UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E COMPUTAÇÃO CIENTÍFICA
DEPARTAMENTO DE MATEMÁTICA APLICADA

JEFFERSON MONÇÃO DA SILVA

Análise automática dos casos de COVID-19: Análise estatística

Campinas
11/01/2021

JEFFERSON MONÇÃO DA SILVA

Análise automática dos casos de COVID-19: Análise estatística

Monografia apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos para obtenção de créditos na disciplina Projeto Supervisionado, sob a orientação do(a) Prof. Alberto Saa.

Resumo

Nesse texto será explicado o tratamento estatístico usado para fazer a analise da pandemia do COVID-19 neste projeto. Será discutido as definições e métodos matemáticos usados para esse projeto acontecer. Serão abordados apenas os detalhes matemáticos e estatísticos relevantes para o projeto.

Abstract

In this text, the statistical treatment used to analyze the pandemic of COVID-19 in this project will be explained. The definitions and mathematical methods used for this project will be discussed.. Only the mathematical and statistical details relevant to the project will be covered.

Conteúdo

1 Introdução

O objetivo do projeto é a elaboração de um sistema de código em Python para análise automática da COVID-19 por meio do modelo epidemiológico mais simples, o modelo SIR. Os dados da pandemia da COVID-19 são publicados diariamente pelo Ministério da Saúde no site [3]. Para mais informações do projeto ide a referência [5].

A realização desse projeto necessitou de diversas ferramentas matemáticas que serão explicadas nas secções seguintes, e esse ferramental foi necessário para entender os dados da pandemia da COVID-19. Esse trabalho pretende de esclarecer a matemática usada no projeto e as escolhas de métodos para tratamento de dado e análise dos resultados.

2 Tratamento de Dados

As funções R,I e S do modelo SIR, a definição dessas funções esta na parte Modelo SIR do projeto referência [8], são funções contínuas no tempo t, que é por si é uma variável contínua. Os dados divulgados pelo Ministério da Saúde são séries temporais de casos acumulados e casos novos, que são de pessoas doentes que já recorreram ao hospital ou estão internadas, então devem ser parte a função $R(t)$ do modelo SIR. Aqui será abordado o tipo de dado que será usado para a analise e as principais definições prévias para o cálculo e análise dos dados.

2.1 Caracterização dos dados

Os dados numéricos usados para análise automática foram extraídos do site do ministério da saúde (MS) através de um código escrito em Python e com a biblioteca Requests dessa linguagem de programação. Os dados assim obtidos são na forma de uma sequência ou série temporal com distância temporal de 1 dia entre um valor e outro, assim a unidade da variável t será de 1 dia. O modelo SIR usa como variável independente o tempo (t), uma variável contínua. Afim de analisar a pandemia da COVID-19 através da série temporal obtida de MS é necessário discretizar a variável,i.e., $t \in \mathbb{Z}$. Dessa forma, a derivada temporal deve ser calculada como uma diferença de valores da série. E optou-se por diferenças atrasadas:

$$\frac{\partial S(t)}{\partial t} = \dot{S}_t = S_t - S_{t-1} \tag{1}$$

De modo análogo podemos calcular a segunda derivada da série:

$$\frac{\partial^2 S(t)}{\partial t^2} = \ddot{S}_t = \dot{S}_t - \dot{S}_{t-1} \tag{2}$$

Isso é feito desse modo pois devido a discretização da variável t têm-se que a menor variação possível para o calculo das diferenciais é $dt = 1$, então usando a forma da diferencial de qualquer função diferenciável na variável discreta t tem:

$$dS(t) = S_t - S_{t-1} = \frac{\partial S(t)}{\partial t} dt = \dot{S}_t \tag{3}$$

Os dados obtidos são de casos (mortes) acumulados(as) e novos(as). Dado que o modelo SIR usa as variáveis S (suscetíveis), I (infectados) e R (Removidos), então devemos fazer a analise dos casos acumulados e usar o parâmetro α para relacionar os casos acumulados com os casos removidos, $R = \alpha C$, na análise e calculo do R_0. Para entender melhor sobre parâmetro α vide a parte do Modelo SIR desse projeto, referência [8].

2.2 A derivada como diferença finita

A forma com a qual definimos a derivada pode ser interpretado como truncar a expansão de Taylor em primeira ordem com $h = 1$

$$S(t) = S(t-1) + \dot{S}(t) + \vartheta(h^2) \tag{4}$$

Como é visto na página 12 da referência [1], o erro cometido por truncar a série de Taylor em primeira ordem é da ordem de:

$$\vartheta(h) = h\frac{\ddot{S}(\xi)}{2} = \frac{\ddot{S}(\xi)}{2}, \xi \in \{t-1, t\} \tag{5}$$

Então a aproximação da derivada como diferença finita é confiável pois a ordem dos valores da segunda derivada dos casos acumulados é muito menor que a ordem dos casos acumulados, pode ser visto um exemplo disso na imagem seguinte e comparado aos dados da imagem 2.

Figura 1: Gráfico da segunda derivada dos casos acumulados do estado de São Paulo, a imagem foi extraída da referência [5].

Afim de fazer uma aproximação numérica do erro note como no dia 60 aparece

um extremo de aproximadamente 90 casos da segunda derivada dos casos acumulados. Então o erro cometido pela derivada como diferença finita é menor ou igual a $\vartheta = 90/2 = 45$. No mesmo dia os casos acumulados atingem aproximadamente 20.000 implicando um erro percentual $\eta = 45/20000 \approxeq 0,002$.

3 Suavização

Os dados brutos possuem uma flutuação, i.e., a quantidade de casos novos divulgados pelo MS oscila entre valores altos e baixos. Essa flutuação é devida a diversos motivos que não são do escopo desse projeto. Para que seja feita a análise do real crescimento de casos essa flutuação deve ser eliminada por meio de uma suavização.

3.1 O produto de convolução

A convolução é um operador funcional do espaço de funções e é definido por:

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x-t)dt \tag{6}$$

Afim de trabalhar com funções de domínio discreto é preciso usar a definição de produto convolução discreto:

$$(f * g)(k) = \sum_{i=0}^{k} f(i)g(k-i) \tag{7}$$

Considere que f, g tenham $2n+1$ pontos. E o ponto central seja um dado t fixo e seja $g(k-i) = \frac{1}{2n+1}$ $\forall$ k-i $\in$ {t-n,t+n}

$$(f * g)(k) = \frac{1}{2n+1} \sum_{i=t-n}^{t+n} f(i) \tag{8}$$

Isso é a média móvel e será usada nesse projeto na suavização de ruídos dos casos acumulados, como exemplo será usado os dados do estado de São Paulo no até o dia 21/05/2020. Segue um exemplo de existência do ruído:

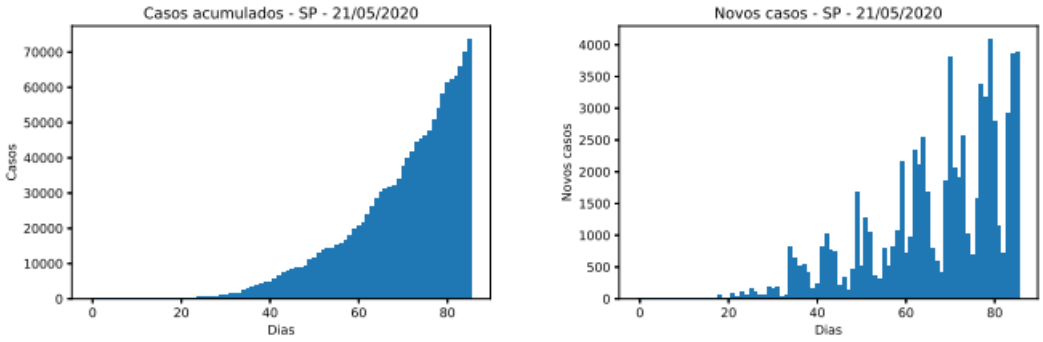

Figura 2: Extraída da referência [5]. Mostra os casos acumulados e novos casos para o estado de São Paulo a partir dos dados publicados pelo Ministério da Saúde em [3]. A evidente existência de "ruído" dificulta à análise da variações direta

3.2 Filtro da média móvel

O ruído mostrado na figura 2, dificulta a análise a partir de diferenças finitas. Parte do ruído se deve a causas que não são do interesse da analise do projeto, como a queda de dados divulgados pelos municípios durante os fins de semana ou outros motivos. Então se faz necessário excluir os ruídos de maneira a manter as tendencias de crescimento e a média semanal de casos. Por isso é feita a média móvel com janela de uma semana ($n = 3$). O objetivo de fazer essa suavização é o de obter, pelo menos a segunda derivada dos casos acumulados (C) minimamente suave para o calculo do R_0. Uma explicação de como isso é feito esse calculo e a definição de R_0 esta em [8] A ideia é fazer a média móvel para suavização e substituir o valor de cada dia do gráfico pelo valor médio de uma janela de uma semana. "Janela"se refere a numero de elementos do intervalo de onde será calculado as médias.Usando a seguinte equação:

$$\overline{C}_t = \frac{1}{2n+1} \sum_{k=t-n}^{t+n} C_k \tag{9}$$

Para a janela de uma semana devemos colocar $n = 3$ nessa ultima equação. Segue exemplo de uma suavização feita pelo código:

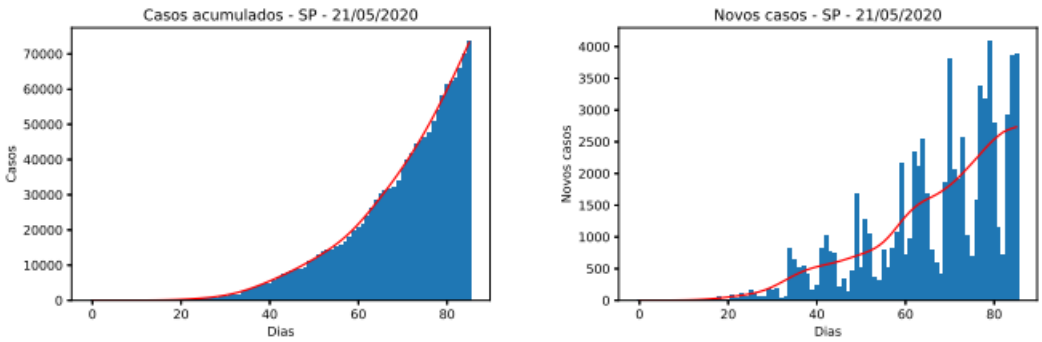

Figura 3: Extraída da referência [5]. Mostra dos mesmos dados da figura 2 com a suavização (n=3) da média móvel representada pela linha vermelha.

Repare na figura 3 que a suavização é representada por uma função contínua que liga os pontos da suavizados. Se essa função ligasse os pontos não-suavizados então possuiria muitas flutuações. Entenda também que a suavização mantém as taxas de crescimento e decrescimento do gráfico. Note que a média móvel usa para, o calculo de cada ponto fixado, os dados n dados anteriores e n dados futuros. Então, o calculo direto da média móvel nas bordas não é possível. Invés disso é acrescentado os $2n$ pontos necessários para se fazer o calculo da média móvel nas bordas, n antes de $t = 0$ e n após o último dia k. Primeiramente acrescentaremos a borda inferior, para $t < 0$. Como $t = 0$ é o inicio da pandemia, é imediado que $C_t = 0, \forall t < 0$, i.e., não deve existir casos antes do primeiro caso (considere que o primeiro caso define $t = 0$). Acrescentar a borda superior, i.e., os casos futuros do último k analisado, é fazer uma previsão de crescimento futuro da pandemia. Então, para $n = 3$ (janela de uma semana na média móvel) é esperado que a taxa de crescimento dos últimos casos acumulados seja mantida para a borda. Tome $k \equiv$ último dia coletado $\therefore$

$$C_{k+j} - C_{k-n+j} = C_k - C_{k-n} \tag{10}$$

Isso significa que para cada $1 \leq j \leq n$, a variação de $C_{k+j} - C_{k-n+j}$, que são os dados após k, é a mesmas que $C_k - C_{k-n}$, dados antes de k. Então o crescimento de C_t antes de k nos indica os valores que devem ser acrescentados para $t > k$.

$$C_{k+j} = C_k + C_{k-n+j} - C_{k-n} \tag{11}$$

Afim de intensificar o efeito da suavização para que seja mais efetivo é aplicada a função de suavizar sucessivamente. É feito 4 interações da suavização, então para cada ponto a equação que calcula o valor ajustado pela suavização é o resultado de aplicar a convolução da equação 8 diversas vezes. Isso resulta em uma janela de aproximadamente 4 semanas ao invés de 1 semana, pois a cada interação da suavização tem que cada ponto da janela anterior necessita de um intervalo simétrico de $2n + 1$ pontos para o cálculo da média móvel. Então o a cada interação a janela da convolução cresce 2n pontos, n em cada borda. Após 4 interações na janela passa a ter $6n + (2n + 1) = 8n + 1$ pontos, i.e., com $n = 3$ a janela passa a ser de 25 dias. Ao todo a série necessita de $6n = 18$ acréscimos nas bordas para que seja feita a suavização, 9 dias em cada extremidade. E a forma que esses acréscimos são feitos obedece as mesma intuição da primeira interação do filtro de média móvel. I.e., os 9 dias acrescentados na borda inferior são identicamente 0. Já os 9 dias acrescentados na borda superior obedecem a equação 10. Como pode ser visto nas referências [5] e [7], a recursão da suavização pode ser vista como fazer a convolução discreta com outras funções especiais. A média móvel pode ser vista da seguinte maneira recursiva:

$$y_1[i] = y_1[i-1] + \frac{1}{2n+1}(x[i+p] - x[i-q]) \tag{12}$$

onde o subíndice em y_1 representa o número de interações da suavização, no caso 1 interação e $p = (M-1)/2 = n$ e $q = p + 1$. Essa equação pode ser verificada na referência [7]. Então para j interações a relação de recursão se torna:

$$y_j[i] = y_j[i-1] + \frac{1}{2n+1}(y_{j-1}[i+p] - y_{j-1}[i-q]) \tag{13}$$

Como é dito na referência [5], 2 interações da suavização pode ser interpretado como a convolução por uma função triangular, enquanto a 4 interações da suavização pode ser interpretado como a convolução por uma função cúbica. Essa discussão não tem real importância para elaboração do projeto, as suavizações são feitas por iterar a função que realiza a suavização então não tem necessidade do calculo direto da convolução para as iterações, essa parte do algoritmo de computação esta explicada em [2]. Mas vale lembrar que a suavização altera os dados e muda a quantidades deles, principalmente próximo das bordas os dados passam a não ser confiáveis. Então não é interessante fazer muitas

aplicações da suavização, por isso a quantidade de iterações que é aplicada na função $R_0(t)$ é diferente dependendo da forma que ele é determinado.

4 Ruídos do R_0 e as formas de calcula-lo

O R_0 é o parâmetro que indica o estágio da pandemia da COVID-19 e a velocidade de propagação do vírus. Por isso para fazer a análise matemática da pandemia é necessário determinar o R_0 a partir dos dados coletados. R_0 não será considerado constante para os fins desse projeto, o que é chamado de R_0 efetivo. Contudo os parâmetro α e γ são considerados constantes. Para entender mais sobre a definição dessas quantidades vide a parte "Modelo SIR"desse projeto. Existem no mínimo duas formas de determinar o R_0 como pode ser visto na referência [5] as equações que resultam no R_0 são:

$$R_0(t) = \frac{1}{1-\mu} + \frac{\ddot{C}}{\gamma \dot{C}(1-\mu)} \tag{14}$$

onde

$$\mu(t) = \frac{\alpha}{\gamma N}(\gamma C + \dot{C}) \tag{15}$$

Onde $N = I + S + R$ é a população total. O R_0 também pode ser determinado pelo seguinte método que leva em conta o teorema do valor médio para integrais no intervalo $[t_0, t]$:

$$ln\frac{S_0}{S(t)} = \bar{r}_0 \frac{R_0(t) - R_0}{N} \tag{16}$$

onde R_0 nessa última equação representa $R(t_0)$ e $\bar{r}_0$ é o valor médio de $R_0(t)$ em $[t_0, t]$. A dedução desses dois métodos es na referência [5]. Notando $R_0(t)$ para o resultado da equação 14 e $\tilde{R}_0(t)$ para o resultado da equação 16 para comparar o ruído em $R_0(t)$ com em $\tilde{R}_0(t)$. Primeiro note que como $R_0(t)$ pode ser entendido como um calculo instantâneo do parâmetro, enquanto $\tilde{R}_0(t)$ é calculado como um valor médio. Então o esperado é que $\tilde{R}_0(t)$ seja mais suave que $R_0(t)$ e portanto necessite de menos iterações da suavização.

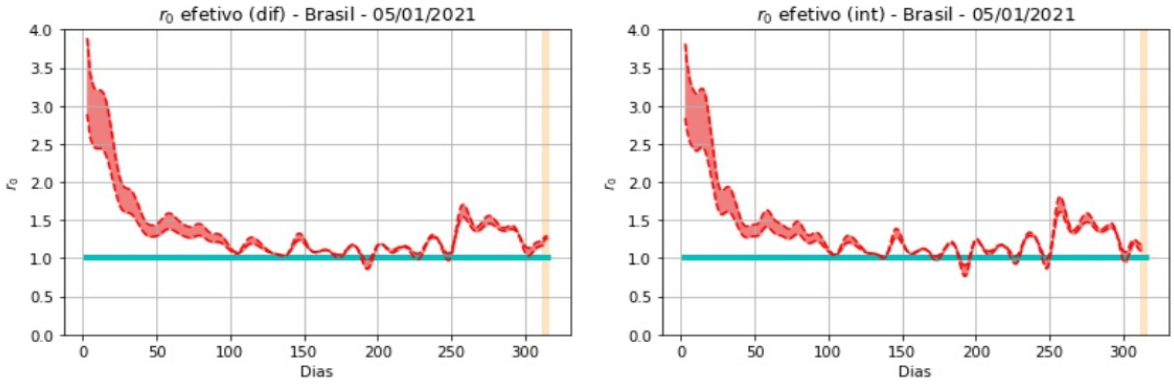

Figura 4: Extraída da referência [6]. Compara $R_0(t)$ e $\tilde{R}_0(t)$ do Brasil calculados pelo algoritmo do projeto no dia 05/01/2021.

$\tilde{R}_0(t)$ foi suavizado com uma iteração de 2 vezes da função de suavizar, enquanto $R_0(t)$ foi suavizado com uma iteração de 4 vezes. Olhando a figura 4 note que não existe grandes diferenças nas flutuações de $\tilde{R}_0(t)$ e $R_0(t)$, então como o esperado $\tilde{R}_0(t)$ necessita de menos iterações da suavizações para excluir os ruídos.

5 Conclusão

Nesse trabalho foi apresentado os principais métodos matemáticos-estatísticos usados no projeto 'Análise automática da COVID-19'. Foi por meio do recurso ferramental apresentado nas secções anteriores que foi possível gerar a análise automática que esta disponível no site da referência [6]. As outras partes do projeto estão explicadas nas referências [8] e [2], correspondente a uma explanação sobre a modelagem matemática do modelo SIR e sobre o algoritmo computacional usado para a analise e na geração do arquivo HTML onde é divulgado a análise.

Os códigos e dados coletados estão disponíveis na pagina do GITHUB da referência [4]

Referências

[1] Wagner H. Bonat. PDF sobre derivação numérica disponibilizado pelo Departamento de Estatística da Universidade Federal do Paraná no endereço http://www.leg.ufpr.br/lib/exe/fetch.php/disciplinas:numericalderint.pdf.

[2] Davi Alves Monteiro Carvalho. Análise automática dos casos de covid-19: Arquivo html, 2020.

[3] Ministério da Saúde, 2020. Site do Ministério da saúde onde é divulgado os dados da COVID-19, endereço: https://covid.saude.gov.br/.

[4] A. Saa, 2020. Códigos disponíveis abertamente na pagina do github https://github.com/albertosaa/COVID.

[5] A. Saa. Análise automática do painel coronavírus, 2020. Texto integralmente disponível em https://vigo.ime.unicamp.br/COVID/covid.pdf.

[6] D. Carvalho; J. Monção; S. Zani; A. Saa, 2020. Painel Coronavirus, endereço eletrônico: http://vigo.ime.unicamp.br/COVID/.

[7] Luciano Scandelari. PDF Filtros Digitais da disciplina Laboratorio de PDS da Universidade Tecnológica Federal do Paraná.

[8] Sabrina Camargo Zani. Análise automática dos casos de covid-19: Modelo sir, 2020.