Uncertainty reduction in logistic regressions: a COVID-19 case-study using surrogate locations' asymptotic values

Bruno Hebling Vieira^{a,c}, Nathalia Hanna Hiar^b, George Cunha Cardoso^{a,*}

^a*Department of Physics, FFCLRP, University of São Paulo, Ribeirão Preto, 14040-901, Brazil*

^b*Department of Biology, FFCLRP, University of São Paulo, Ribeirão Preto, 14040-901, Brazil*

^c*InBrain Lab, FFCLRP, University of São Paulo, Ribeirão Preto, Brazil*

Abstract

Logistic regressions are subject to high uncertainty when the data are not past the inflection point. For example, for logistic regressions estimated with data up to or before inflection point, uncertainties in the upper asymptotic value K can be of the same order of magnitude of the population under analysis. This paper presents a method for uncertainty reduction in logistic regression using data from a surrogate logistic process. We illustrate the procedure using the Richards' growth function (Generalized Logistic Function) to make predictions for COVID-19 evolution in Brazilian cities at stages before and during their epidemic inflection points. We constrain the logistic function regression with K calculated from selected surrogate international cities where the epidemic is clearly past its inflection point. Information gained with this constraint stabilizes the logistic regression, reducing the

*Corresponding author

Email address: gcc@usp.br (George Cunha Cardoso)

uncertainty in the curves' parameters, including the rate of growth at the inflection point. The uncertainty is reduced even when the actual surrogate K is used just as an anchor to simulate different epidemic scenarios. Results predicted for COVID-19 trajectories within Brazil agree with actual data. These results suggest that in the absence of big data, a simple logistic regression may provide low uncertainty if surrogate cities have been identified for estimates of K , even if the specifics of the evolution in the surrogate cities are different. The method may be used for other logistic models and for other logistic processes in other areas such as economics and biology, if surrogate processes can be identified.

Keywords: Decision-making, COVID-19, logistic functions, Richards' logistic function, Generalized Logistic Function

1. Introduction

In real-world epidemic situations, compartmental models, which are based on average observable parameters, often fail to predict epidemic peak timing, case rate and total number of cases [1]. Published COVID-19 predictions in 2020 make clear that compartmental models are inadequate to deal with an emerging epidemic in an unknown complex network [2]. One of the causes of failure of compartmental models, such as the Susceptible-Exposed-Infected-Recovered (SEIR), is the assumption of a homogeneous basic reproduction number (R_o) and homogeneous population behavior [3]. Agent networks can model the propagation of disease, but that would require thorough epidemiological knowledge of both the pathogen and the network of agents [4, 5, 6]. Logistic models are an intermediate alternative that has been shown to represent epidemic evolution reasonably well [7, 8, 9].

Logistic models describe monotonic growth processes that start growing exponentially or sub-exponentially, reach an inflection point, and asymptotically saturate to a maximum value. Examples of logistic models with exponential growth include the Verhulst function and the Richards' growth *function* [10, 8], also known as generalized logistic *function* (GLF). Sub-exponential growth logistic models include the generalized Richards' model [11, 12, 7], which is defined by its differential equation without an analytical solution. For example, the number of accumulated cases $C(t)$ in an epidemic is a type of problem that has been described by logistic functions [7, 8, 11]. As time passes, the availability of susceptible agents decreases, and $C(t)$ grows asymptotically to the limit K . This is the upper asymptotic value of the logistic model, often referred to as carrying capacity when addressing pop-

ulation growth. In epidemics, K is the maximum number of accumulated cases expected for the population under study. It depends on the pathogen transmission mechanism, the social network structure and the underlying contagion dynamics [1]. The key benefit of logistic models is their ability to make predictions with little insight of the underlying specifics of the process, as is often the case in real epidemic surges.

In epidemic problems, the characterization of the logistic model around its inflection point enables estimation of its time derivative. Its peak, on the other hand, informs the time of maximum epidemic growth and the corresponding maximum rate of new cases per day. Logistic models accommodate different choices for the logistic function, and Verhulst's is the most traditional one. It has been used to predict the evolution of COVID-19 $C(t)$ in countries, by finding the number of cases $C(t_{\text{critical}}) = K/2$ at its inflection point, which is fed into a machine learning algorithm to predict further evolution of the epidemic curve [13]. The procedure works as long as the inflection point has been reached in the data. Likewise, the GLF, also known as Richards' growth model, was used to successfully predict the evolution of COVID-19 of various localities, as long as the inflection point has been identified [7, 12].

However, logistic models overfit the data if regression is done before its inflection point has been reached [8, 14]. In the exponential growth stage for $C(t)$, regressions for forecasts are an ill-posed problem that leads to unacceptable uncertainties in the predictions; even estimates of the final number of accumulated cases K at the end of the epidemic have uncertainties of the order of the total population under consideration [12]. This is unfortunate

51 because one of the reasons for modeling is to estimate both the time and the
52 number of cases at the peak of the epidemic, either for preparedness or for
53 the study of different strategies to deal with the disease.

54 Here, we use independent estimates of the logistic model's upper asymp-
55 tote K to constrain the GLF, allowing for stable non-linear regression on data
56 series that end before or around the inflection point. We use surrogate cities
57 from the United States, Italy, the United Kingdom, and Spain where the
58 inflection point of the first COVID-19 wave has been identified, to estimate
59 plausible values for K , that we use to constrain the GLF to predict scenarios
60 in Brazilian cities. We need at least one surrogate city as an anchoring point.
61 The choice of the GLF is for convenience, since it has been shown in the liter-
62 ature to work well for COVID-19 prediction [7]. To increase reliability of the
63 data used, we use the accumulated fatalities count curve $D(t)$, and estimate
64 $C(t)$ using the infection fatality rate (IFR). Our predictions use least squares
65 curve-fitting of the constrained GLF on the cumulative fatalities curve data
66 $D(t)$, for different plausible values of K , for each Brazilian city under study.

67 This paper uses COVID-19 data to demonstrate the proposed strategy
68 of using the asymptotic value of surrogate cities as an anchoring point for
69 the development of scenarios of logistic growth in cities of interest. Our key
70 contribution in this paper is offering an accessible early-predictions tool to
71 gauge ranges of plausibility of the epidemic peak timing and magnitude in late
72 onset cities. Our methodology gives a relatively narrow range of optimistic
73 and pessimistic scenarios, using only $C(t)$ data available before the epidemic
74 peak has happened in the city of interest. Applications of our proposed
75 methodology includes estimates of growth curves of less connected cities and

76 countries for any process that obey logistic growth, such as epidemics.

77 2. Methodology

78 2.1. Generalized logistic function

79 We have used the GLF to model the evolution of fatalities [8, 7]:

$$Y(t) = \frac{K}{(1 + \alpha e^{-\alpha r(t-t_c)})^{(\frac{1}{\alpha})}}, \quad (1)$$

80 where K is the asymptotic value, r is the initial growth rate, α depends on
 81 the epidemic dynamics in the network, and t_c is the abscissa for the inflection
 82 point. The parameters r and α are known to trade-off such that $r\alpha$ equals
 83 the slope of the curve at the inflection point [15]. Two examples of Richards'
 84 curves and their time derivatives are shown in Figure 1.

Figure 1: Examples of generalized logistic function regression to COVID-19 data of select regions, and respective time derivative. The proportion of fatalities divided by the IFR is shown in black dots.

85 To stabilize the logistic regressions of Equation 1 to data of cities of
 86 interest where the inflection point has not been reached yet, we will use K
 87 values derived from surrogate cities. The asymptote K for a city of interest
 88 could, in principle, be estimated from an appropriate epidemic model. For
 89 a system that obeys to the Susceptible-Infected-Recovered (SIR) model, the
 90 relationship between R_o and K is given by [16].

$$R_o = -\frac{1}{K} \ln(1 - K). \quad (2)$$

91 However, in most real-world scenarios compartmental models overestimates
 92 K especially for R_o with considerable heterogeneity [3] – which is a quite
 93 common occurrence. In addition, the relative uncertainty in K grows as R_o
 94 approaches one. More sophisticated models could be used for estimation of
 95 K , all suffering from inaccuracies due to lack of knowledge of the variance
 96 and higher moments of R_o . Nevertheless, R_o may be used to establish an
 97 upper limit to K , irrespective of the existence of surrogate cities.

98 2.2. Data

99 We have analyzed fatalities data from select regions from the U.S.A., Italy,
 100 Spain, the U.K., and Belgium, which had gone through the peak in fatalities.
 101 We collected the data from the 2019 Novel Coronavirus Visual Dashboard op-
 102 erated by the Johns Hopkins University Center for Systems Science and Engi-
 103 neering (JHU CSSE; <https://coronavirus.jhu.edu/>), except for Queens
 104 borough, NY, USA ([https://www.census.gov/quickfacts/fact/table/](https://www.census.gov/quickfacts/fact/table/queenscountyqueensboroughnewyork/P0P645218)
 105 [queenscountyqueensboroughnewyork/P0P645218; https://www1.nyc.gov/](https://www1.nyc.gov/site/doh/covid/covid-19-data-boroughs.page)
 106 [site/doh/covid/covid-19-data-boroughs.page](https://www1.nyc.gov/site/doh/covid/covid-19-data-boroughs.page)).

107 We have also selected fatalities data from cities in Brazil for analyses.
 108 (<https://www.ibge.gov.br/cidades-e-estados>; <https://bigdata-api.fiocruz.br/relatorios>). Details about the data and regions used can be
 109 found in <https://github.com/bhvieira/SurrogateLocalitiesCovid>.
 110

111 For all regions, daily data comprised the period between the beginning of
 112 the epidemic and June/21/2020.

113 *2.3. Modeling*

114 A major issue with the early modeling of pandemic outbreaks is that the
 115 total capacity K in Equation 1 must be determined by data regression alone.
 116 This is a major source of uncertainty in the model. A possibility is using
 117 exponential growth modeling, thus forsaking the modeling of K , for early
 118 modeling [7]. We instead proposed to model different scenarios for a city
 119 based on plausible values for K derived by curve fitting $Y(t)$ on surrogate
 120 cities data where the epidemic is clearly past the inflection point.

121 For the regression analyses we have used nonlinear least-squares opti-
 122 mization to estimate the free-parameters (K, α, r, t_c) in Equation 1 for the
 123 surrogate cities, or (α, r, t_c) for the cities of interest, where in the latter case
 124 K was kept fixed. Optimization was performed with a trust-region-reflective
 125 algorithm, implemented in the function “lsqcurvefit”, from the Optimization
 126 Toolbox in MATLAB. This optimizer, albeit more complex than traditional
 127 Levenberg-Marquardt based ones, can better cope with negative curvature
 128 of the estimate of Hessian of the objective (for a quick primer, see [17]). α
 129 and r are both constrained to the $[0, 1]$ interval in the model, for stability [7].
 130 Because t_c is not constrained, and will often be orders of magnitude larger
 131 than α and r , we divide its gradient by 10^3 . This is so because, otherwise,

132 its gradient magnitude would dominate over other terms.

133 Since our objective function is non-convex on the parameters, one hun-
 134 dred random initialization were tried. Local minima initializations were not
 135 counted, but their resulting models remained on the pool of candidate mod-
 136 els. Model selection was done by minimizing the Akaike information criterion.
 137 In our particular case, this resulted in picking the model with the highest
 138 likelihood, *i.e.* the one with the smallest mean squared error (MSE) given
 139 the Gaussian error assumption of nonlinear least-squares.

140 Instead of estimating our model from the actual cumulative number of
 141 reported cases $C(t)$, we opted instead to estimate the number of cases from
 142 the cumulative number of fatalities $D(t)$. Notoriously, the number of cases
 143 is susceptible to several testing and sampling biases. The number of fa-
 144 talities $D(t)$, while not completely free from its own biases, is more robust
 145 to them. We assume an IFR of 1% (See [https://github.com/bhvieira/](https://github.com/bhvieira/SurrogateLocalitiesCovid)
 146 **SurrogateLocalitiesCovid** for details on demographics age corrections made)
 147 and got the number of cases by dividing the number of fatalities by the IFR.
 148 This had the side effect of inducing a lag to our time estimates, since there
 149 is a time-shift between the cases and fatalities curve, such that $C(t - t_D) =$
 150 $D(t)/\text{IRF}$, where t_D is the average time between diagnosis and death, in case
 151 of fatality. The delay t_D is also a source of uncertainty. To minimize this
 152 issue, we used $C(t - t_D) = D(t)/\text{IRF}$ for determination of K in surrogate
 153 cities and to illustrate $C(t)$ in cities of interest, but use fatality rate $D(t)$
 154 directly in our estimations for t_c and the corresponding peak fatality rate.

155 2.4. Estimating the inflection point: the epidemic peak

156 In Equation 1, t_c is the time where the derivative $Y'(t) = \frac{dY(t)}{dt}$ is maxi-
 157 mal. From this, we estimate the actual peak in fatalities and its uncertainty.
 158 However, derivatives introduce uncertainty; to decrease uncertainty in dY/dt ,
 159 we average the number of fatalities per day in an interval T approximately
 160 equal to the duration of the disease: $T = 15$ days around t_c , as shown in
 161 Equation 3. The approximation to the nearest-integer is denoted by the $\lfloor \cdot \rfloor$
 162 function.

$$\Delta Y_{\text{peak}} = \frac{Y(\lfloor t_c + T/2 \rfloor) - Y(\lfloor t_c - T/2 \rfloor)}{15}. \quad (3)$$

163 The uncertainty of ΔY_{peak} can be approximated by the variance formula,
 164 resulting in Equation 4.

$$\sigma_{\Delta Y_{\text{peak}}}^2 = \frac{\sigma_{Y(\lfloor t_c + 7.5 \rfloor)}^2 + \sigma_{Y(\lfloor t_c - 7.5 \rfloor)}^2}{15^2} \quad (4)$$

165 where the uncertainty $\sigma_{Y(x)}^2$ can be obtained by the formula for prediction
 166 intervals, using the conventional Delta method for uncertainty propagation.

167 3. Results

168 Results for K values for surrogate cities are shown in Table 1 for our data,
 169 corresponding to the first wave of COVID-19.

170 Figure 2 shows GLF regressions for different fixed values of K , and also
 171 for K as a free parameter (automatic), for select Brazilian cities.

172 Figure 3 shows the prediction of the day of the epidemic peak, given by
 173 t_c in Equation 1, and the prediction of the number of fatalities averaged in
 174 an interval $T = 15$ days centered on t_c , as described in Equation 3.

Figure 2: Prediction of epidemic evolution of the estimated number of cases for Brazilian cities. The thin black line denotes a model where the carrying capacity K was automatically optimized as a parameter (denoted “auto” in the legend).

Country	City	Asymptotic value (K)
United States of America	Los Angeles	0.0378 (0, 0.2180)
	New York	0.2617 (0.1326, 0.3907)
	Rockland	0.2011 (0, 0.4554)
	Philadelphia	0.1073 (0, 0.2853)
	Fairfield	0.1456 (0.0578, 0.2334)
	Essex (MA)	0.1434 (0.0377, 0.2491)
Italy	Lombardia	0.1627 (0.0504, 0.2750)
Spain	Madrid	0.2302 (0, 1)
United Kingdom	London	0.0672 (0.0006, 0.1338)
Belgium	-	0.0832 (0, 0.1683)

Table 1: Asymptotic logistic regression values for international localities. The point-estimated for K is shown alongside the 95% confidence interval between parentheses.

Figures for other cities studied are found in Table A.2 can be found in
<https://github.com/bhvieira/SurrogateLocalitiesCovid>.

4. Discussion

The COVID-19 pandemic has revealed that sophisticated epidemic models, in the absence of granular agent-level knowledge of the disease, might not be enough to make good epidemic predictions [18]. Real-world epidemic outbreaks usually have poor quality data. Using the GLF has been shown to fit well the cumulative fatality curve in COVID-19 and other epidemics. However, when the total number of cases is not well past the inflection point, the GLF over-fits the data leading to uncertainties of the order of 100% [7]. Here we discuss the results of using the GLF to predict the evolution of cases

Figure 3: Prediction of the peak for Brazilian cities. The black dot (auto) represents a model where the carrying capacity K was automatically optimized as a parameter (denoted “auto” in the legend). To decrease uncertainty levels, results for the number of cases are averaged over 15 days centered on the peak day.

186 $C(t)$ in cities where the epidemic comes late. We use plausible ranges for the
 187 carrying capacity K , which are estimated from surrogate cities where the
 188 epidemic is beyond the inflection point. This alternative approach seeks to
 189 mitigate the over-fitting problem at the expense of increasing the uncertainty
 190 about the actual scenario that will develop in the city being forecast. Nev-
 191 ertheless, this strategy substantially decreases the range of predicted peak
 192 weeks and predicted peak number of cases per day, and decreases their sta-
 193 tistical uncertainties compared to unconstrained fittings.

194 Using surrogate values for K , we performed GLF regression to Brazil-
 195 ian cities' fatalities curves starting from the onset of COVID-19 until June,
 196 21/2020. Comparison with the actual observed curve by October/2020,
 197 shows that the method gives ranges of expectation for the inflection point up
 198 to 100 times narrower than when K is left as a free parameter.

199 Here we will discuss the major results. In surrogate cities, where the
 200 pandemic had already passed the inflection point of the first wave, GLF
 201 regressions give K in the 5% to 25% interval (New York, London, Madri
 202 and other – see online material). For a Brazilian city still in the exponential
 203 phase of the epidemic (Figure 2a), we observe that the GLF regression with
 204 K as a free parameter leads to unacceptably high uncertainties (Figure 3a –
 205 auto). This high uncertainty is characteristic of a city where the inflection
 206 point of $C(t)$ has not passed yet. The minimum possible value for K for
 207 Aracaju, given the data, is $\simeq 10\%$, that represent the last data point in
 208 Figure 2a. Re-running the GLF regression with constrains $10\% < K < 25\%$,
 209 the uncertainties are reduced (Figure 3a) and agree with actual observed data
 210 (July/2, 14 fatalities/day) [19], that is compatible with $K \simeq 12\%$. Notice

the narrow predicted range, especially for the infection fatality rate, even if $K = 25\%$ is chosen.

Ribeirão Preto is another city shown still in the exponential phase (Figure 3b); when $K \simeq 12\%$ is chosen, the predicted epidemic peak is compatible with the observed data (peak between July/7 and July/27, with ~ 10 fatalities at the peak [19]. Later we will discuss the choice for $K = 12\%$, which is based on other Brazilian cities that are around or past the inflection point. Ribeirão Preto is a city where the epidemic curve is at the beginning. Notice that a tentative GLF regression leads to a carrying capacity of 100%, implying that the whole population would be infected at some point, which could not be correct because it would not take into account herd immunity. The uncertainties for Ribeirão Preto are higher than the ones for the other cities discussed in this paper because the data provided is at the very beginning of the exponential phase of the epidemic (Figure 2b).

Now let us look at Brazilian cities that had an earlier peak. These are probably more appropriate surrogates for other Brazilian cities, including the ones discussed above. Belém, Fortaleza, and Rio de Janeiro, Figure 2c, 2d and 2e, are past the epidemic peak in our time series, which is characterized in the GLF regressions by the small uncertainties in “auto”-fit for K (Figure 3c, 3d and 3e). The carrying capacity for the first waves of Belém, Fortaleza and Rio de Janeiro are $K \simeq 13\%$, $K \simeq 16\%$, and $K \simeq 10\%$ respectively, which are compatible with the corresponding carrying capacities found for Aracaju and Ribeirão Preto, in the previous paragraph. We will now use $10\% < K < 25\%$ as a surrogate in the GLF regression to estimate the inflection point for the city of São Paulo.

236 The GLF regression for the city of São Paulo shows $K = 9.2\%$ when K is
 237 one of the free parameters (Figure 3f). Yet, the uncertainty in the inflection
 238 point's position for $K = 9.2\%$ shows that the estimate may be improved. A
 239 carrying capacity of 9.3% would correspond to an epidemic peak at the end
 240 of May (in the past). For cities where the inflection point has happened in
 241 the past, the uncertainties were smaller. Let us use plausible values of K
 242 range from 10% to 16% , based on surrogate Brazilian cities. Figure 3f shows
 243 regressions with carrying capacities constrained to $K = 10\%$ and $K = 15\%$,
 244 showing that the epidemic peak happened around the second week of June,
 245 which is compatible with official data for the complete fatality curve for São
 246 Paulo. The peak number of fatalities found is also compatible with 123
 247 deaths/day for the 7-day moving average reported in the official data [19].
 248 We can also observe in Figure 3f worse case scenarios, where K would grow to
 249 25% . In such scenario the epidemic peak is predicted to happen a month later
 250 but with at least 50% more casualties — this could happen if the population
 251 that followed an initial trend relaxes their prophylactic measures.

252 It is important to have in mind the possibility of overloading the health
 253 system in a high K scenario. With overloading, it would be necessary to
 254 know the fraction of $C(t)$ that need hospitalization, and what fraction of this
 255 number would convert into fatalities without hospitalization. Let us say, for
 256 example, that $12\%C(t)$ needs hospitalization, and that without saturation
 257 of hospitals $D(t) = 1\%C(t)$. With saturation, a 10% increase in $C(t)$ would
 258 raise fatalities to $D(t) = (1\% + 12\%10\%)C(t)$, which is more than a 100%
 259 increase in fatalities.

260 The method presented in this paper is useful in predictions for peripheral

261 cities or countries with pandemics or epidemic-like phenomena, if an adequate
 262 surrogate city or country has been identified. As the pandemic evolves, cities
 263 geographically and socio-economically closer to the cities of interest can be
 264 used as surrogates, increasing the plausibility of adequacy of the surrogate
 265 K .

266 Our method may be compared to the one described in [15], where a
 267 Bayesian hierarchical model integrates multiple countries' data for infection
 268 trajectory prediction of the spread of COVID-19 based on the Richards'
 269 growth curve. The results reported, despite good rational and sophistication
 270 of the model [15], underestimated the number of cases in main COVID-19
 271 countries by a factor of at least five, even if predictions were done after
 272 the inflection points had occurred for the countries of interest. Our method
 273 could suffer from a similar problem, especially if some Asian cities were used
 274 as surrogates for Brazilian cities. Many cities in Asia have government and
 275 population behavior that would not serve as surrogate for Brazilian cities'
 276 management of COVID-19. It might become clear from the data that the
 277 asymptotic value K for Asian cities are not appropriate surrogates for a
 278 Brazilian city when $K_{asian} \approx C(t)_{Brazilian}$ even when the Brazilian cities
 279 below their epidemic inflection point.

280 One challenge of our method is the identification of a proper surrogate
 281 city. The biggest source of uncertainty in the evaluation of the peak number
 282 of cases per day is the IFR. The IFR, in conditions where the health system
 283 is not yet saturated, we have used as surrogates the IFRs for Italian, Korean
 284 and American cities or regions, corrected for the demographics of Brazilian
 285 cities, and concluded that a 1% IRF is a plausible value (Supplementary

286 Material).

287 Changes in adherence to prophylaxis, such as social distancing, mask
288 usage and self-isolation due to case tracking & contact tracing, will alter the
289 value of K . Our model is agnostic to such possible changes, modeling instead
290 a static K . Thus, it is more adequate for short to intermediary timescales.

291 As presented, the method seem to be more appropriate for the first wave
292 or for single wave epidemics in a city. The first wave is the most important
293 because it will get the health system unprepared. Care must be taken be-
294 cause for an arbitrary epidemic case it is not known if the chosen logistic
295 function will be adequate for the data regression. This uncertainty creates
296 an unknown bias to confidence intervals reported. However, this issue is
297 common to all mathematical models. An intrinsic limitation to the GLF
298 used is regarding its inflection point: it is limited between $K/2$ (for $\alpha = 1$)
299 and K/e (for $\alpha \rightarrow 0$). Yet, the literature shows that GLF represents a good
300 balance between number of free parameters and ability to describe well epi-
301 demic curves [8]. Indeterminacy of α vs. r remained unresolved in our study,
302 but different from [14], here K is pre-determined, allowing for continuous
303 transitions between curves for different K 's. Finally, we also observed that
304 the product αr is proportional to the slope of the logistic function [15]. Be-
305 cause of the difficulty in determining both α and r independently, in some
306 references $\alpha r \propto r_o$, where r_o is not represent the initial growth of the pro-
307 cess [8]. More sophisticated logistic models may be used for the constrained
308 regression for the creation of alternative predictive scenarios, for example,
309 using generalized exponentials [20, 21].

310 In conclusion, the proposed method gives a complementary prescription

311 to the traditional growth prediction methods. It can apply to any early lo-
 312 gistic process, where there is an earlier surrogate process for estimation of
 313 the carrying capacity. This strategy overcomes the intrinsic inability logistic
 314 models have for prediction of the inflection point. It can be used not only for
 315 epidemics but also for commerce, economics, viral information dissemination
 316 in a population. Our method lowers the uncertainty in the prediction for
 317 optimistic and pessimistic epidemic scenarios without the need for sophisti-
 318 cated models or big data type of resources, and gives decision-makers anchor
 319 points specific for their situation.

320 **Data and code availability statement**

321 Code and data are available online. See: <https://github.com/bhvieira/CovidRichards/>

322 **Acknowledgments**

323 We thank Professor Alexandre S. Martinez for critical reading of the
 324 manuscript.

325 **References**

- 326 [1] Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epi-
 327 demic processes in complex networks. *Reviews of modern physics*.
 328 2015;87(3):925.
- 329 [2] Holmdahl I, Buckee C. Wrong but useful—what covid-19 epidemiologic
 330 models can and cannot tell us. *New England Journal of Medicine*. 2020.

- 331 [3] Hébert-Dufresne L, Althouse BM, Scarpino SV, Allard A. Beyond R0:
332 Heterogeneity in secondary infections and probabilistic epidemic fore-
333 casting. medRxiv. 2020.
- 334 [4] Nakamura GM, Monteiro ACP, Cardoso GC, Martinez AS. Efficient
335 method for comprehensive computation of agent-level epidemic dissem-
336 ination in networks. Scientific reports. 2017;7(1):1–12.
- 337 [5] Nakamura GM, Monteiro ACP, Cardoso GC, Martinez AS. Finite sym-
338 metries in agent-based epidemic models. Mathematical and Computa-
339 tional Applications. 2019;24(2):44.
- 340 [6] Nakamura GM, Cardoso GC, Martinez AS. Improved susceptible-
341 infectious-susceptible epidemic equations based on uncertainties and au-
342 tocorrelation functions. Royal Society Open Science. 2020;7(2).
- 343 [7] Vasconcelos GL, Macêdo AMS, Ospina R, Almeida FAG, Duarte-Filho
344 GC, Souza ICL. Modelling fatality curves of COVID-19 and the effec-
345 tiveness of intervention strategies. medRxiv. 2020:2020.04.02.20051557.
346 Available from: [https://www.medrxiv.org/content/10.1101/2020.](https://www.medrxiv.org/content/10.1101/2020.04.02.20051557v1)
347 [04.02.20051557v1](https://www.medrxiv.org/content/10.1101/2020.04.02.20051557v1).
- 348 [8] Wang XS, Wu J, Yang Y. Richards model revisited: Validation by
349 and application to infection dynamics. Journal of Theoretical Biology.
350 2012;313:12–19.
- 351 [9] Pongkitivanichkul C, Samart D, Tangphati T, Koomhin P, Pimton P,
352 Dam-O P, et al. Estimating the size of COVID-19 epidemic outbreak.
353 Physica Scripta. 2020;95(8).

- 354 [10] Richards F. A flexible growth function for empirical use. *Journal of*
355 *experimental Botany*. 1959;10(2):290–301.
- 356 [11] Chowell G. Fitting dynamic models to epidemic outbreaks with quan-
357 tified uncertainty: A primer for parameter uncertainty, identifiability,
358 and forecasts. *Infectious Disease Modelling*. 2017;2(3):379–398.
- 359 [12] Aviv-Sharon E, Aharoni A. Generalized logistic growth modeling of the
360 COVID-19 pandemic in Asia. *Infectious Disease Modelling*. 2020;5:502–
361 509.
- 362 [13] Wang P, Zheng X, Li J, Zhu B. Since January 2020 Elsevier has created
363 a COVID-19 resource centre with free information in English and Man-
364 darin on the novel coronavirus COVID- 19 . The COVID-19 resource
365 centre is hosted on Elsevier Connect , the company ’ s public news and
366 information ; 2020.
- 367 [14] Clark F, Brook BW, Delean S, Reşit Akçakaya H, Bradshaw CJ. The
368 theta-logistic is unreliable for modelling most census data. *Methods in*
369 *Ecology and Evolution*. 2010;1(3):253–262.
- 370 [15] Lee SY, Lei B, Mallick BK. Estimation of COVID-19 spread curves
371 integrating global data and borrowing information. *arXiv preprint*
372 *arXiv:200500662*. 2020.
- 373 [16] Heffernan JM, Smith RJ, Wahl LM. Perspectives on the basic reproduc-
374 tive ratio. *Journal of the Royal Society Interface*. 2005;2(4):281–293.

- 375 [17] Berghen FV. Levenberg-Marquardt algorithms vs Trust Region algo-
376 rithms; 2004. Available from: [http://www.applied-mathematics.](http://www.applied-mathematics.net/LMvsTR/LMvsTR.pdf)
377 [net/LMvsTR/LMvsTR.pdf](http://www.applied-mathematics.net/LMvsTR/LMvsTR.pdf).
- 378 [18] Chin V, Samia NI, Marchant R, Rosen O, Ioannidis JP, Tanner MA,
379 et al. A case study in model failure? COVID-19 daily deaths and ICU
380 bed utilisation predictions in New York State. *European Journal of*
381 *Epidemiology*. 2020;35(8):733–742.
- 382 [19] dos Registradores de Pessoas Naturais AN. Especial COVID-19;.
383 Available from: [https://transparencia.registrocivil.org.br/](https://transparencia.registrocivil.org.br/especial-covid)
384 [especial-covid](https://transparencia.registrocivil.org.br/especial-covid).
- 385 [20] Martinez AS, González RS, Terçariol CAS. Continuous growth models
386 in terms of generalized logarithm and exponential functions. *Physica A:*
387 *Statistical Mechanics and its Applications*. 2008;387(23):5679–5687.
- 388 [21] Tsallis C, Tirnakli U. Predicting COVID-19 peaks around the world.
389 *Frontiers in Physics*. 2020;8:217.

390 Appendix A. Data from Brazilian and international cities

City	Population	Date of first reported fatality
Aracaju	657013	02/04/2020
Belém	1492745	06/04/2020
Brasília	3015268	29/03/2020
Campinas	1204073	31/03/2020
Curitiba	1933105	06/04/2020
Florianópolis	500973	02/04/2020
Fortaleza	2669342	26/03/2020
Joinville	590466	30/03/2020
Manaus	2182763	30/03/2020
Porto Alegre	1483771	25/03/2020
Recife	1645727	25/03/2020
Ribeirão Preto	703293	26/03/2020
Rio de Janeiro	6718903	23/03/2020
Salvador	2872347	29/03/2020
São Luís	1101884	29/03/2020
São Paulo	12252023	17/03/2020
Uberlândia	691305	03/04/2020

Table A.2: Data from representative Brazilian cities. Population and fatalities statistics for Brazil come from <https://www.ibge.gov.br/cidades-e-estados/> and <https://bigdata-api.fiocruz.br/relatorios/>, respectively.

Country	City	Population	Date of first reported fatality
United States of America	Los Angeles*	3979576	11/03/2020
	New York	8336817	12/03/2020
	Rockland*	325789	14/03/2020
	Philadelphia*	1584064	25/03/2020
	Fairfield*	943332	18/03/2020
	Essex (MA)*	789034	26/03/2020
Italy	Lombardia†	10088484	24/02/2020
Spain	Madrid‡	6663394	22/04/2020&
United Kingdom	London*	8961989	06/03/2020
Belgium	-§	11589485	11/03/2020

Table A.3: Data from representative international cities. *: county-level data; †: region-level data; ‡: Community-level data; §: country-level data; &: date report started. Population and fatalities statistics for the United States of America come from <https://www.census.gov/data/tables/time-series/demo/popest/2010s-total-cities-and-towns.html> and https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series, respectively. <http://demo.istat.it/bilmens2019gen/index.html> and <https://statistichecoronavirus.it/regioni-coronavirus-italia/lombardia/> for Lombardia. <http://www.madrid.org/iestadis/fijas/estructu/demograficas/padron/estructupopc.htm> and <https://www.comunidad.madrid/servicios/salud/2019-nuevo-coronavirus> for Madrid. <https://www.citypopulation.de/en/uk/greaterlondon/> and <https://www.england.nhs.uk/statistics/statistical-work-areas/covid-19-daily-deaths/> for London. <https://www.worldometers.info/world-population/belgium-population/> and <https://coronavirus.jhu.edu/map.html> for Belgium.