

# Estimativa da taxa de subnotificação de casos de COVID-19 na cidade do Rio de Janeiro no início da pandemia

Lucas Machado Moschen<sup>1</sup>

EMAp/FGV, Rio de Janeiro, RJ

Orientadora: Maria Soledad Aronna<sup>2</sup>

EMAp/FGV, Rio de Janeiro, RJ

**Resumo.** A doença COVID-19 causada pelo vírus SARS-CoV-2 espalhou-se rapidamente pelo mundo desde o início de 2020 e entender sua dinâmica na população é importante para tomar medidas que contenham a disseminação. Nesse relatório, o modelo epidemiológico SEIAQR para a COVID-19 é considerado para compreender o início da epidemia no município do Rio de Janeiro e, em especial, a taxa de subnotificação, isto é, a proporção de indivíduos infectados que não foram registrados pelo sistema [2]. As curvas de casos confirmados e óbitos foram ajustadas aos dados reais da cidade usando o método de mínimos quadrados ponderados dos erros. A transmissibilidade e a mortalidade da doença são aproximadas por B-splines cujos parâmetros também foram estimados. Foi analisada a identificabilidade estrutural do modelo e a identificabilidade prática do ajuste para verificar a viabilidade das estimativas. Utilizamos o método Bootstrap para quantificar a incerteza sobre as estimativas. Para o período março-julho de 2020, obtemos a estimativa pontual 0.9 para a subnotificação com intervalo de confiança 95% (0.85, 0.93).

**Palavras-chave.** Modelo COVID-19, estimativa de parâmetros, identificabilidade, mínimos quadrados, B-splines, Bootstrap.

## 1 Introdução

No final de dezembro de 2019, na cidade de Wuhan (Hubei, China), foram identificados diversos casos de pneumonia [38] causados por um novo coronavírus. Ele foi chamado de SARS-CoV-2 e a doença por ele provocada COVID-19, que se espalhou rapidamente pelo mundo. A maioria dos casos resultam em quadros assintomáticos ou com sintomas leves, mas em casos mais severos, falta de ar e dor no peito são mais frequentes [40] e podem levar à morte.

O Brasil declarou a doença como emergência de saúde pública em fevereiro de 2020 com o objetivo de proteger a população e aplicar medidas como isolamento social, quarentena e testagem [9]. O primeiro caso foi registrado em São Paulo no final daquele mês. No mês seguinte (março), o governador do estado do Rio de Janeiro declarou emergência na saúde pública e fechamento de ambientes que favorecem aglomerações, tais como universidades, escolas e teatros. Em seguida, diversas medidas foram tomadas como o controle de restaurantes, de praias, de shopping centers e do comércio não essencial [10]. Atualmente, em abril de 2021, o Brasil é um dos países com maior crescimento da epidemia, com mais de treze milhões de casos e 350 mil mortes.

Em [2], introduzimos um modelo compartmental tipo SEIAQR, que leva em conta isolamento, quarentena de casos confirmados e casos assintomáticos (mais detalhes na Seção 2). O presente

---

<sup>1</sup>lucas.moschen@fgv.edu.br

<sup>2</sup>soledad.aronna@fgv.edu.br

trabalho procura utilizar o modelo mencionado para estimar a taxa de casos não reportados na cidade do Rio de Janeiro, dado que, como já expusemos anteriormente, a maior parte das infecções não resultam em sintomas graves e, portanto, existe uma dificuldade em entender a real extensão da evolução na população. Os valores reportados aqui não procuram ditar a verdadeira quantidade, mas sim, utilizar as ferramentas matemáticas para compreender a dinâmica, o que se torna mais complicado, uma vez que essa é uma doença recente e com informação sendo adquirida ao longo do processo.

O texto se organiza na seguinte forma: na Seção 2 é apresentada uma simplificação do modelo que será utilizada na estimação; na Seção 3 são apresentados os dados utilizados para o problema da cidade do Rio de Janeiro; na Seção 4 são estimados os parâmetros e é utilizado um método para quantificar a incerteza. Por fim, na última seção apresentamos a conclusão com algumas discussões e possíveis melhorias nos procedimentos adotados. Os códigos para os algoritmos e experimentações podem ser encontrados no Github [24].

## 2 Apresentação do modelo

Definimos um modelo compartmental [2] a fim de descrever o espalhamento do vírus SARS-CoV-2 que comporta medidas não farmacêuticas adotadas como forma de combate ao surto. Repartimos a população nos compartimentos  $S$ ,  $E$ ,  $I$ ,  $A$ ,  $Q$ ,  $R$  e  $D$  de forma que um indivíduo suscetível ao vírus inicia em  $S$  e após entrar em contato com um infectado, passa para o compartimento  $E$ , onde apesar de infectado, não infecta outros por um período latente. Após esse tempo, o indivíduo se torna infectado e vai para o compartimento  $I$ , tal que, com um certo tempo, ele pode ser reportado, e ir para o compartimento  $Q$ , ou pode não ser reportado e ser encaminhado para  $A$ . Por fim, esses indivíduos se recuperam da doença no compartimento  $R$  e aqueles casos que estão em quarentena podem ir para o compartimento  $D$  se morrerem.

Consideramos a população normalizada, portanto cada compartimento representa a proporção correspondente. Também removemos as taxas de nascimento e mortes naturais devido ao espaço curto de tempo que pretendemos modelar. A dinâmica é descrita da seguinte forma:

$$\begin{aligned}\dot{E} &= \beta(t)S(I + A) - \rho(t)\delta E - \tau E \\ \dot{I} &= \tau E - \sigma I - \rho(t)I \\ \dot{A} &= \sigma\alpha I - \rho(t)A - \gamma_1 A \\ \dot{Q} &= \sigma(1 - \alpha)I + \rho(t)(\delta E + I + A) - \gamma_2 Q - \mu Q \\ \dot{S} &= -\beta(t)S(I + A) \\ \dot{R} &= \gamma_1 A + \gamma_2 Q \\ \dot{D} &= \mu Q\end{aligned}\tag{1}$$

Os parâmetros relacionados com o patogênico e com a doença induzida se encontram na Tabela 1. Assumimos que entre todos os infectados, uma proporção  $\alpha \in (0, 1)$  representa os casos não reportados que são, em geral, assintomáticos ou com sintomas leves. Um indivíduo assintomático infecta tanto quanto outro sintomático e, se ele não for detectado, pode prolongar a duração do surto. Essa é uma simplificação razoável, dado que estimar a contribuição dessa parcela é complicado, segundo [25]. No modelo,  $\beta(t)$  é a taxa de contato efetiva da doença no tempo  $t$  e leva em conta a taxa média de contatos - diretamente afetada por medidas como isolamento social, proteção pessoal (uso de máscaras e higiene) e a cultura da região - e a transmissibilidade da doença - probabilidade de infecção dado um contato entre indivíduo infectado e outro suscetível.

Outra medida muito importante no combate ao espalhamento do vírus é a detecção através da testagem e posterior quarentena dos casos positivos. Em nosso modelo,  $\rho(t)$  é a taxa de testagem

de pessoas assintomáticas ou com sintomas leves no tempo  $t$ . Assumimos que uma pessoa nos compartimentos  $I$  ou  $A$  sempre testam positivo, em  $S$  sempre negativo e em  $E$  positivo com uma probabilidade  $\delta$ . Os falsos negativos são desconsiderados do modelo, apesar de prejudicarem a estimativa de  $\rho$ . Essa consideração foi feita como forma de simplificação, mas pode ser alterada considerando a incerteza sobre a testagem (ver [2, Remark 2.3] para mais detalhes).

Par.	Descrição
$\tau^{-1}$	tempo latente da exposição ao início da infeciosidade.
$\sigma^{-1}$	tempo entre o início da infeciosidade e o possível início dos sintomas
$\omega^{-1}$	tempo de incubação (i.e. $\omega^{-1} = \tau^{-1} + \sigma^{-1}$ )
$\gamma_1$	taxa de recuperação de casos menos graves
$\gamma_2$	taxa de recuperação de casos mais graves
$\mu$	taxa de mortalidade entre os casos confirmados

Tabela 1: Parâmetros COVID-19

Adicionamos um contador de testes positivos  $T(t)$ , através da equação

$$\dot{T} = \sigma(1 - \alpha)I + \rho(t)(\delta E + I + A) \quad (2)$$

Como a curva  $T$  é a que temos acesso nos dados, além da curva de mortes  $D$ , usaremos ela como referência para a estimativa dos parâmetros. Também assumiremos que a política de testagem na cidade é constante ao longo do tempo, e, portanto  $\rho(t) \equiv \rho$ , como justificado na Observação 4.1. Outra aplicação possível do modelo, explicitada no artigo de referência [2], é verificar o que a variação desse parâmetro pode causar na epidemia.

## 2.1 O número reprodutivo básico

Como discutido em [2], considerando os coeficientes constantes, podemos determinar o número básico reprodutivo  $\mathcal{R}_0$  associado ao modelo.

$$\mathcal{R}_0 = \frac{1}{2} \left( \varphi + \sqrt{\varphi^2 + \frac{4\sigma\alpha}{\rho + \gamma_1}\varphi} \right), \quad (3)$$

com

$$\varphi = \frac{\beta\tau}{(\rho\delta + \tau)(\sigma + \rho)}. \quad (4)$$

Com a evolução da epidemia, a porção da população recuperada e imune à doença torna-se relevante, o que diminui o número reprodutivo. Assim definimos o número reprodutivo dependente do tempo  $\mathcal{R}(t)$  que decresce conforme a população suscetível ( $S(t)$ ) decresce. Em particular, a expressão de  $\varphi$  in (4) é alterada para

$$\varphi(t) = \frac{\beta(t)\tau S(t)}{(\rho\delta + \tau)(\sigma + \rho)}. \quad (5)$$

## 3 Dados

Utilizamos os dados de casos confirmados de COVID-19 e de mortes causadas pela doença do município do Rio de Janeiro que podem ser encontrados no site da prefeitura do Rio de Janeiro [36] e servem de base para o Painel Rio COVID-19 [37], com atualização diária. Na captura de dados,

	dt_notific	dt_inicio_sintomas	bairro_resid_estadia	ap_residencia_estadia	sexo	faixa_etaria	evolucao	dt_evolucao	raca_cor
0	09/18/20	09/03/20	PACIENCIA	5.3	M	De 50 a 59	OBITO	09/17/20	Preta
1	11/25/20	11/02/20	BARRA DA TIJUCA	4.0	M	De 80 a 89	OBITO	12/01/20	Branca
2	05/06/20	05/06/20	CACHAMBI	3.2	M	De 70 a 79	OBITO	05/07/20	Ignorado
3	11/12/20	11/02/20	BARRA DA TIJUCA	4.0	M	De 70 a 79	OBITO	12/12/20	Branca
4	06/13/20	04/26/20	MARECHAL HERMES	3.3	M	De 60 a 69	OBITO	05/16/20	Ignorado

Tabela 2: Cinco casos confirmados e os dados individuais: data de notificação, data de início dos sintomas, bairro de residência, área de planejamento em saúde, sexo, faixa etária, evolução, data de evolução e raça.

podemos encontrar uma série de dificuldades relacionadas ao rastreio de casos, tais como a baixa testagem na cidade, o atraso na notificação e os ciclos semanais causados pela forma como são feitos os registros. Esses fatores dificultam o processo de aprendizado sobre os parâmetros.

A partir da obtenção e tratamento desses dados, podemos obter a curva do número de indivíduos infectados confirmados que declaram início dos sintomas no tempo  $t$ , representado pela curva  $T$  no nosso modelo, e também o dia de morte, o que permite recuperar a curva  $D$ . Está claro que pessoas podem informar a data de início dos sintomas com uma imprecisão de alguns dias, mas assumimos que esses erros são independentes, como explicado na Seção 4. A análise de dados mais detalhada pode ser conferida no repositório do Github [24].

### 3.1 Análise de dados

Após a obtenção dos dados em formato *CSV*, observamos que os dados como na Tabela 2. A data de início de sintomas indica o dia que o paciente reportou ter começado a sentir os sintomas e a data de evolução marca o dia em que a pessoa se recupera ou falece e deixa de ser um caso ativo. Essa coluna é a que mais possui campos vazios (em torno de 6%), em que apenas dois eram indivíduos com óbito confirmado e os outros se dividiam entre recuperados e ativos. As outras colunas são auto-explicativas e não terão atenção nesse trabalho.

Podemos visualizar a quantidade de novas notificações diariamente na Figura 1. Uma das características da curva em azul é a existência de uma grande variabilidade semanal, como afirmado no início da seção e, por esse motivo, como descrito na Seção 3.2, é visualizada a curva da média de 7 dias em vermelho. Na Figura 2, comparamos as curvas médias de novas notificações e de indicações de início dos sintomas. Por exemplo, observamos que o primeiro pico do início dos sintomas não é repercutido na média de notificações (apesar de sabermos que a integral em todo o percurso é a mesma). Isso mostra que existe uma distribuição de quando as pessoas começam a sentir os sintomas a partir da infecção e de quando elas são notificadas pelo sistema.

Em agosto, o pico na média de notificações não é antecipado por um pico na curva de sintomas, isto é, as pessoas que notificaram esse mês devem ter começado a sentir os sintomas em um período anterior. Isso sugere que a curva em azul (média por início dos sintomas) é melhor a ser analisada, pois não sofre pelos problemas de atraso e inconstância. No final dessa curva, existe um decréscimo que não faz sentido com o pico anterior, pois as pessoas que indicarão o início dos sintomas nesse período vão ser notificadas nas semanas seguintes.

Essa diferença de quando uma pessoa foi notificada e quando ela sentiu os primeiros sintomas pode ser visualizada em formato de histograma na Figura 3. É importante destacar que valores negativos são claramente erros de digitação ou informação nos dados, como pode ser examinado que existiam casos que foram notificados em 2021, mas com início dos sintomas em janeiro de 2020. Apesar disso, esses erros são em menor quantidade quando comparados à massa de dados. Registrhou-se que em torno de 19,34% dos dados têm tempo de notificação maior do que 30 dias e em torno de 1,29% maior de 100 dias o que indica o grande atraso nas notificações.

Uma análise com relação aos dados faltantes com comparação aos bairros, faixa etária e sexo

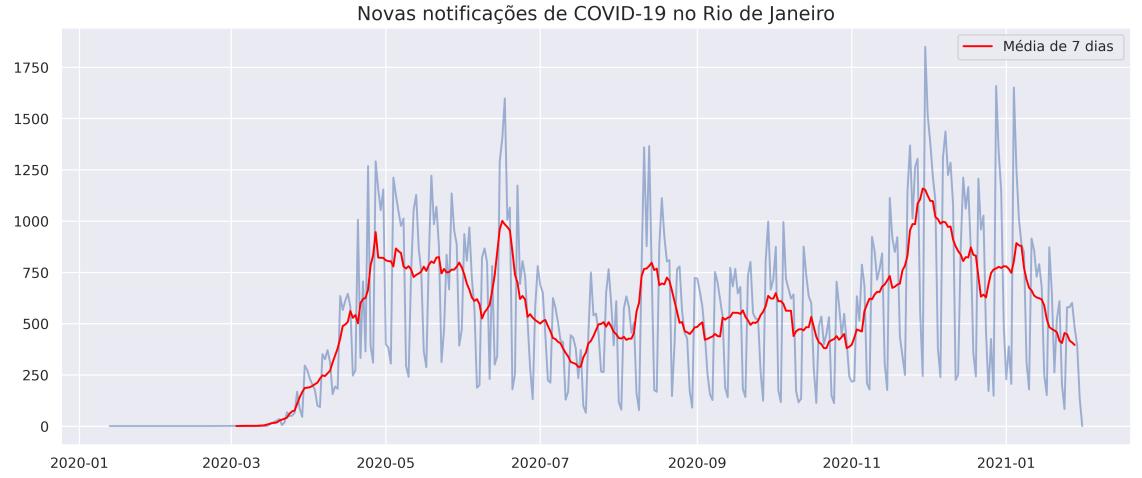


Figura 1: Notificações a cada dia ao longo da pandemia no Rio de Janeiro.

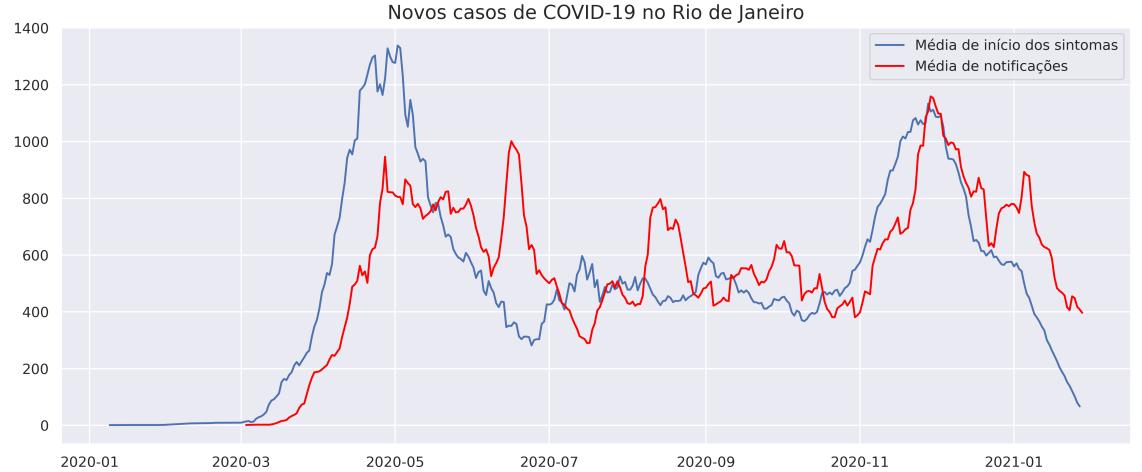


Figura 2: Comparaçāo das curvas de média de 7 dias de novas notificações e indicação de início dos sintomas.

também foi feita e pode ser vista no Github, mas apresentou pouca evidência para contribuir nesse trabalho. O mesmo tipo de exame pode ser feito para a evolução dos casos de óbitos, como pode ser visualizado na Figura 4. Muitos casos de mortes tiveram sua notificação a posteriori também, e omitimos o gráfico por razões de similaridade com o histograma anterior. Por fim, para os campos com dados faltantes em data de início de sintomas e data de evolução daqueles com óbito registrado, houve uma imputação aleatória baseada na distribuição dos histogramas correspondentes.

### 3.2 Suavização das curvas

Constatamos ao longo do surto que existe uma sazonalidade semanal nos casos confirmados e de mortes, dado que nos fins de semana há um desvio negativo, enquanto na terça-feira um positivo. Essa variação semanal prejudica o modelo, pois ele não possui ajuste sazonal e, por esse

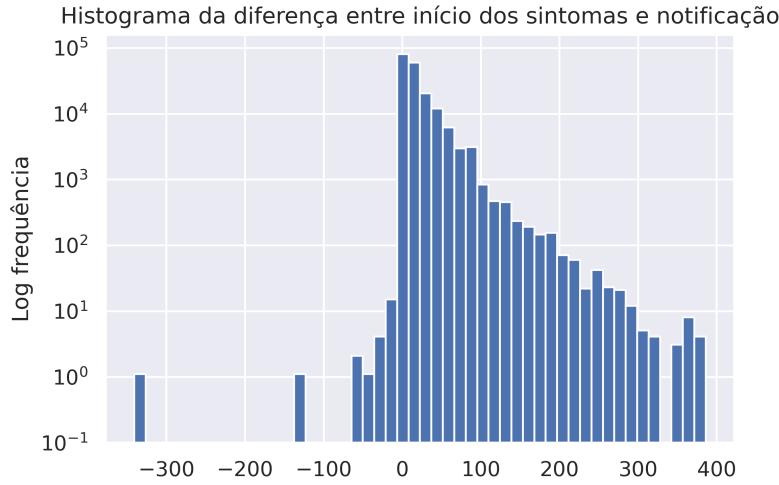


Figura 3: Histograma da diferença entre o dia de notificação e o dia de início dos sintomas.

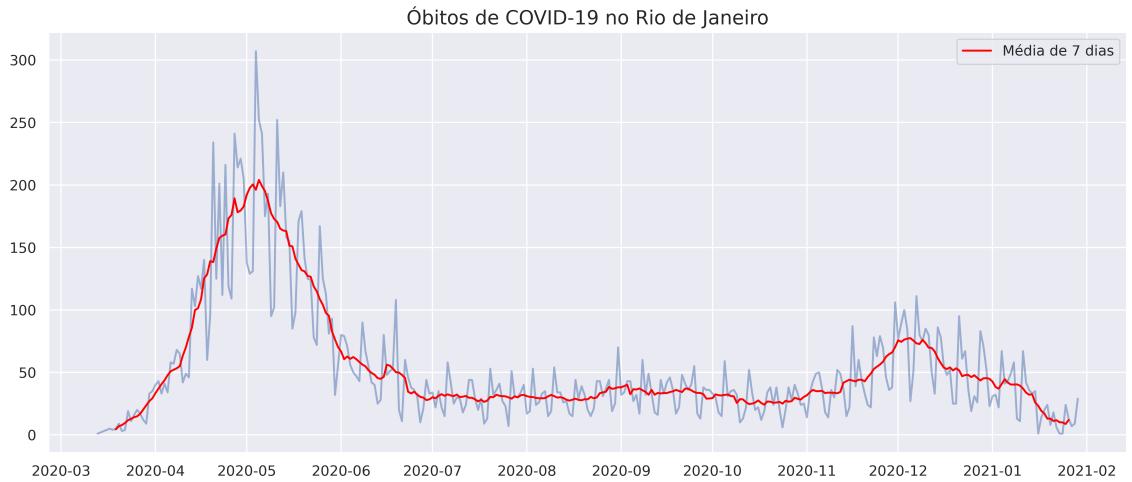


Figura 4: Evolução de óbitos no Rio de Janeiro.

razão, a aplicação de um suavizador se faz necessária. Nesse sentido, utilizaremos uma *média móvel* centrada com 7 dias. Desta forma, se  $\{x_s\}_{1 \leq s \leq n}$  são nossos dados iniciais, estamos interessados em

$$\hat{x}_t = \frac{1}{2k+1} \sum_{s=-k}^k x_{t+s}, \quad (6)$$

em que  $2k+1 = 7$ . Essa escolha é arbitrária, mas deriva diretamente da problemática semanal.

### 3.3 Dados acumulados

Por fim, obtemos as curvas acumuladas de novos casos e óbitos, a fim de comparar com as curvas do modelo  $T$  e  $D$ . Dessarte, nosso dado será da forma

$$y_t = \sum_{i=-14}^t \hat{x}_i, \quad (t = 0, \dots, n) \quad (7)$$

## 4 Estimativa dos parâmetros

A construção do método de estimação presente nesse trabalho foi baseada em [8, 19, 31]. Como explicitado na Seção 3, os dados representam a informação diária do número de casos novos e de mortes a partir do começo de março, sendo que o primeiro dia desse mês tem muito mais casos do que o esperado e, portanto, foi excluído da análise. Consideraremos o início do modelo sendo o dia 16, quando as primeiras medidas de contenção foram tomadas [32] e a evolução do vírus deixa de ser tão irregular. No modelo, essas curvas são acumuladas [Seção 2] e, portanto, observamos

$$\hat{x}_i^{(1)} = (T(i) - T(i-1)) + \varepsilon_i^{(1)}, i = -14, \dots, n \quad (8)$$

$$\hat{x}_i^{(2)} = (D(i) - D(i-1)) + \varepsilon_i^{(2)}, i = -14, \dots, n \quad (9)$$

em que  $\hat{x}_0^{(1)}$  se refere ao aumento do número de casos notificados no dia 16 de março e  $n$  é o número de dias considerados. Assumimos que  $\{\varepsilon_i^{(1)}\}_{-14 \leq i \leq n}$  e  $\{\varepsilon_i^{(2)}\}_{-14 \leq i \leq n}$  são sequências de variáveis independentes e normalmente distribuídas com variâncias  $\sigma_k^2, k = 1, 2$  desconhecidas<sup>3</sup>. Para facilitar, vamos trabalhar com as observações

$$y_i^{(1)} = T(i) + \xi_i^{(1)}, i = 0, \dots, n \quad (10)$$

$$y_i^{(2)} = D(i) + \xi_i^{(2)}, i = 0, \dots, n \quad (11)$$

onde, para  $k = 1, 2$ , definimos  $\xi_i^{(k)} = \sum_{j=-14}^i \varepsilon_j^{(k)}$ . Se tormarmos  $i \leq j$ , calculamos

$$Cov(\xi_i^{(k)}, \xi_j^{(k)}) = (i+15)\sigma_k^2 \quad (12)$$

e o vetor  $\xi^{(k)}$  tem distribuição normal multivariada com média 0 e matriz de covariância dada pela equação (12). Normalizarmos os dados obtidos pelo tamanho da população, que assumimos ser de 6.7 milhões [14] no Rio de Janeiro. A estimativa dos parâmetros foi dividida em (1) identificabilidade, (2) ajuste nos dados e (3) quantificação da incerteza.

### 4.1 Escolha e modelagem dos parâmetros

Afirmamos que os parâmetros  $\tau, \sigma, \gamma_1, \gamma_2$  e  $\delta$  são epidemiológicos e, portanto, utilizamos estimativas da literatura que podem ser encontradas na Tabela 3. O parâmetro de testagem  $\rho$  é aproximado segundo a Observação 4.1. Por fim, os parâmetros  $\beta(t), \alpha$  e  $\mu(t)$  são estimados, de forma que a transmissibilidade e mortalidade são modeladas conforme explicado na Seção 4.1.1.

**Observação 4.1** (Estimativa parâmetro  $\rho$ ). *Vamos considerar  $\rho(t) \equiv \rho$  e utilizamos os dados de testagem do estado do Rio de Janeiro obtidos pelo IBGE através do PNAD COVID-19 [13] como resumido na Tabela 4. Em particular, percebemos que 2% da população é testada por mês,*

---

<sup>3</sup>A hipótese de normalidade tem a intuição de ser a soma de vários pequenos erros individuais independentes.

Par.	Valor	Referência
$\omega^{-1}$	5.74 dias	[30]
$\tau^{-1}$	3.69 dias	[18]
$\sigma^{-1}$	$\omega^{-1} - \tau^{-1}$	[18]
$\gamma_1^{-1}$	7.5 dias	[6]
$\gamma_2^{-1}$	13.4 dias	[6]
$\delta$	0.01	[17]

Tabela 3: Valores dos parâmetros epidemiológicos estimados pela literatura.

e entre os testados, 20% eram positivos. Sabemos que parte dessa testagem ocorre em pessoas que foram identificadas pelo sistema com o aparecimento de sintomas e gostaríamos de separar entre esses e aqueles que não possuíam sintomas, o que infelizmente não é informado. Então, por dia, a proporção de 0,00013 da população foi testada, isto é,  $\rho \leq 1.3 \cdot 10^{-4}$ . Tomaremos  $\rho = 10^{-5}$  e uma análise da influência dessa escolha é feita na Tabela 6.

	Julho	Agosto	Setembro	Outubro
Percentual (%)	6,8	8,6	10,2	11,9
Percentual que testaram positivo (%)	1,2	1,5	1,9	2,4

Tabela 4: Percentual de pessoas que fizeram algum teste para saber se estavam infectadas pelo SARS-CoV-2 no total da população.

#### 4.1.1 Estimação dos parâmetros que variam com o tempo

Na cidade, ao longo da epidemia, ocorreram diversas medidas de isolamento, como, por exemplo, a obrigatoriedade do uso de máscaras [33] e o sistema de bandeiras [12]. Por consequência, a taxa de transmissibilidade da doença, o parâmetro  $\beta$  do modelo, varia conforme a aceitação dessas medidas pela sociedade. A modelagem em si da resposta da sociedade é bem complexa e não será estudada nesse texto. Então selecionamos o uso da aproximação por *B-splines*, que pode ser expressa da seguinte forma:

$$\beta(t) \approx \sum_{j=1}^s \beta_j B_{j,k}(t) \quad (13)$$

onde  $\beta_j$  são os coeficientes a ser estimados e  $B_{j,k}(t)$  formam a base de funções de ordem  $k$ . Além disso, foi observado que ao estimar o modelo em questão, apesar da curva de novos casos ter bons resultados, a curva de mortes teve um comportamento diferente: ela apresentou um pico bem mais concentrado e uma subida e descida muito mais intensas que as previstas. Por esse motivo, tratamos  $\mu = \mu(t)$  da mesma forma que  $\beta$ , mas com  $r$  coeficientes.

Definimos, portanto, o vetor de parâmetros a ser estimado  $\theta = (\alpha, \beta_1, \dots, \beta_s, \mu_1, \dots, \mu_r)$  com  $s + r + 1$  parâmetros. Por hipótese assumiremos que os *knots*, pontos onde os polinômios se ligam, são igualmente espaçados. Poderíamos procurar os pontos ótimos conforme os momentos da epidemia em que a resposta do público foi modificada segundo às ações tomadas pelo poder público.

## 4.2 Identificabilidade

O primeiro passo após a concepção do modelo é analisar sua *identificabilidade*, porque a estimativa pode variar dependendo de quando o problema é ou não bem-posto. De forma geral, essa

análise identifica se os parâmetros desconhecidos podem ser estimados de forma única [1,35]. Existem duas formas de fazer essa análise: a forma estrutural (a priori) e a forma prática (a posteriori). A primeira é uma propriedade teórica que reside na própria estrutura do modelo. Os parâmetros são *estruturalmente identificáveis* se eles podem ser unicamente estimados a partir do experimento desenvolvido e são chamados *localmente estruturalmente identificáveis* se essa característica é verificada em torno do ponto ótimo. A segunda é feita após o processo de ajuste de dados, quando outros problemas podem aparecer em relação ao ruído encontrado na informação. Em particular, utilizaremos a *matriz de correlação* dos parâmetros.

Conforme descrito em [21], considere um sistema dinâmico

$$\begin{aligned}\dot{x} &= f(x(t), \theta), \quad x(0) = x_0 \\ y(t) &= h(x(t), \theta)\end{aligned}\tag{14}$$

em que  $x(t) \in \mathbb{R}^n$ ,  $y(t) \in \mathbb{R}^m$  e  $f$  e  $h$  são vetores de funções racionais da variável  $x$  e  $\theta \in \Theta \subset \mathbb{R}^p$ . A variável  $y$  é observável e no nosso modelo é indicada pelos casos confirmados e mortes. Quando  $x(0) = x_0$ , denotamos  $y = \psi_{x_0}(\theta, u)$  como a observação com esse valor inicial. O objetivo é contar o número de soluções da equação:

$$\psi_{x_0}(\theta, u) = \psi_{x_0}(\theta^*, u).\tag{15}$$

Dizemos que esse modelo é *globalmente identificável* na solução  $\theta^*$  se a equação (15) tem solução única  $\theta = \theta^*$ , para todo tempo  $s$ . Ele será *localmente identificável* se essa propriedade valer em uma vizinhança de  $\theta^*$ . Aplicamos um método computacional de identificabilidade estrutural através do programa DAISY [4].

#### 4.2.1 DAISY - Differential Algebra for Identifiability of SYstems

DAISY é um *software*, com base na linguagem de programação *Reduce*, específico para o problema de identificabilidade em sistemas dinâmicos, com algumas condições brandas para a utilização. Se o modelo possui essas condições, a codificação e manuseio são fáceis, o que é uma vantagem quando queremos aplicar os algoritmos algébricos envolvidos. Tratamos todos os parâmetros do modelo indicados na Seção 2 como invariantes no tempo. Supondo o conhecimento da curva de recuperados  $R$ , o problema foi considerado globalmente identificável. Infelizmente esses dados não são completamente disponíveis, conforme destacado na Seção 3.1 e na ausência deles, o *software* não conseguiu resolver em pelo menos 3 dias, e, portanto, decidimos não continuar por esse caminho. O código pode ser encontrado no Github [24].

### 4.3 Ajuste nos dados

#### 4.3.1 Hipóteses iniciais

Vamos considerar que no início da epidemia  $S \approx 1$ . Com essa aproximação e restringindo o modelo para as equações de  $E, I, A$  e  $T$ , temos um sistema com quatro equações lineares (16), cuja solução é combinação linear de exponenciais. Podemos assumir que a testagem diária no início da pandemia era de  $\rho \approx 0$  e, como já mencionado na sessão anterior, os parâmetros  $\gamma_1, \tau$  e  $\sigma$  são conhecidos. Os parâmetros  $\alpha$  e  $\beta$  são fixos no período.

$$\begin{bmatrix} \dot{E} \\ \dot{I} \\ \dot{A} \\ \dot{T} \end{bmatrix} = \begin{bmatrix} -\tau & \beta & \beta & 0 \\ \tau & -\sigma & 0 & 0 \\ 0 & \sigma\alpha & -\gamma_1 & 0 \\ 0 & \sigma(1-\alpha) & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} E \\ I \\ A \\ T \end{bmatrix}.\tag{16}$$

O valor  $T(-14)$  corresponde aos dados de confirmados em 2 de março e  $T(-15) = 0$ . Assim estimamos  $\theta_0 = (\alpha, \beta, E(-14), I(-14), A(-14))$  de forma a minimizar a expressão pgf

$$\sum_{i=-14}^0 w_i \left( y_i^{(1)} - \hat{T}(i, \theta_0) \right)^2, \quad (17)$$

em que  $w_i = \frac{i+14}{14}$ . Com essas estimativas, obtemos os valores  $(E(0), I(0), A(0))$  que determinam o início da pandemia. Não há mortes no início da pandemia, o que permite colocar  $\mu = 0$ . Usando que  $R(-14) = 0$ , obteremos que  $Q(-14) = T(-14)$  e, assim, podemos integrar as curvas  $Q$  e  $R$  até obter os valores  $Q(0)$  e  $R(0)$ . Por fim  $S(0) = 1 - E(0) - I(0) - A(0) - Q(0) - R(0)$ .

#### 4.3.2 Método de estimação

Utilizaremos o método de *mínimos quadrados ponderados* para estimar os parâmetros desconhecidos, considerando a observação de casos confirmados e mortes diárias como indicado nas equações (10) e (11). Nessa abordagem, resolvemos um problema de minimização não linear com restrições e a cada iteração integramos o sistema de equações diferenciais. Definimos a função objetivo como:

$$F(\theta) = (y^{(1)} - \hat{T}(\theta))^T \Sigma^{-1} (y^{(1)} - \hat{T}(\theta)) + \psi (y^{(2)} - \hat{D}(\theta))^T \Sigma^{-1} (y^{(2)} - \hat{D}(\theta)) \quad (18)$$

tal que  $\hat{T}(\theta)$  e  $\hat{D}(\theta)$  são os vetores soluções numéricas das equações diferenciais através do método Runge-Kutta,  $\sigma_k^2 \Sigma$  é a matriz de covariância dada pela equação (12) ( $k = 1, 2$ ) e  $\psi$  é um peso de importância das mortes na minimização e está relacionado à razão das variâncias  $\sigma_1^2 / \sigma_2^2$ . A escolha dessa função está relacionada com as propriedades estatísticas dos dados. Sabemos que, pela hipótese de normalidade, minimizar a expressão 18 é equivalente a maximizar a verossimilhança. Utilizamos o algoritmo L-BFGS-B [5], baseado no método de projeção gradiente e no algoritmo BFGS com uso limitado de memória computacional, implementado na biblioteca Scipy [39] da linguagem de programação Python e o código está disponível no Github [24].

#### 4.3.3 Análise residual e $\mathcal{R}_t$ estimado

Escolhemos, de modo arbitrário, fixar a data final para análise dos dados em 31 de julho de 2020 e primeiro ajustamos a curva aos dias iniciais da epidemia, conforme explicado na Seção 4.3.1. Obtemos o resultado na Figura 5. Definindo a priori  $\psi = 119.57$  (valor aproximado entre as variâncias dos dados de casos e óbitos) e os hiperparâmetros ligados à otimização, como os limites superior e inferior e chutes iniciais de cada parâmetro, vamos analisar os resíduos do ajuste de dados e o  $\mathcal{R}_t$  induzido pela estimação dos parâmetros segundo:

- (i) o número de coeficientes a serem estimados para  $\beta$  e  $\mu$ ; e
- (ii) a ordem das B-splines para os parâmetros  $\beta$  e  $\mu$ .

Por exemplo, colocando 4 coeficientes para as B-splines de ordem 3 para ambos os parâmetros que variam no tempo, obtemos as curvas descritas na Figura 6. Em particular  $\alpha$  foi estimado em 0.899.

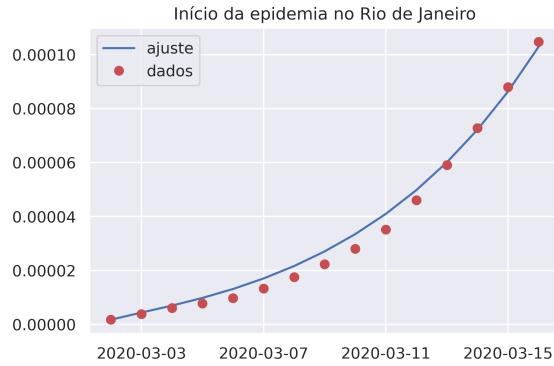


Figura 5: Ajuste à curva acumulada de casos no início da epidemia para obtenção dos valores iniciais, como explicado na Seção 4.3.1.

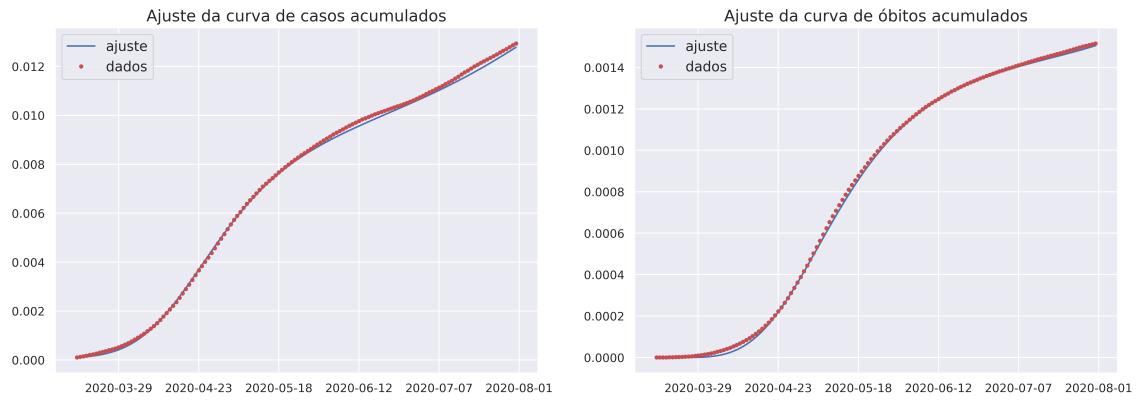


Figura 6: Ajuste das curvas acumuladas de casos e óbitos, respectivamente.

Com esse resultado, podemos verificar se os resíduos do modelo aproximam os erros, que por hipótese têm distribuição Gaussiana. Primeiro diferenciamos para obter resíduos diários, de forma que sua variância seja constante (ver equações (8) e (9)) e, então, desenhamos os *histogramas* (indicador da frequência de cada resíduo) e os *gráficos Q-Q* (comparação dos quartis da distribuição Gaussiana com os quartis da distribuição amostral dos resíduos) que são apresentados na Figura 7. Já visualizamos que ambas as curvas não têm um ajuste com resíduos como esperado, principalmente a curva de mortes. Além dessa análise visual, também podemos aplicar testes estatísticos para verificar correlação e normalidade.

O teste *Ljung-Box* [20] é um teste estatístico que verifica a autocorrelação em uma série temporal, cuja hipótese nula afirma que as k-correlações são nulas, isto é, a série temporal é independentemente distribuída. Aplicando-o nos resíduos do ajuste para as duas curvas, obtivemos o p-valor 0 computacional, o que indica que os resíduos não são descorrelacionados como assumimos. Já o teste *Jarque-Bera* [15] avalia a assimetria e a curtose da distribuição amostral e compara com as da normal. No nosso caso, a hipótese nula não foi rejeitada à nível 5%, o que é um bom indicativo de normalidade dos resíduos. Concluímos, em particular, que o modelo, apesar de um ajuste interessante, com resíduos normalmente distribuídos, tem resíduos correlacionados, o que precisa ser melhor estudado em trabalhos futuros. Podemos fazer essa análise para todas as combinações de

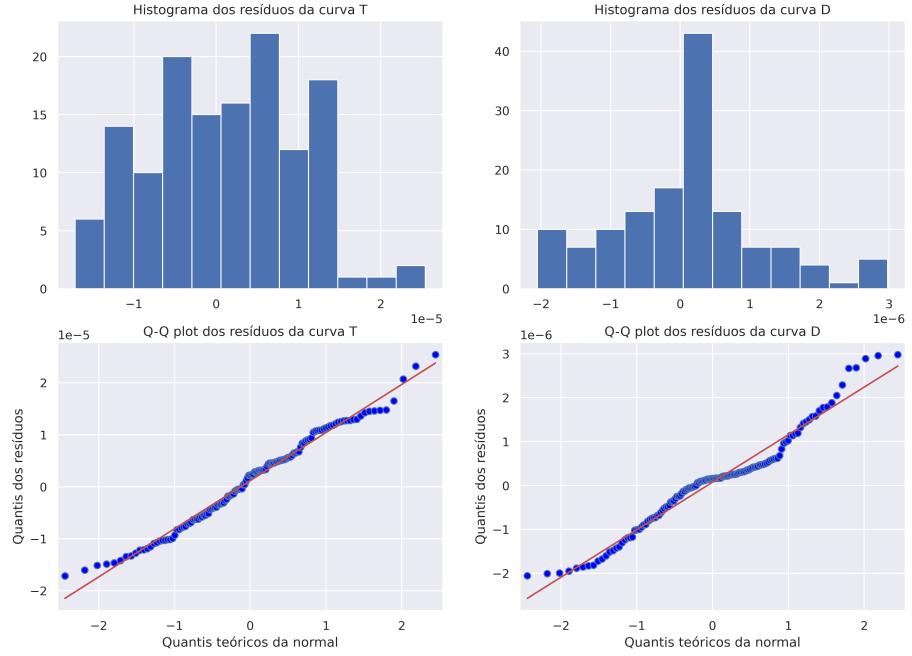


Figura 7: Análise gráfica dos resíduos com histograma e Q-Q plot.

parâmetros que desejarmos. Para mais detalhes, confira no Github [24].

Outra característica importante a ser analisada é o número reprodutivo básico, descrito na Seção 2.1. Para os parâmetros estimados no exemplo, obtemos o  $\mathcal{R}_t$  descrito na Figura 8. Considerando as estimativas do  $\mathcal{R}_0$  para o Brasil em [3], encontramos no dia 23 de março o valor de 2.325 para o modelo SIRD, e 2.897 para o modelo SIRASD. Em [22], a estimativa pontual para o dia 9 de maio de 2020 do estado do Rio de Janeiro foi de 1.1 com 95%-intervalo de confiança [0.9, 1.3], que inclui nosso valor. Por fim, a curva estimada por [26] para a cidade do Rio de Janeiro tem um formato muito similar à da Figura 8, em que no começo de maio, o  $\mathcal{R}_t$  tem estimativa menor do que 1 e volta a crescer até ficar maior do que 1 no final de julho. Esses valores corroboram nossa estimativa.

Para determinar a ordem  $r$  e  $s$ , e o número de coeficientes para a aproximação por B-splines, usamos o *critério de informação AIC* descrito em [19], que sob a hipótese de normalidade dos erros, pode ser calculado segundo a fórmula

$$AIC = n \ln \left( \frac{RSS}{n} \right) + 2K, \quad (19)$$

onde  $K$  é o número de parâmetros desconhecidos e  $RSS$  é a soma dos quadrados dos resíduos do modelo. Os resultados podem ser conferidos na Tabela 5. Para esse experimento, escolhemos comparar modelos com 3 e 4 coeficientes, pois modelos com menos do que 3 coeficientes se mostrarem incapazes de capturar a variação temporal, enquanto modelos com mais do que 4 parâmetros são difíceis de estimar. Com isso, a ordem das splines pode variar de  $k = 0$  a  $k = \text{número de coeficientes} - 1$ .

Segundo o critério, o modelo escolhido possui 4 coeficientes para os dois parâmetros, com ordem das B-splines 2 para  $\beta$  e 1 para  $\mu$ . Os resultados para o modelo escolhido não são muito diferentes dos gráficos já apresentados, incluindo a estimação de  $\alpha = 0.898$  e podem ser conferidos no Github.

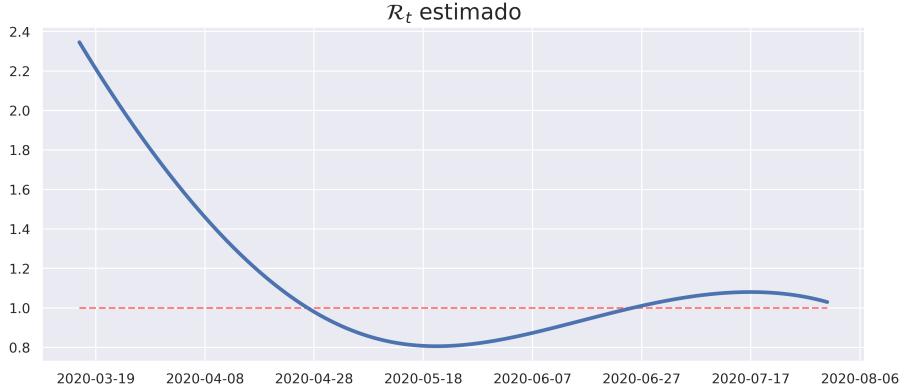


Figura 8: Número reprodutivo ao longo do primeiros cinco meses de pandemia para os parâmetros do exemplo.

B-splines	Parâmetros $\mu$						
	(3,0)	(3,1)	(3,2)	(4,0)	(4,1)	(4,2)	(4,3)
(3,0)	-2.021	-2.029	-2.023	-2.025	-2.028	-2.032	-2.001
(3,1)	-2.228	-2.241	-2.266	-2.288	-2.298	-2.311	-2.298
(3,2)	-2.246	-2.230	-2.253	-2.294	-2.337	-2.331	-2.309
(4,0)	-1.967	-1.979	-1.979	-1.998	-2.002	-2.000	-1.987
(4,1)	-2.262	-2.304	-2.277	-2.302	-2.372	-2.348	-2.338
(4,2)	-2.346	-2.243	-2.262	-2.300	-2.374	-2.357	-2.332
(4,3)	-2.213	-2.233	-2.251	-2.294	-2.363	-2.347	-2.323

Tabela 5: Seleção de modelo segundo o AIC (escala  $10^3$ ): os parâmetros que variam no tempo aproximados por 3 ou 4 parâmetros e ordem de 0 ao número de coeficientes - 1.

#### 4.3.4 Identificabilidade prática

O processo de identificabilidade estrutural desenvolvido na Seção 4.2 é feito com base em duas hipóteses: a estrutura do modelo é precisa e os erros de mensuração são ausentes. Essas hipóteses não são válidas na prática e, por isso, é necessário avaliar se os parâmetros podem ser estimados de forma confiável e com precisão a partir dos dados ruidosos [23]. Seja  $\hat{\theta} = (\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\mu}_r)$  o vetor de estimativas a partir do ajuste aos dados. A *matriz de correlação* é um método que examina as correlações entre os parâmetros do modelo e pode ser calculada com base na *matriz Informação de Fisher (FIM)* da seguinte maneira,

$$FIM = \frac{1}{\hat{\sigma}_1^2} \left( \frac{\partial \hat{T}}{\partial \theta} \right) \Bigg|_{\theta=\hat{\theta}}^T \Sigma^{-1} \left( \frac{\partial \hat{T}}{\partial \theta} \right) \Bigg|_{\theta=\hat{\theta}} + \frac{1}{\hat{\sigma}_2^2} \left( \frac{\partial \hat{D}}{\partial \theta} \right) \Bigg|_{\theta=\hat{\theta}}^T \Sigma^{-1} \left( \frac{\partial \hat{D}}{\partial \theta} \right) \Bigg|_{\theta=\hat{\theta}} \quad (20)$$

e a *matriz de covariâncias C* é igual a inversa de *FIM* (equação (20)). Finalmente, o elemento  $r_{ij}$ ,  $1 \leq i, j \leq r + s + 1$  da *matriz de correlações* pode ser definido como

$$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii}C_{jj}}} \quad (21)$$

que mensura a correlação entre as estimativas  $\hat{\theta}_i$  e  $\hat{\theta}_j$ . Uma correlação próxima a 1 indica que os parâmetros são praticamente indistinguíveis e um depende fortemente do outro. Em particular, ao

calcularmos  $R = [r_{ij}]$  para os parâmetros estimados, obtemos uma matriz de dimensão  $K \times K$  onde  $K$  é o número de parâmetros. Podemos visualizar através de um mapa de calor na Figura 9 que  $\alpha$  e o segundo coeficiente da B-spline de  $\beta$  são fortemente correlacionados, mas negativamente. Isso não é um bom sinal quando se trata de identificabilidade.



Figura 9: Matriz de correlação dos parâmetros estimados do modelo.

Outra questão que pode ser levantada é sobre a influência da escolha dos parâmetros epidemiológicos, isto é, se eles variassem um pouco, o quanto isso afetaria a estimativa de  $\alpha$ . Para fazer essa análise, escolhemos uma rede de valores para cada parâmetro baseada nos intervalos de confiança estimados nos respectivos artigos citados na Tabela 3, estimamos os parâmetros do modelo e reportamos os menores intervalos que englobam os valores estimados de  $\alpha$  (com precisão até a terceira casa decimal), que podem ser conferidos na Tabela 6. A fixação dos parâmetros influí pouco na estimativa final de  $\alpha$ , portanto.

Parâmetro	Intervalo	Faixa de valores de $\alpha$
$\tau^{-1}$	[2, 4]	[0.897, 0.902]
$\sigma^{-1}$	[2, 4.5]	[0.897, 0.9]
$\rho$	[0, $10^{-4}$ ]	[0.898, 0.899]
$\gamma_1^{-1}$	[6.5, 9.5]	[0.894, 0.903]
$\gamma_2^{-1}$	[11, 16]	[0.896, 0.898]

Tabela 6: Rede de valores dos parâmetros fixados e o menor intervalo que inclui os valores estimados de  $\alpha$

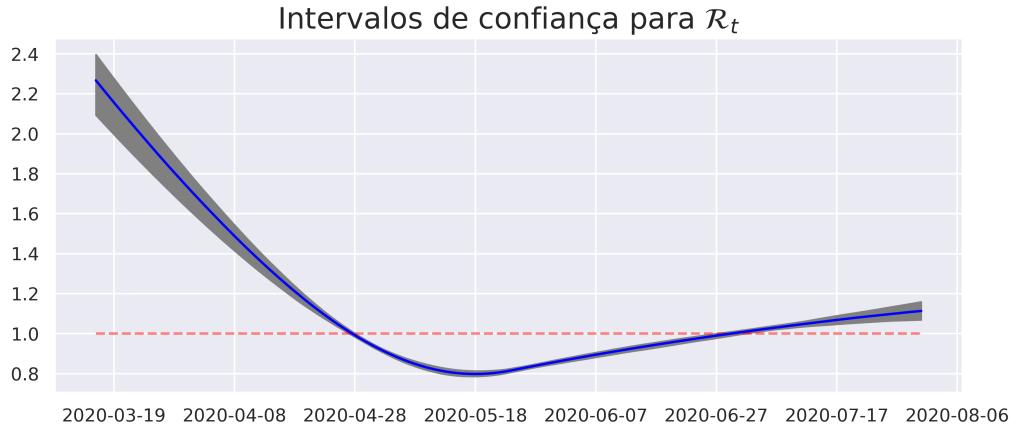


Figura 10: A curva em azul é a mediana das curvas estimadas e em cinza o intervalo de confiança. Em vermelho é o limiar 1 para o  $\mathcal{R}_t$ .

#### 4.4 Quantificando incerteza sobre os parâmetros

Os parâmetros estimados em um sistema dinâmico estão sujeitos à incerteza presente na captação dos dados, tal como erros de mensuração ou a variabilidade natural dos dados, e nas hipóteses adotadas pelo método de estimação. Assim, nessa seção, o objetivo será construir intervalos de confiança para os parâmetros desconhecidos de forma que, no longo prazo, a proporção dos intervalos calculados que contêm o parâmetro seja o nível desejado. Utilizaremos o *método Bootstrap* [11] que tem uma abordagem por simulação baseada nos dados. Ele gera, a partir de uma série  $Y$ , dados replicados  $Y_1^*, \dots, Y_N^*$  e realiza as estimativas para cada um. Os intervalos de confiança para os valores de interesse são, então, os percentis correspondentes das  $N$  amostras replicadas [16].

Como consequência da estrutura dos erros apresentada nas equações (8) e (9), as curvas são geradas com o valor inicial  $\hat{T}(0) = y_0^{(1)}$  e para cada  $i \geq 1$ ,  $\hat{T}(i+1) = \hat{T}(i) + \varepsilon_{i+1}$ , em que  $\varepsilon_{i+1} \sim \mathcal{N}(\hat{x}_{i+1}^1, \hat{\sigma}_1^2)$  e  $\hat{\sigma}_1^2$  é a estimativa para  $\sigma_1^2$ . A construção é análoga para a curva de óbitos. Os parâmetros, então, foram estimados para cada uma das simulações. Além disso, para evitar que todas as simulações resultem em uma mesma região com mínimo locais, aleatorizamos o chute inicial do algoritmo de otimização. Para cada  $j$  entre 1 e  $r+s+1$ ,  $\theta_j^{inicial}$  é tomado aleatoriamente com distribuição uniforme no intervalo  $(l_j, u_j)$ , isto é, entre os limites inferior e superior estabelecidos. Realizamos esse processo  $m$  vezes e tomamos a estimativa com menor erro na função objetivo (equação (18)).

Fizemos  $N = 500$  simulações com  $m = 10$  (ou seja, 5000 simulações no total) e os outros parâmetros definidos como na Seção 4.3.3. Após as simulações, estimamos os intervalos de confiança para os parâmetros com nível 95%, assim como as curvas  $\mathcal{R}_t$  induzidas a cada tempo  $t$ . O intervalo estimado para a taxa de subnotificação  $\alpha$  foi  $(0.849, 0.931)$ . Na Figura 10, é possível visualizar os intervalos de  $\mathcal{R}_t$ . É interessante visualizar o gráfico que exibe as relações dois a dois entre os parâmetros, o que mostra a correlação nas estimativas. As correlações estimadas na Figura 9 podem ser observadas nos gráficos de dispersão na Figura 11, como, por exemplo, a relação linear entre  $\alpha$  e  $\beta_2$ . Os histogramas dos parâmetros também podem ser observados. Na Tabela 7 é sumarizada a informação para todos os parâmetros. Informações e gráficos adicionais podem ser visualizados no Github [24].

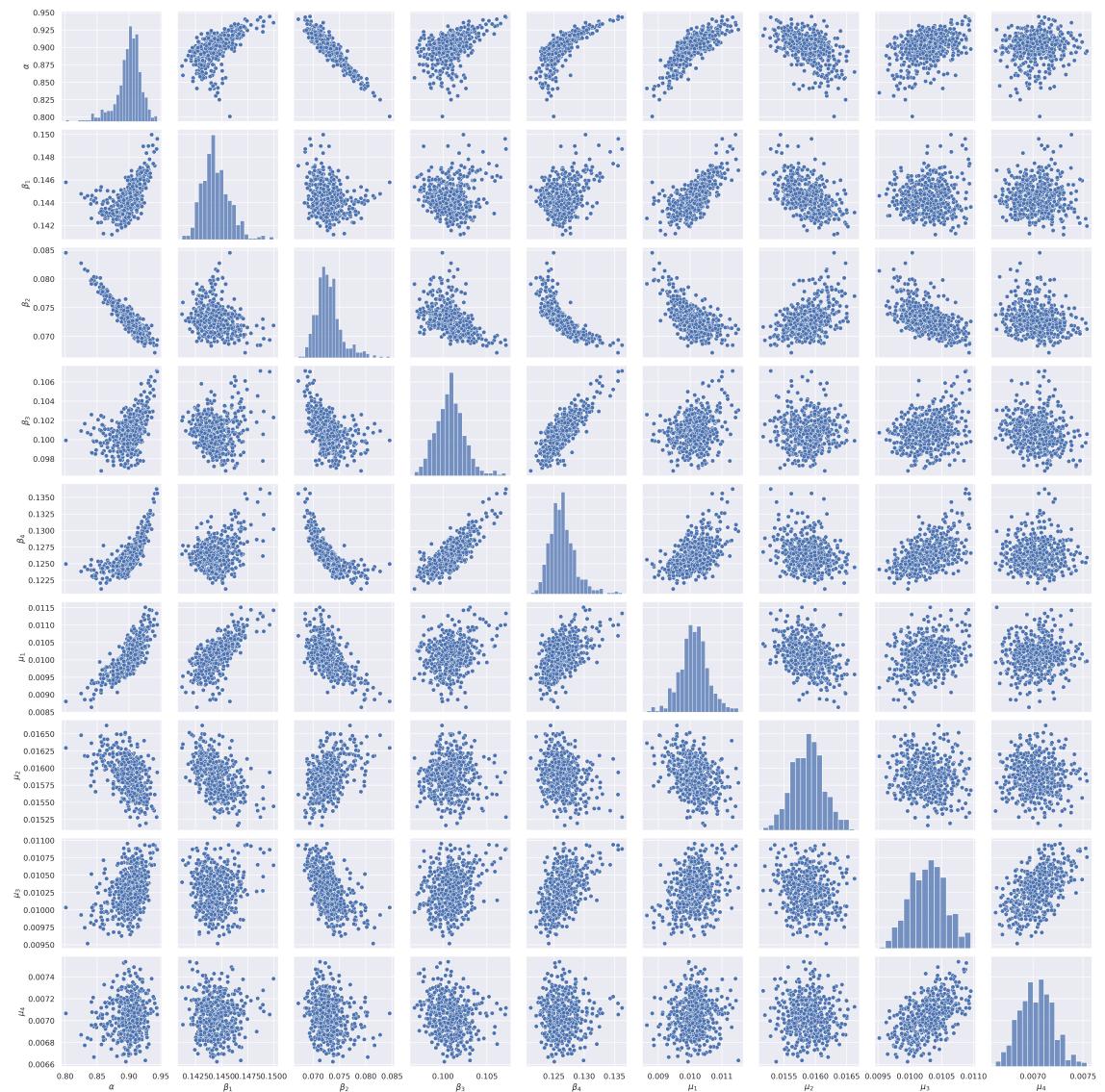


Figura 11: Gráficos de dispersão a cada dois parâmetros indicando a relação entre eles e os histogramas das estimativas.

Parâmetro	Mediana	Intervalo calculado
$\alpha$	0.903	[0.849, 0.93]
$\beta_1$	0.144	[0.142, 0.147]
$\beta_2$	0.073	[0.069, 0.079]
$\beta_3$	0.101	[0.098, 0.104]
$\beta_4$	0.126	[0.123, 0.132]
$\mu_1$	0.0101	[0.0092, 0.011]
$\mu_2$	0.0159	[0.0154, 0.0164]
$\mu_3$	0.0103	[0.0098, 0.0108]
$\mu_4$	$7.03 \cdot 10^{-3}$	[ $6.73 \cdot 10^{-3}$ , $7.38 \cdot 10^{-3}$ ]

Tabela 7: Estimativas da mediana e dos intervalos de confiança para cada parâmetro desconhecido do modelo.

## 5 Conclusão

Nesse trabalho, utilizamos ferramentas matemáticas de equações diferenciais, estatística e otimização para estimar a taxa subnotificação de COVID-19 na cidade do Rio de Janeiro no início da pandemia. Obtivemos que a proporção  $\alpha$  de indivíduos não identificados pelo sistema é em torno de 90%, com intervalo de confiança entre 85% e 92%. Isso significa que a cada indivíduo notificado pelo sistema, em torno de 9 ou 10 não foram observados, por serem assintomáticos, apresentarem sintomas leves ou não serem testados. Esse valor está de encontro com várias cidades do Brasil, de acordo com [7]. Em uma nota técnica de abril de 2020, [28] estimou a notificação de casos no Brasil entre 7.8% e 8.1% e [27], também em abril de 2020, em 7%. No Rio de Janeiro, [29] estimou a notificação em 7.2%. Esses resultados vão de encontro com o desse trabalho.

A análise de dados na Seção 3 permitiu o melhor entendimento de problemas em relação ao tempo de notificação de um indivíduo infectado. Nesse sentido, além da falta de testagem, existem atrasos e erros em todas as etapas do processo que, sem o devido cuidado, podem gerar más inferências. Gráficos adicionais sobre as outras variáveis estão em formato *notebook* no Github [24]. Obtivemos um ajuste interessante do modelo, com resíduos distribuídos normalmente, apesar de estarem correlacionados. Os resultados de identificabilidade estrutural (Seção 4.2) reforçam a necessidade da obtenção da curva de recuperados, enquanto na identificabilidade prática (Seção 4.3.4), sugerem que outra abordagem para estimar as curvas de transmissibilidade e mortalidade deve ser tomada. A despeito disso, a adoção de limites nos parâmetros permitiu uma estimativa mais precisa, inclusive quando os parâmetros previamente fixados variavam (ver Tabela 6).

O número reprodutivo básico como função do tempo mostrou ter um comportamento similar às outras estimativas, como mencionado em 4.3.3. A pouca variabilidade deve-se à suavidade induzida pelas B-splines. Para construir uma curva mais suscetível a mudanças diárias, uma aproximação por outro método deve ser adotada. Por fim, a aparente subestimação da incerteza nessa curva deriva dessa problemática. Outras fontes de incerteza podem ser adicionadas ao Bootstrap em trabalhos futuros.

## 6 Agradecimentos

Gostaria de agradecer à Orientadora do trabalho Maria Soledad Aronna (FGV/EMAp, Rio de Janeiro) pela compreensão e aconselhamento ao longo do processo. Agradecer também aos professores Roberto Guglielmi (Universidade de Waterloo, Canadá) e Luiz Max de Carvalho (FGV/EMAp, Rio de Janeiro) pela contribuição em tópicos importantes. Essa iniciação científica foi apoiada pela

FAPERJ (Brasil) através do Programa “Jovem Cientista do Nossa Estado” e pelo CNPq (Brasil) através do Programa “Iniciação Científica e Mestrado” (PICME).

## Referências

- [1] Audoly, S. *et al.* Global identifiability of nonlinear models of biological systems. *IEEE Transactions on Biomedical Engineering* 48(1):55-65, 2001, DOI: 10.1109/10.900248.
- [2] Aronna, M. S., Guglielmi, R., Moschen, L. M. A model for COVID-19 with isolation, quarantine and testing as control measures. *Epidemics*. 34, 2021. DOI: 10.1016/j.epidem.2021.100437
- [3] Bastos, S. B., Cajueiro, D. O. Modeling and forecasting the early evolution of the covid-19 pandemic in brazil. *Scientific Reports*, 10(1):19457, 2020. DOI: 10.1038/s41598-020-76257-1.
- [4] Bellu, G., Saccomani, M. P., Audoly, S., D'Angiò, L. DAISY: a new software tool to test global identifiability of biological and physiological systems. *Comput Methods Programs Biomed.* 88(1):52-61, 2007. DOI:10.1016/j.cmpb.2007.07.002.
- [5] Byrd, R. H., Lu P., Nocedal, J., Zhu, C. A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific and Statistical Computing* 16(5): 1190-1208, 1995. DOI: 10.1137/0916069
- [6] Byrne A. W. *et al.* Inferred duration of infectious period of SARS-CoV-2: rapid scoping review and analysis of available evidence for asymptomatic and symptomatic COVID-19 cases. *BMJ Open*. 10(8), 2020. DOI: 10.1136/bmjopen-2020-039856.
- [7] Canzian, F. Estados e municípios no país relatam subnotificação gigantesca de casos. *Folha de São Paulo*. <https://www1.folha.uol.com.br/equilibrioesaude/2020/04/estados-e-municipios-no-pais-relatam-subnotificacao-gigantesca-de-casos.shtml>, abril de 2020.
- [8] Cao, J., Huang J. Z., Wu H. Penalized Nonlinear Least Squares Estimation of Time-Varying Parameters in Ordinary Differential Equations. *Journal of Computational and Graphical Statistics*, 21:1, 42-56, 2012. DOI: 10.1198/jcgs.2011.10021
- [9] Croda, J. *et al.* COVID-19 in Brazil: advantages of a socialized unified health system and preparation to contain cases. *Rev. Soc. Bras. Med. Trop.* 53, 2020. DOI: 10.1590/0037-8682-0167-2020
- [10] Dantas, G. *et al.* The impact of COVID-19 partial lockdown on the air quality of the city of Rio de Janeiro, Brazil. *Science of The Total Environment*, 729, 2020. DOI: 10.1016/j.scitotenv.2020.139085.
- [11] Efron, B., Tibshirani, R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statist. Sci.* 1(1):54-75, 1986. DOI: 10.1214/ss/1177013815
- [12] Estadão. Governo do Rio cria classificação em 3 bandeiras para flexibilizar isolamento. <https://revistapegn.globo.com/Noticias/noticia/2020/05/pegn-governo-do-rio-cria-classificacao-em-3-bandeiras-para-flexibilizar-isolamento.html>, maio de 2020.
- [13] IBGE – Instituto Brasileiro de Geografia e Estatística. Pesquisa Nacional por Amostra de Domicílios - PNAD COVID19. <https://www.ibge.gov.br/estatisticas/sociais/trabalho/27946-divulgacao-semanal-pnadcovid1.html?=&t=downloads>. 2020.

- [14] IBGE – Instituto Brasileiro de Geografia e Estatística. Cidades e Estados. <https://www.ibge.gov.br/cidades-e-estados/rj/rio-de-janeiro.html>. 2021.
- [15] Jarque, C. M., Bera, A. K. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3): 255-259, 1980. DOI: 10.1016/0165-1765(80)90024-5
- [16] Joshi, M., Seidel-Morgenstern, A. Kremling, A. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering*, 8(5): 447-455, 2006. DOI: 10.1016/j.ymben.2006.04.003
- [17] Kucirka L. M. *et al.* Variation in false-negative rate of reverse transcriptase polymerase chain reactionbased SARS-CoV-2 tests by time since exposure. *Ann Intern Med*. 173(4):262–7, 2020. DOI: 10.7326/M20-1495
- [18] Li R, Pei S, Chen B, *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*. 368(6490):489-493, 2020. DOI:10.1126/science.abb3221
- [19] Liang, H., Miao, H., Wu, H. Estimation of constant and time-varying dynamic parameters of HIV infection in a nonlinear differential equation model. *Ann. Appl. Stat.* 4, 1: 460-483, 2010. DOI 10.1214/09-AOAS290
- [20] Ljung, G. M., Box, G. E. P. On a measure of lack of fit in time series models. *Biometrika*, 65:297-303, 1978. DOI: 10.1093/biomet/65.2.297
- [21] Ljung, L. , Glad, T. On global identifiability for arbitrary model parametrizations. *Automatica*, 30: 265-276, 1994. DOI:10.1016/0005-1098(94)90029-9.
- [22] Mellan, T., *et al.* Subnational analysis of the COVID-19 epidemic in Brazil. *Cold Spring Harbor Laboratory Press*, 2020. DOI: 10.1101/2020.05.09.20096701.
- [23] Miao, H., Xia, X. Perelson, A. S., Wu, H. On Identifiability of Nonlinear ODE Models and Applications in Viral Dynamics. *SIAM Review*, 53(1): 3-39, 2011. DOI: 10.1137/090757009
- [24] Moschen, L. M. Repositório Covid-19. *Github*. Disponível em <https://github.com/lucasmoschen/covid-19-model>.
- [25] Nogradi, B. What the data say about asymptomatic COVID infections. *Nature*. 587: 534-535, 2020. DOI: 10.1038/d41586-020-03141-3
- [26] Observatório COVID-19 BR. R efetivo no Rio de Janeiro. Disponível em [https://covid19br.github.io/municipios.html?aba=aba3&uf=RJ&mun=Rio\\_de\\_Janeiro](https://covid19br.github.io/municipios.html?aba=aba3&uf=RJ&mun=Rio_de_Janeiro). Acesso em abril de 2021.
- [27] Portal COVID-19 Brasil. COVID-19 BRASIL [acessado 2021 Abril]. Disponível em: <https://ciis.fmrp.usp.br/covid19/>
- [28] Prado, M. F. do *et al.* Análise de subnotificação do número de casos confirmados da COVID-19 no Brasil. Disponível em [https://drive.google.com/file/d/1\\_whlqZnGgvqHuWCG4-JyiL2X9WXpZAe3/view](https://drive.google.com/file/d/1_whlqZnGgvqHuWCG4-JyiL2X9WXpZAe3/view).
- [29] Prado, M. F. do *et al.* Análise da subnotificação de COVID-19 no Brasil. *Rev. bras. ter. intensiva [online]*. 32(2): 224-228, 2020. DOI: 10.5935/0103-507x.20200030.

- [30] Rai, B., Shukla, A., Dwivedi, L.K. Incubation period for COVID-19: a systematic review and meta-analysis. *J Public Health (Berl.)* 2021. DOI: 10.1007/s10389-021-01478-1
- [31] Ramsay, J. O., Hooker, G., Campbell, D. e Cao, J. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:741–796, 2007. DOI:10.1111/j.1467-9868.2007.00610.x
- [32] Rio de Janeiro. Decreto nº 46.973, 16 de março de 2020. Reconhece a situação de emergência na saúde pública do estado do Rio de Janeiro em razão do contágio e adota medidas de enfrentamento da propagação decorrente do novo coronavírus (COVID-19); e dá outras previdências. **Diário Oficial do Estado do Rio de Janeiro**, Rio de Janeiro, RJ, n. 049-A, 17 de março de 2020. Disponível em <https://pge.rj.gov.br/comum/code/MostrarArquivo.php?C=MTAyMjI>.
- [33] Rio de Janeiro. Lei nº 8859, 03 de junho de 2020. Estabelece a obrigatoriedade do uso de máscaras respiratórias, no âmbito do estado do Rio de Janeiro, enquanto vigorar o estado de calamidade pública em virtude da pandemia do novo coronavírus (COVID-19). **Diário Oficial do Estado do Rio de Janeiro**, Rio de Janeiro, RJ, n. 100, 04 de junho de 2020. Disponível em <http://www.aerj.net.br/file/04-06-2020-leiestadomascara.pdf>.
- [34] Rodriguez-Fernandez, M., Egea, J. A., Banga, J. R. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics* 7:483, 2006. DOI: 10.1186/1471-2105-7-483
- [35] Saccomani M. P., Thomases K. Calculating all multiple parameter solutions of ODE models to avoid biological misinterpretations. *Mathematical Biosciences and Engineering : MBE* , 16(6):6438-6453, 2019. DOI: 10.3934/mbe.2019322.
- [36] Secretaria Municipal de Saúde (SMS), Prefeitura da Cidade do Rio de Janeiro. Dados individuais dos casos confirmados de COVID-19 no município do Rio de Janeiro. <https://www.arcgis.com/home/item.html?id=f314453b3a55434ea8c8e8caaa2d8db5>, março de 2021.
- [37] Secretaria Municipal de Saúde (SMS), Prefeitura da Cidade do Rio de Janeiro. Painel Rio COVID-19.<https://experience.arcgis.com/experience/38efc69787a346959c931568bd9e2cc4>, 2021.
- [38] The 2019 nCoV Outbreak Joint Field Epidemiology Investigation Team and Q. Li. An Outbreak of NCIP (2019-nCoV) Infection in China - Wuhan, Hubei Province, 2019 - 2020.-<http://weekly.chinacdc.cn/en/article/id/e3c63ca9-dedb-4fb6-9c1c-d057adb77b57>, janeiro de 2020
- [39] Virtanen P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*. 17(3): 261-272, 2020.
- [40] World Health Organization. Coronavirus disease (COVID-19). <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19#:~:text=symptoms>, outubro de 2020