Model-based evaluation of transmissibility and re-infection for the P.1 variant of the SARS-CoV-2 - Supplementary Material

Renato Mendes Coutinho[1,7], Flávia Maria Darcie Marquitti[2,7], Leonardo Souto Ferreira[3,7], Marcelo Eduardo Borges[7], Rafael Lopes Paixão da Silva[3,7], Otavio Canton[3,7], Tatiana Pineda Portella[4,7], Silas Poloni Lyra[3,7], Caroline Franco[3,7], Antonio Augusto Moura da Silva[6,7], Roberto André Kraenkel[3,7], Maria Amélia de Sousa Mascena Veras[5,7], and Paulo Inácio Prado[4,7]

[1]*Centro de Matemática, Computação e Cognição - Universidade Federal do ABC, Santo André, SP, Brazil*
[2]*Intituto de Biologia and Instituto de Física "Gleb Wataghin" - Universidade Estadual de Campinas, Campinas, SP, Brazil*
[3]*Instituto de Física Teórica - Universidade Estadual Paulista "Júlio de Mesquita Filho", São Paulo, SP, Brazil*
[4]*Instituto de Biociências - Universidade de São Paulo, São Paulo, SP, Brazil*
[5]*Departamento de Saúde Coletiva - Faculdade de Ciências Médicas da Santa Casa de São Paulo, São Paulo, SP, Brazil*
[6]*Universidade Federal do Maranhão, São Luis, MA, Brazil*
[7]*Observatório COVID-19 BR*

In order to estimate key parameters of the variant of concern (VOC) P.1, we developed a model and fitted it to time-series data of number of hospitalized cases and frequency of the P.1 variation. Section I describes the model, section II relates the values of the parameters taken from the current literature, III describes the contact matrix used, and finally section IV describes the treatment of case data (subsec. IV-A), the choice of initial conditions (subsec. IV-B), and the fitting procedure (subsec. IV-C).

I. Model Equations

The model is an extended *SEIR* model that comprises susceptible (S), pre-symptomatic (E), asymptomatic (A), mild symptomatic (I), severe/hospitalized (H), recovered (R) and deceased (D) compartments. These compartments are duplicated to account for a second variant of SARS-CoV-2, and each of them comprises three age classes: young ($<$ 20 years), adults (20-59 years), and the elderly ($\geq$ 60 years).

We assume that the second variant is capable of reinfecting individuals who have recovered from infection by the wild variant while the inverse is not possible; in the absence of data indicating this possibility, allowing reinfection by the wild variant on recovered of infection by P.1 would have negligible effect due to the small time window (3 months) considered in the present work. We also consider that a variant is not capable of reinfecting individuals recovered from the same lineage. Our model does not include vaccination due to the lack of vaccines in Brazil during the time span of our study.

To model the virus spread in the population, we assume that asymptomatic individuals have equal infectiousness compared to symptomatic ones, while pre-symptomatic

individuals have reduced infectiousness given by ω. To model behaviour, we assume that symptomatic individuals self-isolate themselves to some degree, reducing their contacts by ξ. Individuals with severe disease have greater isolation ξ_{sev} due to hospitalization. The daily contacts between each age class is given by the matrix $\hat{C}$. The force of infection λ_k for each variant k is given below:

$$\lambda_k = \beta_k \hat{C}[A_k + \omega E_k + (1-\xi)I_k + (1-\xi_{sev})H_k]$$

The complete system of equations is given by:

Completely Susceptible

$$\frac{dS}{dt} = -\lambda_1 \frac{S}{N} - \lambda_2 \frac{S}{N} \tag{1a}$$

Wild variant

$$\frac{dE_1}{dt} = \lambda_1 \frac{S}{N} - \frac{E_1}{\gamma_1} \tag{1b}$$

$$\frac{dA_1}{dt} = \frac{(1-\sigma_1)\alpha_1 E_1}{\gamma_1} - \frac{A_1}{\nu_{i,1}} \tag{1c}$$

$$\frac{dI_1}{dt} = \frac{(1-\alpha_1)(1-\sigma_1)E_1}{\gamma_1} - \frac{I_1}{\nu_{i,1}} \tag{1d}$$

$$\frac{dH_1}{dt} = \frac{\sigma_1 E_1}{\gamma_1} - \frac{H_1}{\nu_{s,1}} \tag{1e}$$

$$\frac{dR_1}{dt} = \frac{A_1}{\nu_{i,1}} + \frac{I_1}{\nu_{i,1}} + \frac{(1-\mu_1)H_1}{\nu_{s,1}} - p_r \lambda_2 \frac{R_1}{N} \tag{1f}$$

$$\frac{dD_1}{dt} = \frac{\mu_1 H_1}{\nu_{s,1}} \tag{1g}$$

P.1 variant

$$\frac{dE_2}{dt} = \lambda_2 \frac{S}{N} - \frac{E_2}{\gamma_2} + p_r \lambda_2 \frac{R_1}{N} \tag{1h}$$

$$\frac{dA_2}{dt} = \frac{(1-\sigma_2)\alpha_2 E_2}{\gamma_2} - \frac{A_2}{\nu_{i,2}} \tag{1i}$$

$$\frac{dI_2}{dt} = \frac{(1-\alpha_2)(1-\sigma_2)E_2}{\gamma_2} - \frac{I_2}{\nu_{i,2}} \tag{1j}$$

$$\frac{dH_2}{dt} = \frac{\sigma_2 E_2}{\gamma_2} - \frac{H_2}{\nu_{s,2}} \tag{1k}$$

$$\frac{dR_2}{dt} = \frac{A_2}{\nu_{i,2}} + \frac{I_2}{\nu_{i,2}} + \frac{(1-\mu_2)H_2}{\nu_{s,2}} \tag{1l}$$

$$\frac{dD_2}{dt} = \frac{\mu_2 H_2}{\nu_{s,2}} \tag{1m}$$

Supplementary Equations

$$\frac{dC_1}{dt} = \chi(1-\alpha_1)\sigma_1 \frac{E_1}{\gamma_1} \tag{1n}$$

$$\frac{dC_2}{dt} = \chi(1-\alpha_2)\sigma_2 \frac{E_2}{\gamma_2}\ , \tag{1o}$$

where C_1 ad C_2 are the cumulative hospitalization cases reported, and each variable of the system (S, $E_k, ...,$ C_k) is actually a vector containing each age class, e.g., $E_1 = (E_{1,y}, E_{1,a}, E_{1,e})^T$. The equations were numerically solved by the R package developed by Soetaert et al. (11).

II. Parameterization of the model

In Table I we have the parameters considered for the wild variant. The parameters for the P.1 variant are the same except for those considered in the model fitting.

TABLE I: Epidemiological parameters

Parameter	Description	Value	Source
γ	Average time in days between being infected and developing symptoms	5.8	Wei et al. (13)
ν_i	Average time in days between being infectious and recovering for asymptomatic and mild cases	9.0	Cevik et al. (2)
ν_s	Average time between being infectious and recovering/dying for severe cases	8.4	SIVEP-Gripe for São Paulo State
ξ	Reduction on the exposure of symptomatic cases (due to symptoms/quarantining)	0.1	Assumed
ξ_{sev}	Reduction on the exposure of severe cases (due to hospitalization)	0.9	Assumed
ω	Relative infectiousness of pre-symptomatic individuals	1.0	Assumed
α	Proportion of asymptomatic cases	[0.67,0.44,0.31]	Juvenile (3) Adult and Elderly (12)
σ	Proportion of symptomatic cases that require hospitalization	[0.001,0.012,0.089] [a]	Salje et al. (10)
μ	In-hospital mortality ratio	[0.417,0.188,0.754]	Portella et al. (7)
χ	Case report probability	1.0	Assumed

[a] The proportion is weighted by the age distribution of the population.

III. Contact Matrices

Our model includes three age group categories, namely Young ($[0-19]\, y.o.$), Adults ($[20-59]\, y.o.$) and Elderly (greater than $60 y.o.$). To model contacts between these groups we use estimated contact matrices computed by Prem et al. (8), but since the original matrices use five-year age bins going up to 95+ years, we aggregate classes leading to a 3×3 matrix in the following way:

Let A, B be sets of indexes forming age groups (not necessarily of equal sizes), $x_{i,j}$ denoting contact between age groups i and j in the original matrix, d_i denoting population size of the age group i, then the new contact matrix $\hat{C}$ is given by:

$$\hat{C}_{A^*,B^*} = \frac{\sum_{i \in A} \sum_{j \in B} d_i x_{i,j}}{\sum_{i \in A} d_i} \tag{2}$$

where A^*, B^* denotes a new indexation rule. Note that the contact matrices depend on local demographics and therefore must be computed for each place of study.

IV. Data Analysis Procedures

A. Nowcasting

Data used in parameters estimation were collected from the national public health system of severe acute respiratory illness (SARI) surveillance database, named *Sistema de Vigilância da Gripe - SIVEP-Gripe*. In fact, reporting of cases can be delayed for several reasons, including the notification system itself and confirmation of RT-PCR test results. The nowcasting procedure estimates, based on the past delay distribution, the number of cases that already occurred but were not yet reported. A window of 10 weeks is the acting window on the series, since delays greater than this are rare.

Nowcasting requires a pair of dates: (i) onset date of the event and (ii) report date of the event. The delay distribution is modeled as being best described as a Poisson distribution for days since the onset date to the report date. We considered *the first symptoms date* as the onset date. For the report date, we used the latest between *the test result date* and *the clinical classification date*. The nowcasting algorithm were developed by McGough et al 2020 (5), and implemented in the *NobBS* (Nowcasting by Bayesian Smoothing) package in R.

B. Initial Condition Estimation

The model requires appropriate mid-epidemic initial conditions in order to give relevant results. In the model, the number of new hospitalizations at a given time – h_{new}, is directly proportional to the number of exposed individuals at that time, therefore data was used to get an approximation of the number of exposed people. Also, to quantify the number of people belonging to the recovered class, prevalence was used.

We can estimate the appropriate initial conditions by finding an approximation for our model that relates more directly to the available data in each class. In the absence of the variant P.1, the model has four classes of infected compartments, namely $\mathbf{y} = (E_1,\ A_1,\ I_1,\ H_1)^T$, and another three classes, represented by $\mathbf{z}$, i.e., $\mathbf{z} = (S,\ R_1,\ D_1)^T$.

To that effect, we can write the system as

$$\dot{\mathbf{y}} = F(\mathbf{y}, \mathbf{z}) - G(\mathbf{y}, \mathbf{z}), \tag{3}$$

$$\dot{\mathbf{z}} = J(\mathbf{y}, \mathbf{z}) \ , \tag{4}$$

where F comprises all entries of new Infected, coming from classes $\mathbf{z}$, whilst G accounts for the transitions within infected classes and also recovery and death from the disease. Then, to find a good approximation for a small time window, we perform a linearization of our model around a point $(\mathbf{y}, \mathbf{z})$. Keeping $\mathbf{z}$ fixed, we get

$$\dot{\mathbf{y}} = (\hat{F} - \hat{G})\mathbf{y} \ , \tag{5}$$

where $\hat{F}$ and $\hat{G}$ are the linearized matrices arising from the functions F and G, respectively. The only entrance of new infected comes from the $\beta S\lambda/N$ terms in the $\dot{E}_1 = (\dot{E}_{1,y}, \dot{E}_{1,a}, \dot{E}_{1,e})^T$ equations (sub-indexes are y young, a adults and e elderly), then, the only non-zero elements of $\hat{F}$ are in its first 3 lines. Before proceeding, it's useful to define

$$\hat{b} = \mathrm{diag}(S)\hat{C} \tag{6}$$

which allow us to write

$$\hat{F} = \frac{\beta}{N} \begin{bmatrix} \omega\hat{b} & \hat{b} & (1-\xi)\hat{b} & (1-\xi_{sev})\hat{b} \\ & & \mathbb{0}_{9,12} & \end{bmatrix} \tag{7}$$

$\hat{G}$ contains the terms of Exposed, E_1, the 3 possible forms of the disease considered in the model, that is A_1, I_1 and H_1, as the terms in its first 3 rows, whilst the remainder of its main diagonal contains terms of recovery and death. For simplicity, every constant (or vector for the terms with σ) in $\hat{G}$ expression (8) should be thought as diagonal matrices with its elements given by the constants (or vectors) and every $\mathbb{0}$ is a 3-dimensional square matrix where all entries are null.

$$\hat{G} = \begin{bmatrix} \gamma^{-1} & \mathbb{0} & \mathbb{0} & \mathbb{0} \\ -\alpha(1-\sigma)\gamma^{-1} & \nu_i^{-1} & \mathbb{0} & \mathbb{0} \\ -(1-\alpha)(1-\sigma)\gamma^{-1} & \mathbb{0} & \nu_i^{-1} & \mathbb{0} \\ -\sigma\gamma^{-1} & \mathbb{0} & \mathbb{0} & \nu_s^{-1} \end{bmatrix} \tag{8}$$

The linearization above implies that, for a small time interval, $\mathbf{y}$ has an exponential behavior and that the eigenvalues of $\hat{L} = \hat{F} - \hat{G}$ are related to the exponential growth rates. Therefore, a short time after the beginning of the epidemic, the largest eigenvalue should be the one to dominate. So the exponential growth rate of the wild variant – r, can be matched to the largest eigenvalue of $\hat{L}$ to obtain an estimate for β. The eigenvector associated with the largest eigenvalue gives the proportions of infected classes, which, together with the estimated number of exposed individuals – $E_1 = \gamma_1 h_{new}/\sigma_1$, results in an approximation for the number of people in the other infected classes.

Given a β, the largest eigenvalue of the linearization matrix is computed using the `eigs` function of the R package *rARPACK* (9) and we find the β that gives r as the largest eigenvalue through bisection root finding. Finally, subtracting the number of recovered and infected from the total population gives the number of susceptible individuals.

C. Maximum Likelihood Estimation

Given the cumulative daily curves of hospitalization for wild variant, C_1, and P.1 variant, C_2, we can obtain the daily variation of each curve, namely ΔC_1^t and ΔC_2^t. Those curves are summed up to give the total number of weekly new cases:

$$\Delta C^\tau = \sum_{i=1}^{7} (\Delta C_1^{\tau-1+i} + \Delta C_2^{\tau-1+i}) \tag{9}$$

where τ is a discrete index given in weeks.

To calculate the frequency of P.1 in a given time period T, we use the proportion of new cases in this period from the wild and P.1 variant as follows:

$$P^{t'} = \frac{\sum_{i=1}^{T} \Delta C_2^{T-1+i}}{\sum_{i=1}^{T} \Delta C_1^{T-1+i} + \sum_{i=1}^{T} \Delta C_2^{T-1+i}} \tag{10}$$

where t' is a discrete index given in T periods.

The time period T depends on the dataset of genome sequences: it is daily in Faria et al. (4) and monthly in Naveca et al. (6).

Using maximum likelihood, we fitted the model by estimating five parameters, namely, the relative transmissibility ($\Delta\beta$), the reinfection probability of P.1 (p), initial total prevalence ($\rho^0 = [R/N]_{t=0}$), initial fraction of cases that were caused by the new variant (P^0), and intrinsic growth rate of the wild variant (r). The parameter r incorporates effects related to contact rates for the wild variant, such as non-pharmacological interventions relaxation, elections, and others.

Hospitalization counts were assumed to follow a Poisson distribution, with expected value given by equation (9). The recorded number of P.1 in clinical samples was assumed to follow a binomial distribution with an expected value equal to the product of the total number of genome sequences sampled in each date and the proportion of P.1 cases (equation (10)). The log-likelihood function for the model fitting was then:

$$\mathcal{L} = \sum_i \log \mathrm{Pois}(x^i|\lambda = C^i) + \sum_j \log \mathrm{Bin}(y^j|N = n^j, \theta(\pi^j) = P^j) \ , \qquad (11)$$

where Pois is a Poisson distribution with parameter λ, x^i is the number of recorded hospitalizations in week i, Bin is a Binomial distribution with parameters N (total number of trials) and π^j (probability of success at each trial), n^j is total number of sequences in clinical samples in week or day j, y^j is the number of P.1 sequences in each of these samples, and $\theta(.)$ is the *logit* function.

The model was then fitted by finding the values of the five above mentioned parameters that maximize the log-likelihood function (equation 11), using the function `mle2`, from the R package *bbmle* (1).

To find starting values for the optimization performed by `mle2` we calculated the log-likelihood function for a sample of one million combinations of parameters values within reasonable ranges. The sets of parameters values that provided the two higher values of the log-likelihood function were then used as starting values for the computational optimization.

The confidence intervals for the expected number of cases and frequency were estimated from 10000 parametric bootstrap samples assuming that the estimated parameters follow a mulivariate normal distribution. The parameters of these multivariate distributions were the estimated values and estimated variance-covariance matrix of the parameters. For each sampled combination of parameters, the expected values were calculated and the confidence interval was estimated as the the 2.5% and 95% quantiles of the distribution of bootstrapped expected values.

1) Sensitivity analysis: The model fitting assumed a constant Infection Hospitalization Rate (IHR, parameter σ) for each age group over time for both variants. An increase in IHR caused by P.1 would increase the number of hospitalizations even without any increase in transmissibility or reinfection. Since the pathogenicity of the P.1 variant is unknown, the model fitting was repeated assuming that the odds ratio of the IHR in each age class was twice for P.1 cases compared to wild variant cases (SA1). Moreover, as the collapse of Manaus health system hindered hospitalizations of new severe cases and may have affected case recording in surveillance databases, the model fitting was repeated considering only the period prior to the collapse (10 January 2021) (SA2).

REFERENCES

[1] B. Bolker and R Development Core Team. *bbmle: Tools for General Maximum Likelihood Estimation*, 2020. URL https://CRAN.R-project.org/package=bbmle. R package version 1.0.23.1.

[2] M. Cevik, M. Tate, O. Lloyd, A. E. Maraolo, J. Schafers, and A. Ho. SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *The Lancet Microbe*, 2(1):e13–e22, jan 2021. ISSN 26665247. doi:10.1016/S2666-5247(20)30172-5. URL https://linkinghub.elsevier.com/retrieve/pii/S2666524720301725.

[3] S. M. de Saúde Município de São Paulo. Inquérito sorológico para Sars-Cov-2: Prevalência da infecção em escolares das redes públicas e privada da cidade de São Paulo. http://www.capital.sp.gov.br/arquivos/pdf/2021/coletiva_saude_14-01.pdf, 2021. [Online; accessed 31-January-2021].

[4] N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. d. S. Candido, S. Mishra, M. A. E. Crispim, F. C. Sales, I. Hawryluk, J. T. McCrone, R. J. G. Hulswit, L. A. M. Franco, M. S. Ramundo, J. G. de Jesus, P. S. Andrade, T. M. Coletti, G. M. Ferreira, C. A. M. Silva, E. R. Manuli, R. H. M. Pereira, P. S. Peixoto, M. U. Kraemer, N. Gaburo Jr, C. d. C. Camilo, H. Hoeltgebaum, W. M. Souza, E. C. Rocha, L. M. de Souza, M. C. de Pinho, L. J. T. Araújo, F. S. V. Malta, A. B. de Lima, J. d. P. Silva, D. A. G. Zauli, A. C. d. S. Ferreira, R. P. Schnekenberg, D. J. Laydon, P. G. T. Walker, H. M. Schlüter, A. L. P. dos Santos, M. S. Vidal, V. S. Del Caro, R. M. F. Filho, H. M. dos Santos, R. S. Aguiar, J. L. P. Modena, B. Nelson, J. A. Hay, M. M. Monod, X. Miscouridou, H. Coupland, R. Sonabend, M. Vollmer, A. Gandy, M. A. Suchard, T. A. Bowden, S. L. K. Pond, C.-H. Wu, O. Ratmann, N. M. Ferguson, C. Dye, N. J. L. Loman, P. Lemey, A. Rambaut, N. A. Fraiji, M. d. P. S. S. Carvalho, O. G. P. Pybus, S. Flaxman, S. Bhatt, and E. C. Sabino. Genomics and epidemiology of a novel SARS-CoV-2 lineage in Manaus , Brazil. 2021. URL https://github.com/CADDE-CENTRE/Novel-SARS-CoV-2-P1-Lineage-in-Brazil/tree/main/manuscript.

[5] S. F. McGough, M. A. Johansson, M. Lipsitch, and N. A. Menzies. Nowcasting by bayesian smoothing: A flexible, generalizable model for real-time epidemic tracking. *PLoS computational biology*, 16(4):e1007735, 2020.

[6] F. Naveca, V. Nascimento, V. Souza, A. Corado, F. Nascimento, G. Silva, Á. Costa, D. Duarte, K. Pessoa, M. Mejía, M. Brandão, M. Jesus, L. Gonçalves, C. da Costa, V. Sampaio, D. Barros, M. Silva, T. Mattos, G. Pontes, L. Abdalla, J. Santos, I. Arantes, F. Dezordi, M. Siqueira, G. Wallau, P. Resende, E. Delatorre, T. Gräff, and G. Bello. COVID-19 epidemic in the brazilian state of amazonas was driven by long-term persistence of endemic SARS-CoV-2 lineages and the recent emergence of the new variant of concern p.1. *Preprint*, Feb. 2021. doi:10.21203/rs.3.rs-275494/v1. URL https://doi.org/10.21203/rs.3.rs-275494/v1.

[7] T. P. Portella, S. R. Mortara, R. Lopes, A. Sánchez-Tapia, M. R. Donalísio, M. C. Castro, V. R. Venturieri, C. G. Estevam, A. F. Ribeiro, R. M. Coutinho, M. A. de Sousa Mascena Veras, P. I. Prado, and R. A. Kraenkel. Temporal and geographical variation of COVID-19 in-hospital fatality rate in brazil. Feb. 2021. doi:10.1101/2021.02.19.21251949. URL https://doi.org/10.1101/2021.02.19.21251949.

[8] K. Prem, K. van Zandvoort, P. Klepac, R. M. Eggo, N. G. Davies, A. R. Cook,

M. Jit, et al. Projecting contact matrices in 177 geographical regions: an update and comparison with empirical data for the covid-19 era. *medRxiv*, 2020.

[9] Y. Qiu, J. Mei, and M. Y. Qiu. Package 'rARPACK'. 2016.

[10] H. Salje, C. T. Kiem, N. Lefrancq, N. Courtejoie, P. Bosetti, J. Paireau, A. Andronico, N. Hozé, J. Richet, C.-L. Dubost, Y. L. Strat, J. Lessler, D. Levy-Bruhl, A. Fontanet, L. Opatowski, P.-Y. Boelle, and S. Cauchemez. Estimating the burden of SARS-CoV-2 in france. *Science*, 369(6500):208–211, May 2020. doi:10.1126/science.abc3517. URL https://doi.org/10.1126/science.abc3517.

[11] K. Soetaert, T. Petzoldt, and R. W. Setzer. Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, 33(9):1–25, 2010. ISSN 1548-7660. doi:10.18637/jss.v033.i09. URL http://www.jstatsoft.org/v33/i09.

[12] W. W. Sun, F. Ling, J. R. Pan, J. Cai, Z. P. Miao, S. L. Liu, W. Cheng, and E. F. Chen. Epidemiological characteristics of COVID-19 family clustering in Zhejiang Province. *Chinese journal of preventive medicine*, 54(6):625–629, 2020. ISSN 02539624. doi:10.3760/cma.j.cn112150-20200227-00199.

[13] Y. Wei, L. Wei, Y. Liu, L. Huang, S. Shen, R. Zhang, J. Chen, Y. Zhao, H. Shen, and F. Chen. A systematic review and meta-analysis reveals long and dispersive incubation period of covid-19. *medRxiv*, 2020.