medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

# Covid19, 2020 - a tutorial of sorts on reading data

Isaac Meilijson
*E-mail:* isaco@tauex.tau.ac.il
*School of Mathematical Sciences*
*Tel-Aviv University, 6997801 Tel-Aviv, Israel*

September 11, 2020

**Abstract**

The SIR differential equations in Epidemiology are re-examined in the context of Covid19, 2020. The number of recovered cases is calibrated in time. Methods for estimating all pertinent parameters are described. A notion of implied susceptible population size ISPS is introduced, as the potential target population size for which the solution to the SIR equations would yield the current number of new affected cases. Analysis is applied to the Covid19 2020 data of a number of countries, as reported by the Johns Hopkins University data repository. It will be observed that some countries under-report consistently the number of recovered cases.

## 1 Introduction

The SIR (Susceptibles, Infected, Removed) model introduced by Kermack & McKendrick in 1927 [7] for the progress of an epidemic describes the interdependence between the cumulative number $X(t)$ of affected cases, the cumulative number $R(t)$ of removed cases (dead or recovered) and the ensuing current number of infected cases $Y(t) = X(t) - R(t)$. Henceforth, the words "cumulative" and "currently" will be tacitly intended. The version of this model to be applied in the current report conforms with the usual assumption (henceforth the SIR "$\gamma$-equation")

$$dR(t) = \gamma Y(t)dt \tag{1}$$

that the number of new removed cases constitutes a fixed proportion of the infected cases, and the usual assumption that the number of new affected cases is proportional to the product of the number $K - X(t)$ of susceptible cases and an increasing function of the number of infected cases. However, $K$ is fitted from data instead of letting it be the total population size, and the second factor (commonly modelled as the identity function) is taken (after Grenfell, Bjørnstad & Filkenstädt [4] to be a fractional power $Y(t)^\alpha$ $(0 \le \alpha < 1)$ of the number of infected cases. In other words (the SIR "$\beta$-equation")

$$dX(t) = \beta(Y(t))^\alpha(K - X(t))dt \tag{2}$$

It is common practice to view $\beta$ as varying with time, and watch closely for the date at which the effective reproduction rate $R_0$, the ratio of $K\beta$ to the removal rate $\gamma$, crosses from above to below 1. Thus, equation (2) (with $Y$ instead of $Y^\alpha$ and $K$ as total population size) is perceived as a local relation, and not as a differential equation to be solved. We intend to establish via Covid-19 data, that the introduction of a power function of $Y$ and of an adaptive target population size $K$ make equation (2) solvable with constant $\beta$.

A companion publication by Alon and Meilijson [1] describes ($i$) methods to estimate the parameters $\alpha, \beta, \gamma$ and $K$ by maximum likelihood estimation under diffusion models based on the SIR differential equations, and ($ii$) a regression pre-processing procedure that modifies minimally the infected and removed components of the empirical data $(X_1, Y_1, R_1), \cdots, (X_n, Y_n, R_n)$, so that the proportionality of new removals and infected cases will better hold. Item ($i$), that ignores the

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

SIR $\beta$-equation (2), is expanded and modified in Section 2. Item (*ii*) is analyzed anew in Section 4 with simpler regular regression methods.

The data analyzed in this report is taken from the COVID-19 Data Repository of the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, from January 22, 2020 to August 2020. The purpose is to propose techniques to read and interpret these data, and to a lesser extent to build scenarios to project values into the near future. Attempts are made to take into account warnings and guidelines, such as those of Holmdahl and Buckee [6].

The (country-dependent) training period considered stationary extends roughly from the beginning of March to sometime between late May and early July. The working paradigm is that the three proportionality parameters $\alpha, \beta, \gamma$ estimated in the training period, stay in effect for the particular epidemic beyond the training period. The validity of this paradigm regarding $\gamma$ will become apparent in Section 2. The model as a whole is not strongly identifiable: once the other parameters are adjusted, a relatively wide range of $\alpha$ values fits data adequately well, and perhaps one compromise value could fit all countries, but the analysis will be done per country. Validity regarding $\beta$ can and will be assessed only in stationary periods where $K$ is kept fixed. Under constancy of $\beta$, all variability will be phrased in terms of $K$, as follows. Equation (2) is rewritten as

$$K(t) = X(t) + \frac{\Delta(X)}{\beta(Y(t))^{\alpha}} \tag{3}$$

to be interpreted as the *implied susceptible population size* ISPS. Letting $\Delta(X)$ be a smooth version of $X(t+1) - X(t)$, K(t) is evaluated daily to provide an idea of the current target population of the epidemic. Due to the persistent missing data on weekends, the smooth version is taken throughout to be a moving average on the week around the date in question.

The notion of implied susceptible population size ISPS mimics implied volatility in Finance. The Black & Scholes ([3]) formula for option pricing does not quite hold when the asset estimated volatility is substituted, but it is such a convenient language tool that reverse engineering as in (3) is applied, and implied volatility is defined as the value for which formula and empirical price are the same. This value changes with the maturity date and the moneyness, and this non-constant shape is termed "volatility smile". This is a word we will not export from Finance into Epidemiology.

ISPS is analyzed in Section 6 and illustrated in Figure 5 on six countries, projecting into the near future the solution to the SIR equations, with $\beta, \gamma$ and $\alpha$ as fitted in the training period.

**A word on the parameter** $\alpha$. Imagine a small-world network topology composed of sparsely connected islands that are internally densely connected. Once an island is affected, contagion will apply locally. Much as in the Harris contact process [5], if the affected cases are considered a growing ball that transmits the disease by its perimeter, $\alpha = \frac{1}{2}$ models the intra-island spread force. The "ball" could have some fractal dimension rather than a solid area. Alon and Meilijson [1] present an argument that on infinite populations, $\alpha = 1$ yields exponential growth of $X$, while $\alpha < 1$ gives rise to linear growth of $X$, with $Y(t)$ increasing asymptotically to oscillate around the finite constant $(\frac{K\beta}{\gamma})^{\frac{1}{1-\alpha}}$. Formula (2) explains why should the proper parameter be $K\beta$ rather than $\beta$, for a large population.

The average daily number of new affected and removed cases would then be $\frac{(K\beta)^{\frac{1}{1-\alpha}}}{\gamma^{\frac{\alpha}{1-\alpha}}}$. These levels, to be reported for every country under study, are so high that it is reasonable to infer that only a fraction of the population is susceptible.

Without claiming to delve deeply into epidemiology, it is suggested that isolation measures achieve stationarity by cutting contagion across islands. Relaxing isolation measures may have the effect of allowing for more contagion between islands, and this is what ISPS is intended to reflect.

Israel went through a Covid19 first wave that seemed to saturate, but relaxation of measures brought a second wave. USA is treated as one unit (perhaps wrongly, but state-by-state data is not detailed enough), and as such never quite stabilized. Unlike Israel and USA, the neighboring European countries Germany, Italy and Switzerland seem to have maintained stability at least

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

until mid July, by a combination of official isolation measures and personal responsibility that has prevented for some time a return to the pandemic core. Figure 4 shows the dates considered stable in each of these countries. The epidemic is still growing in Brazil in mid August.

The purpose of this report is to show a technique for parameter estimation and recovery and infection calibration, and to suggest possible uses of ISPS as a measure of growth and perhaps as a driving wheel of epidemic management, without taking sides on discussions about management. With this said, under the strong evidence that $\alpha < 1$, the author support the postulate that Covid19 grows in homogeneous sub-populations slower than exponential. However, increasing ISPS may indicate infection into new horizons, calling for isolation measures.

**A word on data quality**. Each country is analyzed on its own sake. The rate $\gamma$ at which infected cases are removed should be (by the SIR $\gamma$-equation (1)) the slope of the linear regression of the newly removed cases on the number of infected cases. The plots of these linear regressions in Figure 2 are reasonable, but slopes $\gamma$ vary too much from country to country. It is worthwhile to point out in this context that data is taken at face value. Although analysis is based on affected and infected cases only, the proportion of dead cases among removed cases will be evaluated and displayed in Figure 6. Israel and Europe data show this proportion to oscillate around a few percentage points and stabilize in the last two months of monitoring between 0.5% and 1%, with daily removal rate $\gamma$ between 4% and 10%. In contrast, USA data shows lethal proportion oscillating around 50% and stabilizing at 3.5%, with daily removal rate $\gamma$ below 1%. Brazil stands somewhere in between. There seem to be big differences between countries on what fraction of what data is actually reported. A case will be built throughout this paper that USA and Italy under-report consistently the number of recovered cases.

## 2 Removed cases - estimation of $\gamma$ and calibration of recovered cases delay

*This Section hinges on the common observation that recovered cases are reported later than dead or affected cases. As a result, the common naive subtraction operation that assesses the current number of infected cases, is positively biased. A method will be presented to gradually shift recovered cases backwards in time to the point that the SIR $\gamma$-equation is properly satisfied. At this point the number of infected cases will be minimally perturbed so that the SIR $\gamma$-equation is perfectly satisfied. This will provide an assessed day by day number of infected and removed cases, and an estimate of $\gamma$. The balance between dead and recovered cases will be ignored along the procedure, but monitored on the end result. The output of this procedure, with the original empirical cumulative number of affected and dead cases and the modified number of recovered cases, is the input for the next Section, dealing with the SIR $\beta$-equation: estimation of $\alpha$ and $\beta$ via flexible $K$.*

The clear-cut SIR equation $dR(t) = \gamma Y(t)dt$ should be thought of empirically, towards the formulation of regression equations, as

$$R_{t+1} - R_t = \gamma Y_t + \sigma\sqrt{Y_t}Z_t \tag{4}$$

where $Z_t$ are ideally independent with mean zero and unit variance. Think conceptually of $R_{t+1} - R_t$ as binomially distributed with number of trials $Y_t$ and probability of success $\gamma$.

In other words, a homoscedastic regression model would express (4) as

$$\frac{R_{t+1} - R_t}{\sqrt{Y_t}} = \gamma\sqrt{Y_t} + \sigma Z_t \tag{5}$$

from which the regular Least Squares estimate of $\gamma$ is

$$\hat{\gamma} = \frac{\sum_{t=1}^{n-1} \frac{R_{t+1}-R_t}{\sqrt{Y_t}}\sqrt{Y_t}}{\sum_{t=1}^{n-1}(\sqrt{Y_t})^2} = \frac{R_n - R_0}{\sum_{t=1}^{n-1} Y_t} \tag{6}$$

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

**Exact solution $\hat{Y}$ to the SIR $\gamma$-equation**. The above idea towards estimation is the starting point of the proposed pre-processing step. Recalling that $R = X - Y$, a perfect match of the SIR $\gamma$-equation would entail

$$\hat{Y}_t + \gamma \sum_{m=1}^{t} \hat{Y}_m = X_t \text{ or } A\hat{Y} = X \text{ or } \hat{Y} = A^{-1}X \tag{7}$$

where $A = I + \gamma B$ and $B$ is the lower triangular matrix with zeros above the diagonal and ones on and below the diagonal.

Empirical data are noisy. Reported affected and removed cases are at best a rough description of reality. It is proposed to respect the reported affected and dead cases but adopt a solution $\hat{Y}$ of equation (7) as the number of infected cases. The parameter $\gamma$ will be estimated by the value for which $\sum_{t=1}^{n} \frac{(\hat{Y}_t - Y_t)^2}{\max(1, Y_t)}$ is minimal. I.e., $\hat{Y}$ is the minimal tampering with the number of infected cases that will satisfy the SIR $\gamma$-equation.

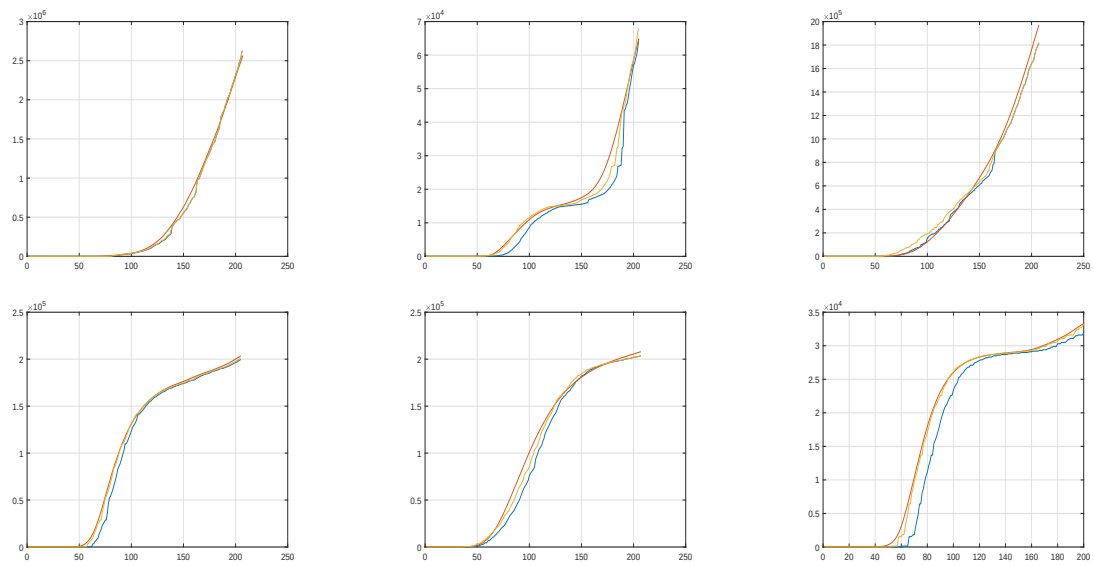

Figure 1: Number of recovered cases. Blue: raw data, yellow: shifted raw data, red (working version): inferred from regression-adjusted infected cases. From top left to bottom right: Brazil, Israel, USA, Germany, Italy and Switzerland.

**DAY - late report of recovered cases**. Actually, the method as such is inappropriate. A plot of $Y$ and $\hat{Y}$ reveals that the two peak at different times. The report of recovered cases seems to be delayed more than the other statistics. In the absence of a model, the parsimonious initial working paradigm is that this additional delay is a constant number of days $DAY$. In other words, the effective number of removed cases $R_t^*$ until date $t$ is to be taken as the number of dead cases until date $t$ plus the number of recovered cases until date $t + DAY$. The effective number of infected cases $Y_t^*$ at date $t$ is then $Y_t^* = X_t - R_t^*$. The parameters $\hat{\gamma}$ and $DAY$ are jointly determined by minimizing $\sum_{t=1}^{n} \frac{(\hat{Y}_t - Y_t^*)^2}{\max(1, Y_t^*)}$.

Figure 1 illustrates the results of this method on each of the six countries. The Brazil and USA data fits well daily with no delay, and the other four manifest delays between 5 and 10 days. It is possible to see that the recovered-cases shifted empirical data $Y^*$ and the regression-fitted $\hat{Y}$ are close to each other (somewhat less so in Italy) and rather far from the unshifted empirical data $Y$. That is, calibration of the recovered cases by the delay DAY, brings the $(S, I, R)$ data to satisfy the SIR $\gamma$-equation more accurately.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

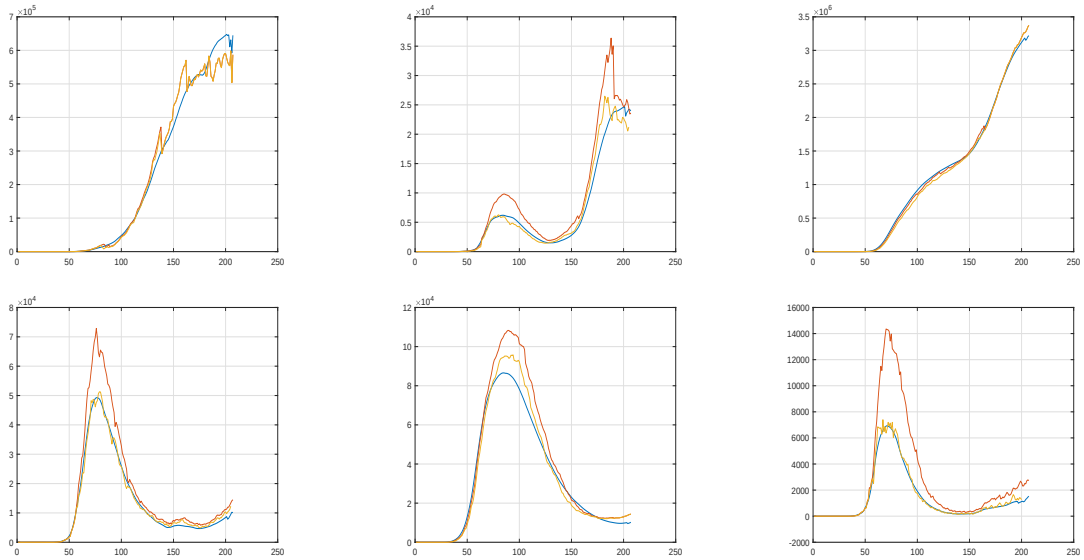

Figure 2: Number of infected cases. Red: raw data, yellow: inferred from shifted raw data, blue (working version): regression-adjusted. From top left to bottom right: Brazil, Israel, USA, Germany, Italy, Switzerland.

But once it is realized that shifting in time the number of recovered cases markedly improves the SIR $\gamma$-equation fit, perhaps it is possible to attempt a gradual shift in time, where the recording of recovered cases turns progressively more efficient.

**Sharp jumps of the number $RC$ of recovered cases**. Suppose that on a particular day the number of newly recovered cases is five times the daily average in the next week. It makes sense to shift the number of recovered cases (by four or less days) backwards in time before this date rather than globally. This procedure is performed in small steps, as follows. Let $\Delta RC_1(t)$ be the number of new recovered cases in day $t$ and let $\Delta RC_2(t)$ be the daily average of new recovered cases in the week following day $t$. Suppose that the ratio $\frac{\Delta RC_1(t)}{\Delta RC_2(t)}$, taken from day 40 or 50 to a week before the end of observation, is maximal at day $t_1$, where it exceeds some threshold like $TH = 3$. The function $RC(t)$ will be modified as follows: It stays as is from day $t_1$ onwards, it is shifted backwards by one day before day $t_1$, and the single day left unassigned is given as value the simple average of the two days next to it. The procedure is applied again and again to the newly defined function, until no ratio exceeds the threshold, or some maximal number of steps, such as 5 or 10, has been performed. The rationale behind choice of $TH$ is that after adding a day when the ratio is $TH$, the new ratio will be $\frac{TH}{2}$. Hence, $TH$ should be somewhere between 2.5 and 3. The choice $TH = 2.5$ is adopted.

If the maximal allowed number of steps is kept small, the modified vector of numbers of recovered cases can be then submitted to the global DAY procedure above to account for a possible delay towards the end of the observation period.

The output of this delay procedure, with the original empirical cumulative number of affected and dead cases and the modified number of recovered cases, satisfying the SIR $\gamma$-equation, should be the input for Section 4, dealing with the SIR $\beta$-equation: estimation of $\alpha$ and $\beta$ via flexible $K$. However, the estimates of $\gamma$ reported in Table 1 are too different. A sanity check is called for. This is the subject of the next section.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

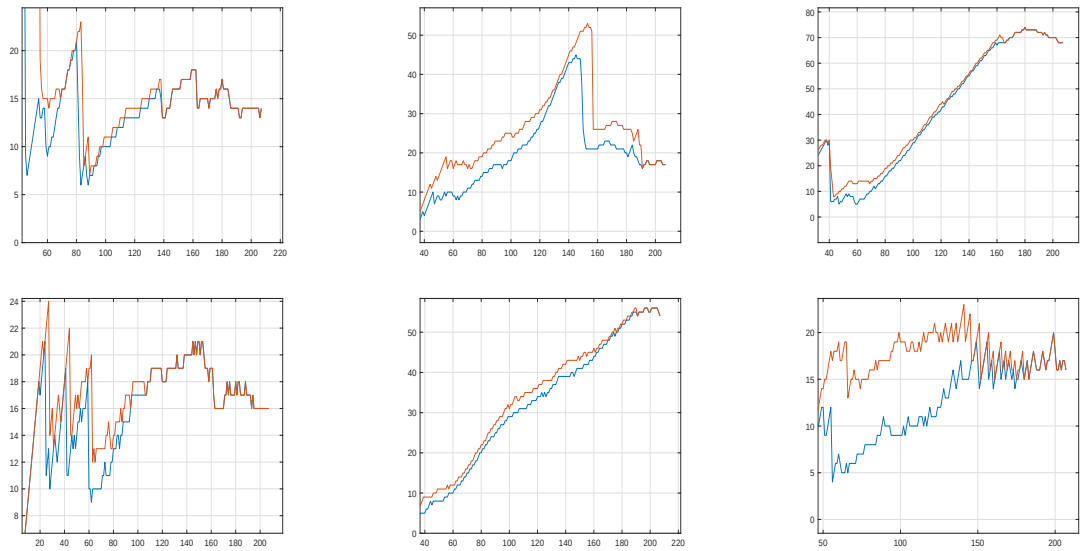

Figure 3: Horizontal distance between the affected and removed curves. Red: raw removed data. Blue: shifted removed data. From top left to bottom right: Brazil (stable around 15), Israel (stable around 18-20, singularity between the twowaves), USA (under-reported recovered cases), Germany (stable around 16), Italy (under-reported recovered cases) and Switzerland (stable around 10-15).

# 3 Instability of $\gamma$: the deficient report of recovered cases

The meaning of $\gamma$ is perhaps carried more crisply by $\frac{1}{\gamma}$, the theoretical average time an infected individual spends as infected until removed (recovered or dead). Table 1 reports values too different for $\gamma$ in the various countries illustrated. It is inconceivable that this average could be 100 days in USA and 27 in Italy on the one hand, and between 8 and 16 days in the other four countries. The ill fit of Italy in Figure 2 *after* the regression correction indicates that something is not right. The horizontal distance between the affected and removed curves in Figure 8 should in principle be this average infection time $\frac{1}{\gamma}$, the time it takes from the moment the number of affected cases reaches some height until an equal number of cases has been removed.

Figure 3 displays these horizontal distances. In a nutshell, the horizontal distance is roughly 16 days. In periods between waves of the epidemic, when the number of affected cases stays roughly constant, this distance is ill defined, with a tendency to appear bigger. This is apparent in Germany, Israel and Switzerland. Brazil never stabilized, and the constancy of the horizontal distance is apparent. Except for the between-waves instability, all four countries conform with the 16-day statement.

Italy and the USA are different. The horizontal distance is a steadily increasing function. The proposed explanation is that in these two countries the recovered cases are only partially reported, and the Data Repository data is defective.

It is clear that data omission in the USA far exceeds that in Italy. There are states of the US, such as California and Florida, for which the entire column of recovered cases is identically zero. Correspondence with CSSE at Johns Hopkins University has corroborated awareness of this phenomenon. A warning message is attached as an Appendix.

The remainder of this section is dedicated to an exercise. The parameter $\gamma$ is fixed as $\frac{1}{16}$ in all countries illustrated, the data on recovered cases is ignored, and the number of infected cases $\hat{Y}$ is derived from the numbers of affected and dead cases only, by means of formula (7). Figure 10 displays a comparison with Figure 9, where each country is assigned its estimated $\gamma$. Equivalently,

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

Table 2 presents parameter estimation under $\gamma = \frac{1}{16}$, compared with Table 1, where each country is assigned its estimated $\gamma$. Due to the observed under-reporting of recovered cases, the number of infected cases in the USA, and to a lesser extent also in Italy, is markedly smaller in the more realistic Figure 10 than in the raw-recovered-data Figure 9.

As the purpose of the current "tutorial of sorts" is to communicate a personal outlook on data exploration rather than presenting a country-by-country analysis as such, this exercise is presented and then ignored for the rest of the paper.

# 4 Affected cases - estimation of $\alpha, \beta$ and $K$

*This Section identifies an interval of time on which epidemic dynamics are as stationary as feasible, to apply as the training period. On this period, the target population size $K$ is viewed as a parameter, estimated together with $\alpha$ and $\beta$, so as to satisfy the SIR $\beta$-equation as closely as possible.*

The SIR $\beta$-equation (3) may be expressed for fixed $K$ as

$$\log \frac{\Delta(X)_t}{1 - \frac{X_t}{K}} = \log(K\beta) + \alpha \log(Y_t) \tag{8}$$

For given $K$, this is a regression equation with slope $\alpha$ and intercept $\log(K\beta) + mean(\log(Y)) = mean(\log \frac{\Delta(X)}{1-\frac{X}{K}})$, from which $\beta$ can be derived as a function of $\alpha$.

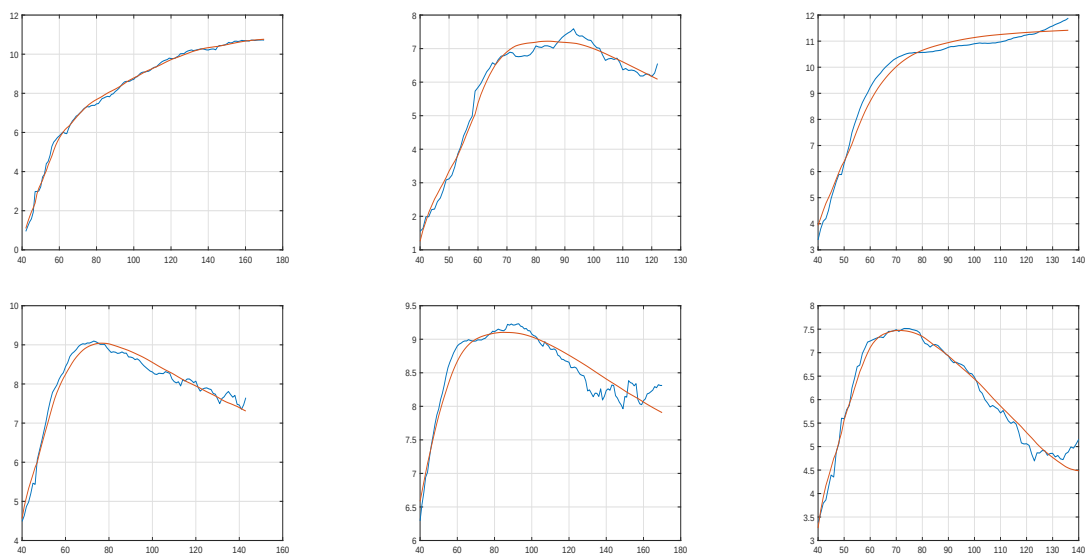

Figure 4: Logarithmic incremental affected cases (blue) and their linear regressions on logarithmic infected cases (red), each country in its training period. From top left to bottom right: Brazil, Israel, USA, Germany, Italy, Switzerland.

**The training period** $[M_1, M_2]$. The first task is to identify a period of stationarity, to be applied as the training period. It starts not earlier than $T_1$, say 40, and ends not earlier than $T_2$, say 90, nor later than $T_3$, say 170. If there are indications that the first two constraints are too wide, the country is left without analysis. Attempts are made to base the definition of the training period on the relatively smooth function $Y$ only.

$M_1$ **- beginning of the training period**. Rather arbitrarily, the training period will be set to start at the first day $M_1$ from day $T_1$ onwards, when the number of infected cases $Y$ is at least some number, like 5.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

| Country | $\alpha$ | SE $\alpha$ | $K\beta$ | $\gamma$ | $K$ | $M_1$ | $M_2$ | Delay |
|---|---|---|---|---|---|---|---|---|
| Brazil | 0.8323 | 0.0737 | 0.8282 | 0.0643 | 9.6 M | 42 | 170 | 3 |
| Israel | 0.9141 | 0.1019 | 0.4675 | 0.0555 | 17100 | 40 | 122 | 11 |
| USA | 0.7516 | 0.0775 | 2.2773 | 0.0099 | 2.24 M | 40 | 136 | 10 |
| Germany | 0.7680 | 0.0768 | 2.1045 | 0.0841 | 218300 | 40 | 143 | 5 |
| Italy | 0.6984 | 0.0605 | 5.0417 | 0.0370 | 255000 | 40 | 170 | 5 |
| Switzerland | 0.8058 | 0.0860 | 1.4178 | 0.1242 | 35500 | 40 | 140 | 8 |

Table 1: Parameter estimation, recovered cases calibration and determination of training period.

$M_2$ **- end of the training period**. The first date to identify is the day $t_1$ when $\log(\frac{Y_{t+1}}{Y_t})$ is minimal. Let $LRY_1$ be this minimal value.

If $LRY_1 \geq 0$, $M_2$ is taken as $t_1$, where the increasing function $\log(Y)$ turned locally from concave to convex.

If $LRY_1 < 0$, identify next the maximal date $t_2$ such that $Y$ is an increasing function on $[M1, t_2]$, and let $LRY_2$ be the minimum of $\log(\frac{Y_{t+1}}{Y_t})$ on $[t_2, T_3]$, attained at day $t_3$. If $t_3 = T_3, M_2 = T_3$. This means that $Y$ increased until day $t_2$ and $\log(Y)$ oscillated thereafter without experiencing a significant local drop.

It remains to define $M_2$ if the significant drop occurred at $t_3 < T_3$. In this case, $M_2$ is the maximal $t \leq T_3$ such that $Y$ is decreasing on $[t_3, t]$.

**Solving the regression equation (3)**. This is a regular regression equation on $[M_1, M_2]$ for fixed $K$. $K$ is chosen by maximizing the correlation coefficient between the independent and dependent variables. Figure 4 displays the dependent variable and its linear predictor.

There are alternative estimation methods. Alon and Meilijson [1] adopted the random time transformation method of Bassan et al [2] for parameter estimation in differential equations, similar to the diffusion methods via Fokker-Planck equations. Whichever method is used, the triple $(\alpha, \beta, \gamma)$ can be then substituted in the ISPS formula (3).

# 5 Country by country results

The maximal number of backward shifts of recovered cases allowed was 10, except for Switzerland where it was taken as 1 because of numerical instability. The global DAY delay added six more days.

Table 1 presents the results of the statistical analysis of the countries under study.

Brazil and Switzerland have $\alpha$ at distance 2.3 SE below 1. For Israel the distance is 0.84 SE, for Germany and USA 3 SE and for Italy 5 SE. The parameter $\alpha$ is thus significantly fractional.

If USA had consistently twice the number of recovered cases, its parameter $\beta$ would have been closer but still smaller than that of other countries. Its proportion of dead cases would then compare with those of Brazil and Italy, with all three stabilizing towards the end of the observation period at twice those of Germany, Israel and Switzerland.

# 6 ISPS - Implied susceptible population size

This notion, defined in (3) and explained in the Introduction, mimics implied volatility in Finance.

Figure 5 displays ISPS during the observation period and its projection into the near future, by the scenario to be described in Section 8).

Ignoring dates beyond the end of the observation period (August 15, day 207), it is possible to see that of the six countries, Brazil stands out in the stability of $K$ around 9 million, with ISPS only slightly above it towards the end of the observation period. All other five countries display to

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

| Country | $\alpha$ | SE $\alpha$ | $K\beta$ | $\gamma$ | $K$ | $M_1$ | $M_2$ | Delay |
|---|---|---|---|---|---|---|---|---|
| Brazil | 0.8313 | 0.0736 | 0.8282 | 0.0625 | 9.4 M | 42 | 170 | 10 |
| Israel | 0.9188 | 0.1024 | 0.4690 | 0.0625 | 17200 | 40 | 122 | 12 |
| USA | 0.7930 | 0.0815 | 1.9079 | 0.0625 | 2.80 M | 40 | 137 | 13 |
| Germany | 0.7371 | 0.0735 | 2.5190 | 0.0625 | 212000 | 40 | 147 | 3 |
| Italy | 0.6941 | 0.0626 | 3.9268 | 0.0625 | 269700 | 40 | 170 | 8 |
| Switzerland | 0.7526 | 0.0761 | 1.6646 | 0.0625 | 32800 | 40 | 148 | 10 |

Table 2: Parameter estimation (other than $\gamma$, fixed at $\frac{1}{16}$), recovered cases calibration and determination of training period.

a larger or smaller extent stability of $K$ between days 90 and 120, and increasing ISPS thereafter. The epidemic appears as gradually affecting an expanding target population.

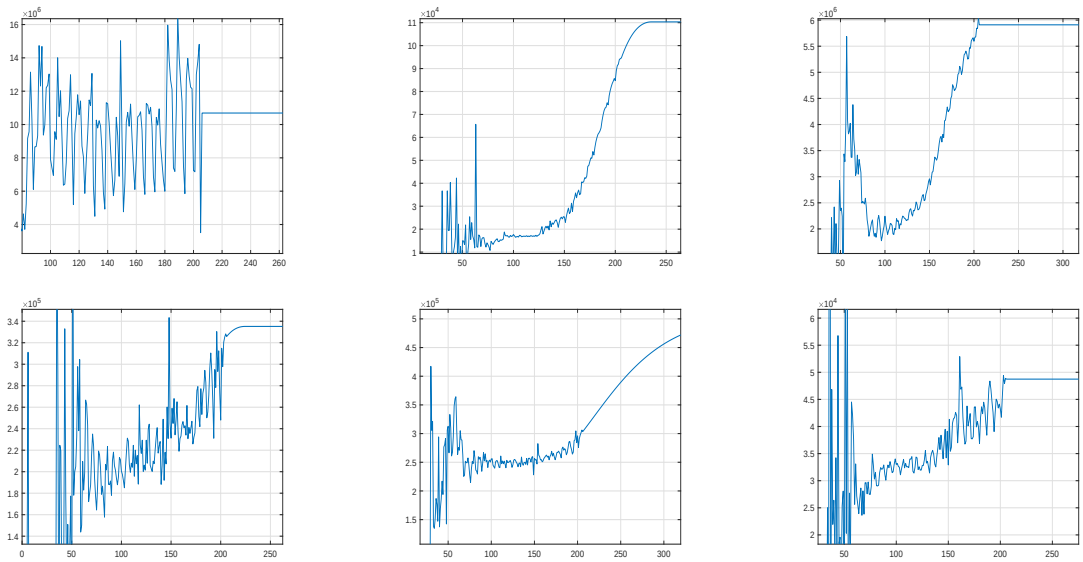

Figure 5: ISPS: Implied susceptible population size ISPS. Empirical measurement until day 207 (August 15) and projected scenario beyond. From top left to bottom right: Brazil, Israel, USA, Germany, Italy, Switzerland.

Since (the ideally constant) ISPS function is viewed as a stochastic process, it is convenient to monitor its logarithmic daily rate of growth function "Delta Log K" $DLK(t) = log(\frac{K(t+1)}{K(t)})$. Under stationary behavior, this function should in principle be zero and in practice it should oscillate around zero. It should be stressed that a period of (positive) constancy of DLK signifies exponential growth of the target population. Let us examine DLK in greater detail.

$$DLK(t) = \log(\frac{K(t+1)}{K(t)}) \approx \frac{K(t+1)}{K(t)} - 1 \approx \frac{\Delta(X(t))}{X(t)} \frac{1}{1 + \frac{\Delta(X(t))}{X(t)} \frac{1}{\beta Y(t)^{\alpha}}} \tag{9}$$

Approximation (9) suggests that $DLK$ should oscillate markedly for small $t$ and barely oscillate for large $t$. This is indeed the case, as would have been seen if Figure 7 extended back in time.

On the other hand, $XDLK(t) = X(t)DLK(t)$ should be nearly homoscedastic. The more regular behavior of $XDLK$ lends itself for fitting a trend curve for the purpose of setting a reliable estimate of $DLK$ towards the end of the observation period. The chosen tool has been linear regression of $XDLK$ on time during the last month of the observation period, to avoid giving undue weight to non

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

smoothed data. These linear functions were submitted to "testimation". If the slope is significant, the linear fit on XDLK is adopted and translated to $DLK$. If the slope is not significant but the intercept is, the constant fit is adopted. If none is significant, $DLK$ is deemed zero.

This procedure was not applied to modify the observed data. It was performed for the sole purpose of building future scenarios, as described in Section 8.

# 7 Removed cases divided into recovered and dead cases

This division is not part of the standard SIR equations. However, it is important on its own sake, and can shed some light on data quality. Figure 6 displays the proportion of incremental dead cases out of incremental removed cases, smoothed by a two-week-window moving average. As a rule, all countries displayed show a sharp decrease of this proportion along time. USA shows a decrease from 40% to 3.5% with $\gamma$ estimated as 0.01. The parallel decrease in Germany and Switzerland is from 4% or 5% to less than 1%, with respective $\gamma$ estimated as 0.085 and 0.125. Israel shows a decrease from 2.5% to less than 1%, with $\gamma$ estimated as 0.056. Brazil and Italy stand in between, with $\gamma$ assessed respectively as 0.064 and 0.037 and lethality proportion decreasing from 10%-20% to 2%.

This disparity may suggest that the identification of Covid19 as infection and cause of death differ markedly between countries. This issue needs attention.

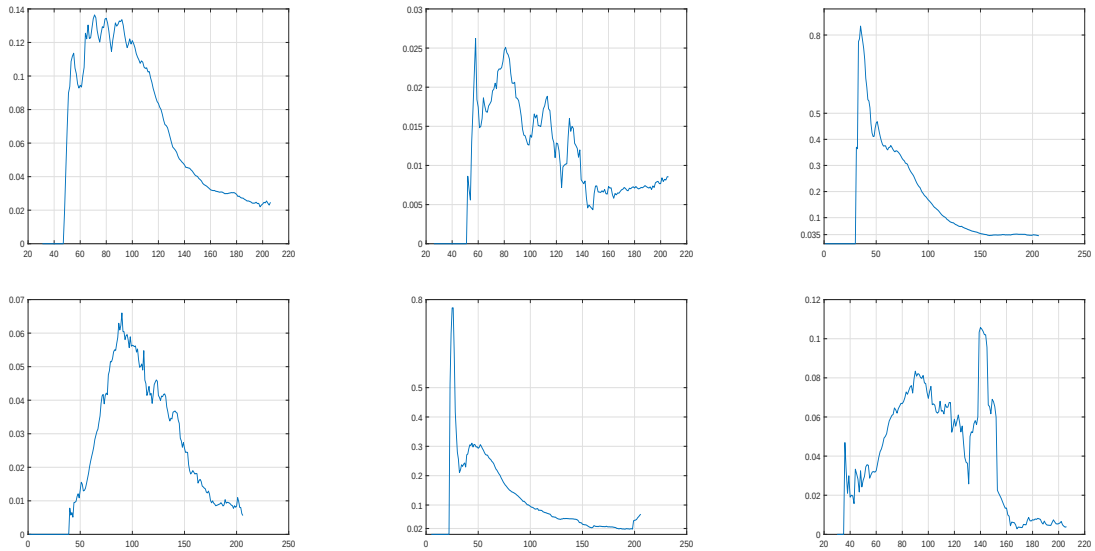

Figure 6: The proportion of incremental dead cases out of incremental removed cases. From top left to bottom right: Brazil, Israel, USA, Germany, Italy, Switzerland.

# 8 Scenario analysis - ISPS beyond the observation period

The plan is to extend the process $DLK$ into the future as a scenario. This invented projection - not prediction - is the solution to the SIR equations with the parameters $\alpha, \beta, \gamma$ fitted in the observation period, and variable $K$ derived as ISPS from the DLK extension. Consider the linear function fitted to XDLK in the last month of the observation period, translated to DLK scale as the red lines in Figure 7. Replace this translation by a linear function going through the two endpoints, and extend this linear function into the future, discounted exponentially at a daily rate of 2.5%, but constrained to be non-negative. The result is the surrogate DLK function plotted in Figure 7 beyond the observation period. The corresponding ISPS functions are plotted in Figure 5, and the

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

solutions to the SIR equations, during and beyond the observation period, are plotted in Figure 8. The projected behavior of the number of infected cases becomes apparent. The dates and heights of the local maxima are marked.

The projected numbers of affected, removed and infected cases have been plotted for two or three months beyond the observation period for illustration purposes. There is no solid basis for claiming predictive power.

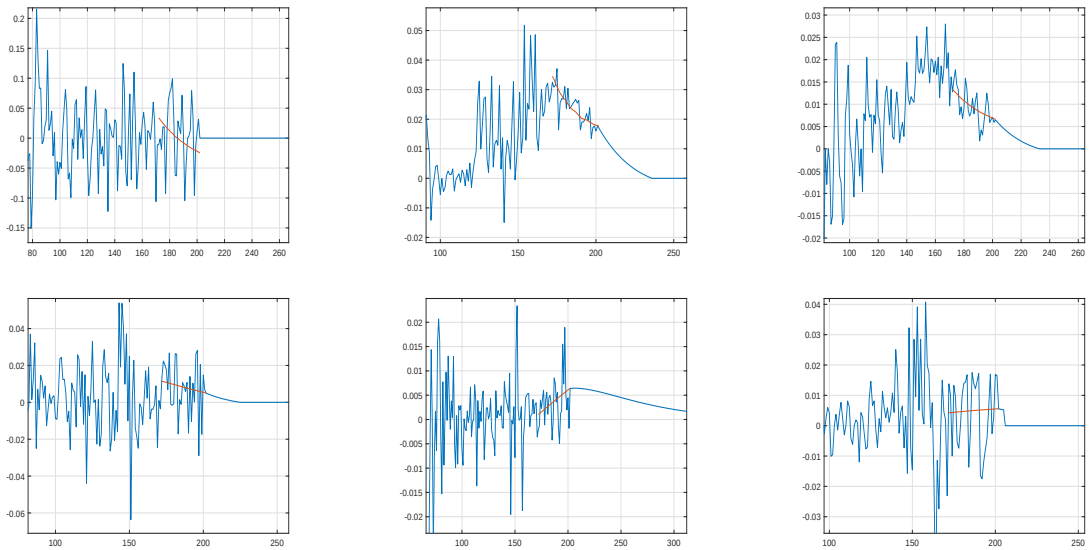

Figure 7: DLK: Incremental logarithmic ISPS. Empirical measurement until day 207 (August 15) and projected scenario beyond. From top left to bottom right: Brazil, Israel, USA, Germany, Italy, Switzerland.

# 9 Conclusions

Analysis of Covid19 dynamics has been performed on and restricted to the most reliable publicly available data, the data repository at Johns Hopkins University. Evidence has been found that recovered cases are only partially recorded in some countries, and the delay in recognizing recovered cases seems longer than the recognition of affection or death. Attempts to adjust the data so as to correct effects of this delay, have led to date-aligned data that satisfies the relation between removed and infected cases claimed by the SIR $\gamma$-equation in epidemiology (Kermack and McKendrick [7]).

The model followed is at variance with the usual interpretation of the SIR $\beta$-equation that relates the number of new affected cases to the number of currently infected cases and the size of the susceptible population. Evidence has been provided that a fractional power $\alpha < 1$ of the number of infected cases (as suggested by Grenfell et. al. [4]) fits data better than the number of infected cases itself ($\alpha = 1$), and that it is justified and useful to allow the data to uncover the extent of the susceptible population.

The size of the susceptible population has been considered a parameter to be estimated together with $\alpha, \beta$ and $\gamma$ in a period of stability, and inferred on daily data thereafter, as ISPS, the implied susceptible population size. ISPS can be instrumental in extending dynamics to the near future following scenario assumptions.

The methods presented have been illustrated on the data of each of six countries.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

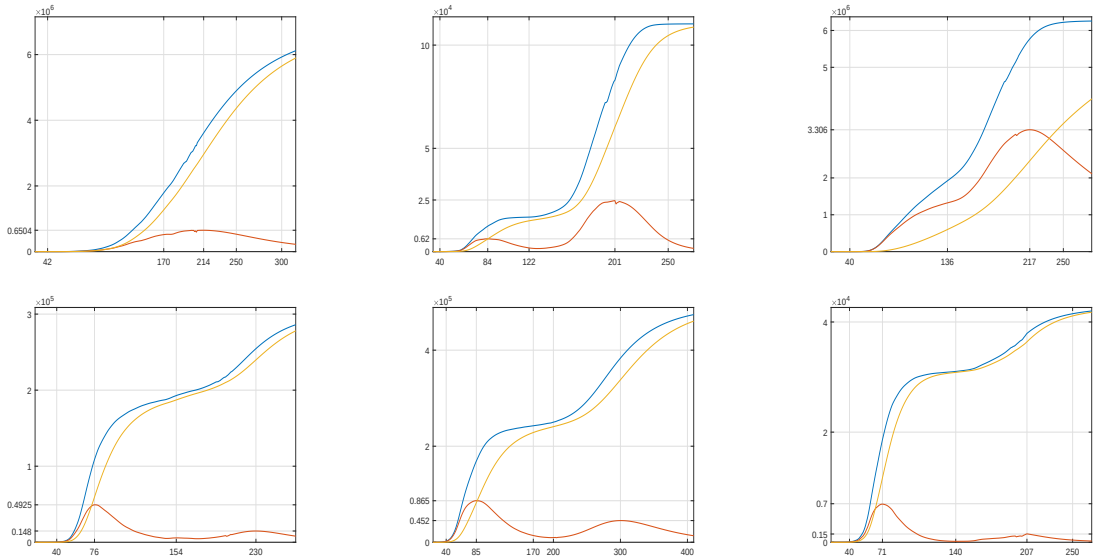

Figure 8: SIR equations solution under the scenario, with marked local maxima of the number of infected cases. From top left to bottom right: Brazil, Israel, USA, Germany, Italy, Switzerland.

# 10 APPENDIX: Covid19, 2020 - WARNING OF ERROR IN USA DATA

**Abstract**

The daily data made available by Johns Hopkins Data Repository since January 22, 2020 includes the cumulative number of affected, recovered and dead cases. The number of currently infected cases is obtained by subtracting removed cases (recovered and dead) from affected cases. The daily number of new removed cases should be roughly proportional to the number of currently infected cases. This rate ranges from 4% to 8% in most countries to be illustrated, but is estimated as 1% in the USA. The current report brings evidence that the number of recovered cases in the USA is heavily under-reported by Johns Hopkins Data Repository, and the number of currently infected cases is much smaller than generally accepted.

The pertinent observation is that there is a bug somewhere that records in some states (e.g., California, Florida) the number of recovered cases as zero, throughout the observation period. The program then subtracts this zero from the number of affected (minus dead) cases, obtaining as a result a number of infected cases very close to the number of affected cases.

The Johns Hopkins and Worldometer data get progressively distant from each other, differing by over one million cases at the writing of this report.

## 10.1 Description of Figure 9

Figure 9 displays the data provided by the Johns Hopkins Data Repository for six countries. The number of recovered cases (yellow) is below but relatively close to the number of affected cases (blue) in all countries except USA. The number of dead cases in each country appears in red. In all of these countries except USA the number of infected cases (purple) is relatively small with respect to the number of affected cases, except perhaps at the peak of the first wave. In the USA it seems that most affected cases in the whole period, are still infected at the end of the observation period August 12.

The natural conclusion is that the number of recovered cases is heavily under-reported. This brought the author to look at the source data.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

### 10.2 Zero recoveries reported in various states of the USA

The problem seems to be a bug in recording data. California, Florida and other states have a column of zeroes for recovered cases, but the program still considers this legitimate and calculates a number of infected cases almost identical to the number of affected cases, since luckily the death rate is small.

Since the data on recovered cases is missing, it is difficult to assess the true extent of infected cases. This extent will hopefully become known once the data damage is repaired.

A scenario was tried, under which the number of recovered cases was taken as twice the reported number. The estimated removal rate increases from 1% to just over 3%, still below the rate in all other counties illustrated here, so the correct number of recovered cases should be more than twice the accepted one.

The data is not provided in full, but the reader is encouraged to consult the Data Repository. A comparison with another common source follows.

### 10.3 Comparison between Johns Hopkins and Worldometer

Among the main public sources, we restrict attention to Johns Hopkins and Worldometer. Generally, the two agree closely. In the data of affected and dead cases, the two agree in the USA too. But they do not on the data for currently infected cases.

On June 1st there is a 162000 cases difference between the 1.120.038 infected reported by Worldometer and the 1.282.530 reported by Johns Hopkins.

The difference between the two becomes 285000 on June 12,
416000 on June 27,
520000 July 12,
775000 July 27 and
989000 August 12, when Worldometer reports 2.384.078 and Johns Hopkins reports an additional million.

For the sake of brevity, the report stops here, omitting further data and analysis. For further details, please consult the reference cited or contact the author. Communication is welcome and encouraged.

## Acknowledgements

Thanks are due to Nitay Alon, Yoav Benjamini, Ilan Eshel, Peter Bickel, Eytan Ruppin, Laura Sacerdote and David Steinberg for helpful suggestions.

The data analyzed in this work is taken from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

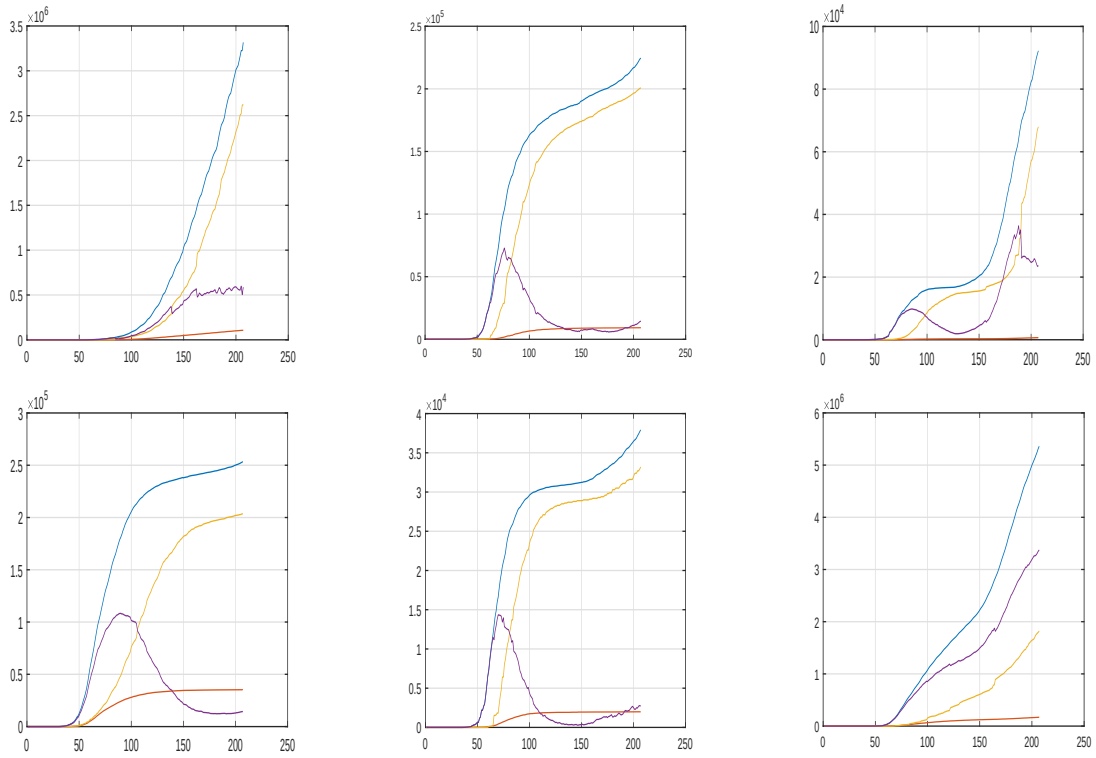

Figure 9: Raw Johns Hopkins data, Jan 22 2020 to August 12 2020. Blue: affected cases, yellow: recovered cases, red: dead cases, purple: infected. From top left to bottom right: Brazil, Germany, Israel, Italy, Switzerland and USA.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

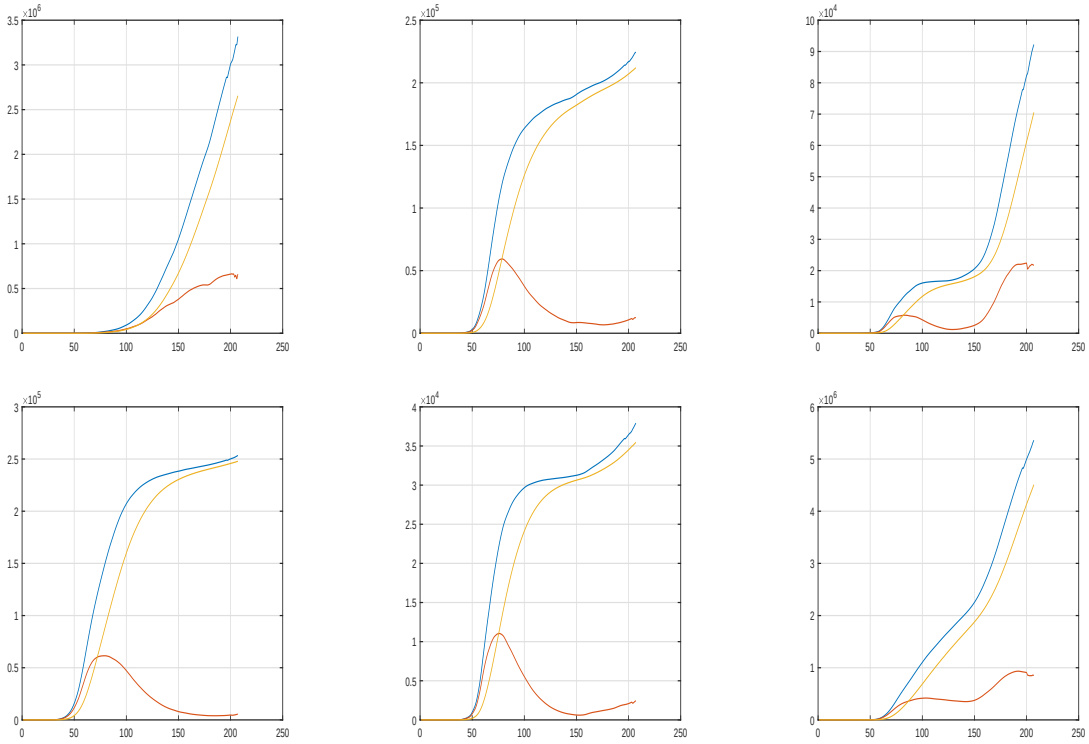

Figure 10: $\gamma = \frac{1}{16}$ modified assessments of infected cases, Jan 22 2020 to August 12 2020. Blue: affected cases, yellow: recovered cases, red: infected. From top left to bottom right: Brazil, Germany, Israel, Italy, Switzerland and USA.

medRxiv preprint doi: https://doi.org/10.1101/2020.07.12.20151860; this version posted September 12, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-ND 4.0 International license .

# References

[1] Alon, N. and Meilijson, I. (2020). Random time transformation analysis of Covid19 2020. MedRxiv.

[2] Bassan, B., Marcus, R., Meilijson, I. and Talpaz, H. (1997). Parameter erstimation in differential equations, using random time transformations. *Journal of the Italian Statistical Society*, **6**, 177–199.

[3] Black, F. and Scholes, M. (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, **81 (3)**, 637-654.

[4] Grenfell, B.T., Bjørnstad, O.N. and Filkenstädt, B. A. (2002) Dynamics of Measles epidemics: scaling, noise, determinism and predictability with the TSIR model. *Ecological Monographs*, **72(2)**, 185-202.

[5] Harris, T.E. (1974). Contact interactions on a lattice. *Annals Probab*, **2 (6)**, 969-988.

[6] Holmdahl, I. and Buckee, C. (2020). Wrong but useful what Covid-19 epidemiologic models can and cannot tell us. *N Engl J Med*, **383**, 303–305. DOI: 10.1056/NEJMp2016822

[7] Kermack, W. O., McKendrick, A. G. (1927). A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society A*, **115 (772)**, 700-721.

[8] Meilijson,I. (2020). Implied susceptible population size: modelling and projecting Covid19 dynamics via the SIR equations. MedRxiv.