Exercício 8: Análise de Dados (parte I)

Ernesto González nº(52857) , Rodrigo Januário nº(53087), André Nunes nº(52868)

Análise de resultados das eleições americanas de 2008 e 2016. Estudo de séries temporais de medição de número de Wolf, associado às manchas solares e correlação com a temperatura média global mensal. Análise da ocorrência de terramotos em Lisboa.

I. ANÁLISE DE DADOS ELEITORAIS

A. Eleições presidenciais americanas 2008

Usando a biblioteca Statistics do python aplicamos as funções: `mean` para calcular a media, `median` para calcular a mediana e `pvariance` para calcular a variância. Obtendo os seguintes valores para cada candidato:

De seguida, construíram-se dois histogramas, o da Figura 1 (a) onde se contabilizaram todos os municípios e o da Figura 1 (b) onde se consideraram apenas os municípios com mais de 20 000 eleitores. Para ambos os gráficos criou-se uma relação entre $F_v = v_O - v_M$ e N_c, onde F_v representa a diferença entre os votos do Obama v_O e os votos do McCain v_M, e N_c a frequência absoluta, ou seja, o número de vezes que observamos diferença relativa (dada por F_v/n, em que n são o numero total de eleitores por município) dentro do intervalo a que corresponde cada barra do histograma. Para fazer os histogramas realizou-se o seguinte processo: para cada município, se o número de votos de Obama fosse superior aos de McCain, a diferença relativa era um valor positivo e, portanto, F_v era guardada numa lista de valores positivos, indicando uma maior probabilidade do Obama sair vencedor. Caso contrário, a diferença relativa F_v era guardada numa lista de diferenças relativas negativas, aumentando as hipóteses de McCain ganhar as eleições.

Tabela I. Média, mediana e variância de votos por condado, para cada candidato, nas eleições presidenciais americanas de 2008.

	Barack Obama	John McCain
Média	21407.07	18726.90
Mediana	4473	6244
Variância	5273664087.51	1848072555.97

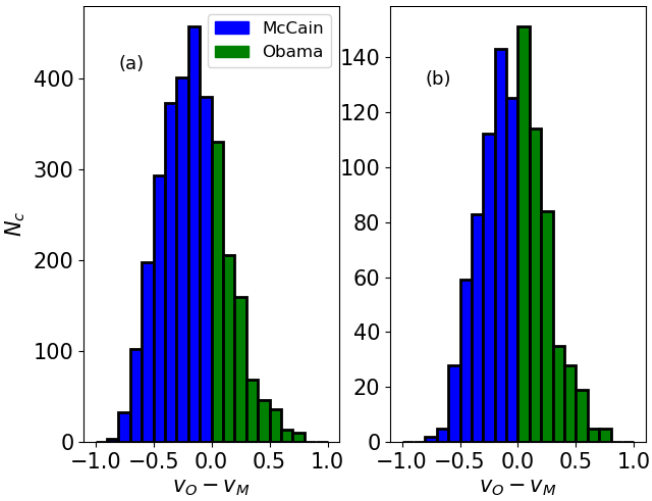

Figura 1. Histogramas da diferença relativa de votos por condado, $v_O - v_M$, para as eleições presidenciais americanas de 2008, considerando todos os condados no histograma (a) e apenas os condados com mais de 20000 no histograma (b).

Pela análise da Figura 1 (a), concluímos que o histograma segue uma distribuição Gaussiana com valor médio de -0.3, resultado de uma maior frequência de municípios em que McCain obteve mais votos que o Obama, o que deveria refletir uma clara vitoria por parte de McCain. No entanto, ao eliminar os municípios com menos de 20000 eleitores, Figura 1 (b), as diferenças relativas que ultrapassam os 0.5 são desprezáveis e deixamos de ter uma distribuição aproximadamente Gaussiana. Levando a uma vitoria por parte de Barack Obama por uma pequena margem, comprovando a teoria dada no enunciado do trabalho.

B. Eleições presidenciais americanas 2016

As eleições presidenciais americanas 2016, na sua fase final, foram concorridas por Hillary Clinton e Donald Trump. Nesta subsecção seguimos o mesmo procedimento seguido nas eleições de 2008.

A Tabela II apresenta a média, mediana e variância de votos por condado, para cada candidato.

A Figura 2 apresenta os histogramas das diferenças relativas de votos por condado, $v_C - v_T$, em que v_C são os

Tabela II. Média, mediana e variância de votos por condado, para cada candidato, nas eleições presidenciais americanas de 2016.

	Hillary Clinton	Donald Trump
Média	20042.195	19635.71
Mediana	3155	7169
Variância	5168264802.05	1632003748.04

votos obtidos por Hillary Clinton e v_T os votos obtidos por Donald Trump no condado. Semelhante à subsecção I.A., na Figura 2 (a) foram considerados todos os condados americanos e na Figura 2 (b) apenas condados com mais do que 20000 votantes.

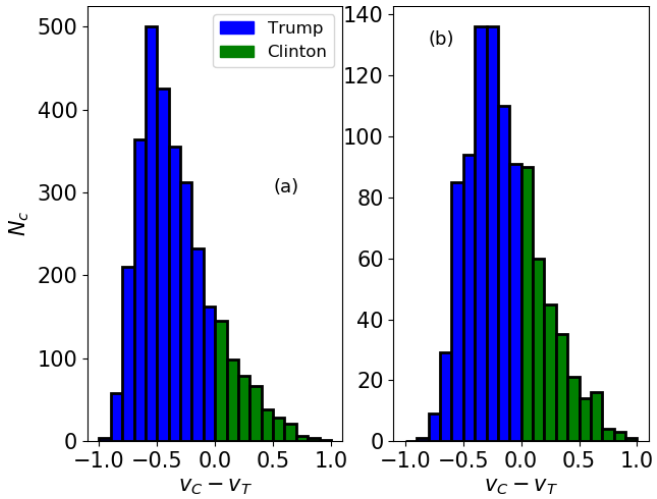

Figura 2. Histogramas da diferença relativa de votos por condado, $v_C - v_T$, para as eleições presidenciais americanas de 2016, considerando todos os condados no histograma (a) e apenas os condados com mais de 20000 no histograma (b).

II. EVOLUÇÃO DAS MANCHAS SOLARES

A. Série temporal da medição do número de Wolf

Nesta parte do trabalho, pretende-se estudar a evolução das manchas solares, com base em dados recolhidos mês a mês entre janeiro de 1749 e dezembro de 1983. As manchas solares são medidas pelo número de Wolf, W, dado por

$$W = k(10g + f), \tag{1}$$

onde k é um fator da escala, f o número de manchas e g o número de grupos. Para tal, começa-se por traçar o gráfico das manchas solares em Wolf em função do mês em que foi medido (Figura 3).

Como podemos observar, a curva aparenta ter uma certa periodicidade. No entanto, devido à natureza dos dados, estes apresentam *ruído*, que pode estar associado à incerteza de medição ou por flutuações características. Assim, para podermos encontrar a periodicidade, precisamos de traçar o gráfico da autocorrelação, isto é, como é que a série de dados se relaciona consigo mesmo no tempo. Para tal, aplicámos a equação

$$r_k = \frac{\frac{1}{N-k}\sum_{t=1}^{N-k}(y_t - \bar{y})(y_{y+k} - \bar{y})}{\frac{1}{N}\sum_{t=1}^{N}(y_t - \bar{y})^2} \tag{2}$$

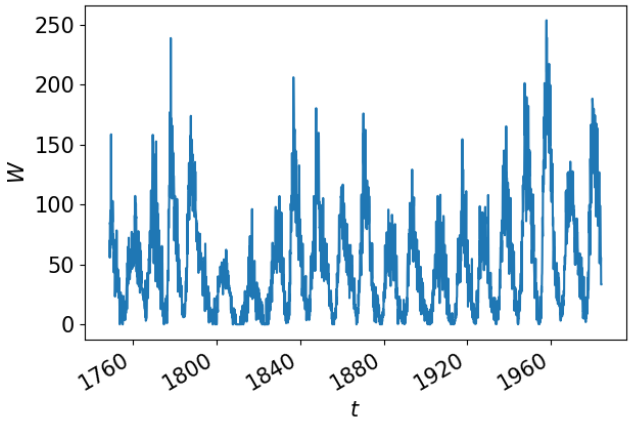

Figura 3. Gráfico do número de Wolf, W, medido em cada mês, para meses entre janeiro de 1749 e dezembro de 1983.

em todos os valores de t, iterando sobre k, para encontrarmos o período que melhor se ajusta aos dados. Obteve-se então o gráfico da Figura 4.

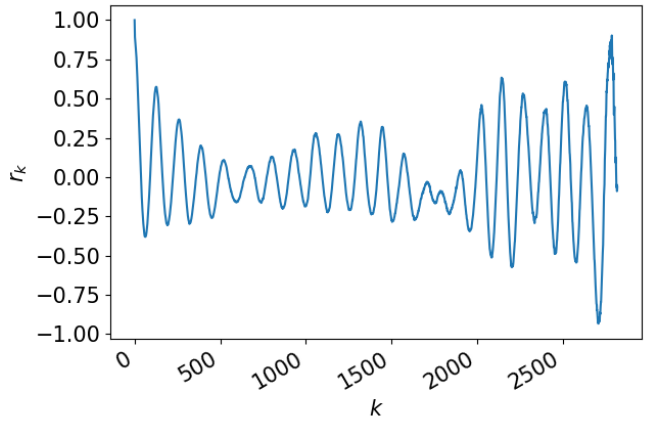

Figura 4. Gráfico da função autocorrelação, r_k, para a medição mensal do número de Wolf, medido entre janeiro de 1749 e dezembro de 1983, e intervalos de tempo k de até 2800 meses.

Da Figura 4, vemos que a medição mensal do número de Wolf tem dependência sazonal: o gráfico da autocorrelação apresenta máximos e mínimos locais de forma periódica. Significa que o número de Wolf $W(t)$ depende do número de Wolf $W(t-T)$ para o período T das *seasons* de máximos e mínimos. Esse período T pode ser obtido medindo a diferença de tempo entre os dois primeiros mínimos locais no gráfico da autocorrelação. Desta forma, obtemos $T = 128$.
De forma a excluir a sazonalidade do número de Wolf, estudamos o comportamento de $W(t) - W(t-T)$. Na

Figura X encontra-se o gráfico desta função.

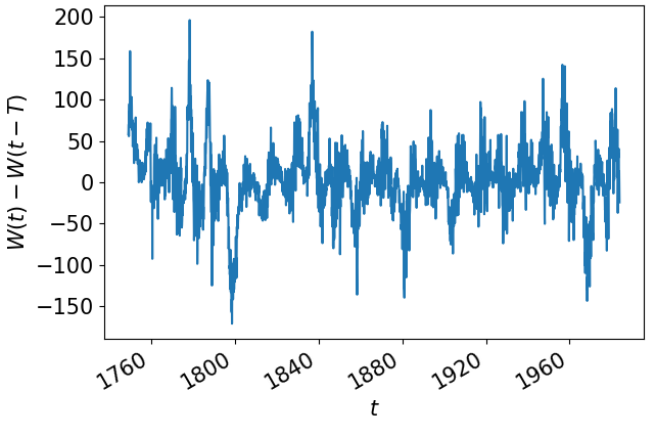

Figura 5. Série temporal da medição mensal do número de Wolf, W, entre janeiro de 1749 e dezembro de 1983, aplicadas diferenças sazonais, $W(t) - W(t - T)$, para um período $T = 128$ meses.

B. Temperatura média global e número de Wolf

Considere-se a variação da temperatura média global. A Figura 6 apresenta a série temporal da variação da temperatura média global mensal à superfície.

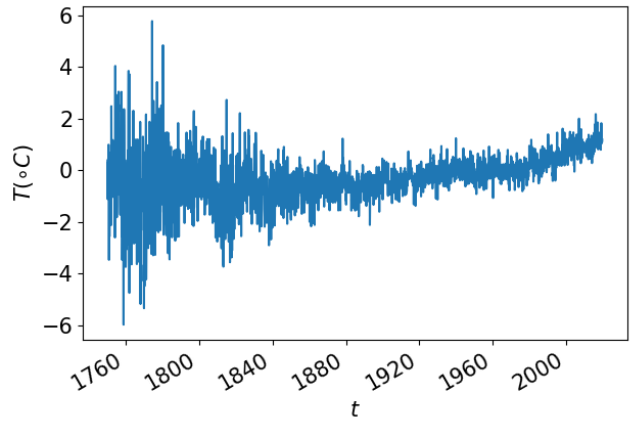

Figura 6. Série temporal da temperatura média global mensal à superfície, T, em graus centígrados, entre janeiro de 1951 e setembro de 2019.

Consideremos a trajetória $W(T(t))$ da medição mensal do número de Wolf, $W(t)$ e da temperatura média global mensal à superfície, $T(t)$. Encontra-se na Figura X os pontos desta trajetória desde janeiro de 1750 até dezembro de 1983, com exceção de dezembro de 1751.

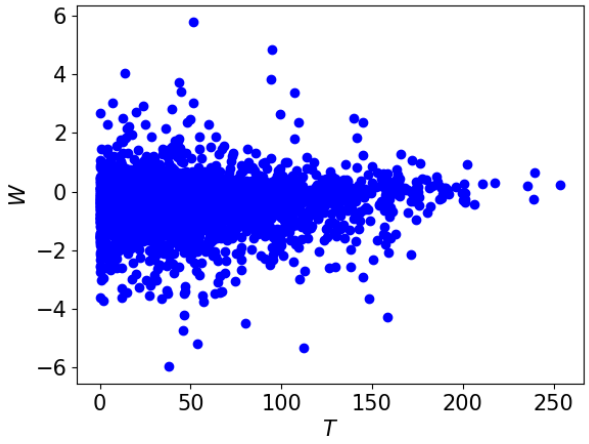

Figura 7. Gráfico dos pontos da trajetória da medição mensal do número de Wolf, W, em função da temperatura média global mensal superficial, T, em centígrados.

Da análise da Figura 7, vê-se que os pontos da trajetória da medição mensal do número de Wolf em função da temperatura média global mensal superficial, $W(T)$, estão concentrados num triângulo isósceles com base em $T = 0$ e e altura ao longo de $W = 0$, logo há algum tipo de correlação entre W e T medidos num determinado mês. O coeficiente de correlação de Pearson entre W e T é

$$R_P = \frac{cov(W,T)}{\sigma_W \sigma_T} = 0.06667595918293301. \qquad (3)$$

O coeficiente de correlação de Pearson, R_P, neste caso é bastante baixo, não refletindo a correlação observada na Figura 7. Deve-se a que este coeficiente apenas considera correlação linear entre as duas variáveis, que não é o nosso caso.

III. FREQUÊNCIA DE TERRAMOTOS

Consideremos a ocorrência de terramotos na Zona Metropolitana de Lisboa. Recorreu-se ao *Mathematica* para obter os dados sobre os terramotos com epicentro nesta área. Na Figura 8 encontra-se a série temporal da intensidade na escala de Richter, I, de terramotos encontrados.

Com destaque para os terramotos depois de 1980 na Figura 9, para melhor visualização.

Considere-se agora o histograma de intensidade dos

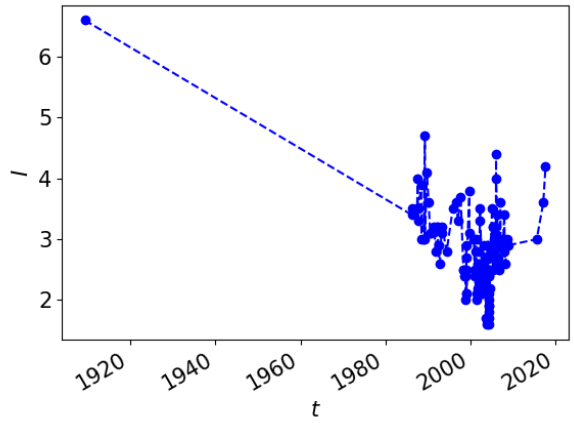

Figura 8. Série temporal da intensidade, I, na escala de Richter, de terramotos com epicentro na Zona Metropolitana de Lisboa nos últimos 120 anos. Algumas ocorrências poderão não ter sido consideradas devido à falta de dados.

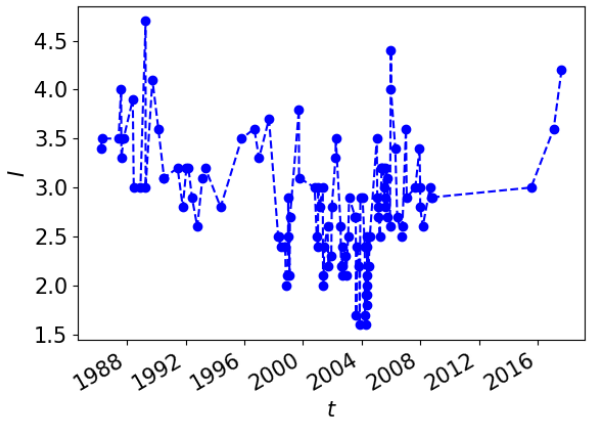

Figura 9. Série temporal da intensidade, I, na escala de Richter, de terramotos com epicentro na Zona Metropolitana de Lisboa nos últimos 50 anos. Algumas ocorrências poderão não ter sido consideradas devido à falta de dados.

terramotos encontrados desde 1960, na Figura 10.

A Lei de Gutenberg-Richter relaciona a magnitude, I, e a frequência absoluta de terramotos de pelo menos intensidade I, N, e é expressa por

$$N = 10^{a-bI} \tag{4}$$

em que a e b são constantes. Na Figura 11 encontra-se o histograma do número de terramotos, N, com pelo menos intensidade I.

Na Figura 11, vemos como os nossos dados dos sismos

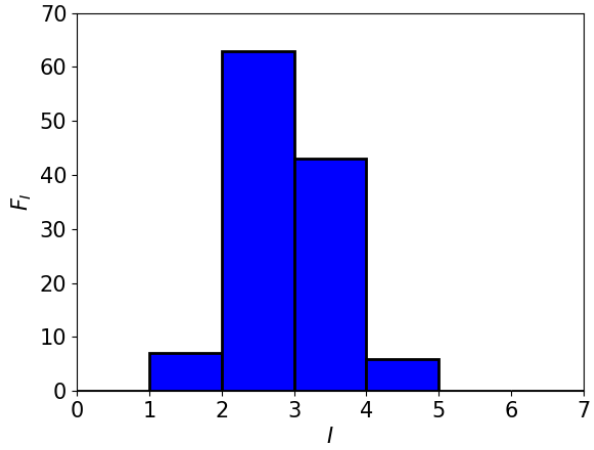

Figura 10. Histograma da frequência de terramotos dentro do intervalo de intensidade associado à barra respetiva, F_I, para os terramotos encontrados com epicentro na Zona Metropolitana de Lisboa.

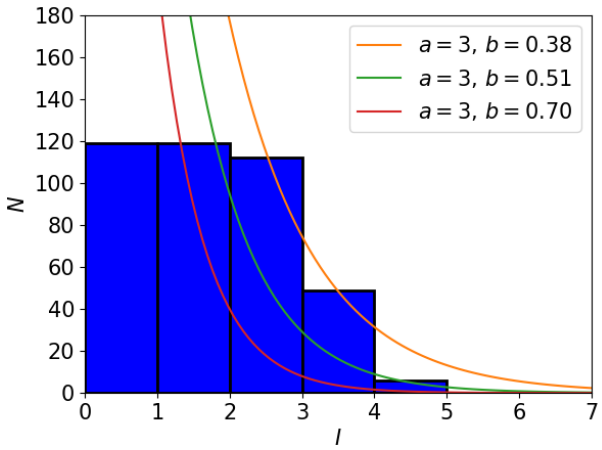

Figura 11. Histograma do número de terramotos, N, com pelo menos intensidade I e previsão feita pela Lei de Gutenberg-Richter, $N = 10^{a-bI}$, para algumas combinações de a e b, para terramotos encontrados com epicentro na Zona Metropolitana de Lisboa nos últimos 120 anos.

em Lisboa não respeitam a Lei de Gutenberg-Richter. Uma das razões pode ser que tivemos de descartar muitas ocorrências de terramotos nos dados iniciais porque não continham a intensidade do sismo, que é um dos parâmetros que pretendemos estudar. Ainda mais, os dados foram originalmente retirados do `Earthquake` do *Mathematica*, podendo esses mesmos estar incompletos.