Supplementary information to Publication:

Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil

Geospatial analysis

We adopted a Bayesian hierarchical model to compute relative risk for each census tract, due to the following reasons: (i) there is a large number of census tracts (*n*=30,815), (ii) there is substantial heterogeneity in the size of census tracts, and (iii) small counts in each tract obscure the spatial distribution of observed cases. The number of observed cases in census tract i is modelled using a Poisson distribution $Y_i = Poisson(\lambda_i)$ with mean $\lambda_i = E_i\,\mu_i$ where E_i is the expected number of cases under a null model in which cases are uniformly distributed among the population. For example, the total number of cases in the MRSP multiplied by the proportion of the population in the census tract $E_{it} = \frac{\sum_i Y_i}{\sum_i pop_i} \times pop_i$. The factor of-$\mu_i$ describes tract specific risk and models the additional variation in the observation process[1]. A log-linear model is used to estimate the relative risk μ_i. For example, the log relative risk is expressed as a sum of an intercept α, which represents the overall relative risk (in our case, the global relative risk is zero), and random effects (Z_i):

$$\log(\mu_i) = \alpha + Z_i$$

We used a Besag-York-Mollié model (BYM)[2] to separate the random effects into a spatially structured U_i, and independent random effects, V_i, so ($Z_i = U_i + V_i$). In the BYM model, a conditional autoregressive (CAR) process is used to introduce correlation among the U_i for each tract. Given the U_i of neighbouring tracts, the U_i has a normal distribution with mean equal to the average of the neighbours' U_i, and variance $s_i^2 = \frac{1}{\#N(i)\tau_U}$ where $\#N(i)$ is the number of tracts that share boundaries with tract i and τ_U is a precision parameter. The random effect, V_i follows a zero mean normal distribution with unknown precision, τ_V=$\frac{1}{\sigma_{v^2}}$ (where σ_{V^2} is the variance). Both random effects in the model capture extra-Poisson variability, and were expressed as the following:

$$U_i|\, U_{j\neq i} \sim Normal\left(m_i, s_i^2\right), \quad V_i \sim N(0, \sigma_v^2)$$

$$m_i = \frac{\sum_{j\in N(i)} U_j}{\#N(i)} \quad , \quad s_i^2 = \frac{\sigma_U^2}{\#N(i)} = \frac{1}{\#N(i)\tau_U}$$

The log of the precision parameters, τ_U and τ_V, follows a gamma distribution with shape 1 and rate 0.0005. These are the default priors used by R-INLA and are minimally informative[3]. The prior default distributions in R-INLA were used for the precision parameters of both U_i and V_i. These are minimally informative and are the recommended settings [4].

To quantify the uncertainty in the point estimates of the mean relative risk estimates, we mapped the posterior probability of elevated relative risk in each census tract (**Extended Data Fig. 9**). This is the posterior probability, which a tract has an elevated risk of observing cases, formally, this is Prob(μ_i > 1| data). For instance, a probability of 0.6 in a census tract indicates a 60% chance that this census tract is at greater risk of observing cases relative to the rest of the MRSP.

Analysis of the relationship between income per capita and final diagnostic category in the Metropolitan Region of Sao Paulo (MRSP)

We evaluated the relationship between final diagnostic category (COVID-19 or SARI cases with unknown aetiology) and socioeconomic status in the subset of cases in the MRSP with geocoded residential information. We focused on the cases in epidemiological weeks 12, 16 and 22, where the census tracts that reported cases varied across weeks. In each of the three weeks, if a census tract reported any COVID-19 or SARI cases with unknown aetiology with unknown aetiology, we calculated the proportion of the number of COVID-19 cases. Since most census tracts reported only one case each week, the proportion of COVID-19 of each census tract were mostly either 0 or 1 in a given week. Based on this observation and let i index the census tracts, we subsequently defined the binary outcome Y_i of census tract i, where (i) $Y_i = 0$ if census tract i only reported SARI cases with unknown aetiology with unknown aetiology, i.e. no COVID-19 cases, (ii) $Y_i = 1$ if census tract i reported at least one COVID-19 case in the week. Logistic regression models were applied to investigate the association between this binary outcome and the log(X+1) transformed income per capita. The analyses were adjusted by the logarithm of the population sizes. In addition, the census tracts were grouped by their geographic locations using cluster analysis, and the groupings were used as the random effect in the logistic regressions to account for potential spatial autocorrelation. The number of clusters was chosen based on the AIC/BIC values of the logistic regression models. The analysis was performed individually for each of epidemiological weeks 12, 16 and 22.

A likelihood ratio test (LRT) is applied to each analysis to examine whether the log(X+1) transformed income per capita provides information in addition to the information from the log population size and the random effects. The regression coefficients and LRT P-values of income are presented in (**Supplementary Table S3**).

Estimating basic reproduction number (R_0)

Since SARS-CoV-2 is a novel virus, and we are subsetting data to avoid the impact of either non-pharmaceutical interventions or depletion of the susceptible pool, we deemed it reasonable to model the incidence of infection with an exponential approximation to the early behaviour of an SIR model, i.e., the incidence grows exponentially [5]. This model makes several strong assumptions about the dynamics of the epidemic: (i) the populations under consideration mix homogeneously, (ii) the proportion of the population that is susceptible stays close to 100%, (ii) the proportion of infections that are observed, and the basic reproduction number are constant throughout time, and (iv) the delay between infection, and notification is a constant. Although there are obvious violations of these assumptions, they provide a convenient starting point for estimating the basic reproduction number. Ignoring the delay between infection and observation will on average only translate the results in time by the incubation period and the delay from infection to diagnosis.

Under the assumptions outlined above, the expected number of daily cases, □(□) on day □ is given by the following equation: □(□) = □R$_0$□□$_0$□$^{(R0-1)□.□}$ where □ is the probability of an infection being counted in the time series, R_0, is the basic reproduction number, □ is the rate at which individuals cease to be infectious and i_0, is the proportion of the population that was infectious at the start of the observations. We assume that the observed number of cases on day n was drawn from a negative binomial observation where the mean is □(□) and the variance, □ = □ + □^{2}/□, with fixed size parameter, □ (*dispersion parameter*). The product of □ and □$_0$ is denoted ξ. Since the probability of being observed and the initial condition only appear as the product ξ in the likelihood, there is an

identifiability problem preventing the estimation of □ and □$_0$ individually, consequently we only consider their product, ξ. Although in this model it is theoretically possible to estimate both R_0 and □, in practice this is difficult so we will use an informative prior to constrain □ to a priori plausible values.

Regarding prior distributions, for R_0 we used a uniform prior over values from 1 to 10. The removal rate, □, was given an informative prior distribution: a normal distribution with mean (1/5 + 1/14) / 2 = 0.1357, leading to an average duration 7.4 days during which an individual is infectious. Moreover, the average duration of infectivity is constrained to be between the extremes of 5 and 14 days. These values for the infective duration were found in the literature [6,7]. The standard deviation of the prior distribution for □ is (1/5 - 1/14) / 4 = 0.03124, this ensures that 95% of the prior probability lay within these bounds. For the parameter ξ, we used a log-normal prior with a log mean of 0.0 and a log standard deviation of 1.0. For the size parameter of the negative binomial, k, a log-normal distribution was used with a log-mean of 0.0 and log-standard deviation of 1.0 to enable this parameter to have a large range of values.

Samples from the posterior distribution were obtained using MCMC running 4 chains from random initial conditions using the mcmc library available on CRAN2 and using coda for diagnostics[8,9]. Trace plots of the posterior samples suggested that the chain had converged and mixed, and there was an effective size of at least several hundred for each of the 4 parameters of this model. The prior and posterior distributions were checked to ensure that (beyond the removal rate) each parameter was being informed by the data. Each data set: Brazil and Europena countries (Italy, the United Kingdom, France, and Spain) or Brazilian states (São Paulo, Rio de Janeiro, Amazonas, and Ceará) were run as independent analyses, the model fit from the point estimate along with the corresponding trace plots and prior/posterior comparisons is shown in **Extended Data Figs. 5 and 6.**

References

1 Lawson, A. B. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiolog*. (2008).

2 Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1-20, doi:10.1007/BF00116466 (1991).

3 Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**, 319-392, doi:10.1111/j.1467-9868.2008.00700.x (2009).

4 Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology* **7**, 39-55, doi:10.1016/j.sste.2013.07.003 (2013).

5 Brauer, F., van den Driessche, P. & Wu, J. *Mathematical Epidemiology*. (Springer-Verlag Berlin Heidelberg, 2008).

6 Wolfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature*, doi:10.1038/s41586-020-2196-x (2020).

7 European Centre for Disease Prevention and Control. Novel coronavirus (SARS-CoV-2). (2020).

8 Geyer, C.J., & Jonson, L. T. mcmc: Markov Chain Monte Carlo. R package version 0.9-6 (https://CRAN.R-project.org/package=mcmc, 2019).

9 Plummer, M., Best, N., Cowles, K. & Vines, K. CODA: convergence diagnosis and output analysis for MCMC. *R News* **6**, 7-11 (2006).