DOI: 10.48011/asba.v2i1.982

Seleção de modelos epidemiológicos via análise de sensibilidade global

Michel Tosin * **Americo Cunha Jr** ** **Flávio C. Coelho** ***

* *Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, RJ (e-mail: michel.tosin@uerj.br).*
** *Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, RJ (e-mail: americo.cunha@uerj.br)*
*** *Escola de Matemática Aplicada, Fundação Getúlio Vargas, RJ (e-mail: fccoelho@fgv.br)*

Abstract: This paper propose a methodology for epidemiological model selection by using Akaike information criteria bringing as novelty the construction of a likelihood function based in the results of a global sensitivity analysis through the Sobol's indices obtained by using polynomial chaos expansion. The main ideia is to incorporate of the information about the influence of the parameters on the response to select a more interesting model inside of a set of candidates. The strategy is applied to a set of compartmental models compatible with those used to analyze the recent COVID-19 pandemic, allowing to compare them without the presence of experimental data.

Resumo: Este paper propõe uma metodologia de seleção de modelos epidemiológicos via critério da informação de Akaike trazendo como novidade a construção de uma função de verossimilhança baseada nos resultados de uma análise de sensibilidade global através dos índices de Sobol obtidos usando expansão em polinômios caos. A ideia geral é incorporar a informação sobre a influência dos parâmetros na resposta para selecionar um modelo mais interessante dentro do conjunto de candidatos. A estratégia é aplicada a um conjunto de modelos compartimentais compatíveis aos usados para analisar a pandemia de COVID-19 recente, permitindo compará-los sem a presença de dados experimentais.

Keywords: compartmental models; COVID-19; Sobol's indices; polynomial chaos expansion; Akaike information criteria.

Palavras-chaves: modelos compartimentais; COVID-19; índices de Sobol; expansão em polinômio caos; critério da informação de Akaike.

1. INTRODUÇÃO

A análise de epidemias pelo uso de ferramental matemático remonta aos trabalhos de W. H. Hamer (Brauer, 2017), mas ganhou grande destaque devido aos surtos recentes dados por diversas doenças ao redor do mundo e que já demonstram sequelas na população mundial (Toscano et al., 2020). A enfermidade mais recente e agressiva tem sido a COVID-19, cuja pandemia foi responsável (até outubro de 2020) pela morte de aproximadamente 1 milhão de pessoas em 216 países (World Health Organization, 2020). Desde então, diversas metodologias foram desenvolvidas para trabalhar com esse tipo de fenômeno, usando de equações diferencias a até redes neurais e aprendizado de máquina (Wiratsudakul et al., 2018; Kuhl, 2020). Em geral, a estratégia recente baseia-se em aprender sobre o comportamento de um surto por meio de dados, seja para validar seu modelo ou aprimorá-lo. Claro que dados, apesar de extremamente úteis, estão sujeitos a erros de coleta e muitas vezes não estão disponíveis facilmente. Por isto, muitas vezes não é simples analisar o comportamento do seu modelo para investigar a sua fidelidade ao fenômeno.

Como o comportamento de doenças infecciosas é descrito por leis dinâmicas que não são tão bem estabelecidas como as leis da física, muitas vezes a construção de um modelo adequado não é uma tarefa tão simples. Uma alternativa é desenvolver estratégias que permitam aprender sobre determinados modelos candidatos baseados em sua estrutura, interpretabilidade, dentre outros fatores. Ou seja, obter informações através do próprio modelo. Dentro desse objetivo, análise de sensibilidade pode ser uma grande fonte de informação sobre o comportamento dinâmico do modelo à luz da região paramétrica analisada e de como esta o influencia (Wu et al., 2013).

Apesar de análises de sensibilidade permitirem identificar os parâmetros que afetam a sua quantidade de interesse de forma mais direta e efetiva, isto não costuma ser levado em consideração na tomada de decisão por um determinado modelo em detrimento de outro. Normalmente busca-se orientar essa escolha se escolhe baseado em dados

⋆ Os autores agradecem o suporte financeiro trazido pelas agências de fomento FAPERJ (Fundação Carlos Chagas Filho de Amparo à Pesquisa), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior).

experimentais como se fossem a única fonte confiável de informação. Dependendo do interesse, poderia-se usar como critério para qual modelo a resposta é mais sensível a um determinado parâmetro chave como o número básico de reprodução (Diekmann et al., 1990), por exemplo. Outra alternativa seria avaliar uma métrica mais ampla de sensibilidade, distribuída nos diversos instantes de tempo dentro de uma janela de observação.

Neste trabalho apresenta-se uma estratégia para seleção de modelos epidemiológicos que utiliza análise de sensibilidade global para quantificar as entradas mais importantes de cada modelo para a quantidade de interesse desejada. Para isso será empregado o método dos índices de Sobol via expansão em polinômio caos, o que permite observar o efeito conjunto dos parâmetros de cada modelo, evitando um custo computacional alto e instabilidades numéricas. Daí, os modelos são selecionados utilizando o critério da informação de Akaike, atribuindo como função de verossimilhança o produto dos índices obtidos para cada parâmetro. Assim, a informação utilizada dentro do critério de decisão é o efeito de cada parâmetro para a resposta de seu respectivo modelo. Para isso, este paper está organizado em quatro seções: Na Seção 2 é descrito todo o ferramental matemático utilizado para compor a metodologia de seleção de modelos adotada. A Seção 3 permite observar essa estratégica em prática ao trazer alguns resultados para modelos encontrados na literatura para analisar a pandemia de COVID-19. Finalmente, a Seção 4 fecha o trabalho ao trazer algumas conclusões iniciais sobre os resultados e algumas direções futuras.

2. METODOLOGIA

2.1 Índices de Sobol

Análise de sensibilidade é uma estratégia interessante para observar o comportamento do modelo a depender das suas entradas. Em geral, se pretende identificar quais parâmetros são mais importantes para a resposta (Wu et al., 2013). Obviamente, esse resultado é altamente dependente da região do espaço parâmetro que foi observada e de que maneira ela foi explorada. Nesse sentido é importante destacar os métodos globais dos locais. Enquanto os últimos têm por característica serem mais dependentes do ponto específico observado, os primeiros tendem a trazer *insights* mais amplos, pois observam o comportamento típico da região paramétrica abordada (Saltelli et al., 2019). Além disso, métodos globais levam em conta o efeito conjunto dos parâmetros, isto é, não só o efeito de se variar cada entrada individualmente como também a consequência de quando essas mudanças são concomitantes.

Dentro da literatura de métodos de análise de sensibilidade global, esse trabalho faz uso da técnica de índices de Sobol, um método baseado em decomposição de variança que basicamente compara a variância gerada na resposta devido à um determinado conjunto de parâmetros em relação a variância total decorrente das mudanças em todos eles (Sobol, 1993). Em maiores detalhes, seja $\mathbf{x} \in \mathbb{R}^n$ o vetor de parâmetros do modelo e $y_t = \mathcal{M}(\mathbf{x}, t) \in \mathbb{R}$ a quantidade de interesse calculada em cada instante de tempo t a depender da entrada $\mathbf{x}$ pelo operador matemático $\mathcal{M}$. Denota-se então por $\mathbf{X} \sim f_{\mathbf{X}}$ o vetor aleatório de entradas, associado a distribuição de probabilidades $f_{\mathbf{X}}$ e de suporte $\mathcal{I}$, de maneira que

$$Y_t = \mathcal{M}(\mathbf{X}, t) \tag{1}$$

passa a ser um processo estocástico cuja variância será assumida como finita. Se, além disso, $\mathcal{M}$ for um operador integrável, a variância total de Y_t pode ser decomposta na forma da expressão

$$\mathrm{Var}(Y_t) = \sum_{\mathbf{u}} \mathrm{Var}\left(\mathcal{M}_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}}, t)\right), \quad \text{for } \varnothing \neq \mathbf{u} \subset \{1, \ldots, n\}, \tag{2}$$

onde $\mathrm{Var}\left(\mathcal{M}_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}}, t)\right)$ representa a variância condicional para o subvetor $\mathbf{X}_{\mathbf{u}}$ que reúne as variáveis cujos índices estão indicados pelo subconjunto $\mathbf{u}$, e cada parcela do conjunto de componentes $\mathcal{M}_{\mathbf{u}}$ pode ser calculada de forma recorrente em cada instante de tempo como descrito em Konakli and Sudret (2016).

Ao dividir-se cada componente de (2) pela variância total, obtém-se como resultados as expressões para os índices de Sobol de ordens 1 a n (a depender da ordem da componente) para cada parâmetro (Sobol, 2001),

$$S_u(t) = \frac{\mathrm{Var}\left(\mathcal{M}_{\mathbf{u}}(\mathbf{X}_{\mathbf{u}}, t)\right)}{\mathrm{Var}(Y_t)}, \quad \text{for } \varnothing \neq \mathbf{u} \subset \{1, \ldots, n\}, \tag{3}$$

permitindo assim que o efeito total gerado pela variação de um parâmetro x_i, seja, individualmente ou em conjunto a outas entradas, caracterizado pelo índice de Sobol total

$$S_i^T(t) = \sum_{\substack{\mathbf{u} \subset \{1,\ldots,n\} \\ i \in \mathbf{u}}} S_{\mathbf{u}}(t), \qquad i = 1, \cdots, n. \tag{4}$$

Novamente, enfatiza-se que, uma vez que essa discussão é feita dentro do contexto de sistemas dinâmicos, cada parâmetro possui um índice de Sobol em cada instante de tempo de análise.

2.2 Expansão em polinômio caos

Como se pode imaginar, uma simulação dos índices de Sobol gerada via método Monte Carlo pode ser muito custosa pelo número alto de simulações do modelo. Além disso, simulações de variância via Monte Carlo são conhecidas por gerarem erros de cancelamento, cujo efeito pode ser muito grave em alta ordem (Higham, 2002). Uma maneira de contornar ambos os problemas é utilizar um metamodelo baseado em expansão em polinômios caos (PCE) (Ghanem and Spanos, 1991; Xiu and Karniadakis, 2002). Considerando as n entradas de $\mathbf{X}$ como independentes, o PCE truncado, $\mathcal{M}^{PC}$, para aproximar a resposta estocástica trazida pelo operador $\mathcal{M}$ é definido por

$$Y_t \approx \mathcal{M}^{PC}(\mathbf{X}, t) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}_n} y_{\boldsymbol{\alpha}}(t)\, \Phi_{\boldsymbol{\alpha}}(\mathbf{X}), \tag{5}$$

com $\Phi_{\boldsymbol{\alpha}}(\mathbf{X})$ sendo as bases polinomiais ortonormais multivariadas, caracterizadas pela distribuição de entrada dos

parâmetros, $y_{\boldsymbol{\alpha}}(t)$ os coeficientes da expansão em cada instante, e o n-dimensional multi-índice $\boldsymbol{\alpha}$ indica a ordem de cada polinômio univariado nos termos $\Phi_{\boldsymbol{\alpha}}(\mathbf{X})$, dentro do conjunto de multi-índices truncado $\mathcal{A}_n \subset \mathbb{N}^n$ (Xiu, 2010). Substituindo a equação (5) em (1) obtém-se que os índices de Sobol via PCE podem ser obtidos analiticamente em cada instante de tempo pela expressão

$$S_{\mathbf{u}}(t) = \frac{\sum\limits_{\boldsymbol{\alpha}\in\mathcal{A}_{\mathbf{u}}} y^2_{\boldsymbol{\alpha}}(t)}{\sum\limits_{\boldsymbol{\alpha}\in\mathcal{A}_n\setminus\mathbf{0}} y^2_{\boldsymbol{\alpha}}(t)}\,, \tag{6}$$

sendo $\mathbf{u}$ o subconjunto de índices adequado e $\mathcal{A}_{\mathbf{u}}$ o conjunto de multi-índices que reúne apenas os termos da expansão truncada pertinentes. Ou seja, em cada instante de tempo, esse cálculo passa a depender unicamente dos coeficientes do modelo aproximado, possibilitando então reduzir tanto o custo computacional quanto os erros de cancelamento.

Existem diferentes maneiras para calcular esses coeficientes tal como uma gama de métricas de erro que permitem estimar a capacidade do PCE obtido em reproduzir a resposta estocástica do operador $\mathcal{M}$. Como uma descrição mais profunda nesse assunto foge ao escopo desse trabalho, recomenda-se Marelli and Sudret (2018) para maiores detalhes. Contudo, independente do método para obtenção do metamodelo, sempre é necessário iniciar calculando uma amostra do comportamento do modelo original. Para isto, é estabelecido um número, N_s, de amostras de cada variável aleatória X_i, para compor o chamado conjunto experimental

$$\boldsymbol{\chi} = \left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(N_s)}\right\}, \tag{7}$$

para o qual a resposta do modelo deve ser calculada de forma a caracterizar o seu comportamento:

$$\begin{aligned} y_t^{(1)} &= \mathcal{M}(\mathbf{x}^{(1)}, t)\,, \\ y_t^{(2)} &= \mathcal{M}(\mathbf{x}^{(1)}, t)\,, \\ &\vdots \\ y_t^{(N_s)} &= \mathcal{M}(\mathbf{x}^{(N_s)}, t)\,. \end{aligned} \tag{8}$$

Essa amostra é então utilizada no cálculo dos coeficientes do PCE e na estimação do erro de aproximação. O custo computacional em obter os índices de Sobol via PCE mora, portanto, nesta etapa de construção do PCE (Tosin et al., 2020b). A implementação computacional aqui então faz uso do template trazido no pacote SoBioS (Tosin et al., 2020a).

2.3 Critério de informação de Akaike

Dentro da literatura de seleção de modelos, o critério da informação de Akaike (AIC) é um classificador bem estabelecido. Firmado na teoria da informação, o AIC é conhecido por estabelecer um bom compromisso entre a qualidade de ajuste do modelo ao seu conjunto de dados experimentais quanto o princípio da parcimônia estatística (Akaike, 1973). Seja, novamente, n o número de parâmetros do modelo e, d a quantidade de dados usada, o critério de Akaike é calculado como

$$AIC = 2n - 2\,\ln\hat{L}\,, \tag{9}$$

com $\hat{L}$ indicando o valor de máxima verossimilhança atingido pelos parâmetros usando os dados. No entanto, caso o número de pontos dos dados seja muito pequeno em relação ao tamanho do vetor de parâmetros (entenda-se como muito pequeno quando $n > d/40$ (Sugiura, 1978)), deve haver uma correção na equação (9) levando em conta a baixa fonte de informação, evitando apontar exclusivamente para modelos com maior número de parâmetros. Obtém-se então a fórmula

$$AIC_c = AIC + \frac{2n(n+1)}{d-n-1} = \frac{2\,n\,d}{d-n-1} - 2\,\ln\hat{L}\,. \tag{10}$$

Como foi dito anteriormente, a ideia da estratégia seguida nesse trabalho é ter um critério de decisão numa etapa anterior ao uso de dados ou quando estes não estão disponíveis. Dentro desse objetivo, será usada a "função de verossimilhança" construída da seguinte forma:

$$\mathcal{L}(t_j) = \prod_{i=1}^{n} S_i^T(t_j)\,, \tag{11}$$

de forma que t_j indica os instantes observados na janela de tempo $\mathcal{T}$. Note que essa expressão é alimentada pela informação sobre a importância de cada parâmetro capturada pela análise de sensibilidade global baseada nos índices de Sobol totais calculados segundo (6) com o tempo cumprindo o papel de hiper-parâmetro para a função de verossimilhança. Em outras palavras, usa-se uma noção distribuída de sensibilidade da resposta ao invés do menor desajuste desta a um conjunto de dados reais. Ademais, o valor de máxima verossimilhança pode ser extraído a partir da equação abaixo:

$$\hat{L} = \max_{t_j\in\mathcal{T}} \mathcal{L}(t_j)\,, \tag{12}$$

isto é, utiliza-se o maior produto de índices de Sobol totais. A classificação se baseia em selecionar o modelo com menor valor para o classificador de Akaike descrito na equação (9) (análogo para os casos com (10)).

Apesar deste cálculo descrito em (12) ser simples, erros numéricos devem ser levados em consideração cuidadosamente, uma vez que se está lidando com o produto de valores possivelmente muito pequenos. Para reduzir o efeito desses no resultado final, pode-se adotar

$$\ln\hat{L} = \max_{t_j\in\mathcal{T}} \left[\,\ln\mathcal{L}(t_j)\right]\,. \tag{13}$$

Desta maneira a multiplicação de diversos valores da ordem de 10^{-1} ou menor é substituída pela soma de valores da ordem de 10^2 ou maior.

2.4 Esquematização do método

Após a descrição teórica de cada ferramenta utilizada durante o processo de classificação de modelos, resta descrever como elas são exploradas. Após escolher um conjunto de modelos candidatos, as distribuições de probabilidade para seus parâmetros devem ser adotadas, assim como os hiper-parâmetros pertinentes à cada distribuição. Daí, seleciona-se as quantidades necessárias para o cálculo do PCE como número de amostras, grau máximo para as bases polinomiais, método de cálculo dos coeficientes, etc. Com isso, pode-se calcular os modelos aproximados e, posteriormente, os índices de Sobol para cada parâmetro de cada modelo. Neste momento, as funções de verossimilhança podem ser obtidas juntamente com os valores de classificação. Finalmente, resta identificar o modelo com menor valor de classificação. A Figura 1 reúne as etapas do processo de classificação via índices de Sobol e critério da informação de Akaike descritas neste parágrafo para um grupo hipotético $\mathcal{P}$ de modelos candidatos, onde cada i-ésimo modelo é referido por $\mathcal{P}_i$.

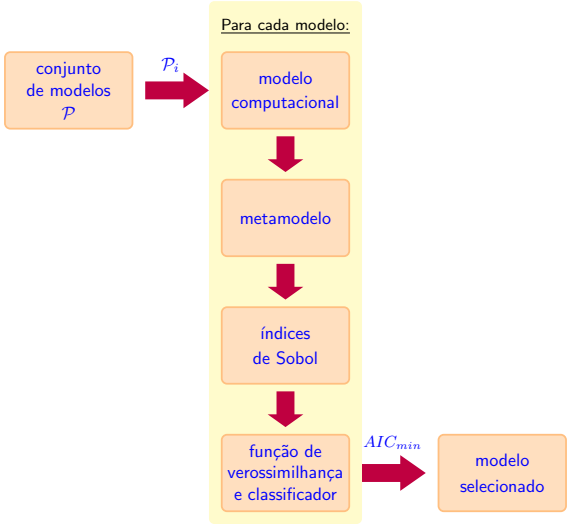

Figura 1. Representação esquemática do método de seleção de modelos adotado neste trabalho.

3. RESULTADOS

Para ilustrar o uso da metodologia de seleção de modelos via critério de Akaike com índices de Sobol, é apresentada uma família de candidatos adaptados de um mesmo modelo compartimental base, encontrado na literatura de estudos recentes sobre a pandemia de COVID-19 ao redor do mundo. A ideia central de modelagem compartimental é separar a população hospedeira da doença em compartimentos que indicam o estado de saúde de cada indivíduo com relação à doença, em cada instante de tempo. Conforme o tempo evolui, os indivíduos migram entre os compartimentos segundo um conjunto de leis evolutivas (neste caso, equações diferenciais).

3.1 Modelo SEAITR base

Os modelos candidatos usados neste trabalho são obtidos tomando como base o modelo compartimental SEAITR explorado por Okuonghae and Omame (2020) para extrair resultados no contexto pandêmico de COVID-19 em Lagos, Nigéria. Os autores referem-se ao modelo como $\mathrm{SEAII}_d\mathrm{R}$, mas aqui a notação para o compartimento I_d é substituída por T para evitar confusão com o compartimento I. Outras mudanças de notação foram feitos com o intuito de trazer maior clareza ao texto ao mesmo passo que se aproxime dos símbolos normalmente encontrados na literatura para representar os mesmos elementos. Apesar desses ajustes notacionais, o modelo SEAITR é versátil aos objetivos deste trabalho por contemplar diferentes aspectos observados na dinâmica de COVID-19 ao mesmo tempo que apresenta mantém um nível de complexidade que não rouba atenção demais para si dentro do trabalho.

O referido modelo SEAITR baseia-se numa adaptação do modelo compartimental SEIR tradicional (Brauer, 2017) pela adição de assintomáticos e testados. Em modelos deste tipo a ideia é separar a população hospedeira da doença em compartimentos que indicam o estado de saúde de cada indivíduo com relação à doença, em cada instante de tempo. Conforme o tempo evolui, os indivíduos migram entre os compartimentos segundo um conjunto de equações diferenciais. Neste caso, temos então 5 compartimentos (ou estados possíveis pela doença): suscetíveis, S, ou seja, aqueles que não tiveram contato prévio com a doença; expostos, E, que são os indivíduos que foram infectados pelo patógeno, mas que ainda não são capazes de infectar; infecciosos assintomáticos, A, definidos pela capacidade de infectar outros indivíduos apesar de não apresentarem os sintomas da doença; infecciosos sintomáticos, I, que se diferenciam dos assintomáticos por apresentarem os sintomas da doença, permitindo-se mais facilmente de serem identificados; testados, T, que são aqueles que, sintomáticos ou não, foram identificados por meio de testagem e se isolaram; recuperados, R, isto é, os que passaram pelas etapas de infecção e se recuperaram da doença. Vale destacar que, apesar das notícias recentes de reinfecção por COVID-19 (Ducharme, 2020; Simonato, 2020), por simplicidade será considerado que os recuperados adquirem imunidade à doença. Isso não é uma hipótese de modelagem muito forte dentro de uma janela de análise pequena. Pelo mesmo motivo, mudanças demográficas serão negligenciadas. Contudo, mortalidade pela doença será levada em conta. Pelo que foi descrito neste parágrafo, a evolução temporal dos compartimentos do modelo SEAITR pode ser analisada por meio do sistema de equações diferencias a seguir:

$$\begin{aligned}
\frac{\mathrm{d}S}{\mathrm{d}t} &= -S\,\frac{(\beta_A\,A+\beta_I\,I)}{N-T}\,,\\
\frac{\mathrm{d}E}{\mathrm{d}t} &= S\,\frac{(\beta_A\,A+\beta_I\,I)}{N-T}-\alpha\,E\,,\\
\frac{\mathrm{d}A}{\mathrm{d}t} &= \nu\,\alpha\,E-(\gamma_A+\theta_A)\,A\,,\\
\frac{\mathrm{d}I}{\mathrm{d}t} &= (1-\nu)\,\alpha\,E-(\gamma_I+\theta_I+\mu_I)\,I\,,\\
\frac{\mathrm{d}T}{\mathrm{d}t} &= \theta_A\,A+\theta_I\,I-(\gamma_T+\mu_T)\,T\,,\\
\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma_A\,A+\gamma_I\,I+\gamma_T\,T\,,
\end{aligned} \tag{14}$$

onde β indica as taxas de transmissão da doença, α a taxa de exposição, γ de recuperação, e μ de mortalidade pela

doença. A proporção de assintomáticos ν regula a diferença na quantidade de entrada entre os compartimentos de infecciosos não testados, enquanto que as taxas de testagem θ comandam a migração para esse compartimento. Os índices nos parâmetros permitem adicionar as mudanças numéricas entre parâmetros com mesma interpretação. Como existem mortes pela doença, existe variação na população total de humanos vivos N dada pela entrada de novas mortes em cada instante de tempo. Assim, podemos escrever a equação que observa a mudança no compartimento complementar de número de mortes, que usaremos como quantidade de interesse (QoI), como

$$\frac{\mathrm{d}D}{\mathrm{d}t} = \mu_I I + \mu_T T\,. \tag{15}$$

Considerando que os parâmetros desse tipo de modelo utilizam $dias^{-1}$ (a menos de ν que é adimensional) como unidade de medida, o número de mortes será calculado por dia. É importante observar que uma vez que os parâmetros serão analisados dentro de intervalos, o que diferencia o cenário de Lagos com o de outra região é o vetor de condições iniciais. De forma a adaptar o estudo ao município do Rio de Janeiro, a população inicial de indivíduos $N(0)$ será aproximado por 6.7×10^6 (Instituto Brasileiro de Geografia e Estatística, 2020). As demais condições iniciais foram assumidas e reunidas na Tabela 1, tal como os intervalos de valores para cada parâmetro pertinente ao modelo tratado. Os valores são semelhantes aos usados por Okuonghae and Omame (2020), ou assumidos baseado nos resultados obtidos por eles. Apesar de que valores particulares de parâmetros obviamente diferenciarem um cenário local de outro, pode-se trabalhar com a ideia de que os intervalos são semelhantes sem grandes perdas. Dentro da ótica dos parâmetros como variáveis aleatórias uma maneira de diferenciar os valores observados em cada região particular seria mudando as distribuições que caracterizam os parâmetros. Contudo não interessa a este paper particular entrar profundamente nesta questão.

Tabela 1. Intervalos para os parâmetros e condições inicias usadas no modelo 1.

Parâmetro	Intervalo de suporte
β_A	0.25 – 0.5
β_I	0.25 – 0.5
α	0 – 1
ν	0 – 1
γ_A	1/30 – 1/3
γ_I	1/30 – 1/3
γ_T	1/30 – 1/3
θ_A	0.0001 – 0.001
θ_I	0.0001 – 0.001
μ_I	0.001 – 0.1
μ_T	0.001 – 0.1
Condição inicial	**Valor**
$N(0)$	6.7×10^6
$E(0)$	25
$A(0)$	25
$I(0)$	25
$T(0)$	0
$R(0)$	0
$D(0)$	0
$S(0)$	N(0) - E(0) - A(0) - I(0) - T(0) - R(0) - D(0)

3.2 Modelos candidatos

A ideia é modificar os modelos removendo ou modificando seus parâmetros. Assim, a estrutura geral será a mesma assim como a quantidade de interesse (dada por (15)). Para diferenciar um modelo candidato dos demais, será indicada a adaptação adotada e o vetor de parâmetros atualizado.

Modelo 1: SEAITR base

$$\mathbf{x} = [\beta_A, \beta_I, \alpha, \nu, \gamma_A, \gamma_I, \gamma_T, \theta_A, \theta_I, \mu_I, \mu_T].$$

Modelo 2: SEAITR adaptado em γ_A

Considera-se que a taxa de recuperação para assintomáticos é a mesma que para sintomáticos:

$$\gamma_A = \gamma_I\,,$$

$$\mathbf{x} = [\beta_A, \beta_I, \alpha, \nu, \gamma_I, \gamma_T, \theta_A, \theta_I, \mu_I, \mu_T].$$

Modelo 3: SEAITR adaptado em θ_A

Considera-se que a taxa de testagem para assintomáticos é a mesma que para sintomáticos:

$$\theta_A = \theta_I\,,$$

$$\mathbf{x} = [\beta_A, \beta_I, \alpha, \nu, \gamma_A, \gamma_I, \gamma_T, \theta_I, \mu_I, \mu_T].$$

Modelo 4: SEAITR com remoção de θ_A

Considera-se que infeciosos assintomáticos não realizam testagem:

$$\theta_A = 0\,,$$

$$\mathbf{x} = [\beta_A, \beta_I, \alpha, \nu, \gamma_A, \gamma_I, \gamma_T, \theta_I, \mu_I, \mu_T].$$

Modelos 5,6: SEAITR adaptado em ν

Considera-se que a taxa de assintomáticos é uma constante previamente conhecida do modelo. Serão observados dois valores pontuais no intervalo [0,1]:

$$\nu \in \{0.25, 0.5\}\,,$$

o primeiro e segundo valores de ν serão vinculados aos modelos 5 e 6, respectivamente.

$$\mathbf{x} = [\beta_A, \beta_I, \alpha, \gamma_A, \gamma_I, \gamma_T, \theta_A, \theta_I, \mu_I, \mu_T].$$

3.3 Classificação

Tomando os valores descritos na Tabela 1, os índices de Sobol totais para cada um dos modelos são obtidos usando conjuntos experimentais de 500 amostras para o cálculo dos coeficientes do PCE via método de regressão de menor ângulo (Marelli and Sudret, 2018). O grau máximo permitido às bases polinomiais foi de 7 e o intervalo de tempo analisado foi de 30 dias após a condição inicial. Como aqui o papel de dados no critério de Akaike é exercido pelos índices de Sobol, têm-se 30 pontos de dados em cada análise, fazendo com que a equação (10) seja empregada. As funções de log-verossimilhança encontradas estão expostas nas Figura 2, em módulo. Pode-se observar que as distribuições possuem comportamentos qualitativo quantitativo bem similares. Pode-se observar uma clara tendência decrescente monotônica em todos os casos. Além disso, uma vez que os índices de Sobol são valores não negativos, e menores ou iguais que um, as medidas de log-verossimilhança são não positivas de forma que o maior valor assumido será o menor em módulo. Logo, o valor para a expressão (13) é sempre obtido no instante de tempo final mostrado. Como já é esperado que os valores de classificação sejam próximos entre si, algumas medidas de dispersão podem ser caracterizadas para permitir um olhar mais cirúrgico sobre as tendências trazidas pelos resultados de classificação encontrados nesta análise:

$$AIC_{mean} = \frac{\sum_{j=1}^{N_{\mathcal{P}}} AIC_{cj}}{N_{\mathcal{P}}}\,, \tag{16}$$

$$AIC_{min} = \min_{j}[AIC_{cj}]\,, \tag{17}$$

$$\sigma_j = \sqrt{(AIC_{cj} - AIC_{mean})^2}\,, \tag{18}$$

$$\Delta_j = \sqrt{(AIC_{cj} - AIC_{min})^2}\,, \tag{19}$$

usando $j \in \{1, 2, 3, 4, 5, 6\}$ como índice de referência aos modelos candidatos descritos anteriormente e $N_{\mathcal{P}} = 6$ para representar o número de modelos candidatos. Os valores calculados por essas expressões podem ser encontrados juntamente aos classificadores de Akaike na Tabela 2. Por questão de referência, cabe informar que os valores de média e mínimo obtidas neste estudo foram de $AIC_{mean} = 74.06$ e $AIC_{min} = 68.09$, respectivamente.

Tabela 2. Resultados de classificação obtidos para os modelos candidatos 1–6.

Modelo	AIC_c	σ	Δ
1	82.99	8.93	14.90
2	72.12	1.94	4.02
3	68.09	5.97	0.00
4	68.26	5.79	0.17
5	78.50	4.45	10.41
6	74.37	0.32	6.29

Pelo critério de mínimo valor para o classificador de Akaike, o modelo apontado será o modelo 3, isto é, o modelo SEAITR adaptado pela igualdade de valores $\theta_A = \theta_I$. Este é um cenário em que a taxa de testagem seria

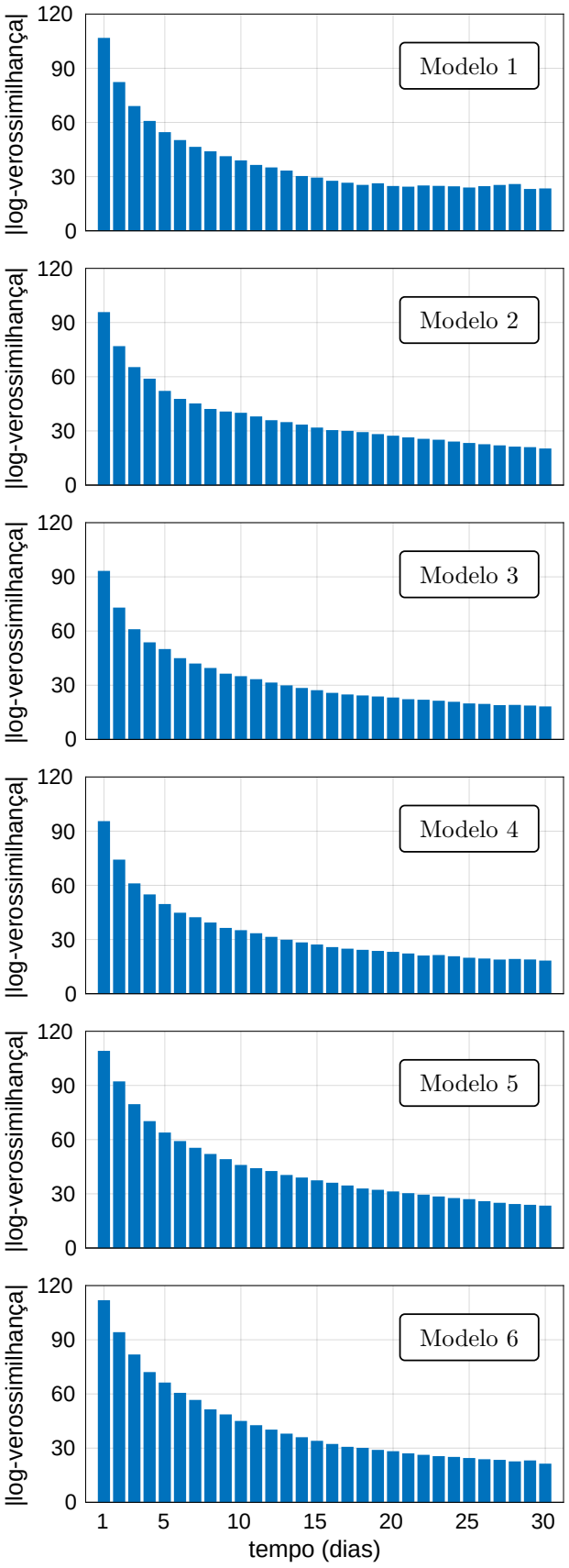

Figura 2. Funções de log-verossimilhança obtidas para o conjunto de modelos candidatos 1–6, tomando os valores presentes na Tabela 1.

indiferente entre sintomáticos e assintomáticos. Isto faria sentido teórico uma vez que os sintomas da COVID-19 tem se mostrado muito diversos, o que dificulta que um paciente possa reconhecê-los em si. Por isso, os indivíduos desses dois grupos de infecciosos teria em média uma mesma tendência a procurar testagem. O segundo modelo a ser selecionado pelo critério seria o modelo 4. Contudo, imaginar um cenário em que a testagem por assintomáticos seja desprezível é pouco provável e não muito desejado. No cenário pandêmico de COVID-19 é esperado que pessoas procurem testagem antes dos sintomas surgirem. Logo, mesmo um assintomáticos iria testar-se quando possível.

Por fim, resta observar-se os resultados à luz das medidas de dispersão. Apesar do modelo 3 ser selecionado, ele possui um desvio da média muito alto. Nota-se também que os modelos mais dispersos do mínimo geralmente são os que dispersam pouco da média. Isto ocorre devido aos modelos 1,5,6 que puxam a média pra cima. Como tem-se três modelos equilibrados numa faixa de valores abaixo da média e três deles reunidos acima da média, a média não é tao representativa nesse caso. Mesmo assim, caso a média fosse levada em consideração, o modelo que mantém um melhor compromisso às duas medidas de dispersão é o modelo 2. Num cenário em que o mínimo e a média são bem próximos, a escolha pelo candidato que melhor equilibra os dois desvios é mais recomendado do que meramente optar pelo de valor mínimo.

4. CONCLUSÃO

Uma metodologia para seleção de modelos usando de critério da informação de Akaike combinado a análise de sensibilidade global por índices de Sobol foi descrita nesse trabalho. O método permitiu selecionar, dentre um grupo de 6 modelos epidemiológicos compartimentais, aquele cuja testagem por sintomáticos ou assintomáticos não era diferenciada. O modelo 2 poderia ser escolhido caso se levasse em conta os desvios, porém o desvio da média não parece ser uma boa métrica neste caso. O modelo separado como resultado mostrou-se consistente à realidade atual da COVID-19 no Rio de Janeiro. Assim, pôde-se observar que esta metodologia de classificação de modelos epidemiológicos é eficiente para casos em que conjuntos de dados não são confiáveis ou inexistentes. Contudo, deve-se tomar o cuidado de não comparar modelos cujas quantidades de parâmetros seja muito distantes para evitar enganos. Obviamente, caso isto seja necessário, as medidas de dispersão devem ser suficientes para revelar esse enviesamento no resultado sem grandes problemas. Em trabalhos futuros pretende-se trazer novos estudos com mais modelos e forma a selecionar um grupo de modelos interessantes ao invés de apenas um de forma a ter opções para análises posteriores à classificação. Com isso, o processo descrito neste trabalho seria uma etapa de um processo maior para analisar cenários epidemiológicos à luz dos modelos que mostrarem-se mais interessantes. Adicionalmente, é também de interesse comparar os resultados obtidos para diferentes quantidades de interesse, sendo uma delas dada por um função de desajuste do modelo a certos dados amostrais. Desse modo a metodologia poderia ser utilizada tanto na ausência da dados reais quanto fazendo um uso diferente destes.

REFERÊNCIAS

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of Second international symposium on information theory*, 267–281.

Brauer, F. (2017). Mathematical epidemiology: past, present, and future. *Infectious Disease Modelling*, 2(2), 113–127. doi:10.1016/j.idm.2017.02.001.

Diekmann, O., Heesterbeek, J.A.P., and Metz, J.A.J. (1990). On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4), 365–382. doi:10.1007/BF00178324.

Ducharme, J. (2020). A new study suggests covid-19 reinfection is possible. here's what to know. `https://time.com/5882907/covid-19-reinfection/`. Acessado: 31/08/2020.

Ghanem, R.G. and Spanos, P.D. (1991). *Stochastic finite elements: a spectral approach*. Sringer, New York. doi:10.1007/978-1-4612-3094-6.

Higham, N.J. (2002). *Accuracy and stability of numerical algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, 2^a edition. doi:10.1137/1.9780898718027.

Instituto Brasileiro de Geografia e Estatística (2020). Cidades e estados: Rio de Janeiro, código: 3304557. `https://www.ibge.gov.br/cidades-e-estados/rj/rio-de-janeiro.html`. Acessado: 01/10/2020.

Konakli, K. and Sudret, B. (2016). Global sensitivity analysis using low-rank tensor approximations. *Reliability Engineering and System Safety*, 156, 64–83. doi:10.1016/j.ress.2016.07.012.

Kuhl, E. (2020). Data-driven modeling of covid-19–lessons learned. *Extreme Mechanics Letters*, 40, 100921. doi:10.1016/j.eml.2020.100921.

Marelli, S. and Sudret, B. (2018). *UQLab user manual – polynomial chaos expansions*. Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich.

Okuonghae, D. and Omame, A. (2020). Analysis of a mathematical model for covid-19 population dynamics in lagos, nigeria. *Chaos, Solitons and Fractals*, 139, 110032. doi:10.1016/j.chaos.2020.110032.

Saltelli, A., Aleksankinac, K., Beckerd, W., P. Fennelle, F.F., Holstf, N., Lig, S., and Wuh, Q. (2019). Why so many published sensitivity analyses are false: a systematic review of sensitivity analysis practices. *Environmental Modelling and Software*, 114, 29–39. doi:10.1016/j.envsoft.2019.01.012.

Simonato, S. (2020). Hc estuda casos de 7 pacientes que podem ter se reinfectado por coronavírus em são paulo: médica do abc relata que voltou a sentir os sintomas meses depois de ter sido infectada pela primeira vez. `https://g1.globo.com/sp/sao-paulo/noticia/2020/08/25/hc-estuda-casos-de-7-pacientes-que-podem-ter-se-reinfectados-por-coronavirus-em-sao-paulo.ghtml`. Acessado: 31/08/2020.

Sobol, I.M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling in Civil Engineering*, 1(4), 407–414.

Sobol, I.M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and Computers in Simulation*, 55, 271–

280. doi:10.1016/j.ress.2016.07.012.

Sugiura, N. (1978). Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics – Theory and Methods*, 7, 13–26. doi:10.1080/03610927808827599.

Toscano, G., Palmerini, F., Ravaglia, S., Ruiz, L., Invernizzi, P., Cuzzoni, M.G., Franciotta, D., Baldanti, F., Daturi, R., Postorino, P., Cavallini, A., and Micieli, G. (2020). Guillain–barré syndrome associated with sars-cov-2. *New England Journal of Medicine*, 382(26), 2574–2576. doi:10.1056/NEJMc2009191.

Tosin, M., Cortês, A.M.A., and Cunha Jr, A. (2020a). SoBioS - Sobol' indices for biological systems. `https://github.com/americocunhajr/SoBioS`.

Tosin, M., Cortês, A.M.A., and Cunha Jr, A. (2020b). *A tutorial on Sobol' global sensitivity analysis applied to biological models*, 93–118. Springer International Publishing. doi:10.1007/978-3-030-51862-2_6.

Wiratsudakul, A., Suparit, P., and Modchang, C. (2018). Dynamics of zika virus outbreaks: an overview of mathematical modeling approaches. *PeerJ*, 6, e4526. doi:10.7717/peerj.4526.

World Health Organization (2020). WHO Coronavirus disease (COVID-19) dashboard. `https://covid19.who.int/`. Acessado: 13/10/2020.

Wu, J., Dhingra, R., Gambhir, M., and Remais, J.V. (2013). Sensitivity analysis of infectious disease models: methods, advances and their application. *Journal of the Royal Society Interface*, 10(86), 20121018. doi:10.1098/rsif.2012.1018.

Xiu, D. (2010). *Numerical methods for stochastic computations: a spectral method approach*. Princeton University Press, New Jersey.

Xiu, D. and Karniadakis, G.E. (2002). The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2), 619–644. doi:10.1137/S1064827501387826.