MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Relatório final
Análise Estatística e Espectral de Processos Estocásticos

Fernando Cossetin, Felipe Menino, Felipe Perin

1 Introdução

Assuntos a serem tratados na introdução:

- Pandemia;
- Análise estatística e espectral de processos estocásticos;
- Objetivo do trabalho;
- Técnicas aplicadas.

2 Conjunto de dados

Para o desenvolvimento do presente Trabalho, foi necessário realizar a criação de um conjunto de dados contendo as informações dos seguintes países: (i) Brasil; (ii) Canadá; (iii) México; (iv) Cuba; e (v) Rússia. Além dos países, foram considerados também os dados do estado de Minas Gerais e do município de Niterói.

A construção do conjunto de dados, foi feita através de duas etapas, a primeira relacionada a definição das fontes de dados e a segunda às formas de aquisição. Para as fontes, buscou-se aquelas com a maior confiabilidade e integridade dos dados fornecidos, com isso, para os dados dos países foi definido o uso dos dados do grupo Our World in Data [1]. Já para os dados regionais, fez-se a escolha dos dados disponibilizados pela iniciativa Brasil.io[2], que recupera os dados de cada município, organiza e publica em um formato

[1] ourworldindata.org
[2] brasil.io

pronto para análise. Ambas iniciativas selecionadas fazem atualizações diárias dos dados, disponibilizando estes através de serviços *web*. Para o consumo dos dados de cada uma dessas fontes, foi realizada a criação de uma ferramenta que recupera os dados das fontes e aplicando os conceitos de Preparação Generalizada de Dados (PGD), salva os dados recuperados em diferentes formados, a citar, SQLite, CSV e JSON.

Os dados que constituem o conjunto de dados criado, possuem múltiplas atributos, porém, como este caso aplica técnicas estatísticas e espectrais para a análise dos dados, é definido que apenas os dados que possuem flutuações diárias serão considerados neste documento, sendo essas, `Número Diário de Casos` (NDC), `Número Diário de Mortes` (NDM) e `Número Diário de testes` (NDT). Porém, caso seja necessário o trabalho com dados acumulados, no repositório[3] deste projeto, estão disponíveis as análises aplicadas para os dados acumulados.

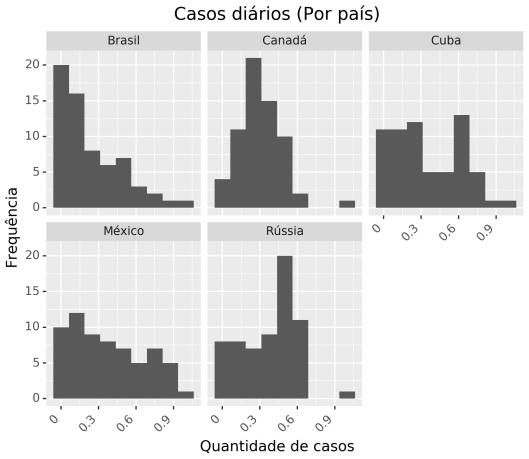
3 Análise dos dados

Esta Seção apresenta a análise das séries temporais dos países citados anteriormente.

3.1 Análise de histograma

A aplicação de histogramas é uma ferramenta crucial para análise e representação de dados quantitativos, agrupados em classes de frequência, que permite distinguir características como a forma, o ponto central, a variação, a simetria e a amplitude da distribuição dos dados. Logo, serão apresentados neste documento os histogramas dos dados do conjunto que possuem flutuação, sendo eles: casos diários, mortes diárias e testes diários.

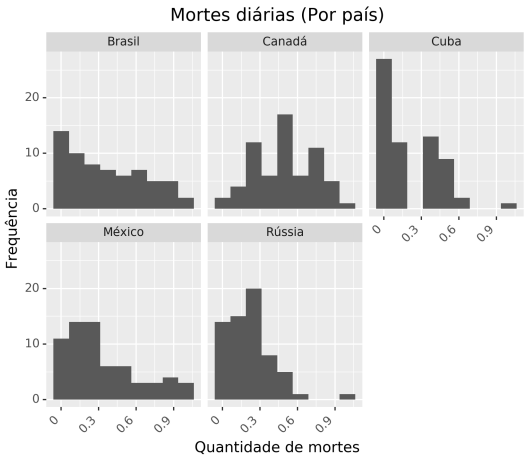
Figura 1: *Histograma dos casos diários por país*



Nota-se pela figura 1 que Canadá e Rússia apresentam maior frequência de maior quantidade de casos diários comparados ao resto do conjunto.

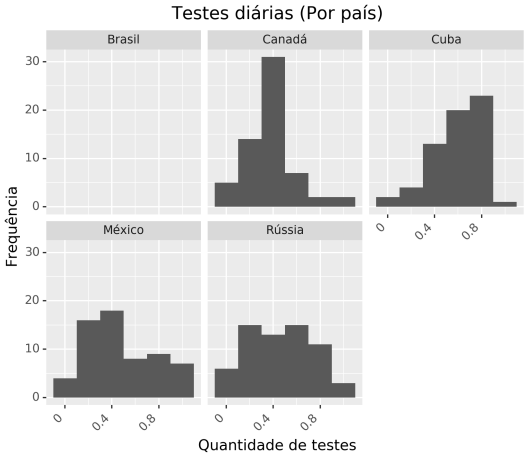
[3] github.com/cmath-covid

Figura 2: *Histograma das mortes diárias por país*



No histograma referente às mortes diárias (Fig. 2, nota-se que Brasil e México apresentam distribuição semelhante da frequência da quantidade de mortes, enquanto que o Canadá apresenta o pico com maior quantidade de mortes diárias, ou seja, segundo a distribuição do histograma ocorreram mais vezes uma quantia consideravelmente alta de mortes diárias comparado aos demais países.

Figura 3: *Histograma dos testes diários por país*



Pela análise do comportamento do histograma de testes diários se observa que Cuba foi o país que realizou mais vezes uma maior quantidade de testes por dia, já o Canadá possui frequência mais constante (em uma quantidade intermediária) de testes realizados por dia. Já o Brasil não possui dados de testes diários.

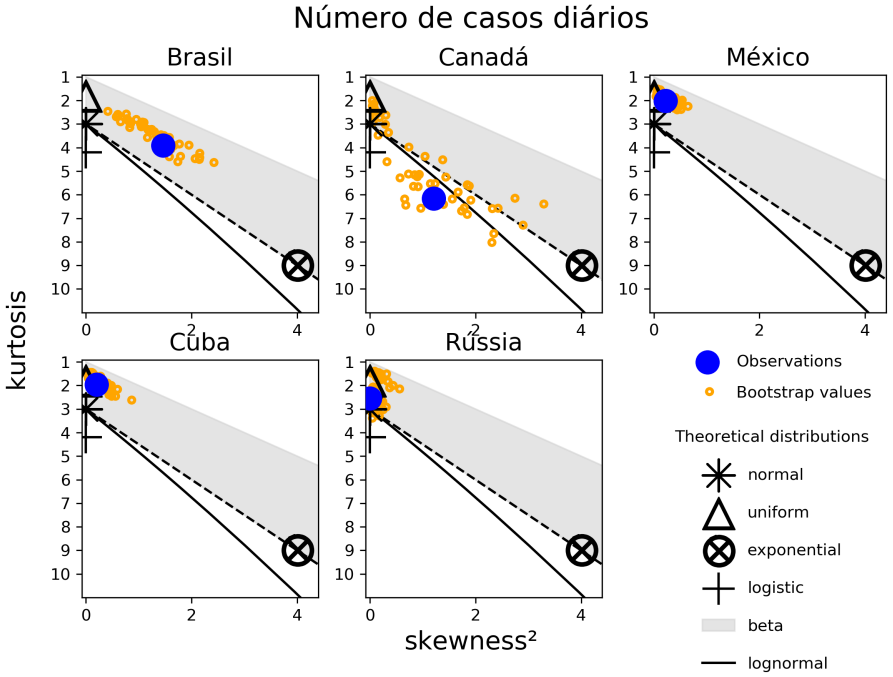
3.2 Identificação da classe estatística

A análise de histograma é relevante principalmente para o levantamento inicial de informações durante o processo de análise, já que, as características identificadas com esta

ferramenta permitem o entendimento da maneira com que as diferentes variáveis estão distribuídas. Neste ponto, vale ressaltar que a identificação de distribuições teóricas, aqui chamadas de Funções de Densidade de Probabilidade (PDF), que se assemelham aos dados é útil, uma vez que através destas é possível realizar mapeamentos, modelagem e predição de possíveis valores e comportamentos que serão assumidos pelo fenômeno estudado. Porém, a depender do método aplicado, a identificação de tais PDFs pode representar uma tarefa demorada, complexa e de difícil definição, neste contexto, existem diversas técnicas e ferramentas que podem auxiliar nesta identificação, dentre elas o Espaço de Cullen-Frey (ECF), que através do mapeamento entre as variáveis `curtose` e `assimetria` apresenta possíveis PDFs teóricas que representam tal comportamento. Esta seção apresenta a aplicação do CF nas séries temporais analisadas neste trabalho.

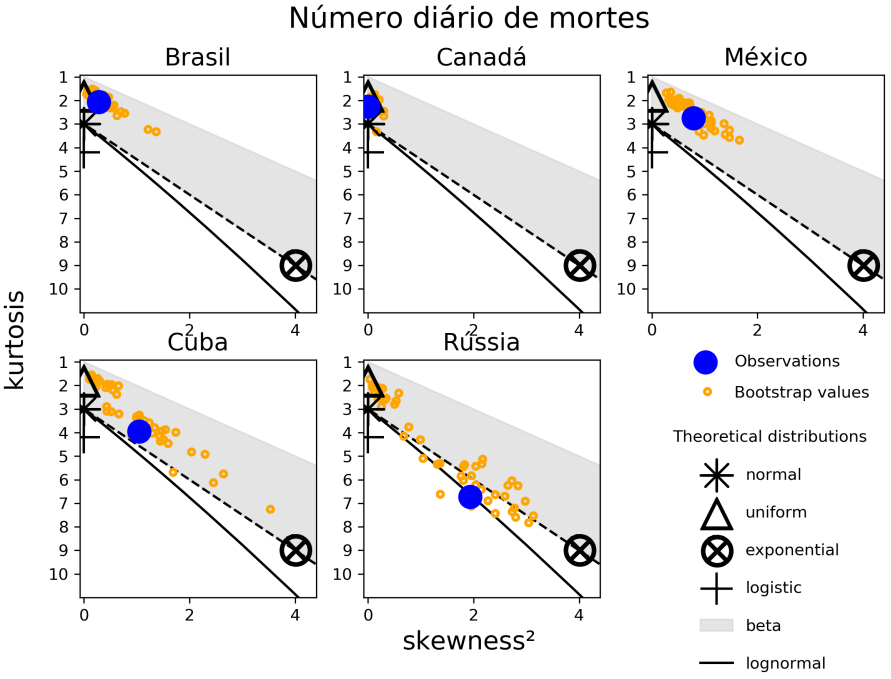
Inicialmente, faz-se a classificação do atributo NDC no ECF (Figura 4), onde é possível perceber que para os países Cuba, Canadá e Rússia, existem atratores normais e uniformes próximos aos dados, o que pode indicar a prevalência deste tipo de distribuição nestes países. Já no caso do Brasil e México, há indicativos de distribuição beta.

Figura 4: *Número diário de casos*



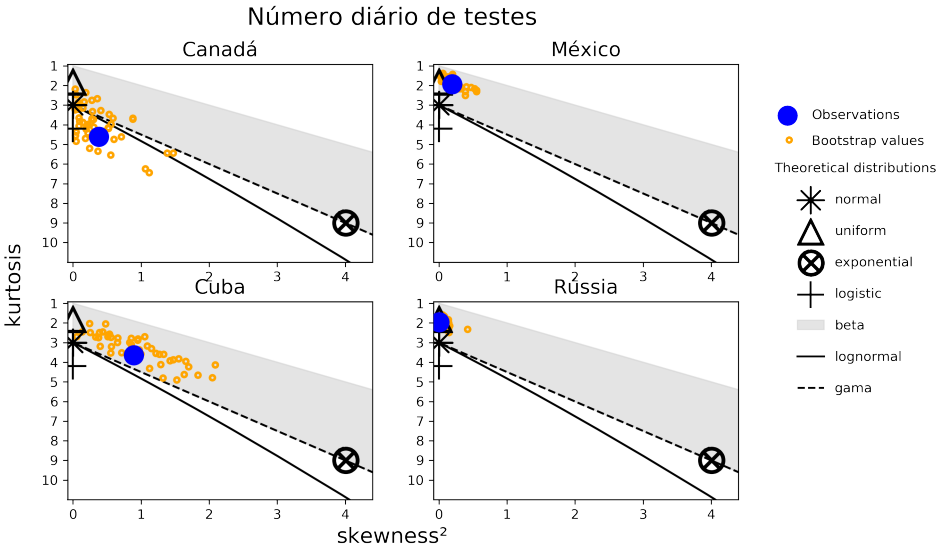
Ao considerar a variável NDM no ECF, tem-se o resultado apresentado na Figura 5, onde, da mesma forma que o comportamento apresentado nos dados de NDC, para os valores de NDM, tem-se indícios de distribuição beta para Brasil, México e Cuba, sendo que países como o Canadá sofrem influência também de atratores de distribuição normal e uniforme. Para o caso da Rússia, é possível perceber a influência do atrator log-normal.

Figura 5: *Número diário de mortes*



Por fim, a análise no ECF é feita com a variável NDT na Figura 6, onde, se comparado com os demais resultados, apresenta uma variação maior nas classes e comportamentos indicados, onde Cuba apresenta indícios de distribuição Beta e Gama. México e Rússia, no espaço de distribuições beta, com atratores normal e uniforme. Já o Canadá apresentou o comportamento mais discrepante dos dados, possuíndo características próximas a gama, porém, com indícios de distribuição normal, uniforme e logística, sendo essas identificadas com o auxílio dos dados gerados com o método *Boostrap*.

Figura 6: *Número diário de testes*



Com o levantamento inicial dos tipos de PDFs que podem explicar o comportamento de cada um dos atributos apresentados, é possível realizar os ajustes e verificar o comportamento. Vale ressaltar que, nem sempre o ECF identifica as distribuições que melhor se encaixam ao problema, mas, como indicação metodológica, o início do estudo dos dados com esta ferramenta reduz o tempo e aumenta a assertividade das decisões tomadas para a realização de ajustes.

3.3 Ajuste de PDF

3.4 Identificação de similaridade

Uma das várias etapas de de análise de dados é a identificação de comportamentos e similaridade no próprio conjunto de dados, o que ajuda no entendimento dos comportamentos e relações assumidas pelos fenômenos registrados. Nas seções anteriores, diferentes técnicas foram utilizadas para a busca de comportamentos dos dados, verificando sua distribuição e suas classes estatísticas, porém, na etapa de análise desses resultados, podem haver características que não foram consideradas ou notas. Desta forma, com o objetivo de identificar e mapear possíveis comportamentos presentes nos dados, é feita a aplicação do análise de agrupamento *K-Means*. Para isto, cada uma das séries é sumarizada considerando as métricas de `curtose`, `assimetria` e `variância`.

Uma das principais etapas para a aplicação do *K-Means* é a identificação do melhor valor de K para o agrupamento dos dados. Em um fluxo padrão de aplicação, esta etapa seria resolvida com métodos como *Elbow* e *Silhouette*, que fazendo avaliações nos grupos gerados com diferentes valores de K e com auxílio de uma função objetivo, definem o melhor valor de K a ser aplicado, porém, neste caso, o conjunto de dados com tamanho limitado faz com que tais técnicas não apresentem bons comportamentos e não ajudem na identificação dos valores de K. Com isso, para a aplicação do método, fez-se a consideração da variação dos valores de K dentro do intervalo $[2,4]$. Este intervalo foi

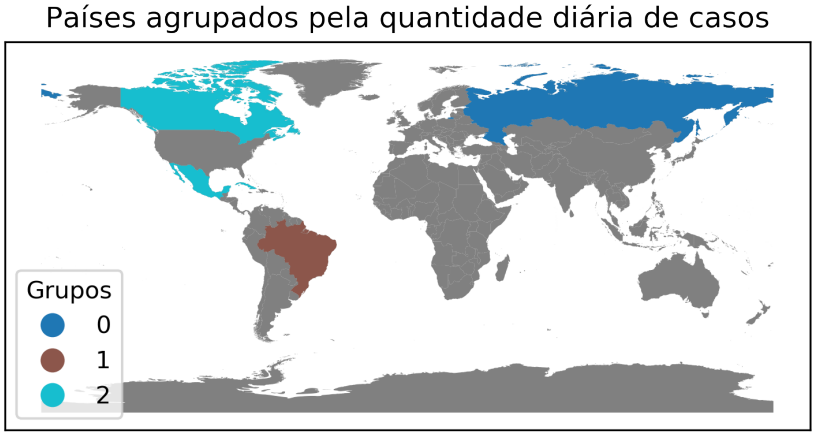
definido seguindo o critério de que, com apenas um grupo e com cinco grupos, não há nenhum agrupamento feito, uma vez que o objetivo é juntar elementos semelhantes.

Para a aplicação da análise de agrupamento, considerou-se os atributos NDC e NDM, por estas serem as únicas que apresentam flutuações, como já informado e por estarem presentes em todas séries do conjunto, uma vez que, para os valores de NDT por exemplo, países como o Brasil não possuem dados. Foram desenvolvidas três baterias de testes, cada uma delas considerando a variação dos valores de K e do atributo considerado. Cada uma das baterias de testes são apresentadas nas subseções abaixo.

3.4.1 Avaliação dos grupos com o atributo NDC

Para o primeiro teste foi realizada o agrupamento dos dados considerando o uso do atributo NDC com o valor de $K = 3$. O resultado do agrupamento é apresentado na Figura 7, onde é possível perceber que o Brasil e a Rússia apresentam comportamentos que acabam sendo diferentes dos apresentados pelos demais países analisados. Ao conferir as características de cada um dos grupos, percebe-se que para um destes dois casos, que ficaram em grupos separados, há valores de variância muito diferente dos demais grupos, o que faz com que, mesmo outras características sendo próximas as demais séries, estes fiquem em grupos isolados dos demais.

Figura 7: *Agrupamento pelo número diário de casos*



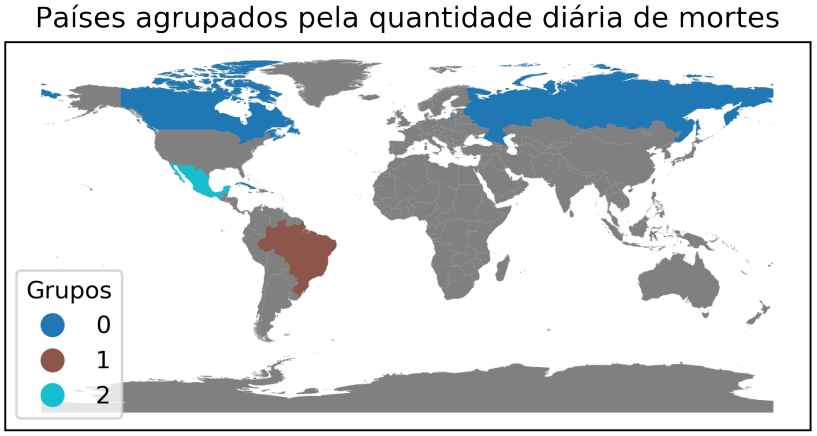
Para este resultado vale notar que, mesmo considerando que o *K-Means* precisa realizar a classificação de cada uma das séries de dados em um grupo, o que pode, a priori ser indicado como apenas um viés do método, considera-se que se tais grupos não fossem tão discrepantes dos demais, eles poderia estar agrupados juntos, dado que sua proximidade os relacionaria e possivelmente outros países tomariam os grupos específicos

3.4.2 Avaliação dos grupos com o atributo NDM

Diferente dos resultados da análise feita anteriormente, para o caso do atributo NDM, tem-se que o México e o Brasil apresentam características discrepantes aos demais membros,

o que faz estes serem inseridos em grupos separados dos demais, como apresentado na Figura 8

Figura 8: *Agrupamento pelo número diário de mortes*



Vale considerar também que, neste caso, o valor de variância encontrado também difere entre os elementos que ficaram em grupos separados dos demais, assim como encontrado no teste anterior. Assim como levantado no teste anterior, a ideia de que, mesmo existindo a necessidade de definição de uma série para cada grupo, fica evidente que os países isolados, possuem características que evitam que estes sejam postos em grupos com outros elementos, como acontece com o Canadá e a Rússia.

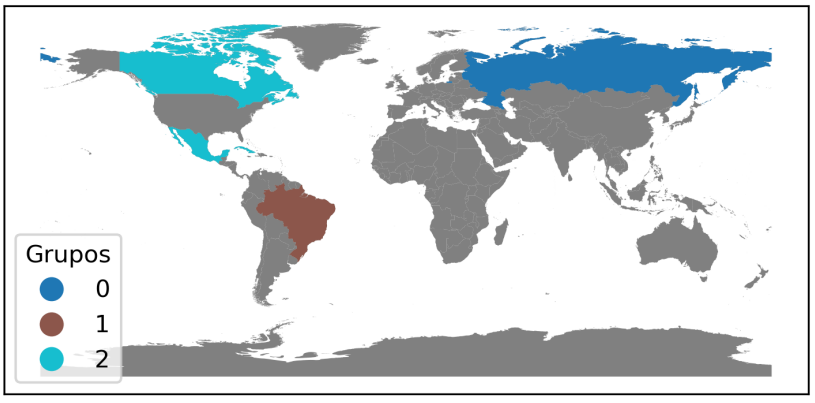
3.4.3 Avaliação dos grupos com os atributos NDC e NDM

Como forma de considerar os dois atributos avaliados anteriormente em uma única análise de agrupamento, esta seção apresenta os testes e avaliação dos resultados obtidos ao realizar o agrupamento considerando estes dois atributos simultaneamente. Para este caso, foi decidido que os agrupamentos realizados serão feitos considerando a variação de K em todos os valores do intervalo apresentado anteriormente. O primeiro teste, feito com valor de $K = 2$ é apresentado na Figura 9, neste é possível notar que o Brasil, mesmo tendo iniciado as curvas da pandemia junto a outros países que estão sendo analisados, como o Canadá, fica isolado dos demais.

Figura 9: *Agrupamento pelo número diário de casos e mortes (K = 2)*

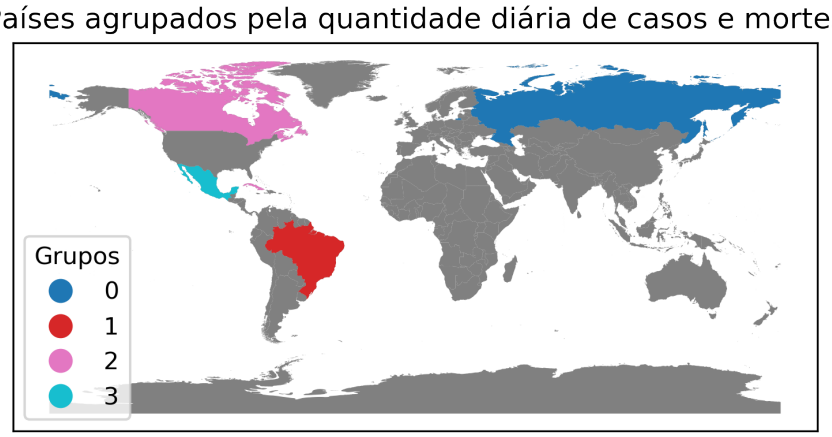
Para este caso, o valor pequeno de K foi definido com o objetivo de identificar o elemento que possuí as características mais discrepantes dentre todos os países que estão sendo analisados. Por outro lado, ao colocar o valor de $K = 3$, percebe-se o mesmo comportamento já apresentado nos casos de testes anteriores, onde, o Brasil e a Rússia ficam separados dos demais elementos do conjunto de dados.

Figura 10: *Agrupamento pelo número diário de casos e mortes (K = 3)*



Por fim, é feita a análise considerando o valor de $K = 4$, para esta, o interessante é notar que, o México, que no segundo teste ficou isolado junto ao Brasil, acabou ficando também em um grupo separado.

Figura 11: *Agrupamento pelo número diário de casos e mortes ($K = 4$)*



É possível considerar que, todo o comportamento mapeado na análise de agrupamento, sofreu influências da variação dos valores de K, onde elementos que se desviam muito dos demais elementos do conjunto de dados foram sendo separados.

3.5 Análise de regressão

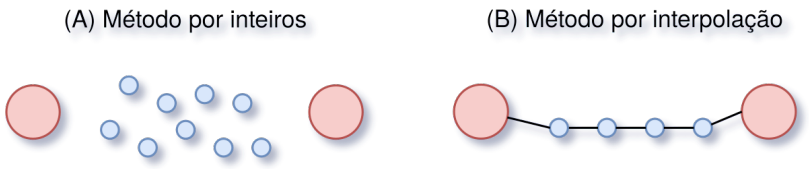
Adicionar informações sobre a análise estatística e espectral frente ao contexto da pandemia de COVID-19 que está sendo apresentada atualmente

3.6 Criticalidade auto-organizada

Identificações de assinatura SOC podem ser feitas em qualquer conjunto de dados, exigindo apenas que este tenha disponível grandes quantidades de amostras. Com isso, a caracterização de SOC no conjunto de dados criado para este trabalho torna-se inviável sem a aplicação de métodos para o aumento do conjunto de amostras. Desta forma, para o o aumento de dados, fez-se a aplicação de interpolações aleatórias nas séries de dados. Neste caso, foi feita a implementação de dois tipos diferentes de interpolação aleatória, uma discreta, nomeada de `Método por inteiros` e outra contínua, nomeada de `Método dos polinômios`.

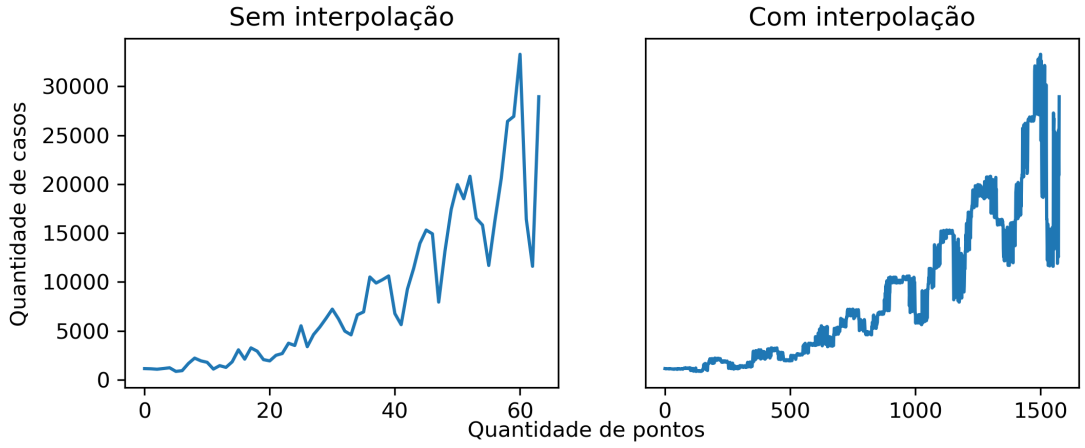
A forma de funcionamento de cada um dos métodos é semelhante, onde, dado um intervalo entre duas datas, faz-se a divisão deste intervalo em N intervalos h e para cada um desses atribui-se um valor. A diferença está somente na maneira com que esta atribuição de valores ocorre. Para o caso do `Método por inteiros`, valores inteiros aleatórios são gerados e atribuídos à cada intervalo h. Já no `Método dos polinômios`, faz-se primeiro o ajuste de um polinômio entre os valores de duas datas, em seguida, este polinômio é utilizado para gerar os valores de cada intervalo h. A representação de cada um desses métodos é feita na Figura 12.

Figura 12: *Métodos de interpolação aleatória*



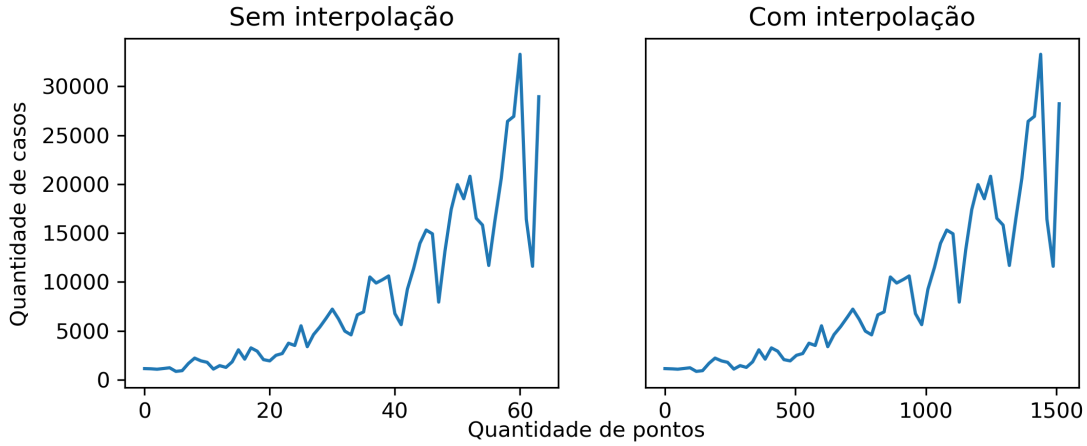
No método de inteiros, a série temporal gerada é pode ser representada como apresentado na Figura 13. Perceba que a série mantém o mesmo formato, porém com um sinal mais grosso, representando múltiplos valores dentro de um intervalo. Neste caso, para cada intervalo entre duas datas é inserido 23 valores, criando a representação das horas.

Figura 13: *Método de interpolação aleatória com valores inteiros*



Por outro lado, na Figura 14 tem-se a apresentação do método por polinômios, onde o sinal aparentemente é o mesmo, isso ocorrendo por conta da natureza contínua dos elementos gerados, o que pode representar uma variação não muito grande dos elementos que estão dentro do conjunto pode ocorrer.

Figura 14: *Método de interpolação aleatória com valores contínuos*

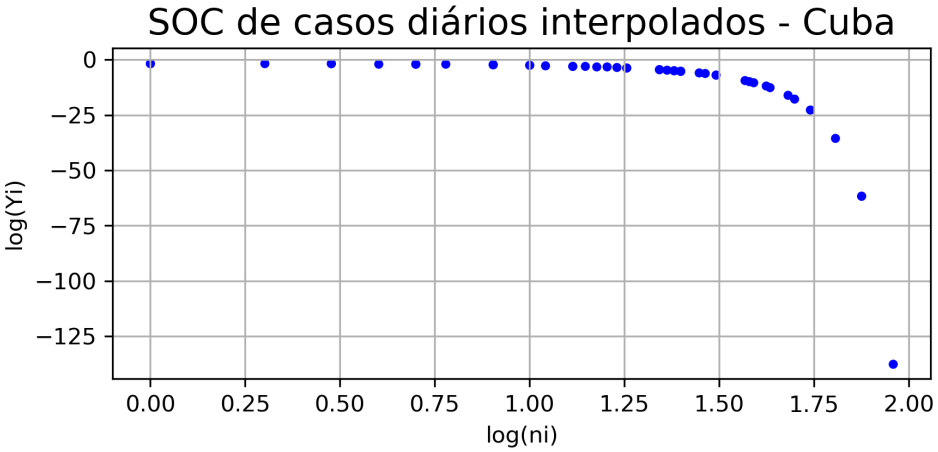


Com estes métodos de interpolação criados, fez-se o aumento de dados de cada uma das séries temporais que constituem o conjunto de dados deste trabalho e então a assinatura SOC pode ser verificada. Como exemplo de resultado desta operação, tem-se os dados de Cuba, que seu o aumento de dados, apresentada algo que não tinha uma boa representação de SOC, como apresentado na Figura 15.

Figura 15: *Assinatura SOC para os dados de Cuba sem interpolação*

Por outro lado, ao aplicar a interpolação, uma possível assinatura SOC é apresentada (Figura 16). Para este caso o método de inteiros foi aplicado, porém, o método de polinômios poderia ser aplicado também.

Figura 16: *Assinatura SOC para os dados de Cuba com interpolação*



Vários outros testes foram realizados e estão disponíveis no repositório[4].

4 Predição de casos diários

Para a etapa de predição de casos diários do COVID-19, fez-se a implementação do modelo desenvolvido e disponibilizado pelo Professor Doutor Reinaldo Rosa. O modelo é baseado em cascatas multiplicativas não-homogêneas, que utiliza como entrada a média de novos casos dos sete dias anteriores ($\langle N_{nb7}$) e o valor do dia atual (N_{kt}) para prever o dia seguinte (N_s) ao N_{kt}. A equação mestre do modelo pode ser definida como apresentado abaixo.

$$N_{s_{min}} = g\ (1 \cdot n_1 + 3 \cdot n_2 + 5 \cdot n_3) \tag{1}$$

$$N_{s_{max}} = g\ (2 \cdot n_1 + 4 \cdot n_2 + 6 \cdot n_3) \tag{2}$$

Onde $N_{s_{min}}$ e $N_{s_{max}}$, representam, respectivamente, os valores mínimos e máximos de possíveis infectados para o dia a ser predito. Para isto,

$$n_1 = p_1 \cdot N_{kt} \tag{3}$$

$$n_2 = p_2 \cdot N_{kt} \tag{4}$$

$$n_3 = p_3 \cdot N_{kt} \tag{5}$$

e

$$g = \frac{\langle N_{nb} \rangle_7}{N_{kt}},\ se\ N_{kt} > \langle N_{nb} \rangle_7 \tag{6}$$

[4]github.com/cmath-covid

$$g = \frac{N_{kt}}{\langle N_{nb} \rangle_7}, \, se \, N_{kt} < \langle N_{nb} \rangle_7 \tag{7}$$

O modelo também calcula o fator de supressão, $s(t)$, baseado nas derivativas de g: Δ_g e n: Δ_{nk}, onde, o Δ_g pode ser definido como:

$$\Delta_g = (g_0 - g) - q_g, \; se \; g_0 < g \tag{8}$$

$$\Delta_g = (g_0 - g) + q_{g0}, \; se \; g_0 \geq g \tag{9}$$

Assim, o derivativo de Δ_{nk} é definido como:

$$\Delta_{nk} = \frac{(\langle N_{nb} \rangle_7 - N_{kt})}{N_{kt}} \tag{10}$$

Onde,

$$s = \frac{2\Delta_g + \Delta_{nk}}{3} \tag{11}$$

Com isto, para testar a implementação realizada, foram gerados casos de testes para todas as localidades que estão sendo analisadas neste trabalho. As subseções abaixo, porém, apresentam dois destes casos de teste, onde são considerados os dados do Estado de Minas Gerais e os dados gerais do Brasil. Os demais testes estão listados e documentados no repositório do projeto[5]. Todos os testes foram desenvolvidos considerando o perído de 08/04/2020 à 28/04/2020, uma vez que, o comportamento crescente do modelo, fez com que, a aplicação do mesmo neste trabalho, fosse feita de modo a considerar intervalos de predição.

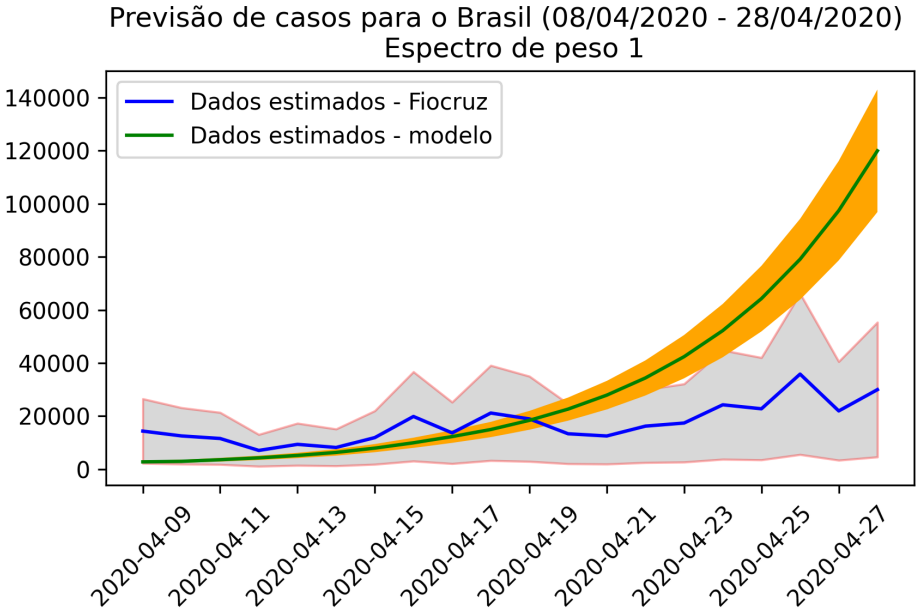
Para a realização de um comparativo da quantidade de elementos que foram estimadas pelo modelo são consideradas as quantidades de possíveis novos casos diários estimados pela Fiocruz.

4.0.1 Dados gerais do Brasil

Para começar, os testes com os dados do Brasil, fez-se a predição dos casos considerando o espectro de peso 1. O resultado apresentado na Figura 17, mostra o comportamento geral de previsão do modelo implementando ao considerar tal espectro de pesos. O comportamento, indica uma crescente contínua, de forma exponencial, do crescimento do número de casos diários no Brasil, porém, vale ressaltar que, o comportamento do modelo, como apresentado anteriormente, considera fatores de subnotificação.

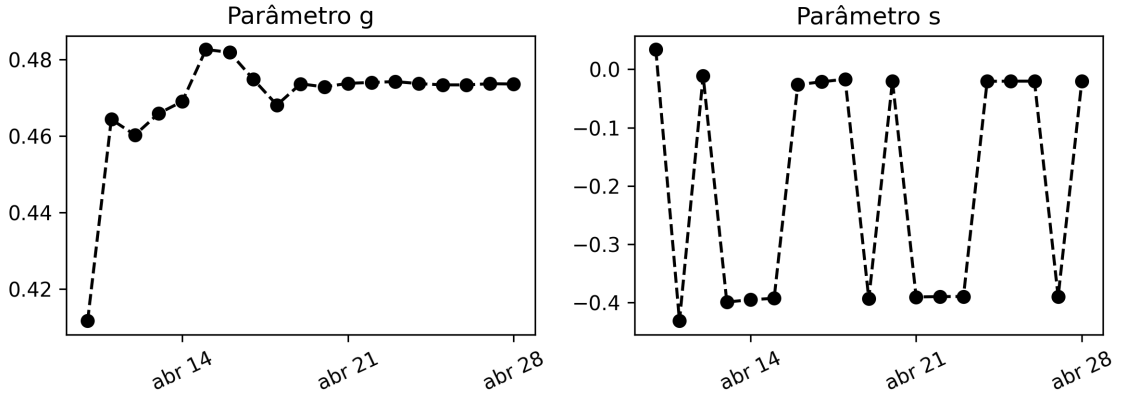
[5]github.com/cmath-covid

Figura 17: *Resultados do modelo com o 1° espectro de pesos*



Durante este teste, para o valor de g, tem-se que o mesmo variou dentro de um intervalo positivo, passando a ter uma baixa variação após algumas iterações do modelo. Por outro lado, os valores do parâmetro s mantiveram um padrão de variação durante toda a simulação de predição realizada, como visto na Figura 18.

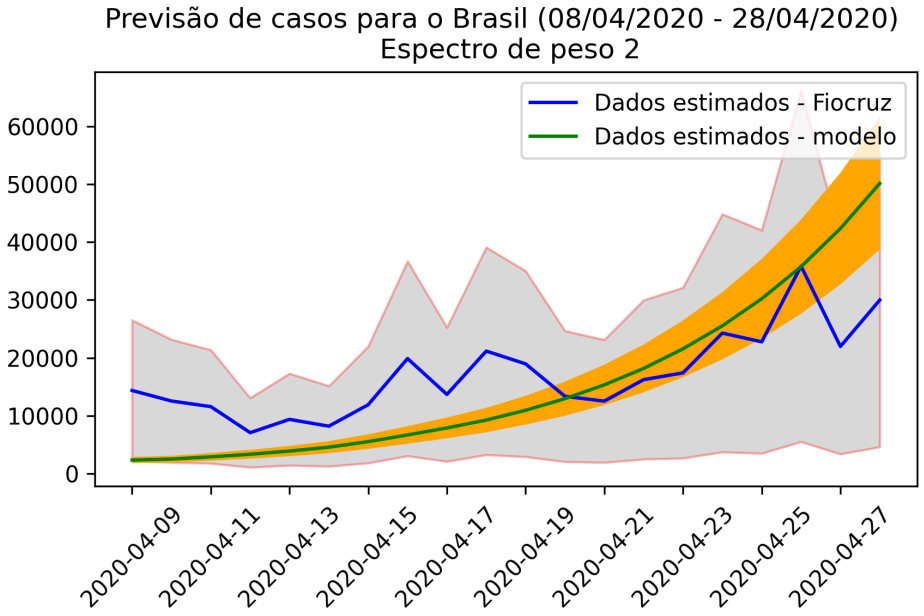
Figura 18: *Variação dos parâmetros g e s com espectro de peso 1*



Após a finalização deste primeiro teste de previsão dos casos diários, um segundo teste foi realizado, este considerando o espectro de peso 2 (Figura 19). Os resultados, diferente do apresentado anteriormente, apresentam um crescimento contínuo, porém, em passos menores e mais suaves, garantindo que a previsão acompanhe a tendência das notificações de casos reais. Este comportamento torna evidente a capacidade do modelo de identificar

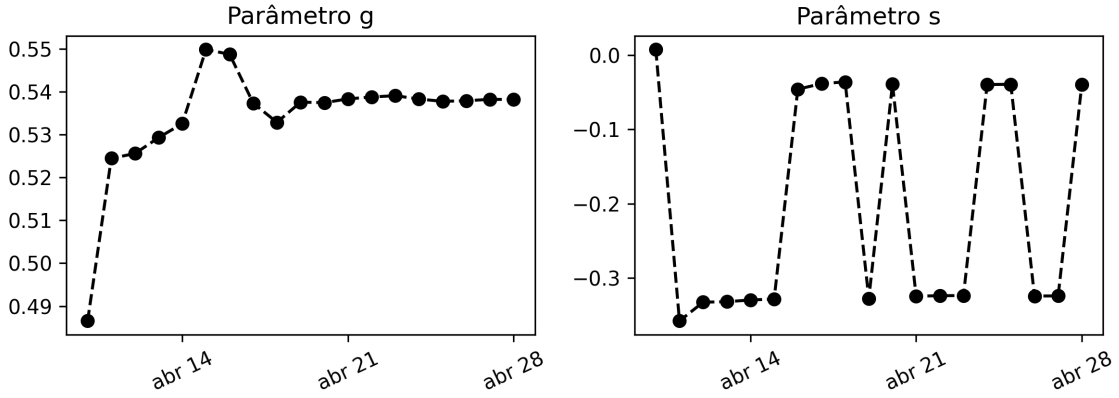
as subnotificações presentes nos dados, mantendo-se próximo aos valores estimados pela Fiocruz.

Figura 19: *Resultados do modelo com o 2° espectro de pesos*



Para os parâmetros g e s, é percebido um comportamento igual ao levantado anteriormente, com a diferença de que as variações ocorrem em um intervalo de valores menor.

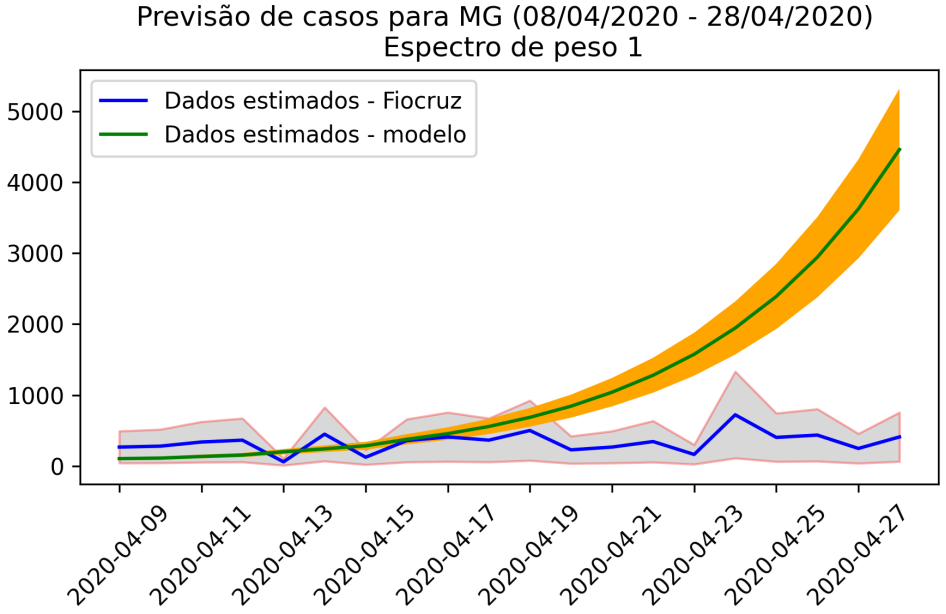
Figura 20: *Variação dos parâmetros g e s com espectro de peso 2*



4.0.2 Dados do Estado de Minas Gerais

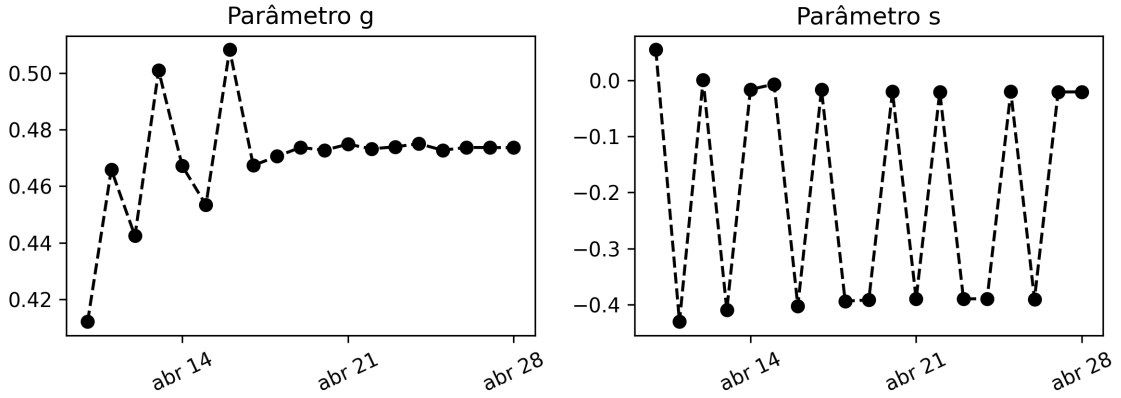
Para a avaliação do comportamento do modelo, foi criada uma segunda bateria de testes, esta considerando os dados do Estado de Minas Gerais, os passos tomados aqui seguem o mesmo padrão dos testes realizados para o conjunto de dados completo do Brasil, com isto, inicialmente faz-se a aplicação do modelo considerando o 1° espectro de pesos e da mesma maneira como apresentado no teste anterior, o comportamento das previsões feitas pelo modelo com estes espectros de peso acabam sendo muito diferentes da realizada, obtendo comportamento de crescimento exponencial.

Figura 21: *Resultados do modelo com o 1° espectro de pesos*



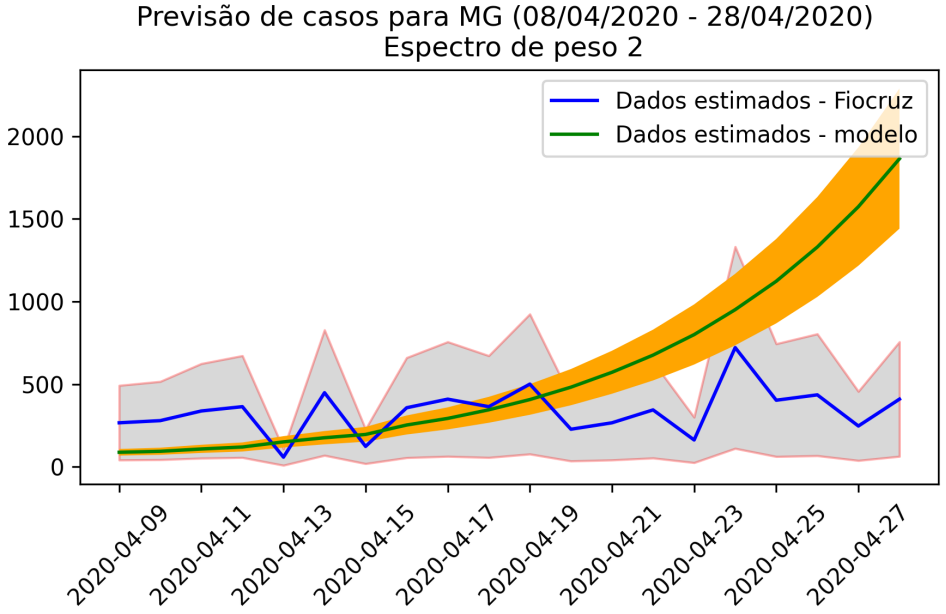
Neste teste, os valores de g e s, como pode ser observado na Figura 22, possuem o comportamento similar ao apresentado na Figura 18, sendo diferente somente nas escalas de variação.

Figura 22: *Variação dos parâmetros g e s com espectro de peso 1*



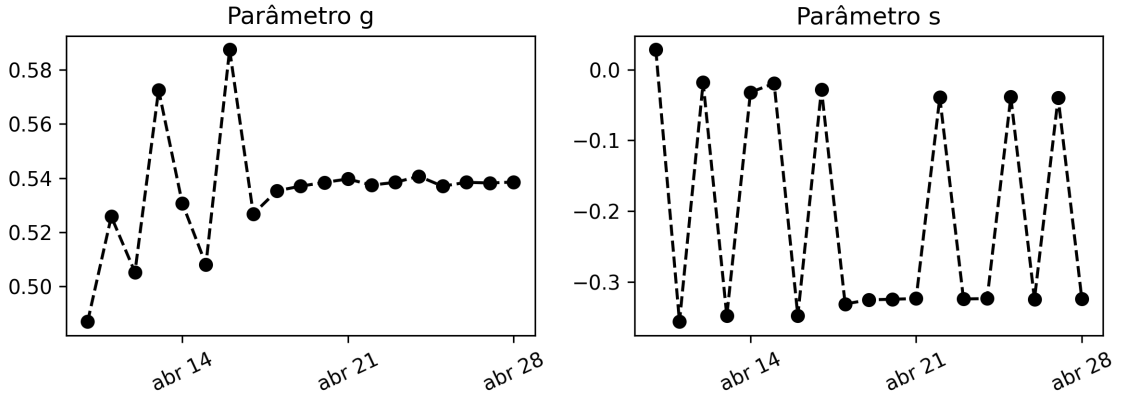
Ao considerar o 2° espectro de pesos para estes dados, o mesmo comportamento observado na Seção 4.0.1 é apresentado, onde uma curva que se ajusta de maneira muito mais clara a realizada é gerada.

Figura 23: *Resultados do modelo com o 2° espectro de pesos*



Nos parâmetros *g* e *s* comportamentos similares ao observado na Figura 20 são identificados.

Figura 24: *Variação dos parâmetros g e s com espectro de peso 2*



5 Conclusão

Referências