Lista de exercícios Matemática Computacional
Parte B – Prof. Dr. Reinaldo Rosa - 2020

Denis M. A. Eiras

Exercício 6 - Descrição

6.1. Considere as séries temporais listadas na tabela dataset_signal e obtenha, para cada série, os valores respectivos do seguintes atributos: S 2 , K, β (via PSD) e α (via DFA). Confira para todas as series se β (via PSD) está bem ajustado a partir da formula WKP: $\beta = 2\alpha - 1$. Construa dois espaços de parâmetros EPSB-K-means: S 2 x K x β e EDF-K-means: S 2 x K x α .

6.2. Classifique, nos espaços de parâmetros do exercício anterior, as séries temporais: (a) ST-Sol3GHz, (b) ST-surftemp504 e (c) ST-OWS_NDC_C

6.3. Aplique k-means para todas as series ST-OWS_NDC_Covid1 considerando os seguintes Espaços de atributos: S 2 x α e K x α. Obtenha os melhores agrupamentos, identifique os grupos e discuta os resultados.

Exercício 6.1 – Detalhes da implementação

Foi criada a função calcula_df_estatistico_por_familia_e_sinal, para gerar arquivos csv das tabelas estatísticas dos espaços de parâmetros requisitados do enunciado.

As seguintes funções de outros exercícios foram utilizadas para gerar os dados da tabela data_set_sinal, reaproveitando assim toda a lógica:

Exercicio1.exercicio1_1 - gerador_de_sinais_aleatorios
Exercicio2.exercicio2 - gerador_de_sinais_colored_noise
Exercicio3.exercicio3 - gerador_de_sinais_pmodel
Exercicio5.exercicio5_1 - gerador_de_sinais_logisticos, gerador_de_sinais_henon

Exercício 6.2 – Detalhes da implementação

Para a leitura de arquivos gerada no exercício anterior e a leitura das séries Covid, Sol e SurfTemp, foi reutilizado o leitor genérico implementado no exercício 4.2. Para calcular as estatísticas, foi reaproveitada uma função do exercício 6.1. Para realizar o agrupamento das séries dentro do espaço de parâmetros, foi utilizado o k-means, do exercício 1.3. Em resumo, foram reaproveitados:

Exercicio4.exercicio4_2_2 - ler_serie_generica_de_arquivo_ou_url
Exercicio6.exercicio6_1 – calcula_df_estatistico
Exercicio1.exercicio1_3 - k_means_e_metodo_do_cotovelo

Exercício 6.3 – Detalhes da implementação

O programa implementado cria séries estatísticas à partir de uma coluna agrupadora e uma coluna de valores da série, configuráveis no programa. Também é possível configurar valores da coluna agrupadora a serem removidos.

Exercício 6.1 – Análise

Para verificar se β (via PSD) está bem ajustado a partir da formula WKP: β = 2 α – 1, foram gerados gráficos contendo os pontos, em azul, no plano α x β, para cada série, como mostram os gráficos da figura 1. Em seguida, foi gerada uma reta que interpola esses pontos, em azul claro, e uma outra reta gerada calculada a partir da equação β = 2 α – 1, em rosa.

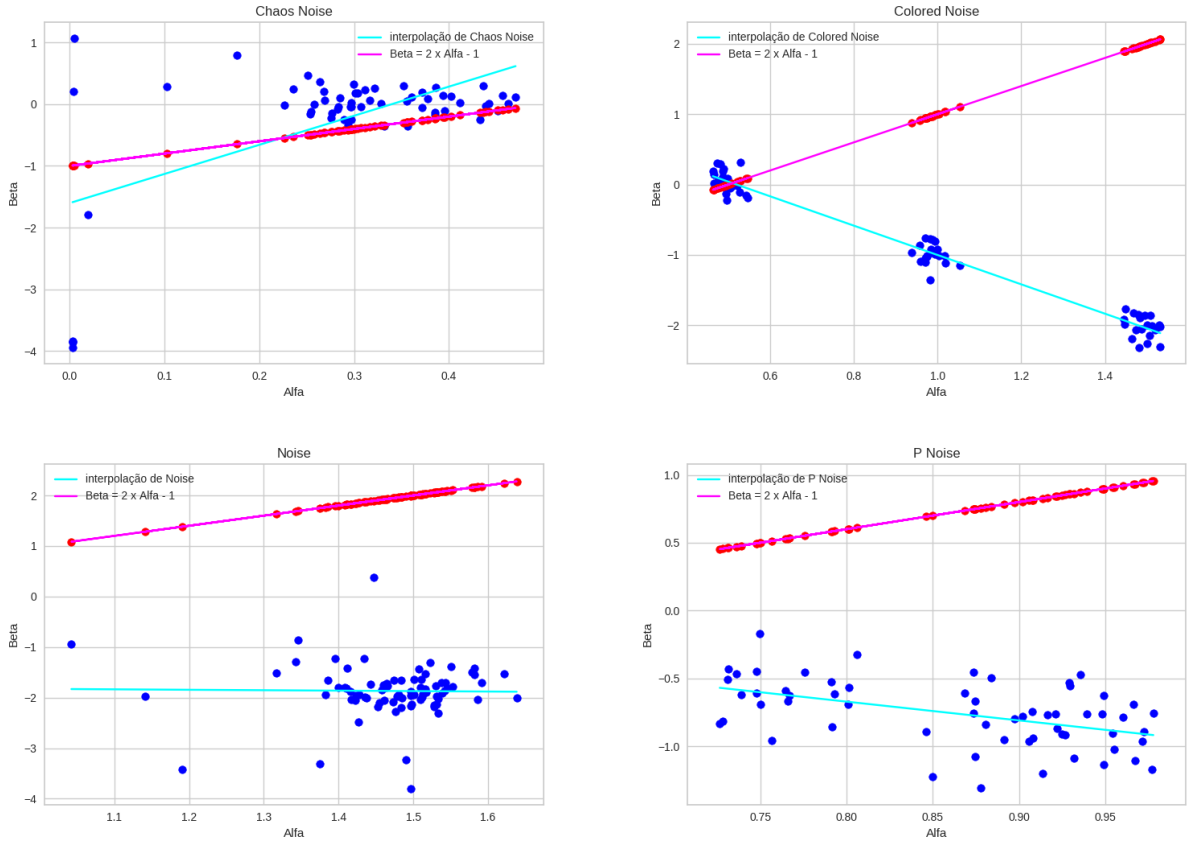

Figura 1. Retas interpoladoras dos pontos, em azul, e retas calculadas À partir da equação β = 2 α – 1, em rosa.

Observando os gráficos, a reta Chaos é a que mais se aproxima da reta calculada pela equação, o que só ocorreu devido à alguns pontos mais fora da reta fora do agrupamento.

Os pontos Colored Noise possuem uma inclinação de reta que parece ser a negativa da equação, isto é, β = -2 α + 1.

Para se ajustar ao sinal Noise, a equação precisaria de um descréscimo de aproximadamente mais 3 pontos em y, e algum ajuste na inclinação.

O sinal P Noise também requer um ajuste em y e na inclinação da reta.

Exercício 6.1 - Conclusão

Observando os gráficos, a reta Chaos é a que mais se aproxima da reta calculada pela equação.

Exercício 6.2 – Análise

Na tentativa de verificar em qual espaçõ de parâmetros, ESPB ou EDF, onde cada uma das séries poderia estar contida, foram gerados dois datasets para cada uma das séries Covid, Sol e

Surf_Temp: Um dataset contém o espaço ESPB e espaço (um ponto) de uma série, e outro dataset contendo o espaço EDF e o espaço (um ponto) de uma série. Isto é:

- Espaço ESPB (um ponto para cada sinal) + Espaço Covid (um ponto)
- Espaço ESPB (um ponto para cada sinal) + Espaço Sol (um ponto)
- Espaço ESPB (um ponto para cada sinal) + Espaço Surf_Temp. (um ponto)
- Espaço EDF (um ponto para cada sinal) + Espaço Covid (um ponto)
- Espaço EDF (um ponto para cada sinal) + Espaço Sol (um ponto)
- Espaço EDF (um ponto para cada sinal) + Espaço Surf_Temp. (um ponto)

Para verificar se as séries poderiam estar em agrupamentos dos espaços ESPB e EDF, o K-means foi executado para cada um dos itens acima, utilizando k = 2, na tentativa de se verificar se o ponto de cada série se encontrava dentro do mesmo agrupamento ou em um agrupamento separado.

Série COVID-19

As estatísticas de novos casos foram geradas para o país Estados Unidos, até o dia 26/05.

Estatísticas da série da Covid:

variancia_ao_quadrado	curtose	beta	alfa
2.982562e+16	-1.21331	0.078597	1.30768

A figura 2 mostra que, o ponto verde representa a série da COVID, como pode ser observado nas estatísticas, nos espaços ESPB e EDF. É possível verificar que o ponto verde não se agrupa no espaço de parâmetros ESPB e EDF, principalmente devido à alta variância, na casa dos 16 dígitos (apesar de haver um ponto da série noise na casa dos 9 dígitos) . O ponto da covid se enquadra dentro de beta mas não se enquadrar dentro de alfa.

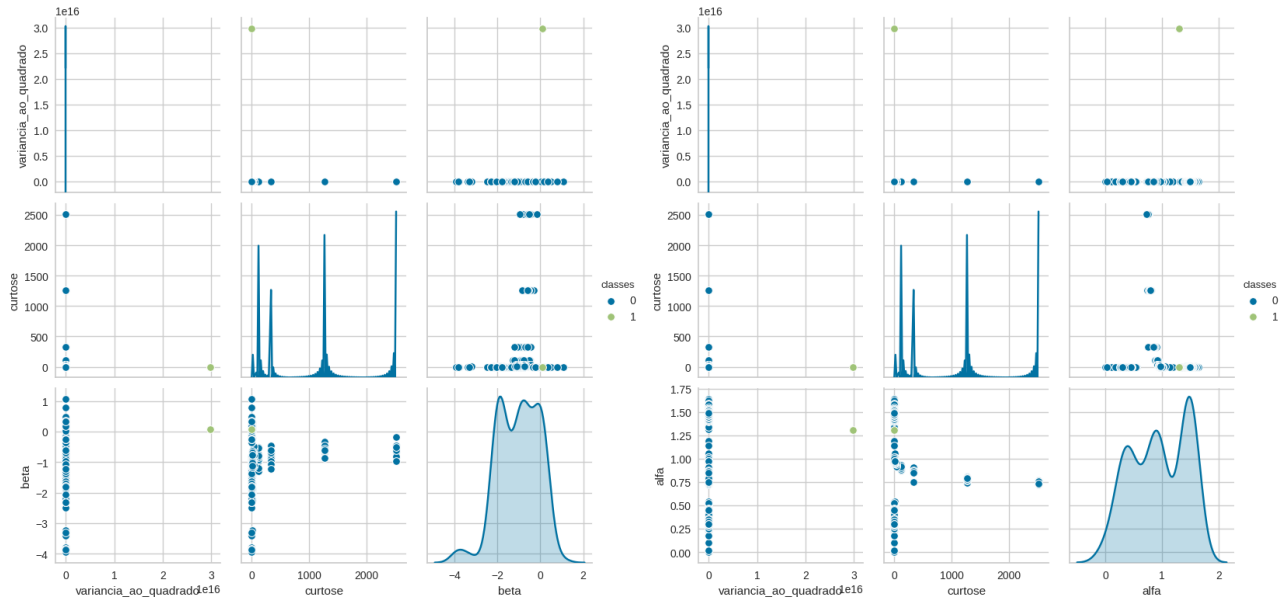

Figura 2. Ponto da série Covid no espaço ESPB, à esquerda, e no espaço EDF, à direita

Série Sol 3ghz

Estatísticas da serie sol:

variancia_ao_quadrado	curtose	beta	alfa
2.110144e+07	3.809948	-1.851297	1.367474

A variância da série Sol é alta, mas acabou se agrupando aos pontos azuis, provavelmente devido a grande variância da série noise, a qual registra variância na casa dos 9 dígitos. O valor de beta da série está dentro do espaço ESPB, e o valor alfa fica fora do espaço EDF.

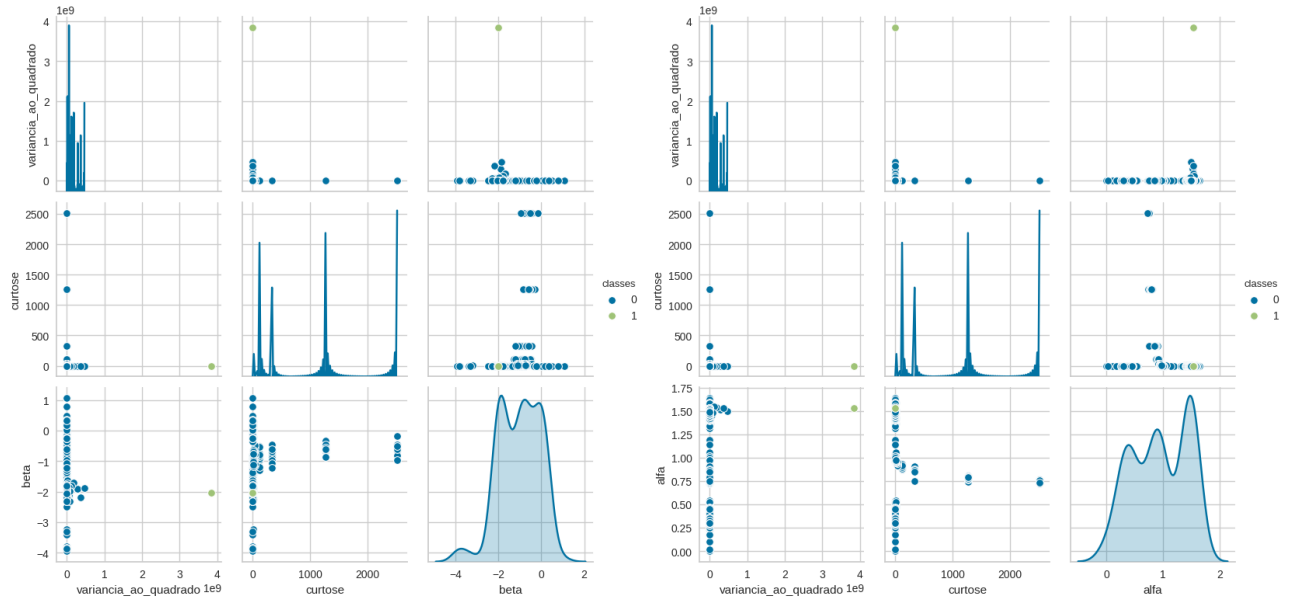

Figura 3. Ponto da série Sol no espaço ESPB, à esquerda, e no espaço EDF, à direita

<u>Série Surface Temperature</u>

Estatísticas da serie surf_temp:

variancia_ao_quadrado	curtose	beta	alfa
0.290777	0.392488	-1.5375	0.740507

A variância da série Surface Temperature é a que mais se aproxima dos espaços ESPB e EDF, assim como a curtose. A série se agrupa aos espaços. Os valores de alfa e beta da série estão dentro do espaço ESPB.

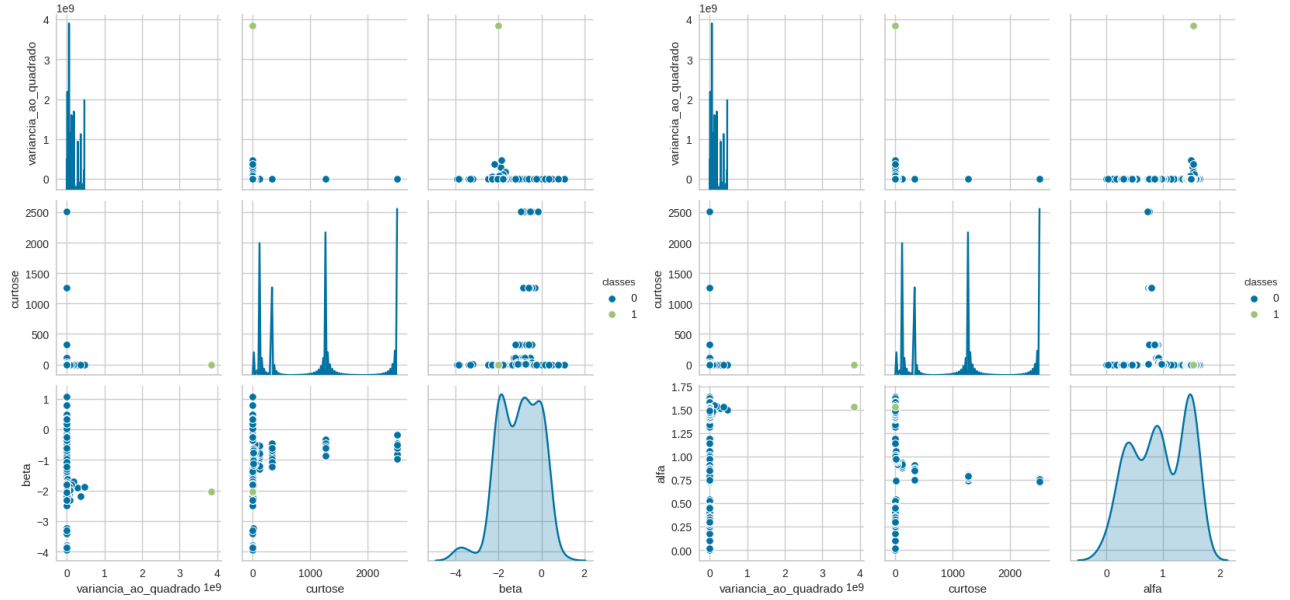

Figura 4. Ponto da série Surface Temperature no espaço ESPB, à esquerda, e no espaço EDF, à direita

Para avaliar a influência dos pontos noise na geração dos espaços de parâmetros ESPB e EDF, os espaços foram refeitos sem o espaço noise, e a figura 5 foi gerada. Esta alteração fez com

que alguns pontos da série Surface Temperature continuassem agrupados ao agrupamento azul, mas alguns outros pontos da série p-noise foram para o agrupamento verde, devido à variância, como mostra a figura 5.

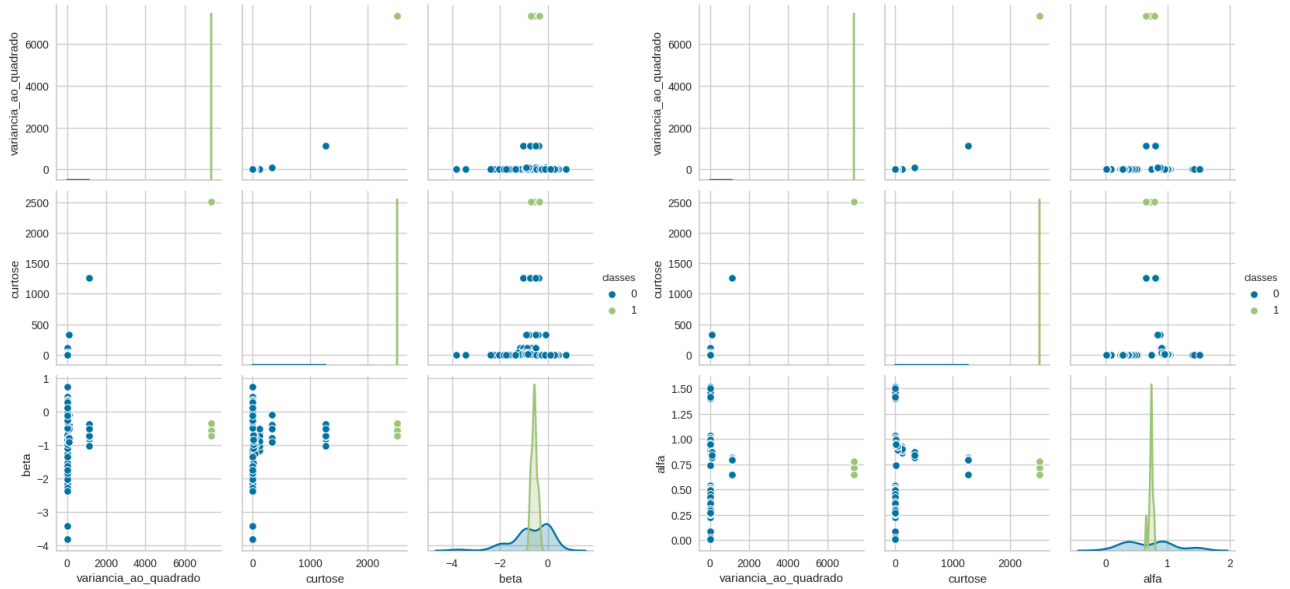

Figura 5. Pontos da série p-Noise agrupados em verde, após remoção da série Noise.

Exercício 6.2 – Conclusão

Devido aos altos valores de variância e ao valor de alfa da série Sol, esta série não se enquadra nos espaços de parâmetros ESPB e EDF.

A série COVID-19 têm uma variância ainda maior que as outras séries. Devido à variância e ao valor de alfa, considera-se que a série está fora do espaço ESPB e EDF.

A série Surface Temperature é a única série que está mais próxima do espaço de parâmetros ESPB e EDF, pois seus valores estatísticos estão dentro dos limites máximos e mínimos de ESPB e EDF.

Quando a série noise é incluída no espaço de parâmetros ESPB e EDF, devido à sua alta variância, bem como a série p-Noise é incluída, com variância não tão alta quando a série noise, os resultados do K-means tendem a agrupar todas a séries Sol e Surface Temperature, dentro do mesmo espaço de parâmetros ESPB e EDF, inadequadamente. Por esse motivo, a utilização de somente uma técnica de análise, como a do K-means, pode não ser muito confiável, ao se tentar separar as séries em dois agrupamentos, dentro dos espaços de parâmetros, pois a análise feita é visual e, nem sempre, é possível afirmar que o ponto da série está dentro de um ou outro agrupamento.

Exercício 6.3 – Análise

As séries estatísticas sobre o número de casos diários da COVID-19 ‘new_cases’ foram criadas utilizando o agrupador ‘location’, isto é, um agrupamento por país. Foram removidos os valores perntencentes à localização ‘World’, que representa o total de todos os países, além de valores com iso_code vazios.

Foram criados os arquivos .csv momentos_estat_var2_alfa.csv e momentos_estat_curtose_alfa.csv para analisar os valores das estatísticas por país, afim de identificar quais países pertencem a quais agrupamentos.

Na figura 6 e nas duas primeiras colunas da tabela 1, são exibidos 11 países com maior variância quadrática. Observando todas as colunas da série COVID, em ordem decrescente de

valores em 26/05/2020, podemos observar que a variância quadrática é um bom indicador do total de casos, como correlacionado em texto colorido na tabela 1, ou de países mais atingidos pela pandemia, em texto na cor preta (Espanha e China).

Tabela 1. Comparativo de maior variância quadrática com total de casos

País	Variância Quadrática	País	Total de casos
United States	2,98E+16	United States	1.662.302
Brazil	5,44E+14	Brazil	374.898
Russia	2,14E+14	Russia	353.427
Spain	3,25E+13	United Kingdom	261.184
United Kingdom	2,27E+13	Italy	230.158
Italy	1,18E+13	Germany	179.002
India	1,00E+13	Turkey	157.814
Germany	9,64E+12	India	145.380
Peru	7,31E+12	France	145.279
China	5,39E+12	Iran	137.724
France	4,83E+12	Peru	123.979

Observando os cinco países com menor curtose, na figura 7 e na tabela 2, verificamos que são os países que apresentam uma estabilibdade no número de casos diários.

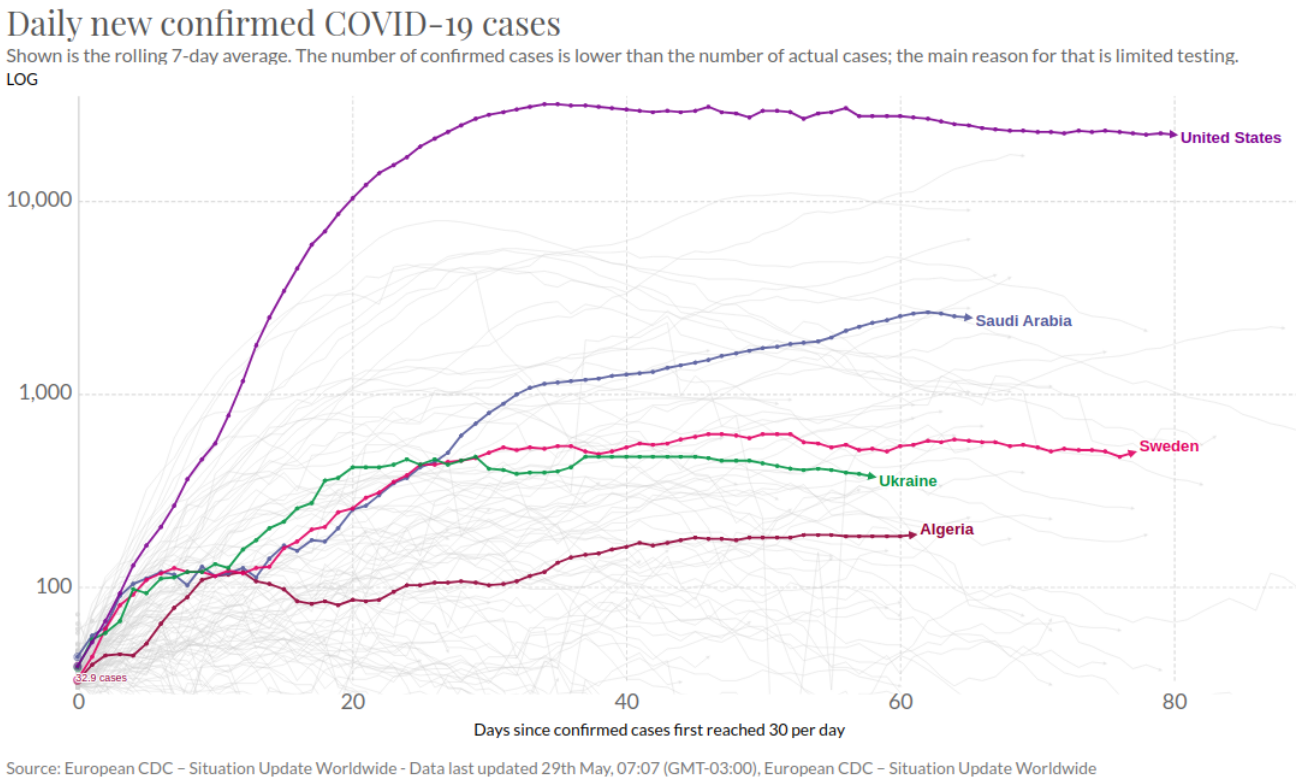

Figura 7. Cinco países com menor curtose.

Tabela 2. Cinco Países com menor curtose

País	Curtose
United States	-1,21
Sweden	-1,1
Algeria	-1,08
Ukraine	-1,06
Saudi Arabia	-1,03

Observando 11 países com menor número Alfa, na figura 8 e na tabela 3, verificamos que são países em que a doença está em início de desenvolvimento, ainda com uma quantidade pequena de total de casos.

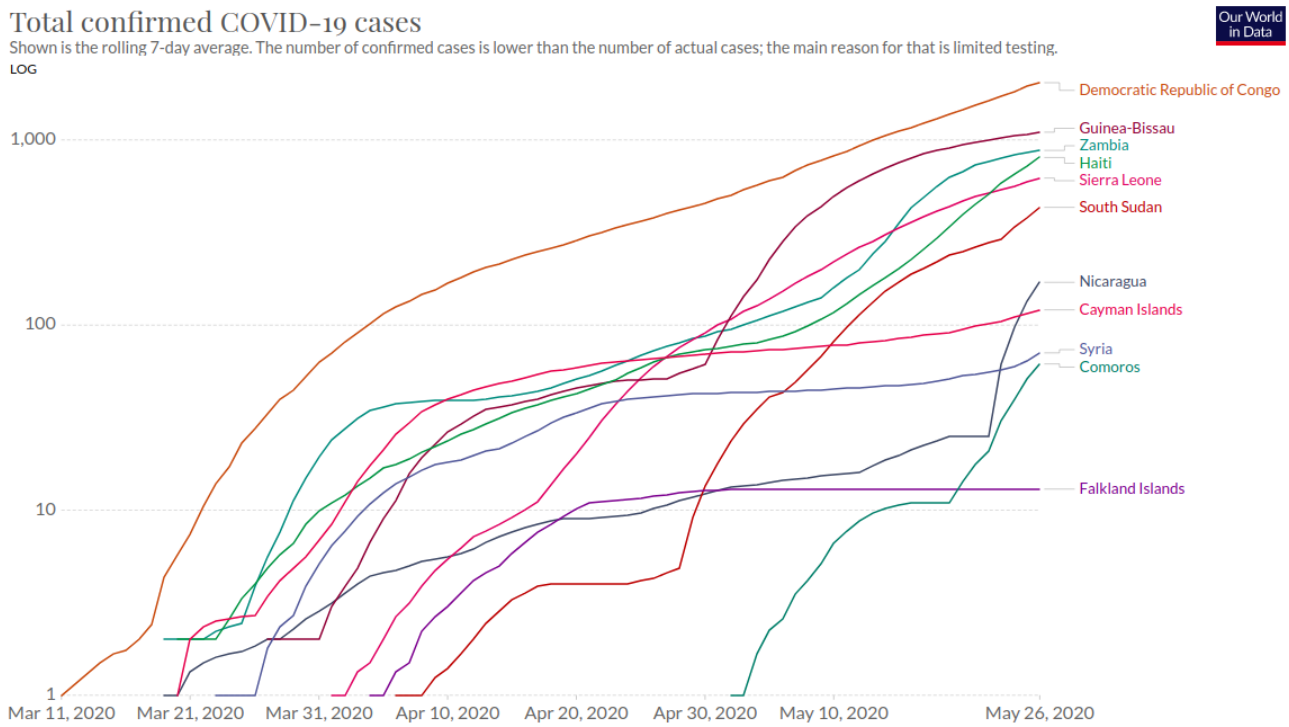

Figura 8. Total de casos em onze países com menor número Alfa

Tabela 3. Países com menor número Alfa

País	Alfa
Cayman Islands	0,37
Democratic Republic of Congo	0,35
Syria	0,35
Haiti	0,34
Guinea-Bissau	0,33
Sierra Leone	0,31
Falkland Islands	0,29
Zambia	0,26
South Sudan	-0,17
Comoros	-0,59
Nicaragua	-1,77

O baixo de número Alfa seleciona sinais mais próximos de um ruído branco, com valores próximos da média, por isso o baixo número Alfa também pode afetar países com mais casos, como é o caso do Equador, que têm uma variação de casos diários em torno da média, como pode ser observado na figura 9.

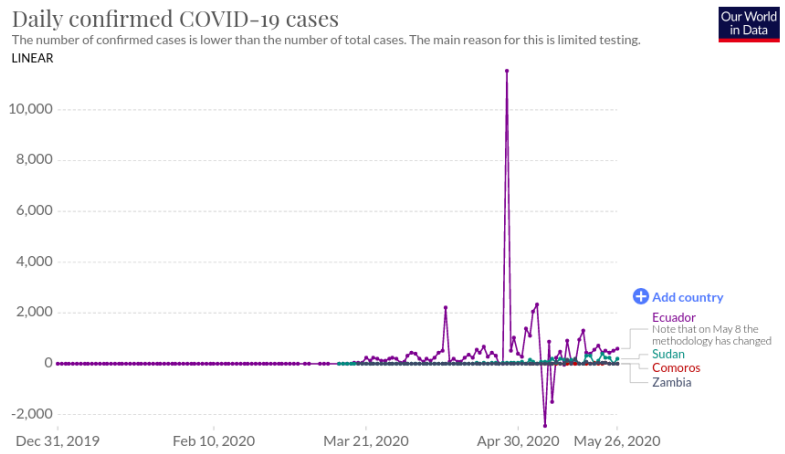

Figura 9. Sinais de ruído Branco

Para identificar o melhor agrupamento, o método do cotovelo indica que o melhor k é igual a 4, para o melhor agrupamento da Variância Quadrática x Alfa, e melhor k igual a 5, para o agrupamento Curtose x Alfa, como pode ser visto na figura 9.

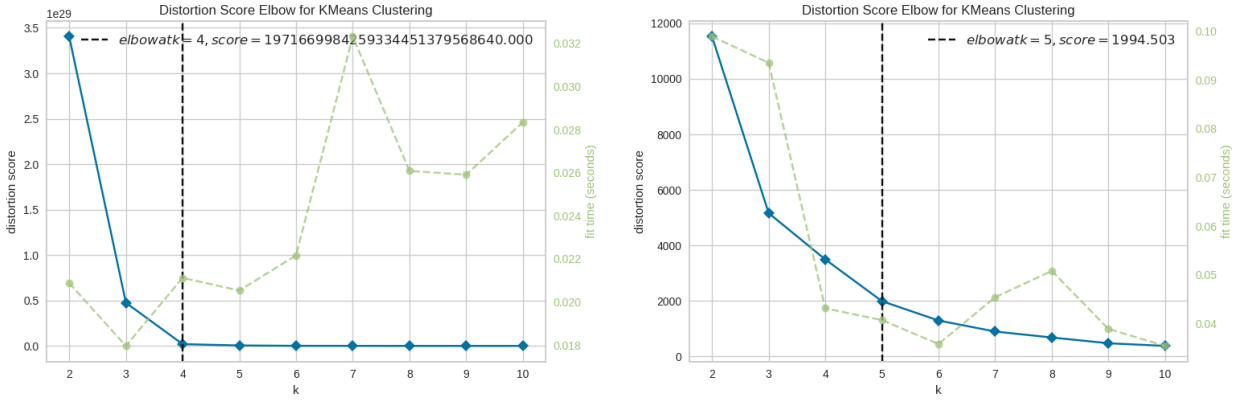

Figura 9. Método do cotovelo. a) Variância quadrática x Alfa; b) Curtose x Alfa

A figura 10 exibe os agrupamentos K-means da Variância Quadrática x Alfa, à esquerda, e o agrupamento Curtose x Alfa, à direita.

Utilizando o melhores agrupamento do K-means para Variância Quadrática x Alfa, pode-se verificar que a alta variância dos Estados Unidos, com valor 2,98E+16, faz com que exista um agrupamento contendo seomente este país, representado por um ponto verde; O ponto em vermelho representa o Brasil, com valor 5,44E+14, e o ponto vinho representa a Rússia, com valor 2,14E+14, como pode ser observado no gráfico da figura 10.a. Todos os outros países estão agrupados como pontos azuis.

No gráfico da figura 10.b, observa-se uma grande quantidade de países agrupadas com a cor vermelha, que apresentam, em sua maioria, mais países com a doença mais desenvolvida e com número de casos mais estabilizado, devido à baixa curtose.

Os pontos com alfa abaixo de 0,5, são países que têm baixa flutuabilidade no número de casos, comumente países emergentes. O agrupamentos com alfa em torno de 1,0, representam uma maior flutuação do número de casos diários, e os com alfa próximos de 1,5, representam o movimento browniano, com uma alta variabilidade de casos.

O ponto em verde com valor de alfa próximo de -2 representa a Nicarágua, onde número de casos contabilizados foi registrado em apenas dois dias, e o ponto verde com valor de alfa próximo de 4 representa a Anguilla, que tem apenas 3 casos

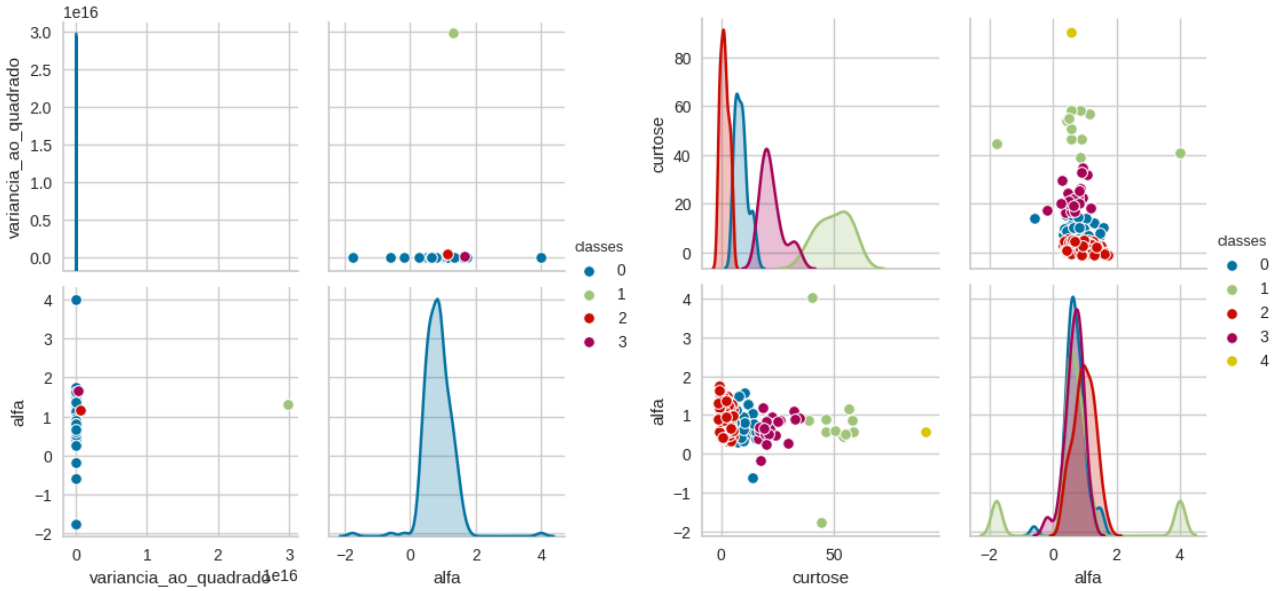

Figura 10. Agrupamentos K-means. a) Variância Quadrática x Alfa; b) Curtose x Alfa.

Utilizando k = 10, ainda obtemos clusters com poucos países, aqueles com alta variância, como os da tabela 1. O cluster de centróide com menor variância, para k=10, está na faixa da 9° casa decimal, ainda bastante alto, e possui os seguintes países mais próximos do cluster: Ireland, Sweden, Colombia, Kuwait, United Arab Emirates, Portugal.

Os melhores valores k para o agrupamento utilizando entre Curtose e Alfa, ficaram entre 5 e 6, utilizando execuções do K-means com k entre 2 e 10 e k entre 2 e 20, respectivamente. Para k=5, os seguintes países ficaram mais próximos dos clusters, em ordem decrescente de curtose: Ecuador (cluster isolado), Timor, Sri Lanka, Ghana e Finlândia.

Interessante observar que, mais países africanos se agruparam próximo do cluster de Ghana: Eritrea, Madagascar, Uganda, Mozambique, Ilha Seychelles e outros países do oriente médio como Afeganistão, e Síria.

Exercício 6.3 – Conclusão

Um baixo valore de alfa, é um bom indicador de países emergentes, pois a curva de casos ainda está em ascenção. O contrário nem sempre é verdadeiro, conforme observado nos valores.

A curtose pode ajudar a identificar os países que apresentam uma estabilidade no número de casos diários, quando o valor é baixo. Esse número também pode estar relacionado à quantidade de testes diários, pois o número de casos tende a ser mais estável em países que fazem mais testes.

A alta variância está relacioanda a países que estão com um número maior de casos, ou países que já passaram pelo pico da pandemia.