

# Análise automática do Painel Coronavírus

Alberto Saa

UNICAMP

(09/07/2020)

## Resumo

A ideia pequeno projeto é a elaboração de um sistema em código Python para uma análise automática, no contexto de um modelo SIR simples, dos dados da epidemia de COVID-19 publicados diariamente pelo Ministério da Saúde. As análises diárias serão publicadas no endereço [1]. O objetivo deste projeto é puramente educacional, com ênfase na análise de dados e programação em Python, e não em epidemiologia. Não obstante, todos os dados tratados aqui são reais e, portanto, os resultados talvez possam ter alguma relevância para se entender a dinâmica real da epidemia de COVID-19, a qual está muito bem analisada, por exemplo, em [2]. Se for citar este sistema ou documento, por favor faça-o como:

- A. Saa, “*Análise automática do Painel Coronavírus*”, 2020. Texto integralmente disponível em <https://vigo.ime.unicamp.br/COVID/covid.pdf>

Todos os códigos e arquivos de dados pertinentes para este projeto estão disponíveis no repositório [3]. Infelizmente, o autor não pode dar nenhum tipo de suporte para a utilização do sistema, mas incentiva todos interessados a utilizar, como quiserem, todo o material disponível no repositório.

# 1 Introdução

Diariamente, o Ministério da Saúde (MS) publica<sup>1</sup> no site [4] alguns dados agregados sobre a epidemia de COVID-19 no País. Infelizmente, contrariando-se as práticas mais elementares para tratamento público de dados, a divulgação é feita diariamente em formato proprietário, especificamente numa planilha Microsoft Excel. No entanto, este ponto pode ser facilmente sanado, pois há várias ferramentas públicas, como o pacote LibreOffice, que permitem converter a planilha em Excel, por exemplo, em um arquivo de texto CSV (*comma-separated values*), possibilitando sua ampla utilização. Utilizaremos aqui o formato CSV.

De todos os dados publicados pelo MS, utilizaremos apenas o número de casos detectados, tanto em sua versão diária como total acumulado até uma certa data. Estes dados são interpretados no contexto do modelo epidemiológico mais simples: o modelo SIR. O número total de óbitos também é utilizado para a construção de alguns gráficos, mas não são utilizados na interpretação no modelo SIR. É interessante notar que, mesmo no contexto bastante simples destes modelos, as tendências identificadas para o número básico reprodutivo  $r_0$  são compatíveis com as análises mais detalhadas apresentadas, por exemplo, em [2].

# 2 Modelo SIR

Há uma vasta literatura sobre o modelo SIR e suas variantes, ver [5], por exemplo, para mais referências. Em sua versão mais simples, que é a que será empregada aqui, uma população de  $N$  indivíduos é dividida em três classes: os suscetíveis a infecção ( $S$ ), os infectados ( $I$ ) e os recuperados, ou removidos, ( $R$ ). A ideia é simples. Os elementos suscetíveis  $S$  são aqueles que podem vir a ser infectados a partir do contato com os infectados  $I$ . Já os elementos da classe  $R$  são aqueles que não mais se infectam e nem vão infectar outros indivíduos, seja por cura com imunidade, seja por que foram afastados. Eventuais óbitos, neste tipo de modelo, são contados na classe  $R$ .

---

<sup>1</sup>Houve certa turbulência na divulgação destes dados durante alguns dias, mas o problema parece ter sido superado. Veja nota sobre a importação dos dados em [1].

A dinâmica do modelo é dada pelo sistema de EDO

$$\frac{dS}{dt} = -\frac{\beta IS}{N}, \quad (1)$$

$$\frac{dI}{dt} = \frac{\beta IS}{N} - \gamma I, \quad (2)$$

$$\frac{dR}{dt} = \gamma I, \quad (3)$$

sendo  $\beta$  e  $\gamma$  parâmetros positivos associados, normalmente, às taxas de infecção e de remoção, respectivamente, no modelo. Ambos têm dimensão  $[T^{-1}]$ . É evidente das equações que os parâmetros  $\beta$  e  $\gamma$ , respectivamente, controlam as conversões das classes  $S \rightarrow I$  e  $I \rightarrow R$ . Esta última conversão representa a evolução natural da doença em questão, e  $\gamma^{-1}$  pode ser interpretado como o tempo médio que um indivíduo passa infectado antes de ser removido do sistema, independente da natureza desta remoção, seja ela uma cura espontânea e assintomática com imunidade, um diagnóstico com isolamento ou hospitalização, ou mesmo um óbito. O parâmetro  $\beta$ , a taxa de infecção, está relacionado com a facilidade de transmissão da doença. Quanto maior o parâmetro  $\beta$ , mais fácil é a transmissão. O parâmetro  $\beta$  não depende apenas da evolução da doença, mas também de diversos outros fatores como intensidade de contato entre os indivíduos da população, efeitos climáticos, etc.

Um dos parâmetros mais importantes neste tipo de modelo é o chamado número básico reprodutivo

$$r_0 = \frac{\beta}{\gamma}. \quad (4)$$

Trata-se de um adimensional cuja interpretação mais simples é o número médio de novos casos gerados por um infectado em uma população não imune. Notem que a Eq. (2) pode ser escrita como

$$\frac{dI}{dt} = \left( r_0 \frac{S}{N} - 1 \right) \gamma I, \quad (5)$$

de onde podemos inferir alguns comportamentos qualitativos interessantes. Notem, primeiro, que  $\frac{S}{N}$  é a fração dos suscetíveis na população e, portanto, está no intervalo  $[0, 1]$ . É evidente de (5) que, para  $r_0 < 1$ , o número de infectados irá decrescer monotonicamente. Trata-se da extinção da epidemia. Para  $r_0 > 1$ , temos um comportamento qualitativo diferente. Para  $r_0 \frac{S}{N} > 1$ ,

o número de infectados cresce, implicando a expansão da epidemia. Porém, com a expansão da epidemia, a fração  $\frac{S}{N}$  tende a diminuir, desacelerando o ritmo de crescimento do número de infectados. Ao chegar ao valor dado por

$$\frac{S}{N} = \frac{1}{r_0}, \quad (6)$$

o número de infectados  $I$  deixará de crescer, e a dinâmica do sistema (1)-(3) implicará na extinção da epidemia. Esta é a situação conhecida como imunidade de grupo, ou imunidade de “rebanho” (*herd immunity*). Iremos considerá-la do ponto de vista dos indivíduos removidos  $R$ . Como  $I = 0$  no estágio dominado pela imunidade de grupo, temos que ele corresponde à situação com

$$n_R = \frac{R}{N} = 1 - \frac{1}{r_0} = \text{constante}. \quad (7)$$

A situação limite  $r_0 = 1$  normalmente está associada a fases endêmicas da infecção, fases nas quais o número de infectados permanece constante no tempo. Para nossos propósitos, basta notarmos que  $r_0 > 1$  implica na expansão da epidemia e, quanto maior for o valor de  $r_0$ , mais rápida será a expansão e maior será o coeficiente  $n_R$  dado por (7) para a imunidade de grupo. Obviamente, quanto maior for  $r_0$ , maior será o impacto da epidemia. Por outro lado, para  $r_0 < 1$  a epidemia sempre se extingue, e quanto menor  $r_0$ , mais rápida será esta extinção.

Admitindo-se que a população total  $N = S + I + R$  é constante, temos que o sistema de equações diferenciais (1) - (3) é efetivamente bi-dimensional e, portanto, basta escolhermos duas classes para termos a descrição completa do modelo. Em nosso caso, escolheremos sempre  $R$  e  $I$  como variáveis dinâmicas. Além disso, o caso de  $\alpha$  e  $\beta$  e, consequentemente,  $r_0$  constantes merece uma atenção especial por sua simplicidade. Neste caso particular, podemos escrever as equações (1) e (3) como

$$\frac{d}{dt} \ln S = -\frac{r_0}{N} \frac{dR}{dt}, \quad (8)$$

que para  $r_0$  constante pode ser imediatamente integrada como

$$S(t) = S_0 \exp\left(-r_0 \frac{R - R_0}{N}\right), \quad (9)$$

sendo  $S_0$  e  $R_0$  constantes arbitrárias de integração, que serão determinadas a partir das condições iniciais do tipo  $S_0 = S(t_0)$  e  $R_0 = R(t_0)$ . Usando-se

novamente (3), podemos escrever (9) como

$$\gamma^{-1} \frac{dR}{dt} = N + R - S_0 \exp\left(-r_0 \frac{R - R_0}{N}\right), \quad (10)$$

que, por ser uma equação separável, também pode ser formalmente resolvida com uma quadratura para o caso de coeficientes constantes, resultando em

$$\int_{R_0}^R \frac{du}{N + u - S_0 \exp\left(-r_0 \frac{u - R_0}{N}\right)} = \gamma(t - t_0). \quad (11)$$

Porém, a primitiva envolvida não pode ser expressa em termos de funções elementares, e portanto esta solução exata não nos ajudará muito em análises quantitativas. A equação (8) será útil mais adiante para estimarmos  $r_0$  a partir dos nossos dados.

Há diversas estimativas disponíveis para os parâmetros do modelo SIR para o caso da atual epidemia de COVID-19. Por exemplo, em [6] estima-se, com base em dados da epidemia na China, que  $r_0 \approx 2.6$ , com intervalo de confiança  $(1.5 - 3.5)$ , e  $\beta \approx 0.34$ , que implicaria  $\gamma \approx 0.13$ . Em [7], obtém-se  $\gamma \approx 0.145$ , com intervalo de confiança  $(0.119 - 0.182)$ , a partir de dados da epidemia no Brasil. Os dados chineses e brasileiros implicam que um infectado, em média, necessita de  $\gamma^{-1} \approx 7$  dias para ser removido do sistema. Durante este período, o infectado tem potencial de disseminar a doença. Estes valores são perfeitamente compatíveis com o que se sabe sobre a evolução natural da doença, ver, por exemplo, [8].

Necessitamos agora interpretar os dados divulgados pelo MS no contexto do modelo SIR. Tomaremos para nossa análise o número acumulado de casos confirmados, que chamaremos de  $C(t)$ . Dada a situação específica da epidemia de COVID-19, sabe-se que estes casos correspondem a indivíduos que, ou estão hospitalizados, ou acudiram a um hospital com sintomas moderados. É razoável supor que, a partir deste diagnóstico, eles terminem internados ou liberados com recomendações estritas de isolamento e quarentena. Nessa situação, parece razoável considerá-los como elementos removidos  $R$  de nosso sistema, pois não se espera que continuem a propagar a doença, independente do resultado final, seja ela cura e imunização, ou óbito. A questão é que há muito mais indivíduos em  $R$  que os dados por  $C$ . No caso específico da epidemia de COVID-19, é razoável supor que estes infectados não detectados são os assintomáticos e também aqueles com sintomas leves que não são atendidos clinicamente e, portanto, desconhecem suas situações de infectados. Na ausência de testes em massa, estes outros elementos de  $R$  devem ser

estimados a partir de outras observações. Esta é uma tarefa extremamente complexa e que está completamente fora do escopo deste projeto. Baseado nas análises apresentadas em [9], usaremos como hipótese que a relação entre os casos confirmados e os reais no País está na faixa entre 1 : 10 e 1 : 20, que corresponde a introduzir  $10 < \alpha < 20$ , com  $R = \alpha C$ . Há consistentes indícios de relações entre casos confirmados e reais desta mesma ordem de magnitude também em outros países [10].

## 2.1 Determinando $r_0$

Nosso propósito é determinar, a partir dos dados disponíveis o número reprodutivo  $r_0$ , o qual não admitimos constante. Por outro lado, admitiremos  $\gamma$  e  $\alpha$  constantes. Há, pelo menos, duas maneiras diferentes para prosseguirmos. O primeiro e talvez mais direto, é determinarmos o valor “instantâneo” de  $r_0$  a partir das equações (2) e (3). Supondo-se  $\gamma$  e  $\alpha$  constantes, podemos obter

$$r_0(t) = \frac{1}{1 - \mu} + \frac{\ddot{C}}{\gamma \dot{C}(1 - \mu)}, \quad (12)$$

com

$$\mu(t) = \frac{\alpha}{\gamma N}(\gamma C + \dot{C}). \quad (13)$$

É interessante notar que, enquanto estivermos longe da imunidade de rebanho, é natural admitir que  $\mu \ll 1$  e, portanto, teremos a aproximação

$$r_0(t) = 1 + \frac{\ddot{C}}{\gamma \dot{C}}, \quad (14)$$

que independe de  $\alpha$ . Esta última expressão ajudará a entender porque a tendência de expansão ou extinção da epidemia não deve depender dos valores de  $\alpha$  e  $\gamma$ , pelo menos nos estágios distantes da imunidade de rebanho. A função  $\mu$  dada por (13) tem também uma outra interessante utilidade. Suponha que consigamos identificar o instante  $t_*$  em que  $\ddot{C} = 0$ , que corresponde ao ponto de inflexão da curva de casos acumulados ou, em outras palavras, o máximo da variação  $\dot{C}$ . Se soubermos o valor de  $r_0(t_*)$ , teremos a identidade

$$\mu(t_*) = 1 - \frac{1}{r_0(t_*)}, \quad (15)$$

de onde podemos determinar o valor de  $\alpha$  conhecendo-se  $C(t_*)$  e  $\dot{C}(t_*)$ .

A segunda maneira para se estimar  $r_0(t)$  a partir dos dados explora a equação (8). Integrando-se ambos os lados, temos

$$\ln \frac{S_0}{S(t)} = \frac{1}{N} \int_{t_0}^t r_0(u) \dot{R}(u) du. \quad (16)$$

É razoável supor que, pelo menos nas fases mais agudas da epidemia, que a função  $R(t)$  é monótona crescente, *i.e.*,  $\dot{R} > 0$ . Nestas fases, podemos mudar a variável de integração e escrever

$$\ln \frac{S_0}{S(t)} = \bar{r}_0 \frac{R(t) - R_0}{N}, \quad (17)$$

que é formalmente semelhante a (9). A diferença é que usamos o teorema do valor médio na integral e  $\bar{r}_0$  deve ser entendido como o valor médio de  $r_0$  no intervalo  $[t_0, t]$ . Como veremos, a expressão (16) oferece algumas vantagens em relação à (12) para a determinação de  $r_0(t)$  a partir dos dados da epidemia.

### 3 Tratamento dos dados

Nossos dados são séries temporais diárias. Assim, iremos discretizar nossa variável independente  $t \in \mathbb{Z}$ . Todas as derivadas serão substituídas por diferenças finitas. Optou-se por diferenças atrasadas, *i.e.*

$$\dot{C}_t = C_t - C_{t-1} \quad \text{e} \quad \ddot{C}_t = \dot{C}_t - \dot{C}_{t-1}, \quad (18)$$

mas este ponto parece ter pouca influência no problema. A Fig. 1 mostra os casos acumulados ( $C_t$ ) e os casos novos ( $\dot{C}_t$ ) para o estado de São Paulo. É evidente destes dados que não conseguiremos realizar nenhuma análise que envolva diferenças finitas, pois há claramente demasiado “ruído” nas séries. Não obstante, é claro que há um evidente regime de crescimento para as duas quantidades. Para qualquer análise que envolva diferenças, como por exemplo as estimativas de  $r_0$ , precisaremos “suavizar” os dados.

Há diversas técnicas para suavização de dados. Utilizaremos a que talvez seja a mais simples de todas, a dos filtros de média móvel. A ideia subjacente destes filtros é substituir o valor de um elemento da série temporal por uma média calculada numa vizinhança simétrica com  $2n + 1$  elementos,  $n \in \mathbb{N}$ , *i.e.*

$$\bar{C}_t = \frac{1}{2n + 1} \sum_{k=t-n}^{t+n} C_k. \quad (19)$$

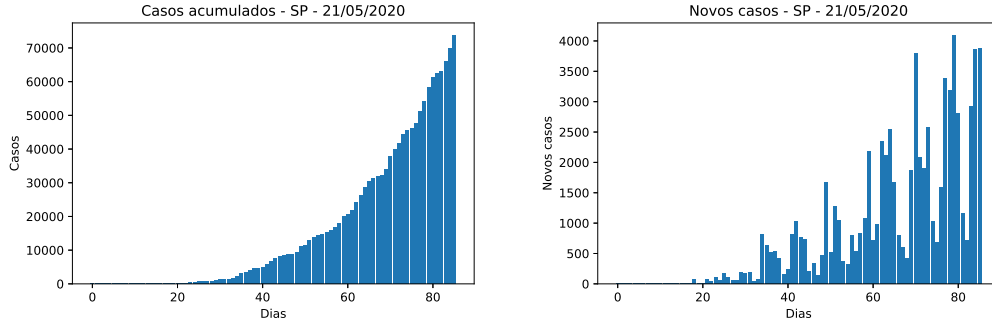


Figura 1: Casos acumulados e novos casos para o estado de São Paulo, obtidos a partir dos dados publicados pelo Ministério da Saúde em [4]. É evidente a presença de “ruído” que impossibilita qualquer análise da variações diretamente destes dados.

A vizinhança sobre a qual se calcula a média é normalmente chamada de “janela” do filtro. Este tipo de filtro é extremamente atraente para o nosso problema específico, já que um dos ruídos mais comuns em nossos dados é a defasagem de alguns dias na incorporação de novos casos. Por certo, esta é a fonte principal do ruído de período semanal presente na Fig. 1. Vários casos do fim de semana acabam notificados apenas na segunda ou terça-feira. Este tipo de problema pode ser conveniente sanado com filtros do tipo média móvel, com  $n = 3$ , ou uma semana de janela. Este será nosso filtro padrão.

Uma observação mais atenta de (19) revela que este filtro, como posto, está definido para uma sequência com  $k$  elementos apenas para  $n < t < k - n$ . Nas regiões próximas às “bordas” da sequência, não conseguimos mais definir a vizinhança simétrica e não podemos calcular (19). Para resolver este problema, devemos impor algumas condições sobre as médias móveis próximas as bordas da sequência. Esta é uma questão delicada em qualquer análise de séries temporais, pois estas condições devem ser escolhidas a fim de não comprometer as tendências que queremos identificar. Na prática, devemos estender a sequência  $C_t$  acrescentando  $2n$  novos elementos, correspondentes a  $1 - n < t \leq 0$  e  $k < t \leq k + n$ . Para o problema em questão, optamos pelas seguintes extensões

$$C_{1-j} = C_1 \text{ e } C_{k+j} = C_k + C_{k-n+j} - C_{k-n}, \quad (20)$$

para  $1 \leq j \leq n$ . A primeira condição é evidente, admitimos a sequência constante na borda esquerda. A segunda condição parece menos clara, mas



é também muito simples. Estamos admitindo que os  $n$  elementos  $C_t$  da sequência com  $k - n < t \leq k$ , devidamente deslocados verticalmente, sucedem o último elemento da esquerda da sequência original. Ambas extensões equivalem a supor que as tendências de crescimento, no contexto de um modelo epidemiológico, são preservadas nas bordas da sequência. Assim, podemos definir o filtro (19) para uma sequência  $C_t$  para qualquer  $1 \leq t \leq n$ .

O filtro (19) pode ser visto também como a convolução da função  $C_t$  com uma função retangular de altura 1 e largura  $2n+1$ . É comum iterarmos a ação destes filtros. Por exemplo, duas aplicações sucessivas de um filtro do tipo média móvel com janela de  $2n + 1$  elementos, corresponderá a convolução da sequência original com uma função triangular com largura  $2n + 4$ , por isso o nome de filtro triangular para dois usos sucessivos do filtro de média móvel. Uma terceira iteração corresponderia a um filtro de convolução com uma função quadrática e janela  $6n + 3$ , e assim sucessivamente. Escolhemos para o tratamento de nossos dados dois filtros: o triangular com janela de 2 semanas e o correspondente a 4 iterações de (19) com  $n = 3$ , que por sua vez corresponde a uma convolução com uma função cúbica sobre uma janela de 4 semanas.

A Fig. 2 mostra os dados da Fig. 1 devidamente suavizados com nosso filtro cúbico, assim como o resultado do cálculo de  $\tilde{C}_t$  e de  $r_0$  de acordo com (12). O gráfico para  $r_0$  merece mais explicações. A linha horizontal corresponde a  $r_0 = 1$ . Acima dessa linha, temos expansão da epidemia. Para debelá-la antes de atingirmos a imunidade de rebanho, devemos sempre ter  $r_0$  abaixo desse limiar. É clara a correlação que existe entre picos de  $r_0$  e regiões de aceleração de casos, basta comparar o gráfico de  $r_0$  com o de novos casos. O gráfico de  $r_0$  apresenta uma região delimitada pelo valor de  $r_0$  calculado a partir de (12) para os limites que consideramos para o parâmetro  $\gamma$  e para  $\alpha = 15$ . É razoável supor que o valor de  $r_0$  esteja contido nessa região. Note que, pela natureza dos filtros utilizados, é sempre prudente descartarmos as regiões das bordas.

## 4 Resultados

Os resultados publicados diariamente em [1] consistem na análise dos dados publicados em [4] para os seguintes casos: Brasil, todos os Estados e o Distrito Federal, e uma série de cidades, incluindo todas as capitais de Estado. Para cada caso, são calculados e apresentados graficamente diversos indicadores,

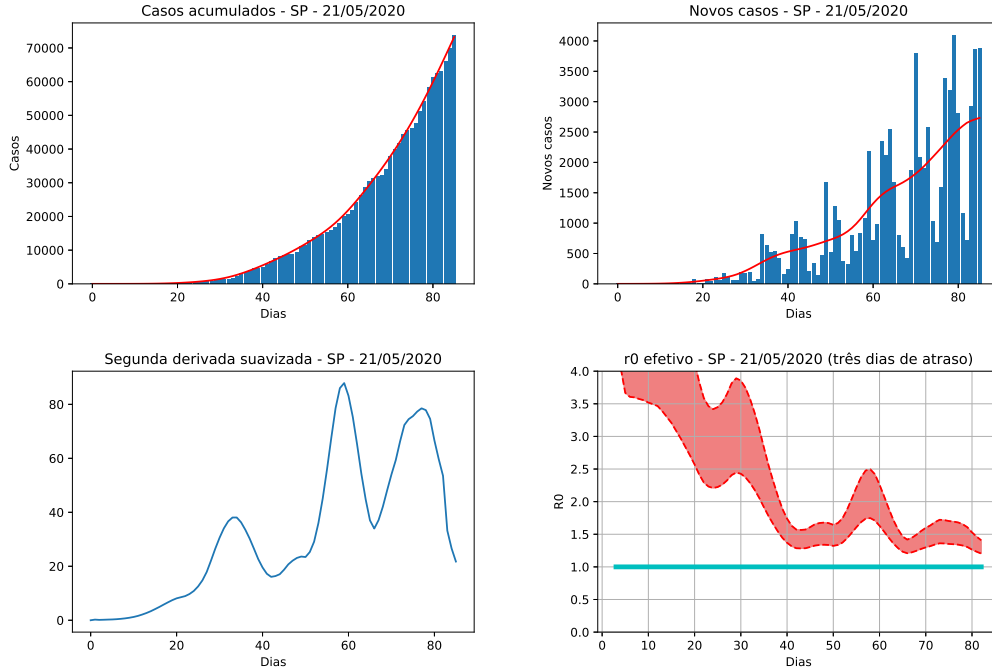


Figura 2: Acima: Casos acumulados e novos casos para o estado de São Paulo, obtidos dos dados publicados pelo Ministério da Saúde em [4], e suavizados com 4 iterações de (19) com  $n = 3$ , correspondendo a um filtro cúbico com janela de um mês. Abaixo: Segunda derivada  $\ddot{C}_t$ , calculada a partir da série suavizada, e  $r_0$ , calculado a partir de (12).

dentre eles: os casos novos e acumulados, com suas respectivas suavizações com o filtro cúbico, os valores de  $r_0$  determinados a partir de (12) (versão diferencial) e de (16) (versão integral), ambos com  $0.119 \leq \gamma \leq 0.182$  e  $\alpha = 15$ , e a previsão para os próximos 5 dias para o número de casos acumulados. A vantagem da estimativa integral para  $r_0$  sobre a diferencial é que a primeira não envolve a segunda derivada  $\ddot{C}_t$  dos dados e, portanto, está menos sujeita às flutuações. Para o cálculo de (16), basta a aplicação do filtro triangular para termos resultados consistentes, enquanto que para (12), necessitamos do filtro cúbico. A previsão de novos casos é feita a partir de uma regressão linear simples dos últimos 10 dias da série  $C_t$ . Também são calculados o número reprodutivo  $r_0$  efetivo das últimas 2 semanas como a média simples dos valores calculados a partir de (16) nesse período, e os respectivos limiares

de imunidade de rebanho associados, a partir de (7). O código Python das análises, assim como todos os dados, podem ser encontrados em [3]. Na próxima seção, alguns detalhes do código são apresentados.

As tendências identificadas nestas análises são muito mais importantes que os valores numéricos em si, já que em última instância estes valores devem ser entendidos dentro do limitado escopo dos modelos SIR com as grandes incertezas nos parâmetros  $\gamma$  e  $\alpha$ . As tendências, por outro lado, podem ser interpretadas como “sinalizadores” para o comportamento da epidemia, e não dependem dos parâmetros  $\gamma$  e  $\alpha$  para estágios distantes da imunidade de rebanho, como já pode ser adiantado de (14). Como ilustração destas análises, as Figs. 3 e 4 em anexo apresentam os resultados, com a mesma faixa de valores para  $\gamma$ , para quatro países que estão em estágio mais avançado da epidemia: Espanha, Itália, Estados Unidos e Reino Unido.

É evidente que a contenção da epidemia (decréscimo do número de novos casos) ocorre nos períodos com  $r_0 < 1$ , independentemente do valor preciso de  $\gamma$ . Na ausência de intervenções farmacológicas, a única estratégia para diminuir efetivamente o valor de  $r_0$  é dificultar a propagação do vírus, objetivo cuja maneira mais eficiente de ser alcançado para os casos de vírus respiratórios é diminuindo-se o contato social entre os indivíduos.

## 5 Alguns detalhes do código Python

O código deste projeto está escrito em Python 3. Para aplicações científicas, recomenda-se fortemente a distribuição Anaconda (versão 3, Python 3.7), disponível para várias plataformas em [11]. O código utiliza apenas recursos padrão e pacotes bem conhecidos. De particular importância para nossos propósitos, são os pacotes `numpy`, `matplotlib.pyplot` e `scipy.stats`, respectivamente, para manipulação numérica de matrizes, análise gráfica e estatística. O código é facilmente legível. Alguns detalhes, porém, merecem alguma atenção.

Iniciemos com a manipulação de arquivos CSV (*comma-separated values*). As instruções no Quadro 1 ilustram a manipulação de arquivos CSV em Python, da maneira que usamos em nosso código. O objeto `linecsv` é varrido no loop e retorna a lista `row` para cada linha do arquivo `20200519.csv`. Os elementos desta lista são as strings correspondentes a cada uma das colunas, que estão sempre separadas por vírgulas no formato CSV padrão. Nesse caso em particular, lê-se o arquivo completamente, seleciona-se as linhas

---

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Tue May 26 09:47:38 2020

@author: asaa
"""

import csv
csvfile = open("20200519.csv")
linecsv = csv.reader(csvfile)

for row in linecsv:
    if (row[1] == "SP") and (row[2] == "Campinas"):
        print(row[7],row[10],1e6*int(row[10])/int(row[9]))

csvfile.close()
```

---

Quadro 1: Exemplo de manipulação de arquivos CSV em Python. Arquivo `csv-example.py` disponível em [3].

correspondentes ao estado de SP (segunda coluna do arquivo) e cidade de Campinas (terceira coluna), e imprime-se a data (coluna 8), o número de casos (coluna 11) e a razão casos por milhão de habitantes. A coluna 10 contém o número de habitantes da cidade. Como todas estas colunas tem formato de string, é necessário convertê-las para valores numéricos, inteiros neste caso, para poder fazer manipulações matemáticas.

No Quadro 2, temos a implementação do nosso filtro de média móvel. A suavização utilizada corresponde a 4 iterações desse filtro. Conforme descrito na Seção 3, a implementação deste tipo de filtro requer a extensão dos dados iniciais em  $n$  elementos, tanto à “esquerda” como à “direita”. No Quadro 2, mostra-se a implementação da extensão discutida na Seção 3. Note que este filtro requer  $k \geq n$ . Testou-se também o filtro de Savitzky-Golay, disponível como `savgol_filter` no pacote `scipy.signal`, mas os resultados foram inferiores.

O último aspecto que merece ser discutido sobre o código são os gráficos interativos. Eles são implementados em JavaScript, utilizando bibliotecas públicas Google. Os arquivos html contendo o scrip são gerados na execução do código em Python. O Quadro 3 tem um exemplo simples do JavaScript

---

```
import numpy as np

def smooth(Y,n):
    k = Y.size
    Y_smooth = np.zeros(k)

    Y_edge = np.concatenate((Y[0]*np.ones(n),Y,Y[k-n:k]+Y[k-1]-Y[k-n-1]))

    for i in range (0,k):
        Y_smooth[i] = np.sum(Y_edge[i:i+2*n+1])/(2*n+1)

    return Y_smooth
```

---

Quadro 2: Filtro de média móvel com a extensão dos dados iniciais, de acordo com o discutido na Seção 3.

necessário para se criar um gráfico interativo como os do projeto.

---

```

<!DOCTYPE HTML>
<html> <head>
<script type="text/javascript"
    src="https://www.gstatic.com/charts/loader.js"></script>
<script>
google.charts.load('current', {packages: ['corechart', 'line']} );
google.charts.setOnLoadCallback(drawChart);

function drawChart() {
    var data = new google.visualization.DataTable();
    data.addColumn('number', 'x');
    data.addColumn('number', 'Curva A');
    data.addColumn('number', 'Curva B');
    data.addRows([
        [0.0,0.0,0.0],[1.0,1.0,2.0],[2.0,2.0,3.0],[3.0,2.0,2.0],]);

    var options = {title: 'Duas curvas', width: 900, height: 700,
series: {0: { lineWidth: 1 }, 1: { lineWidth: 1 }},
    hAxis: {title: 'x'}, vAxis: {title: 'f(x)'} };

    var chart = new
        google.visualization.LineChart(document.getElementById('chart_div'));
    google.visualization.events.addListener(chart, 'select', function()
        { highlightLine(chart,data, options); });
    chart.draw(data, options);
}

function highlightLine(chart,data,options) {
    var selectedLineWidth = 5;
    var selectedItem = chart.getSelection()[0];
    for(var i in options.series) {options.series[i].lineWidth = 1;}
    options.series[selectedItem.column-1].lineWidth = selectedLineWidth;
    chart.draw(data, options); }

</script></head>
<div id="chart_div"></div>

```

---

Quadro 3: Exemplo de JavaScript utilizado para os gráficos interativos. Arquivo graf.html disponível em [3].

## Referências

- [1] <http://vigo.ime.unicamp.br/COVID>
- [2] <https://covid19br.github.io/>
- [3] <https://github.com/albertosaa/COVID>
- [4] <https://covid.saude.gov.br/>
- [5] [https://en.wikipedia.org/wiki/Compartmental\\_models\\_in\\_epidemiology](https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology)
- [6] C. Gu, W. Jiang, T. Zhao, Tianyuan and B. Zheng, *Mathematical Recommendations to Fight Against COVID-19* (March 9, 2020). Available at SSRN: <http://dx.doi.org/10.2139/ssrn.3551006>
- [7] S.B. Bastos and D.O. Cajueiro, *Modeling and forecasting the early evolution of the Covid-19 pandemic in Brazil*, <https://arxiv.org/abs/2003.14288>
- [8] <https://www.ecdc.europa.eu/en/covid-19/questions-answers>
- [9] <https://ciis.fmrp.usp.br/covid19/>
- [10] <https://elpais.com/sociedad/2020-04-07/mas-del-90-de-contagios-estan-ocultos.html>
- [11] <https://anaconda.org/>

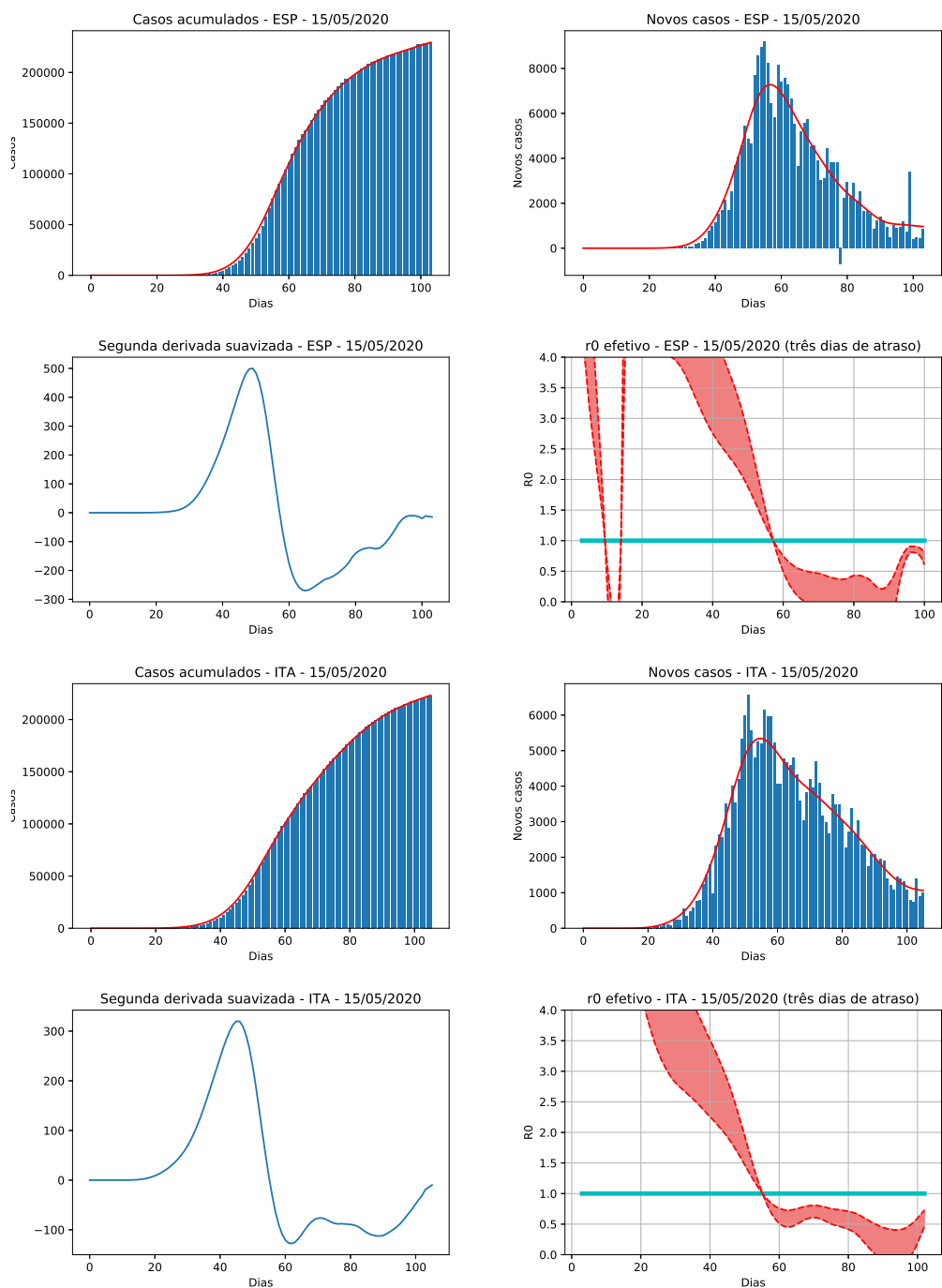


Figura 3: Análises para Espanha e Itália.



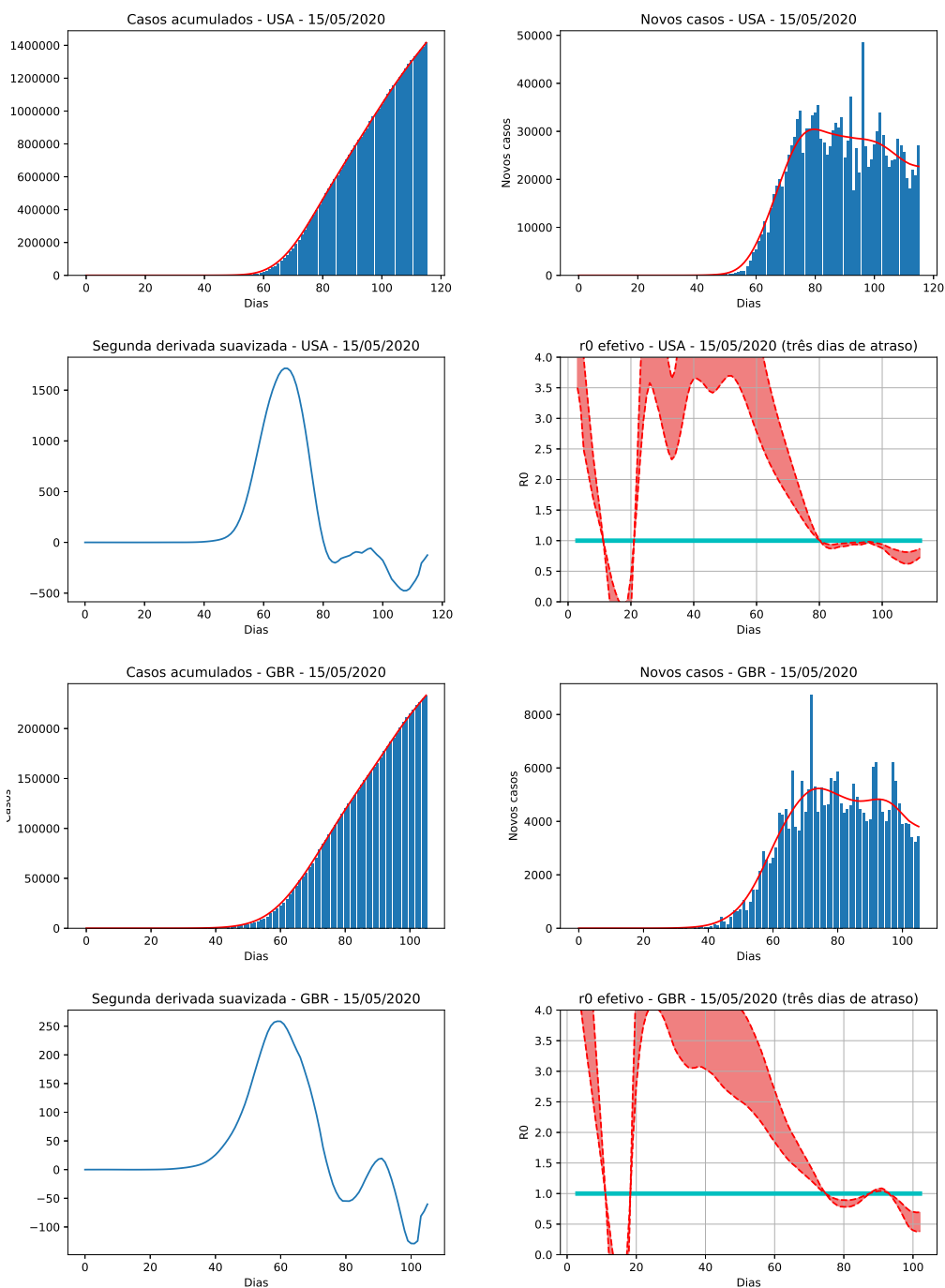


Figura 4: Análises para Estados Unidos e Reino Unido.