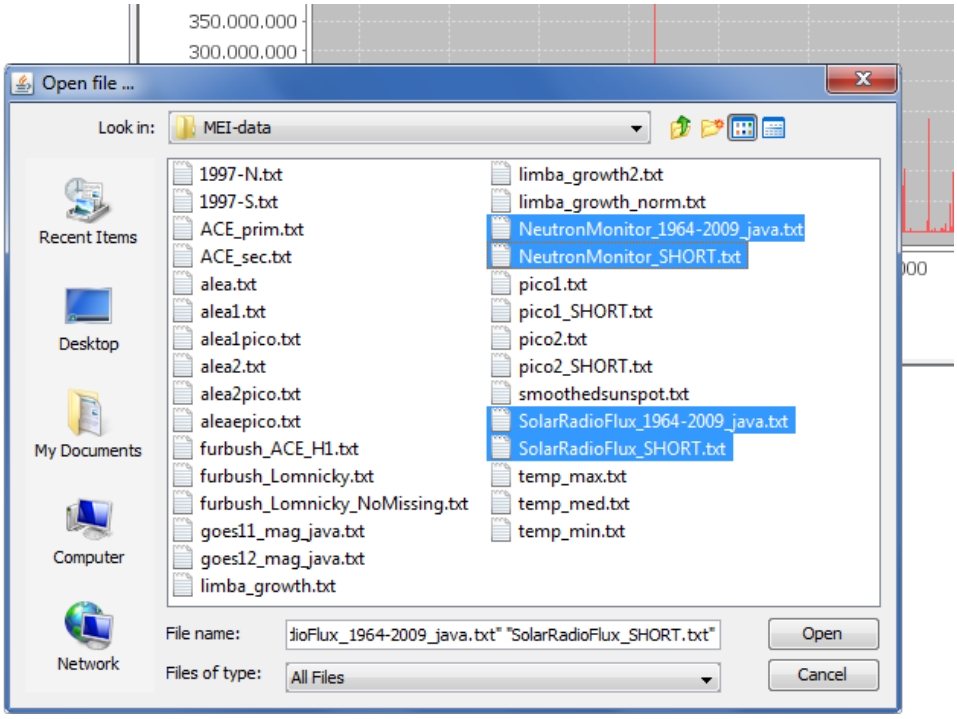


Figura 4.2 Abertura de múltiplos ficheiros

Os ficheiros seleccionados serão então carregados, ficando disponíveis de árvore de listagem. Actualmente a ferramenta suporta ficheiros com apenas um parâmetro, contendo um valor por linha, com ou sem data / hora associada.

4.2.2 Visualização dos Dados

Após carregamento de um ficheiro, é possível visualizar os respectivos dados de forma gráfica na área de visualização, e também de forma textual na área de consola. A visualização gráfica permite funcionalidades de aproximação (*zooming*) seleccionando a área pretendida com o rato.

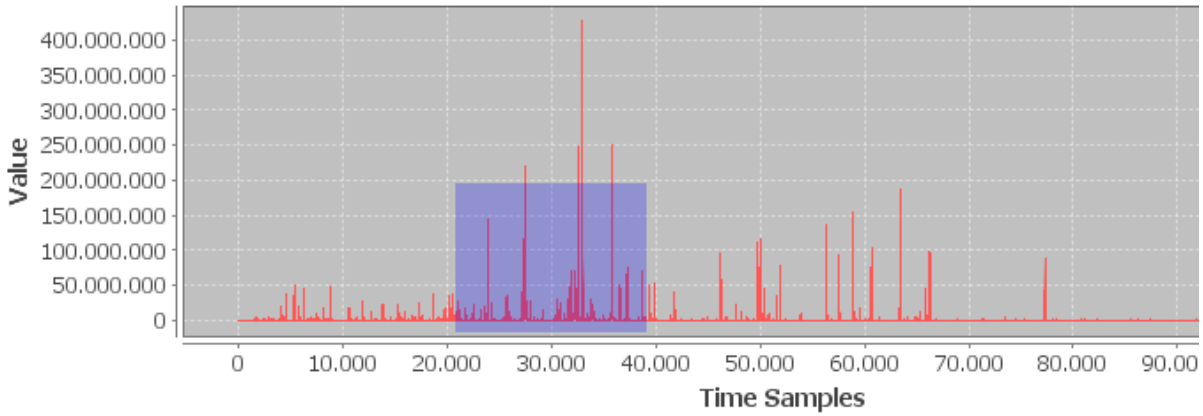


Figura 4.3 Visualização de dados com zoom

Pressionando com o botão direito do rato sobre o gráfico, obtêm-se várias opções, incluindo a possibilidade de copiar a imagem para o *clipboard* ou até gravar como um ficheiro de imagem.

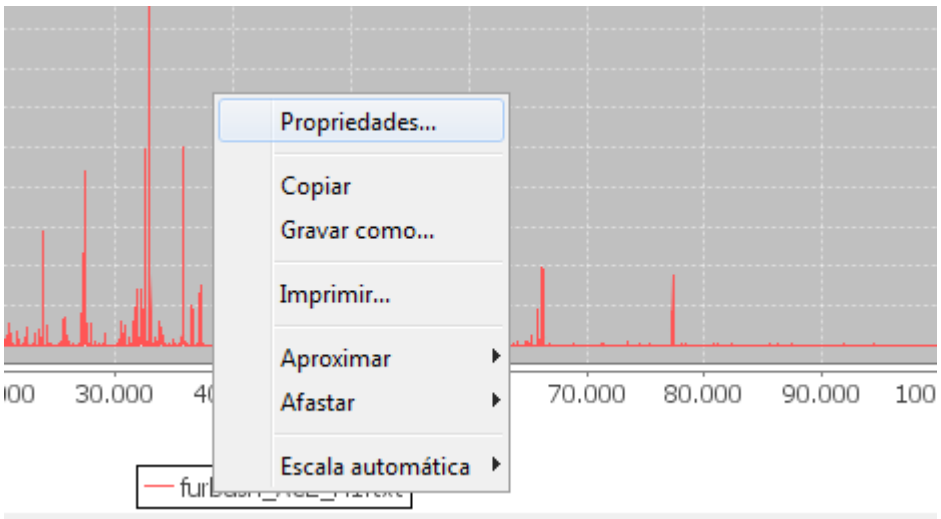


Figura 4.4 Opções no gráfico com botão direito do rato

4.2.3 Funcionalidades por Parâmetro

A maioria das funcionalidades a efectuar sobre um parâmetro está disponível pressionando o botão direito do rato sobre o nome do ficheiro do parâmetro. A Figura 4.5 mostra as opções disponíveis.

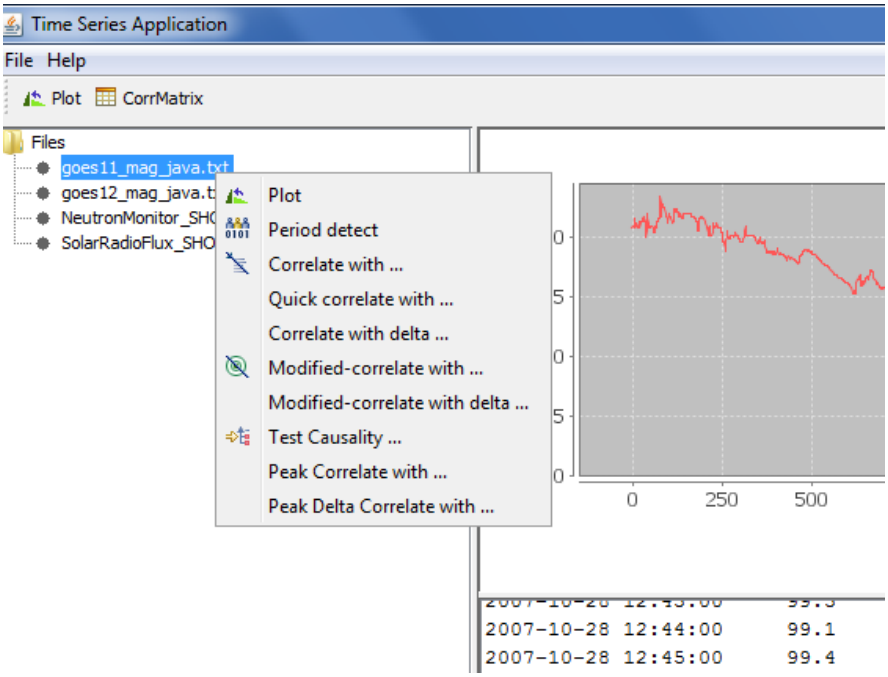


Figura 4.5 Menu de opções por parâmetro

A primeira opção, "**Plot**", permite visualizar os dados de forma gráfica numa janela separada. Isto possibilita a abertura de múltiplas janelas de visualização e de as redimensionar independentemente.

A opção de "**Period detect**" efectua uma análise da periodicidade do sinal do parâmetro seleccionado, recorrendo ao cálculo da auto-correlação. Isto apresenta numa janela separada o resultado do cálculo sucessivo de auto-correlação com desvios temporais associados ao próprio parâmetro (neste caso, usando novamente o parâmetro do número de manchas solares para observar o ciclo solar de 11 anos):

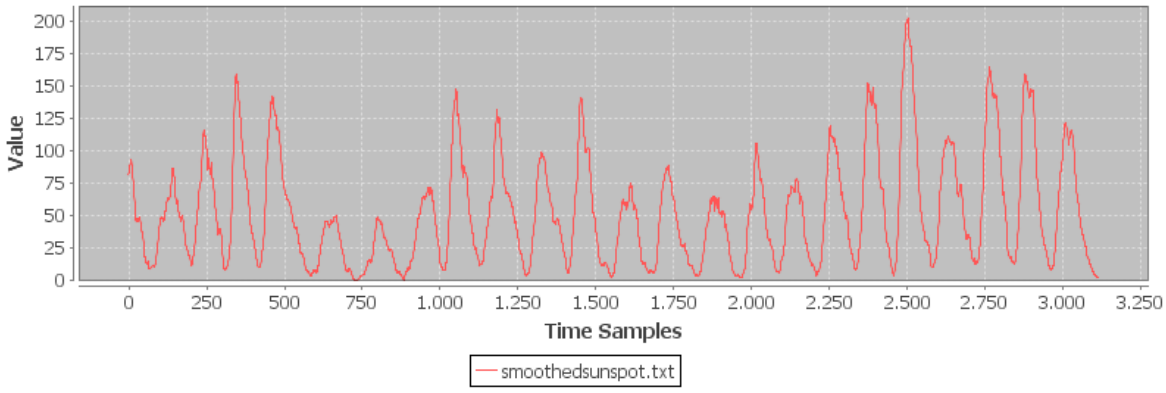


Figura 4.6 Smoothed sunspot number

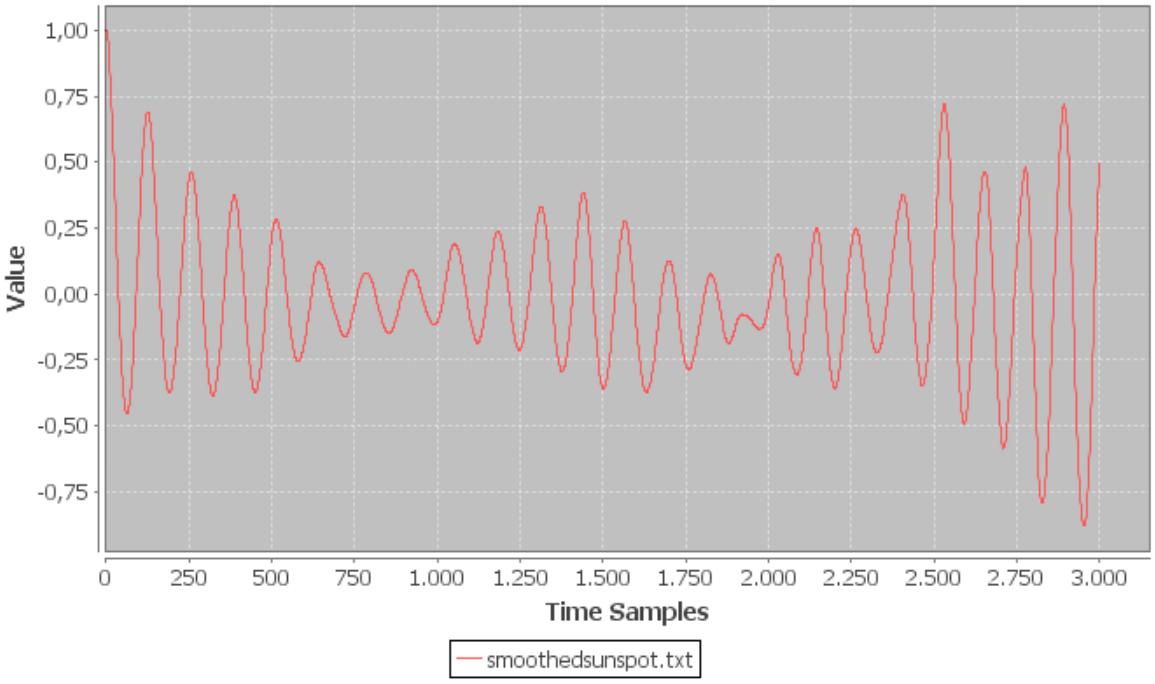


Figura 4.7 Visualização de periodicidades

As restantes opções apresentam uma janela para escolher o parâmetro com o qual efectuar o respectivo cálculo. Seleccionando por exemplo a opção "**Correlate with …**" temos:

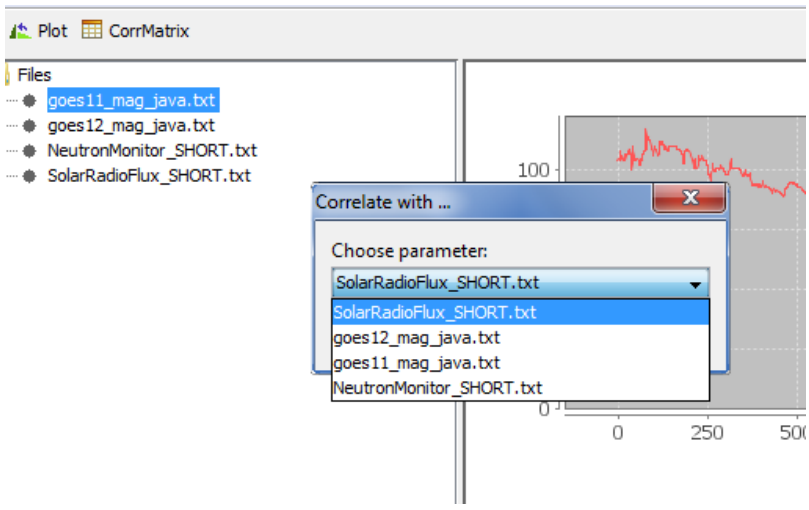


Figura 4.8 Selecção do parâmetro

E após a execução do cálculo, é apresentado o resultado:

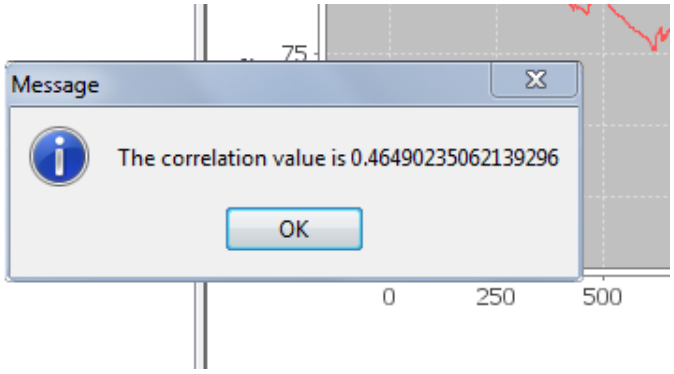


Figura 4.9 Apresentação do resultado

Este mecanismo é também aplicável nas restantes funcionalidades.

4.2.4 Matriz de Correlações

Uma funcionalidade disponível apenas na barra de ferramentas (*toolbar*) é o cálculo da matriz de correlações. Esta opção efectua um cálculo das múltiplas correlações existentes entre os parâmetros carregados na aplicação, e a apresenta-as de uma forma muito prática como se pode ver na figura seguinte:

<>	SolarRadioFlux_SHORT.txt	goes12_mag_java.txt	goes11_mag_java.txt	NeutronMonitor_SHORT.txt
SolarRadioFlux_SHORT.txt	1	-0.22930451244620687	-0.2395524131823749	-0.7931512238867362
goes12_mag_java.txt	-0.22930451244620687	1	0.46490235062139296	0.19735179246918202
goes11_mag_java.txt	-0.2395524131823749	0.46490235062139296	1	0.11373526834015732
NeutronMonitor_SHORT.txt	-0.7931512238867362	0.19735179246918202	0.11373526834015732	1

Figura 4.10 Matriz de correlações

A matriz apresenta uma coloração dependendo da natureza da correlação – verde para correlações positivas, e azul para correlações negativas, sendo que a intensidade da cor será tão maior quanto o valor da correlação, em módulo.

5. Conclusões e Trabalho Futuro

5.1 Conclusões

Esta dissertação aborda os temas da Correlação e da Causalidade entre séries temporais de dados numéricos não categóricos. Neste contexto e neste domínio, procurou-se introduzir algumas alternativas que se julgam inovadoras.

Relativamente ao tema da Correlação, a abordagem usada foi baseada na correlação de Pearson, com vantagem sobre outras como os coeficientes de Kendall e Spearman, por reflectir com maior rigor a relação entre os dados, como foi possível verificar entre séries temporais de natureza conhecida. A métrica usada mostrou-se coerente e suficientemente robusta para não detectar *falsas correlações*. Por extensão da mesma métrica e através do conceito de auto-correlação, foi possível facultar em termos gráficos a leitura simples e intuitiva das periodicidades presentes numa série temporal, facilidade inacessível através do recurso às séries de Fourier, e de obtenção comparativamente custosa através das *wavelets*.

Ainda no domínio da correlação, propôs-se um método de fácil utilização para detecção de correlação local proveniente de picos significativos entre séries temporais aparentemente não correlacionadas.

Na segunda parte desta dissertação, é proposta uma abordagem para detectar a causalidade entre séries temporais de dados numéricos, não categóricos. Trata-se de uma alternativa a outras abordagens que usam modelos probabilísticos, dado que nem sempre é claro como devem ser categorizadas as variáveis numéricas. A ideia central desta abordagem é baseada na assumpção de que um parâmetro representado por um conjunto de dados X determina outro parâmetro Y se, para todas as pequenas variações de valores do domínio em X, correspondem também pequenas dispersões no domínio Y. Com o fim

de medir a dispersão nas partições dos conjuntos de dados, é proposta a métrica *Rd*(.) que mede a dispersão relativa sem a limitação da métrica semelhante *Coeficiente de Variação*.
Esta abordagem permitiu detectar relações causa-efeito a partir de vários pares de séries temporais: foi possível detectar o sentido da relação principal e quão dominante é essa relação relativamente à relação causa-efeito secundária.
Os resultados dos testes mostraram coerência tendo em conta, por um lado, as relações causa-efeito principais propostas pela abordagem e por outro, o conhecimento que se tem das séries temporais usadas para teste.
Esta é também uma alternativa às abordagens que ao usar regressões lineares, estão dependentes da qualidade das funções de regressão para decidir se existem ou não relações de causa-efeito.
Devido à simplicidade desta abordagem, a detecção de relações causa-efeito é rápida, mesmo para séries temporais de grande dimensão.

5.2 Trabalho Futuro

No âmbito da correlação, um requisito a implementar futuramente no protótipo seria a de incluir funções de "alisamento" (*smoothing*) aplicados aos dados antes de iniciar a detecção de correlação. Com isso seria possível alcançar melhorias na detecção de correlações, uma vez que se obtêm as verdadeiras tendências bem como padrões nos dados, por eliminação de ruído no sinal e das mudanças abruptas dos valores. É possível contemplar essa funcionalidade com o protótipo actual, bastando para tal aplicar previamente uma técnica de *smoothing* às séries temporais antes de as carregar na aplicação (apesar de já estar feita uma implementação de teste, usando uma média deslizante *running average*, disponível no protótipo actual).
Outro ponto interessante a explorar seria o de desenvolver um método para detectar de forma automática quais os parâmetros contendo picos relevantes no sinal. Estes seriam candidatos ideais a utilizar como parâmetro de correlação, ou até na pesquisa de potenciais relações causa-efeito. A utilização de transformadas de Fourier para a detecção de picos seria também um tópico a sondar.
Ainda na correlação, poder-se-á abordar o tema de medidas de similaridade ou de proximidade como uma forma de seleccionar potenciais parâmetros candidatos.

No tema da causalidade, um dos aspectos a investigar seria a possibilidade de detectar a causalidade entre conjuntos de parâmetros, e não apenas limitado a detecção entre dois. Em determinados fenómenos, o efeito que se observa num certo parâmetro tem origem na combinação de valores de dois outros parâmetros distintos. A detecção automática da causalidade desses dois parâmetros na alteração observada no terceiro seria um desafio muito interessante.

Bibliografia

Formatação de acordo com norma ISO 690 de Referência Numérica.

1. **Labitzke, K., Matthes, K.** Eleven-year solar cycle variations in the atmosphere: observations, mechanisms and models. *The Holocene.* 2003, Vol. 13, 3, pp. 311-317.

2. **Dorotovič, I., Kudela, K., Lorenc, M., Pintér, T., Rybanský, M.** Evolution of several space weather events connected with Forbush decreases. *Universal Heliophysical Processes.* 2009, Vol. 257, pp. 57-59.

3. **Yousef, Shahinaz M.** The Solar Wolf-Gleissberg Cycle And Its Influence On The Earth. *ICEHM2000.* Cairo, Egypt : s.n., 2000. pp. 267-293.

4. **Falcão, A., Silva, J., Dorotovič, I.** TSCorr: a tool for time-series correlation. *Poster at JENAM, The European Week of Astronomy and Space Science.* Lisbon, Portugal : s.n., 6-10 September, 2010.

5. Introduction to Time Series Analysis. *National Institute of Standards and Technology.* [Online] 2006. http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc41.htm.

6. *SWENET - Space Weather European Network.* [Online] http://www.esa-spaceweather.net/swenet/index.html.

7. *SIDC – Solar Influences Data Analysis Center.* [Online] http://sidc.oma.be/.

8. Space Situational Awareness. [Online] European Space Agency. http://www.esa.int/esaMI/SSA/index.html.

9. ESA Space Situational Awareness. *ESA.* [Online] 2011. http://www.esa.int/esaMI/SSA/SEM0MNIK97G_0.html.

10. **Howson, C., Urbach, P.** *Scientific Reasoning: The Bayesian Approach.* s.l. : Open Court, 1993. ISBN 9780812692341.

11. **Kendall, M.** A New Measure of Rank Correlation. s.l. : Biometrika, 1938. 30, pp. 81-89. doi:10.1093/biomet/30.1-2.81.

12. **Spearman, C.** The Proof and Measurement of Association Between Two Things. *American Journal of Psychology.* 1904, 15, pp. 72-101.

13. **Adler, J., Parmryd, I.** Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytometry Part A.* 2010. 77A, pp. 733–742.

14. **Saad, Ziad S., Glen, Daniel R., Chen,G., Beauchamp, Michael S., Desai, R., Cox, Robert W.** A new method for improving functional-to-structural MRI alignment using local Pearson correlation. 2009, Vol. Volume 44, Issue 3, pp. 839-848.

15. **Chakraborty, S., Rabello-Soares, M. C., Bogart, R. S., Bai, T.** Investigating the correlation between high-frequency global oscillations and solar. *Journal of Physics: Conference Series.* 2011.

16. **Pearson, Karl.** Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. Royal Soc. London Ser. A.* 1896, 187, pp. 253-318.

17. **Pearson, K.** Notes on the History of Correlation. s.l. : Biometrika Trust, Oct. 1920. Vol. 13, 1.

18. **Bloomfield, Peter.** *Fourier Analysis of Time Series - An Introduction.* s.l. : Wiley & Sons, 2000. 0-471-88948-2.

19. **Bracewell, R. N.** *The Fourier Transform and Its Applications.* 3rd Ed. Boston : McGraw-Hill, 2000. ISBN 0071160434.

20. **Grafakos, Loukas.** *Classical and Modern Fourier Analysis.* s.l. : Prentice-Hall, 2004. ISBN 0-13-035399-X.

21. **Chan, Kin-pong, Fu, Ada Wai-chee.** *Efficient Time Series Matching by Wavelets.* Sydney, Australia : s.n., 1999. 0-7695-0071-4.

22. **Hubbard, Barbara Burke.** *The world according to wavelets: the story of a mathematical technique in the making.* Natick, MA, USA : A. K. Peters, Ltd., 1996. 1-56881-047-4.

23. **Sallo, S.** Auto-correlation Functions and Solar Cycle Predictability. Pisa, Italy : s.n., 2000. Retrieved from http://arxiv.org/abs/astro-ph/0010106.

24. **Vega, V., Duarte, C., Ordóñez, G., Kagan, N.** Selecting the Best Wavelet Function for Power Quality Disturbances Identification Patterns. *Harmonics and Quality of Power.* 2008.

25. **Schwabe, H.** Solar Observations During 1843. *Astronomische Nachrichten.* 1843, Vol. 20, 495.

26. **Gleissberg, W.** The Eighty-year Sunspot Cycle. 1958, 68, pp. 148-152.

27. **Dorotovič, I., Kudela, K., Lorenc, M., Rybanský, M.** On 17 – 22 January 2005 Events in Space Weather. *Solar Physics.* 2008. Vol. 250, 2, pp. 339-346.

28. **Polanski, A., Polanska, J., Jarzab, M., Wiench, M., Jarzab, B.** Application of Bayesian networks for inferring cause-effect relations from gene expression profiles of cancer versus normal cells. *Math Biosciences.* October de 2007, 209, pp. 528-46.

29. **Pearl, Judea.** *Causality: Models, Reasoning and Inference.* s.l. : Cambridge Univ. Press, 2000.

30. **Pearl, J.** Causal Inference. *Journal of Machine Learning Research.* 2010, pp. 39-58.

31. **Srikant, R., Agrawal, R.** Mining Quantitative Association Rules in Large Relational Tables. *ACM-SIGMOD Conference on Management of Data.* 1996.

32. **Bayardo Jr, R.J., Agrawal, R., Gunopulos, D.** Constraint-Based Rule Mining in Large, Dense Databases. *Data Mining and Knowledge Discovery Journal.* 2000.

33. **Swanson, N.R., Granger, C.W.J.** Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association.* 1997, Vol. 92, 437, pp. 357-367.

34. **Chu, T., Glymour, C.** Search for Additive Nonlinear Time Series Causal Models. *Journal of Machine Learning Research.* 2008, pp. 967-991.

35. **Lindgren, B.W.** *Statistical Theory, 3rd Ed.* New York : MacMillan Publishing Co., 1976.

Apêndice A – Poster JENAM2010

Apresenta-se de seguida o *poster* apresentado na conferência JENAM2010 – Joint European and National Astronomical Meeting.

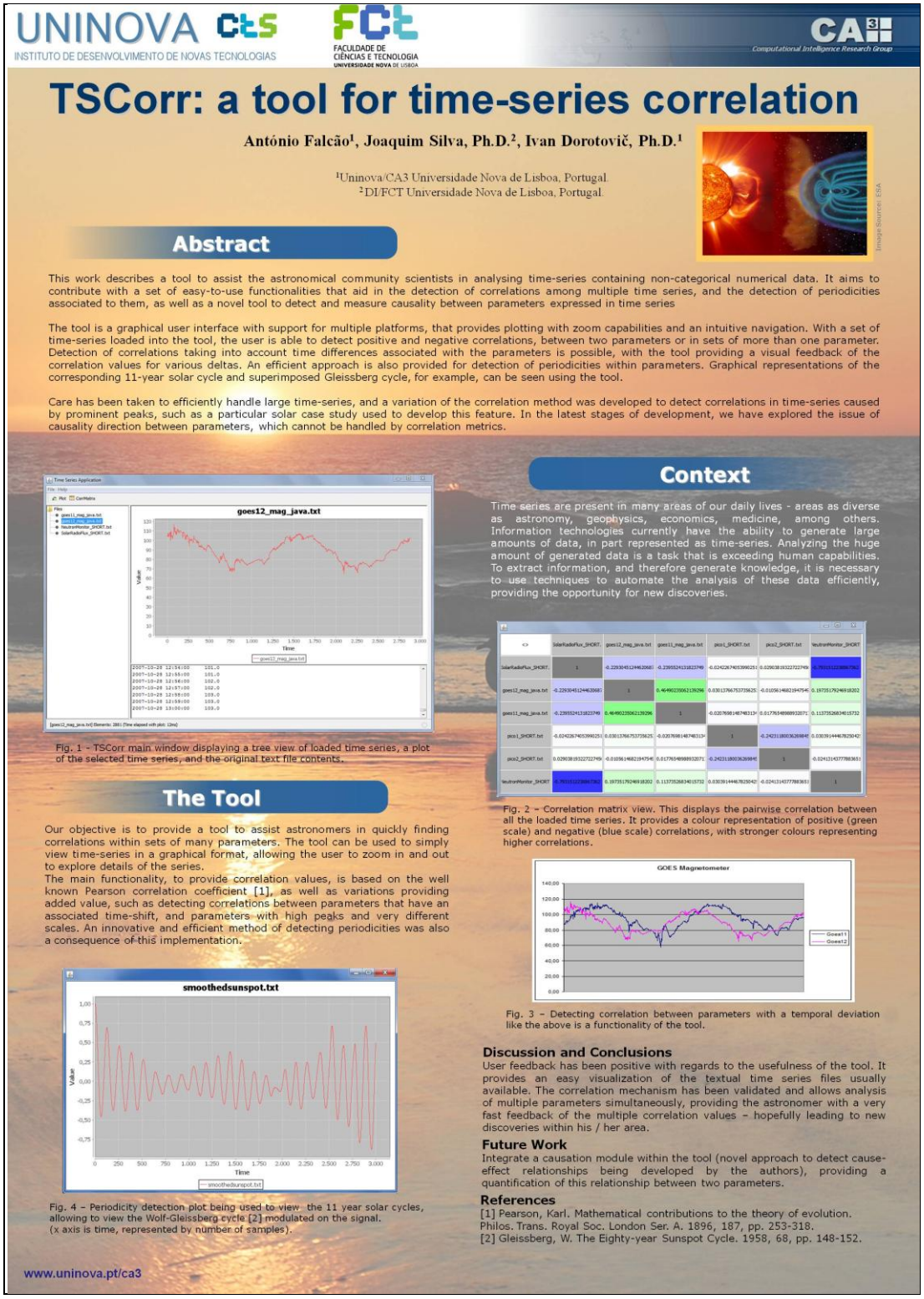


Figura 0.1 *Poster* apresentado no JENAM2010

(fim do documento)