Análise Estatística e Espectral de Processos Estocásticos a partir de dados do COVID-19

Natália Pedroso[1], **Yuri D. M. Nunes** [1]

[1]Computação Aplicada – Instituto Nacional de Pesquisas Espaciais (INPE)

Abstract. *This report describes the process of carrying out the Computational Mathematics 1 work. In this work statistical and spectral analysis of stochastic processes related to COVID-19 were performed, applying techniques that processes data sets and return graphs containing virus behavior information in some countries and regions.*

Resumo. *Este relatório descreve o processo de execução do trabalho de Matemática Computacional 1. Neste trabalho foram feitas análises estatísticas e espectrais dos processos estocásticos relacionados ao COVID-19, aplicando-se técnicas que processam os conjuntos de dados e retornam gráficos que contêm informações sobre o comportamento do vírus em alguns países e regiões.*

1. Objetivo

Este trabalho teve por objetivo realizar um estudo sobre análises estatísticas e estocásticas dos dados do COVID-19 e, a partir dos resultados obtidos, compreender sobre o comportamento do vírus em diferentes países e regiões. Utilizou-se também um modelo para realizar previsões e treinar uma rede neural recorrente, com a qual pode-se prever os dados futuros. Diferentes métodos e técnicas vistas em aula [Rosa(2020)] foram aplicadas.

2. Metodologia aplicada

Para realizar a análise estatística e espectral dos processos estocásticos das séries de pontos relacionados aos COVID-19 foram utlizadas as seguintes técnicas: plotagem da visualização dos dados; plotagem dos respectivos histogramas; ajuste do histograma utilizando PDF; distribuição no espaço de Cullen e Frey; a utilização do Modelo IMCSF-COVID19 e de uma Rede Neural Recorrente.

2.1. Dataset

A partir dos datasets das organizações *Our World in Data* (OWD) e Fundação Sistema Estadual de Análise de Dados (SEADE-SP), criou-se um dataset com os dados que representam

- Número total de casos;
- Número total de mortes;
- Número total de testes;
- Número diário de casos (NDC);
- Número diário de mortes (NDD);
- Número diário de testes (NDT),

dos países e regiões determinados, os quais foram

- África do Sul;
- Brasil;
- Egito;
- Índia;
- Irã;
- São José dos Campos;
- São Paulo.

Partindo da extração dos dados dos países, utilizou-se a biblioteca `pandas` do python para fazer uma consulta dentro do conjunto de dados obtido por meio do repositório do GitHub da organização OWD, realizando uma filtragem pelos nomes dos países desejados. Assim, foi possível criar um arquivo .csv contendo apenas os dados dos países a serem trabalhados. Em relação à extração dos dados regionais (das cidades de São Paulo e São José dos Campos), seguiu-se o mesmo procedimento. No entanto, os dados foram obtidos do repositório do GitHub da SEADE-SP, onde não foram encontrados dados relacionados ao número diário de casos e total de testes aplicados. Para essas categorias não foram aplicadas as técnicas de análise estatística e espectral.

2.2. Visualização

A visualização é uma representação gráfica em linha gerada a partir de um conjunto de pontos, onde os segmentos de retas representam a distribuição dos pontos em classes. Para gerar as visualizações dos pontos foi utilizado o método `plot` da biblioteca *matplotlib* do python que permite montar facilmente uma visualização da distribuição dos casos em classes de equivalência a partir dos dados de entrada.

2.3. Histogramas

Um histograma é a representação gráfica em barras (retângulos) de um conjunto de dados tabulado e divido em classes, onde a base de cada retângulo representa uma classe. No caso de classes uniformes, a altura do retângulo representa a quantidade ou frequência absoluta com que o valor da classe ocorre no conjunto de dados. Para classes não uniformes, a altura representa a densidade de frequência. Para gerar os histogramas foi utilizado o método `hist` da biblioteca *matplotlib* do python, o qual, a partir dos dados de entrada e outros parâmetros, permite montar facilmente um histograma.

2.4. Ajuste dos histogramas utilizando PDF - GEV

A distribuição generalizada de valores extremos (GEV - *Generalized Extreme Value distribution*), é uma família de distribuições de probabilidade contínuas desenvolvidas dentro da teoria de valores extremos. A teoria de valores extremos fornece a estrutura estatística para fazer interferências sobre a probabilidade de eventos muitos raros ou extremos. Já uma função densidade de probabilidade (PDF - *Probability Density Function*), ao ser plotada, mostra a probabilidade relativa (no eixo y) de que uma variável terá valor x (no eixo x). Neste trabalho, a técnica de distribuição GEV foi utilizada, resultando numa PDF que ajusta os histogramas obtidos.

2.5. Regressão linear

A regressão linear é uma equação para estimar o valor esperado de uma variável y (variável dependente) com base nos valores de outras variáveis x (variáveis independentes), que tem como objetivo tratar um valor que, inicialmente, não pode ser estimado.

Neste trabalho, a técnica de regressão linear foi utilizada para comparar os dados diários de casos e testes do COVID-19 relacionados aos países e regiões estudadas. Este procedimento partiu de uma análise entre pares de histogramas ajustados com a técnica de PDF onde o ajuste obtido entre eles estava próximo. Assim, aplicou-se a técnica comparando os seguintes pares de dados:

- Números de testes diários (NDT):
  - Brasil e Egito;
  - Irã e Egito;
- Números de casos diários (NDC):
  - Brasil e Índia;
  - Índia e Egito;
  - Irã e Egito;
  - São José dos Campos e São Paulo;

2.6. Classificação no espaço de Cullen e Frey

A técnica de Cullen e Frey, através do cálculo da assimetria e curtose sobre os dados submetidos, analisa qual a melhor classe de probabilidade em que os dados se encaixam. Ela retorna um mapa que demonstra a distribuição dos conjuntos de pontos submetidos à ela e, a partir disso, é possível visualizar qual a melhor forma de distribuir os valores seguindo as distribuições teóricas definidas. Neste trabalho, a técnica de Cullen e Frey foi utilizada para determinar a distribuição de probabilidade dos dados.

2.7. Modelo IMCSF-COVID19

Este modelo é uma cadeia multiplicativa não homogênea de contatos com N agentes. Foi desenvolvido por Reinaldo R. Rosa e colaboradores [citar modelo] com o intuito de prever o comportamento dos dados (suas flutuações). A partir dele, é possível prever como as séries espectrais do COVID-19 se comportarão, possibilitando a análise da propagação do vírus para um determinado período futuro. Para realizar a previsão dos casos futuros existem variáveis que influenciam sobre o cálculo. Uma dessas variáveis está relacionada ao agente propagador. Outra é associada ao fator geral de propagação.

Em relação ao fator de transmissão proposto no modelo, três tipos de agentes propagadores são definidos: os agentes A1, A2 e A3. O primeiro agente é responsável pelo contágio de no mínimo um indivíduo e no máximo dois. Já o agente número 2 é responsável pela contaminação numa faixa entre três e quatro pessoas. Por último, o transmissor A3 propaga o vírus para cinco até seis cidadãos.

O fator geral de transmissão g é uma variável que faz o ajuste da previsão do valor mínimo e máximo de propagação do dia seguinte, ou seja, que calcula a taxa de supressão. A partir dela é feita uma multiplicação com os agentes de propagação, seguindo as seguintes fórmulas:

$$N_{min} = g(1 \times n_1 + 3 \times n_2 + 5 \times n_3)$$
$$N_{max} = g(2 \times n_1 + 4 \times n_2 + 6 \times n_3)$$

A primeira fórmula calcula o valor mínimo previsto para o dia seguinte, enquanto a segunda calcula o valor máximo da previsão. Assim, obtém-se uma taxa de variação que o dado poderá atingir.

Para verificar se o valor da taxa de supressão g está bem calibrado, de forma que a previsão tenha uma baixa margem de erro, é necessário realizar uma comparação com os valores reais dos novos casos diários do dia seguinte à utilização do modelo (o dia que foi previsto). Assim, caso o novo valor esteja dentro do intervalo de flutuação - entre o N mínimo e máximo -, o g permanece com a taxa de supressão utilizada na previsão. Caso contrário, é necessário inserir uma nova taxa de supressão para obter um ajuste mais preciso.

Para este trabalho foi selecionado, para cada país e região, um intervalo começando no dia em que a categoria de novos casos diários (NDC) atingiu mais de 50 casos, até a data de 20 de maio de 2020. A partir daí foram calculadas as curvas g e s, gerando gráficos que expressam os valores obtidos pela previsão. Além disso, obteve-se também uma visualização dos dados base, das predições e dos valores de N mínimo e máximo. Assim, foi possível verificar como seria a flutuação dos dados de cada país e região, e observar o comportamento que cada país teria, supostamente, nos dias seguintes. Com isso, pode-se estudar se a localidade em questão está próxima de controlar os casos diários do vírus ou não.

2.8. SOC - *Self Organized Criticality*

A criticalidade auto-organizada (SOC - *Self Organized Criticality*) é uma conjectura fenomenológica que admite que a organização de um sistema não-linear com muitos graus de liberdade, quando levado para fora do equilíbrio dinâmico, se dá por troca de forças internas que atuam em todas as escalas de flutuação do sistema. Neste trabalho, tal técnica foi aplicada para os países determinados.

2.9. LSTM - *Long short-term memory*

A memória de longo prazo (LSTM - *Long Short Term Memory*) é um tipo de rede neural recorrente usada em diversos cenários de Processamento de Linguagem Natural. Ela é adequada para classificar, processar e prever séries temporais com intervalos de tempo de duração desconhecido. Diferentemente das redes neurais de avanço padrão, o LSTM possui conexões de feedback, de forma que ele pode processar sequências inteiras de dados, não apenas pontos de dados únicos.

Neste trabalho, foi utilizada uma LSTM desenvolvida por Luis Ricardo Arantes Filho para poder realizar previsões sobre os dados diários do COVID-19. O algoritmo de predição que a rede neural utiliza é o IMCSF-COVID19 descrito previamente. A partir dele, a rede é treinada para poder realizar as previsões através dos seguintes parâmetros: número de épocas, de dias entrados, de dados a serem previstos e de lotes.

Para realizar o treinamento, seguiu-se o critério de inserir sempre o dobro do número de entradas para obter o número de saídas. Esta abordagem foi adotada após, durante um *workshop* sobre a rede neural, chegar à conclusão de que o melhor resultado obtido pelo estudo da LSTM foi gerado com vinte dias entrados para obter os próximos dez dias. Assim, este critério foi seguido com o intuito de visualizar se esta proporção permitia que a rede refinasse seu treinamento e diminuísse a taxa de erros.

Em relação à quantidade de épocas para o treinamento, foram utilizadas as grandezas de 50, 200, 400, 500 e 700. A partir daí, cada uma das épocas citadas foi associada

a um número de quantidades de pontos entrados, uma quantidade de pontos a serem gerados e um lote (que foi 1 para todos os testes). Os testes foram estruturados da seguinte maneira:

- Primeiro teste: 50 épocas, 5 dias entrados e 5 dias a serem previstos;
- Segundo teste: 500 épocas, 30 dias entrados e 25 dias a serem previstos;
- Terceiro teste: 200 épocas, 30 dias entrados e 15 dias a serem previstos;
- Quarto teste: 700 épocas, 40 dias entrados e 20 dias a serem previstos;
- Quinto teste: 400 épocas, 50 dias entrados e 25 dias a serem previstos;

3. Discussão dos resultados

Ao realizar a plotagem dos gráficos de visualização para os países foi possível enxergar a distribuição diária e total dos casos, mortes e testes em escala linear. Em relação aos casos e mortes diárias, notou-se que todos os países começaram a ter novos casos e mortes em datas muito próximas, seguindo ordens de crescimento parecidas, com um grande índice de flutuação, como mostra a figura 1. Tratando-se dos testes diários, os dados de alguns países não são compartilhados com regularidade, o que gera flutuações bruscas, como pode ser observado na figura 2. No caso do Brasil e Egito, tais dados não foram compartilhados.

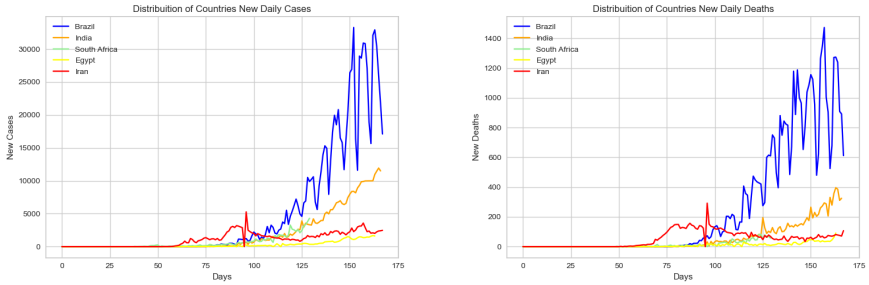

Figura 1. Gráficos de casos e mortes diárias dos países, respectivamente.

Partindo da categoria de casos e mortes totais, os dados são muito parecidos, resultando em visualizações bem próximas umas das outras. Já em relação ao total de testes, considerando os países cujos dados foram fornecidos, Índia e África do Sul possuem números muito parelhos. Notou-se também que o Irã é o país que mais compartilha

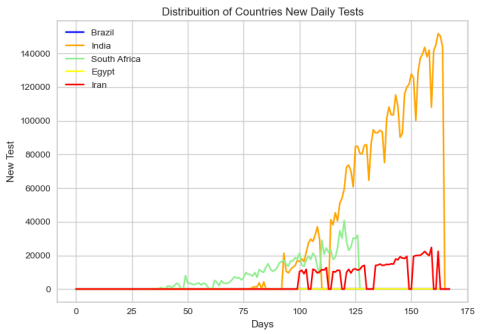

Figura 2. Gráfico de testes diários dos países.

dados sobre o total de testes aplicados. Com base nessas informações, concluiu-se que todos os países desenvolveram casos de COVID-19 em um período próximo, porém alguns compartilham dados com mais frequência que outros. Na análise dos dados regionais, concluiu-se que os dados de São Paulo e São José dos Campos comportam-se de maneira muito parecida, como pode ser visto na figura 3.

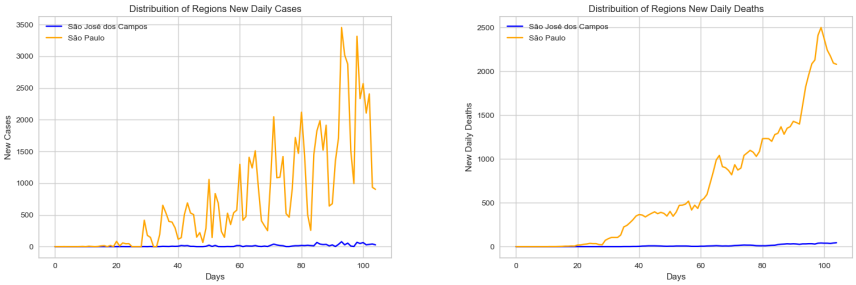

Figura 3. Gráficos de casos e mortes diárias das regiões.

A partir da plotagem dos gráficos dos histogramas ajustados com a PDF, notou-se que, devido aos números dos dados dos países Brasil, Índia e Irã, o ajuste da PDF não foi muito bom. De forma que o histograma não aparece nos gráficos, como é mostrado na figura 4. Já no caso da África do Sul e Egito, os histogramas foram bem ajustados, mostrando o comportamento dos casos diários em ambos os países, o que fica claro na figura 5.

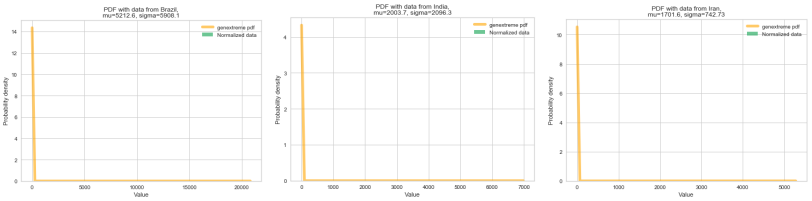

Figura 4. Histogramas ajustados com PDF dos países Brasil, Índia e Irã.

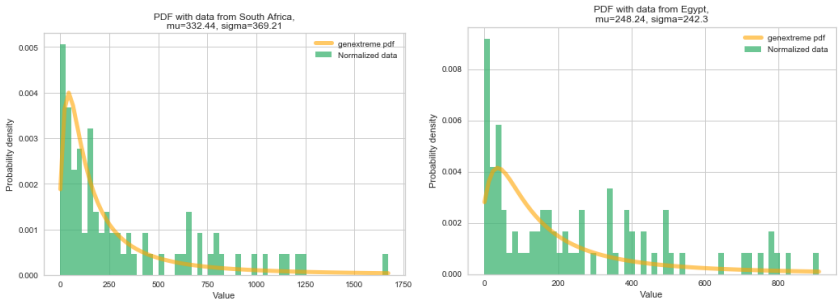

Figura 5. Histogramas ajustados com PDF da África do Sul e Egito.

Em relação à regressão linear aplicada à categoria de novos testes diários, como descrito anteriormente na metodologia, obtiveram-se duas duplas com ajustes de PDF parecidos. Ao analisar o gráfico obtido da dupla Brasil-Egito não foi possível obter uma conclusão, uma vez que tais países não divulgam diariamente sobre os testes aplicados.

A respeito da dupla Irã-Índia, os resultados obtidos refletem que, dentro do intervalo de 0.4 a 0.6 de testes, os valores ficam abaixo da linha ajustada a partir dos coeficientes de regressão (dados do Irã). A partir de 0.6, os valores ficam acima da linha ajustada, porém as flutuações não são muito distintas, levando a conclusão de que existem aleatoriedades em dados bem normalizados, como mostra a figura 6.

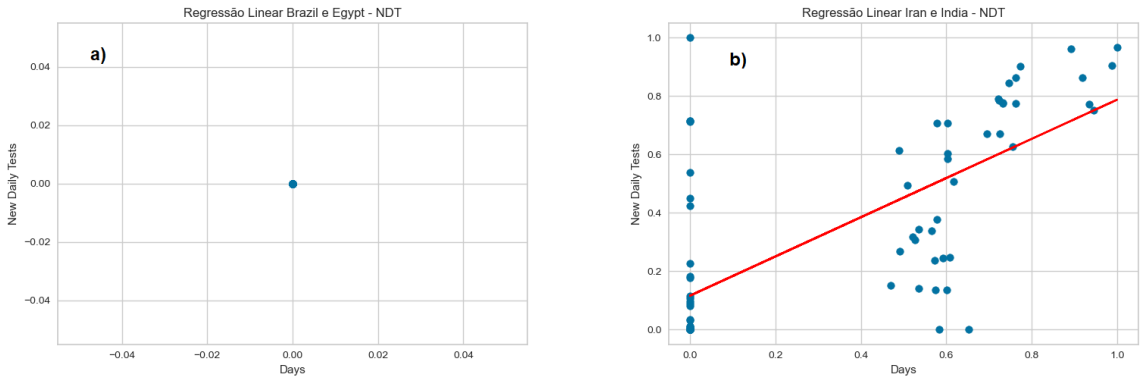

Figura 6. Regressão Linear com número de testes diários, para (a) Brasil e Egito e (b) Irã e Índia.

Ao aplicar a regressão linear para os dados de novos casos diários, obtiveram-se cinco duplas com ajustes de PDF parecidos. Ao analisar o gráfico obtido da dupla Brasil-Índia, percebe-se que os dados estão distribuídos perto da linha de coeficientes, sem aleatoriedades e bem normalizados, o que pode ser visto na figura 7. O mesmo pode-se concluir para a dupla Índia-Egito e São Paulo-São José dos Campos, que são representados nas figuras 7 e 9, respectivamente. No entanto, ao analisar a distribuição dos pontos das duplas Irã-Egito e Irã-Índia, como mostra a figura 8, notou-se que os valores, mesmo com resultados parecidos no histograma ajustado com PDF, possuem aleatoriedades quando comparados com a linha de coeficientes.

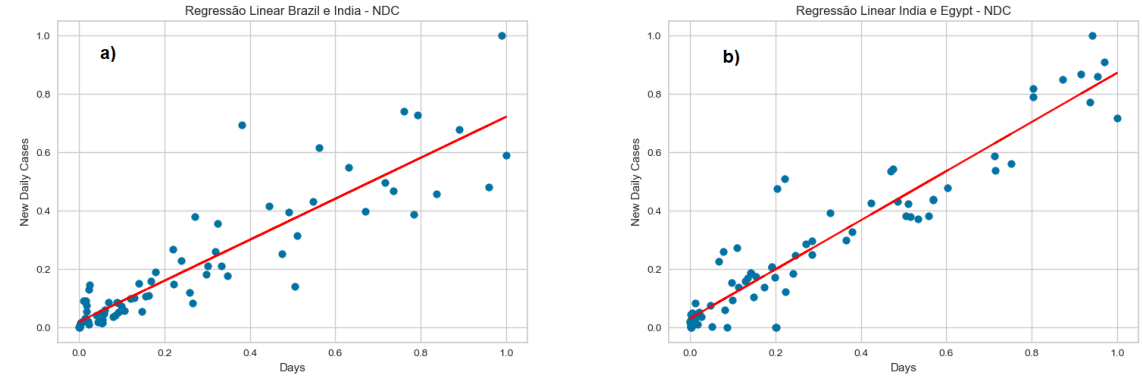

Figura 7. Regressão Linear com número de casos diários, para (a) Brasil e Índia e (b) Índia e Egito.

A partir da aplicação da técnica de Cullen-Frey nos dados, foi possível gerar gráficos que distribuem os países de acordo com a melhor classe de probabilidade em

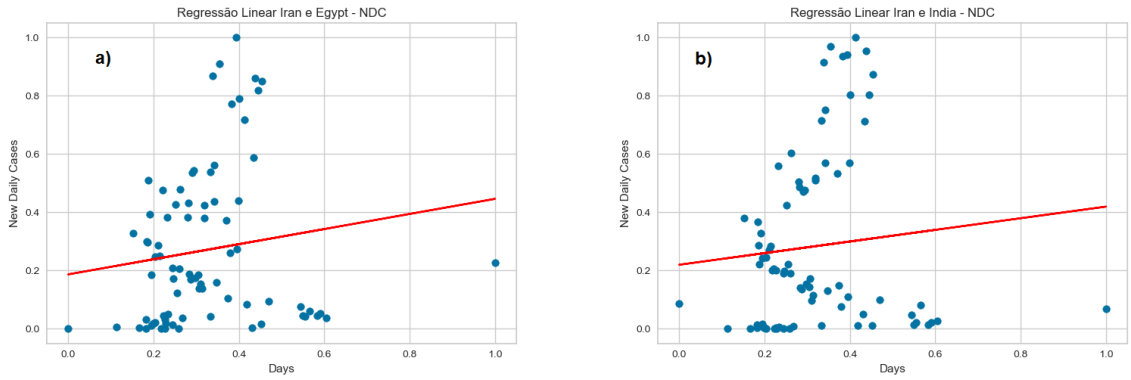

Figura 8. Regressão Linear com número de casos diários, para (a) Irã e Egito e (b) Irã e Índia.

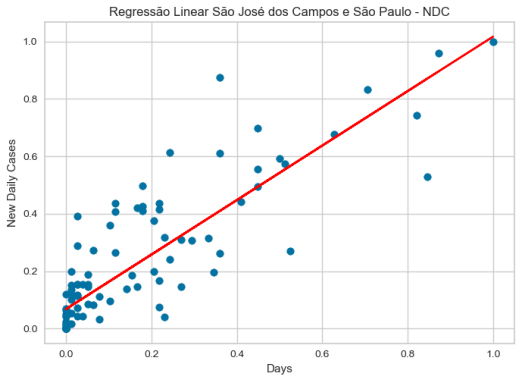

Figura 9. Regressão Linear com número de casos diários para São José dos Campos e São Paulo.

que eles se encaixam, os quais encontram-se nas figuras 10, 11, 12 e 13. Na categoria de novos casos diários, os países se ajustam melhor à classe uniforme, com exceção do Irã, que se encaixa melhor à log-normal. Ao analisar a categoria de novas mortes, percebe-se que Brasil, Índia e Egito classificam-se dentro da classe uniforme (embora os dados do Egito sejam mais flutuantes). Por outro lado, Irã e África do Sul divergem, sendo que o primeiro novamente se ajusta à classe log-normal e o segundo não foi distribuído dentro de nenhuma classe de probabilidade. Com relação ao total de casos e mortes, todos os países estão distribuídos na mesma classe de probabilidade: a uniforme. Na aplicação da técnica foram desconsiderados os dados sobre testes do Brasil e Egito, devido ao fato de que estes não disponibilizam tais dados, e um erro ocorre no código ao realizar divisões por zero. Tratando-se dos dados regionais, é perceptível que para as categorias de novos casos e mortes diárias e total de casos e mortes os dados são classificados dentro da classe de probabilidade uniforme.

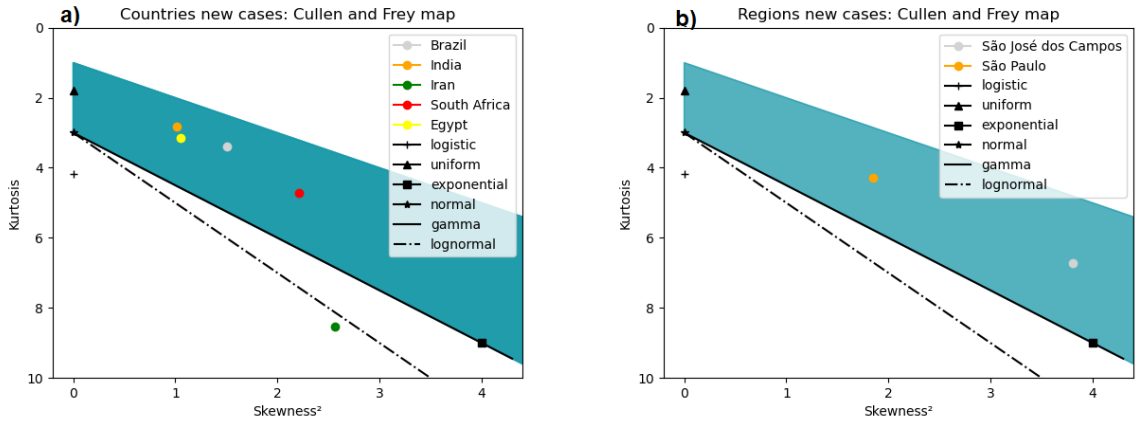

Figura 10. Mapa de Cullen e Frey com número de casos diários, para (a) Países e (b) Regiões.

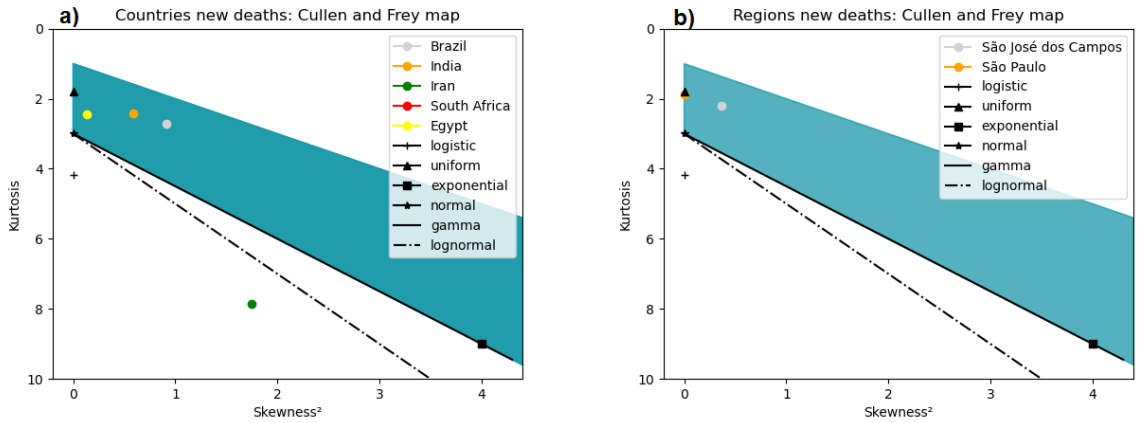

Figura 11. Mapa de Cullen e Frey com número de mortes diárias, para (a) Países e (b) Regiões.

Analisando as previsões do modelo IMCSF-COVID19, de maneira geral, observou-se que os dados previstos sempre apresentam um crescimento maior que o dos

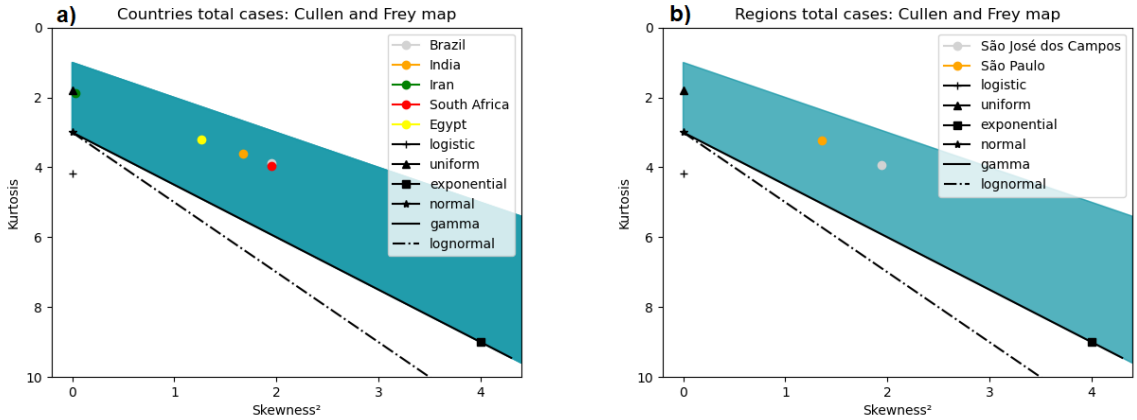

Figura 12. Mapa de Cullen e Frey com número total de casos, para (a) Países e (b) Regiões.

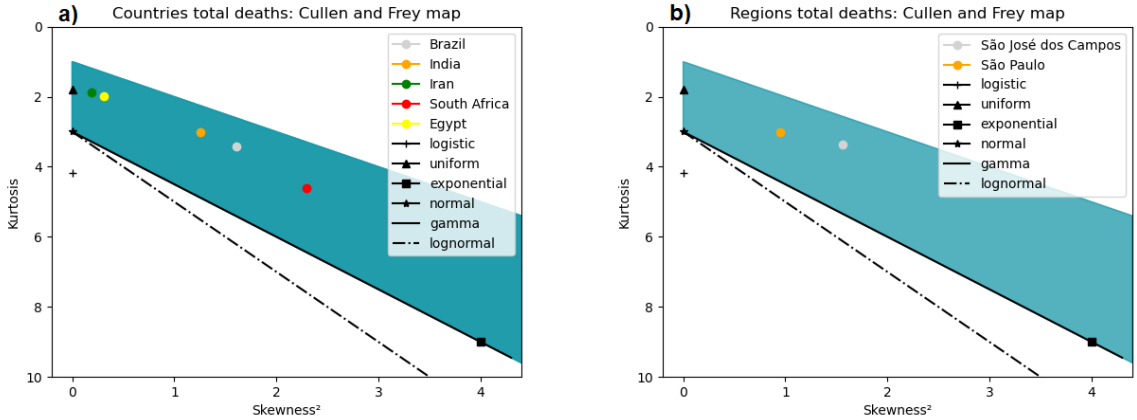

Figura 13. Mapa de Cullen e Frey com número total de mortes, para (a) Países e (b) Regiões.

dados originais, embora tenham o mesmo comportamento que eles. Isso pode ser observado nos gráficos das figuras 14, 15 e 16.

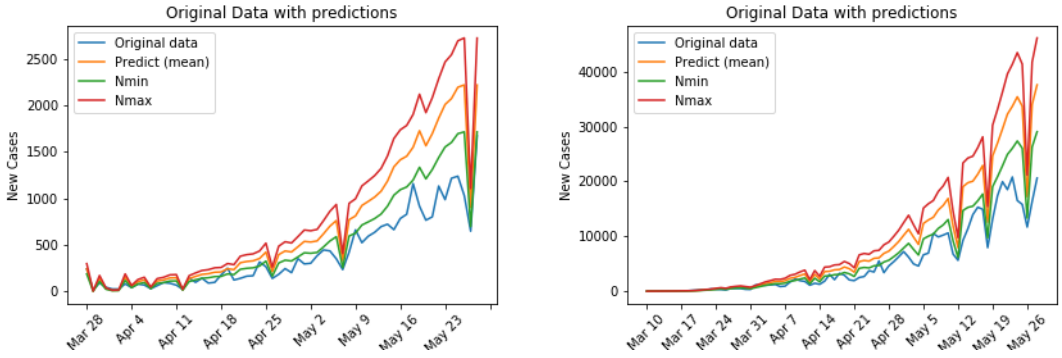

Figura 14. Previsões com o modelo IMCSF-COVID19 dos países África do Sul e Brasil, respectivamente.

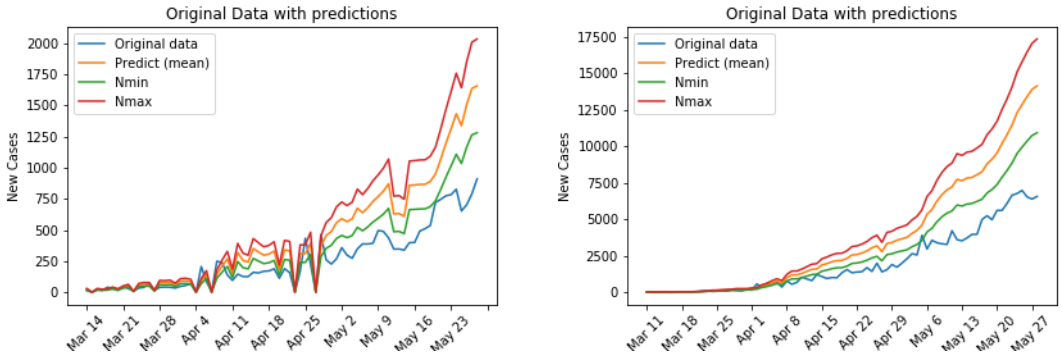

Figura 15. Previsões com o modelo IMCSF-COVID19 dos países Egito e Índia, respectivamente.

Tratando-se da análise de gráficos resultantes da aplicação do SOC, a criticalidade auto-organizada, todos os países, exceto pela África do Sul (como mostra a figura 17), apresentaram um comportamento parecido, de forma que seus pontos atratores encontram-se consideravelmente distantes, indicando uma certa demora para voltar ao seu equilíbrio, isto é, seu estado natural, como mostram as figuras 18 e 19.

Após realizar o treinamento da LSTM seguindo os testes citados e obter os gráficos com as séries de pontos previstos, foi possível observar que, mesmo com um número de eras não tão grande, o teste que tinha como parâmetro 40 dias entrados e 20 a serem gerados obteve uma representação mais próxima à ideal. Conclui-se então que, para ter uma taxa de erro quase zero e, com isso, conseguir fazer uma previsão mais próxima do ideal, é necessário configurar também o parâmetro de lote. Além disso, a abordagem de se utilizar valores dobrados de dias entrados em relação aos dias que serão gerados, pode ser uma solução para aumentar a obtenção dos dados a serem previstos.

4. Considerações finais

Com base nas análises dos gráficos pôde-se concluir que o comportamento dos países de maneira geral mostrou-se parecido.

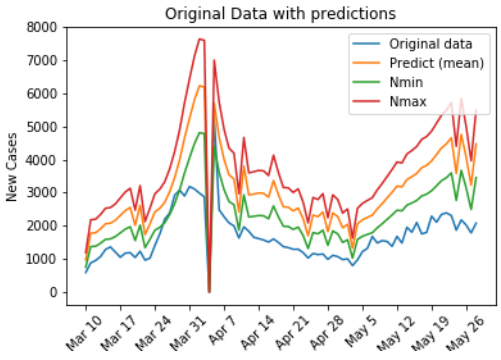

Figura 16. Previsões com o modelo IMCSF-COVID19 do Irã.

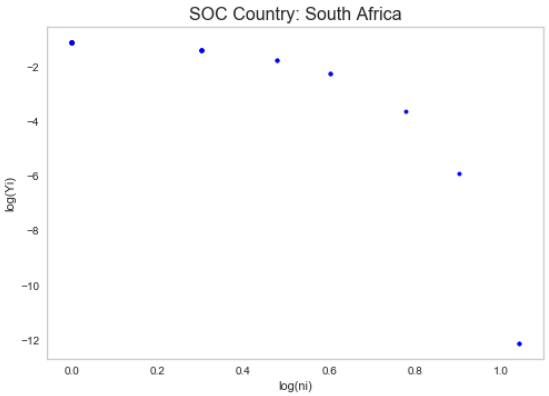

Figura 17. SOC da África do Sul.

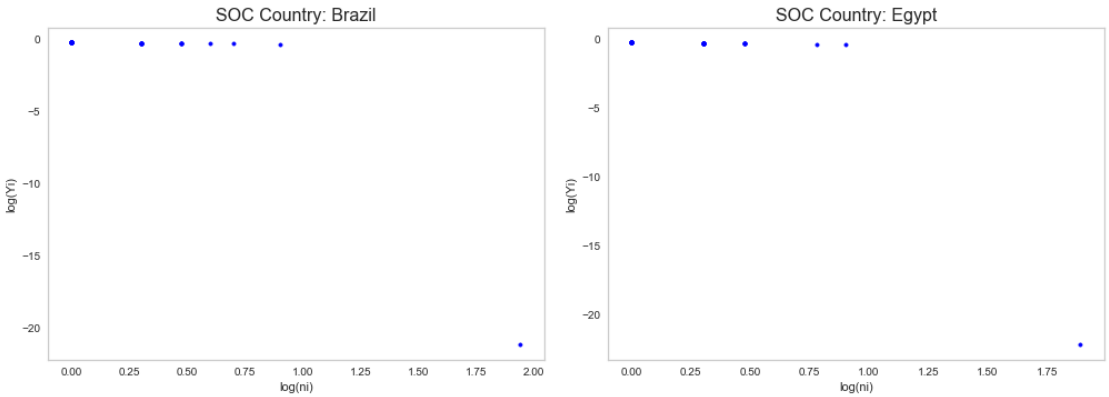

Figura 18. SOC dos países Brasil e Egito.

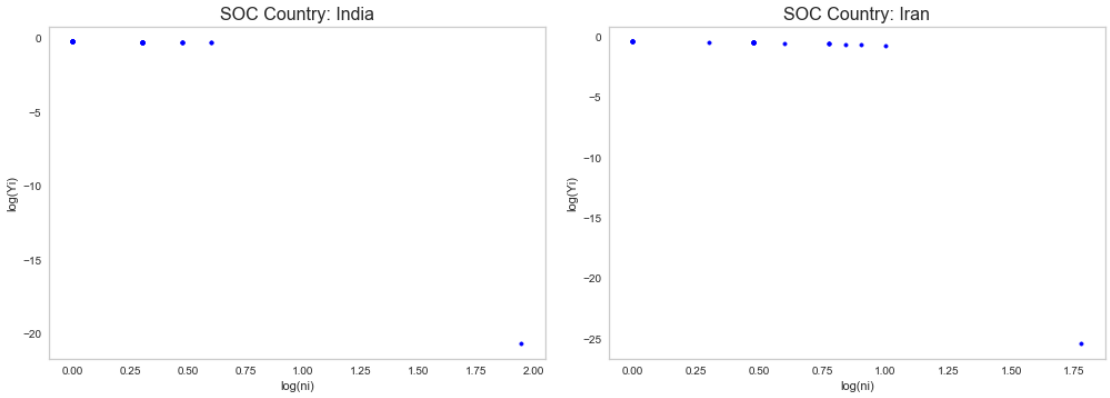

Figura 19. SOC dos países Índia e Irã.

No caso do Brasil, onde o contexto do vírus é consideravelmente pior do que nos demais países, notou-se uma alteração nos dados de histogramas mesmo que ajustados com a GEV. Além disso, pela falta de dados sobre os testes, os resultados gráficos envolvendo regressão linear foram prejudicados. Já na classificação do Espaço de Cullen e Frey, o Brasil ficou dentro da classe uniforme. Assim como os demais, na previsão do modelo IMCSF, os dados previstos seguiram o mesmo comportamento dos dados originais, no entanto, prevendo um crescimento maior.

Tratando-se da África do Sul, pode-se dizer que seus dados foram bem consistentes, de forma que seu histograma foi muito bem ajustado com a PDF. Sua classificação no espaço de Cullen e Frey encaixou-se na classe uniforme e, assim como os demais países, o crescimento previsto com o modelo IMCSF-COVID19 resultou acima do crescimento dos dados originais. Em relação aos resultados obtidos com a aplicação da técnica SOC, pode-se dizer que o comportamento dos dados da África do Sul foi um dos melhores, indicando uma possível normalização em breve.

Assim como a África do Sul, os dados do Egito resultaram num histograma bem ajustado. Além disso, seus dados de testes contribuíram para a análise com regressão linear. Também como os demais países, ele encontra-se dentro da classe uniforme na classificação do espaço de Cullen e Frey, e sua previsão com o modelo IMCSF-COVID19 seguiu o mesmo comportamento dos dados originais, resultando, no entanto, num crescimento maior.

A partir dos resultado obtidos com os dados da Índia, notou-se uma semelhança ao histograma gerado a partir dos dados do Brasil, que relatam uma alteração mesmo que ajustados com a GEV. Abordando a análise de regressão linear, concluiu-se que, dadas as proporções, os dados são equivalentes em relação às flutuações com o Brasil e Egito quando se trata dos números de casos diário. Já em relação aos testes, percebe-se que a Índia flutua muito quando comparada ao Irã, que é o país que compartilha os dados com mais frequência. Já na classificação do Espaço de Cullen e Frey, a Índia ficou dentro da classe uniforme para todas as categorias analisadas. Assim como os demais, na previsão do modelo IMCSF, os dados previstos seguiram o mesmo comportamento dos dados originais, no entanto, prevendo um crescimento maior.

E, por fim, analisando os dados do Irã, percebe-se que o histograma não conseguiu

ser bem ajustado utilizando a técnica de GEV. Sua regressão linear, quando utilizada como linha de coeficientes, demonstrou que os outros países tem uma flutuação de dados, uma vez que o Irã possui uma consistência maior em relação aos valores de casos e testes. Em relação à sua classificação no espaço de Cullen e Frey, percebe-se que os dados sempre tendem a uma log-normal, e, assim como os demais, na previsão do modelo IMCSF, os dados previstos seguiram o mesmo comportamento dos dados originais, prevendo, no entanto, um crescimento maior.

Referências

R. R. Rosa, *Matemática Computacional I*, Notas de aula, 2020.