Exploring COVID-19

Nils Bertschinger

April 8, 2020

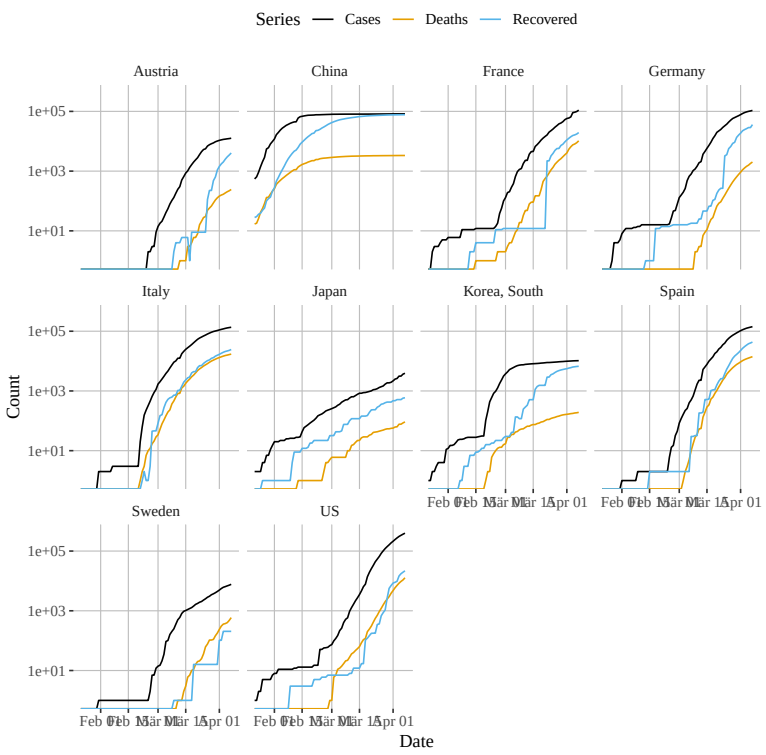
The world stands still ... desperately observing the unfolding of the global COVID-19 pandemic.

Data exploration

The John Hopkins university and other institutes publish daily numbers of cases and death tolls. Here, we build on their data sets and provide some simple explorations and modeling.

Country comparison

Fig. 1 shows the raw data for several countries [I].



[I] Here, we only consider these countries in the following

Figure 1: Data as provided by the John Hopkins university for some selected countries.

As the beginning of the epidemics is different in different countries a direct comparison is difficult. Furthermore, especially the count of cases is highly debated and plaqued with several uncertainties. Here, we assume that the *death counts are reliable* and essentially correct. Thus, in order to compare different countries we align all curves such that day 0 corresponds to the first day that the death count reaches a specified threshold (either absolute or relative per million inhabitants).

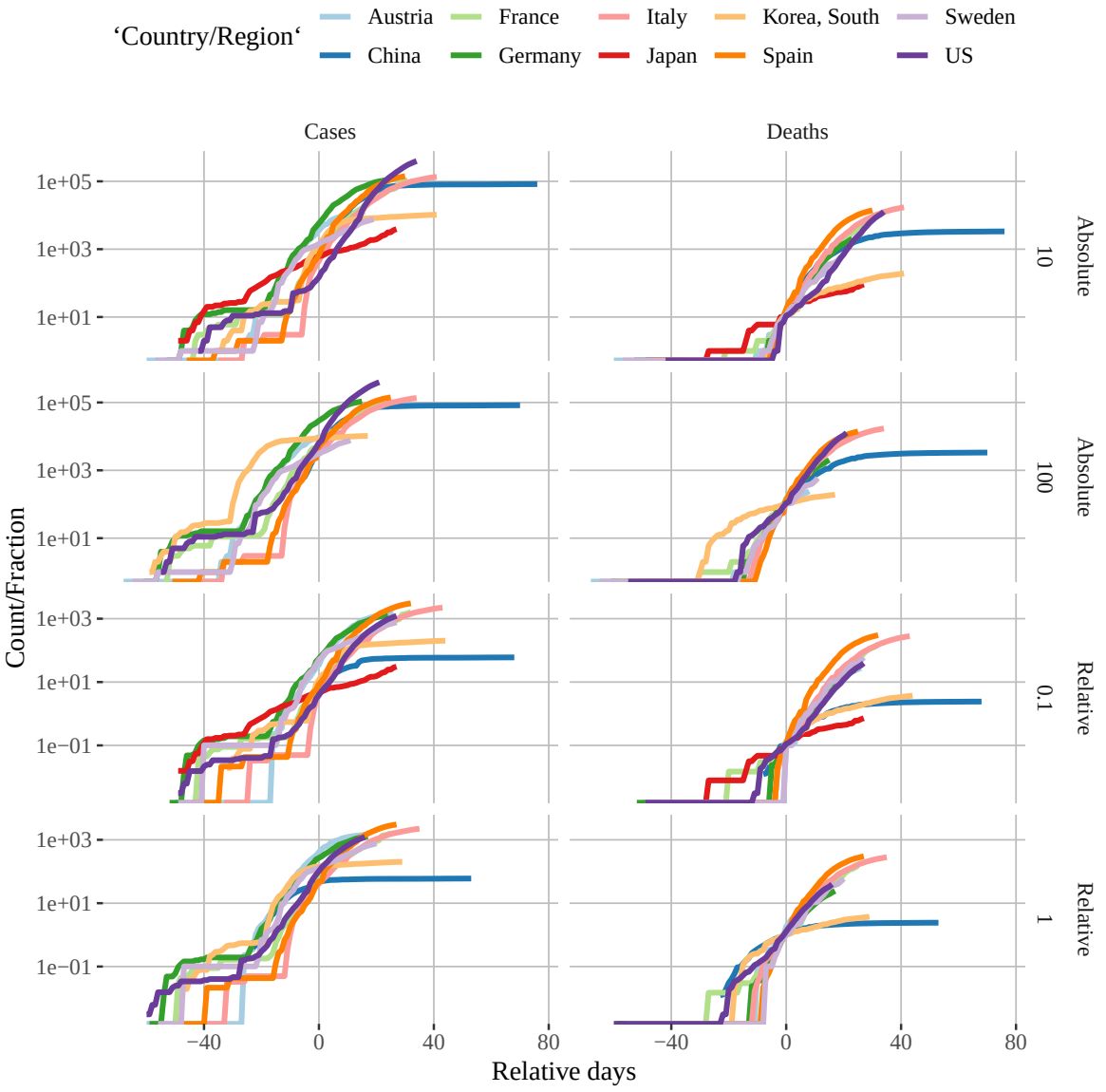


Figure 2: Case and death counts aligned to first day of more than a specified threshold (either absolute or relative per million inhabitants).

Fig. 2 shows a clear data collapse, especially at higher thresholds[II] – which are less noisy – of 100 (absolute) or 1 per million (relative). Furthermore, it is evident that the current growth rate of China, South Korea (which are very similar in relative terms) and Japan is markedly slower than of the other countries. Especially, the European countries appear quite similar and there is little (if any) evidence that Germany reacted faster or better than Italy or Spain.

[II] Note that some countries might not have reached these higher thresholds and are therefore not included in every subplot.

Interestingly, also the case counts are somewhat aligned even though day zero has been defined purely based on the death counts. Furthermore, there appear to be two groups of countries with different systematic delays between case and death counts. In particular, Austria and Germany seem to report deaths consistently later than France, Italy and Spain. This suggests that the surprisingly low death rate reported for Austria and Germany could be an artefact as reported numbers are simply some days older compared to other countries! This *delay effect* makes comparing numbers from different countries difficult and also leads to unreliable estimates when naively comparing numbers from same days only. Similarly, judging the effectiveness of containment measures, e.g. social distances, requires time as well within the second week after the intervention has been established a majority of observed cases had probably been infected already before the intervention.

Statistical modeling

Besides the *case fatality rate* there are two major unknowns in the current pandemic. Namely, the effective transmission or reproduction rate of the virus and the fraction of observed cases. Both are of major importantance in order to judge the state of the pandemic, i.e. how much of the population is already infected, and the effectiveness of mitigation measures such as *social distancing* that have been or are being implemented around the globe. In particular, the future organization of social interactions and restrictions relies on as accurate information as possible.

Epidemic model The basic SIR model [III], assumes that an infection unfolds when susceptible (S) individuals become infected (I) – which in turn infect further susceptible individuals. Finally, infected individuals recover (R) (or die) and are no longer susceptible. In continuous time, the dynamics can be described by the following system of ordinary differential equations (ODEs):

[III] M. Newman. *Networks*. Oxford University Press, 2nd edition, 2018

$$\begin{aligned}\frac{dS}{dt} &= -\beta \frac{I_t}{N} S_t \\ \frac{dI}{dt} &= \beta \frac{I_t}{N} S_t - \gamma I_t \\ \frac{dR}{dt} &= \gamma I_t\end{aligned}$$

where $N \equiv S_t + I_t + R_t$ is constant over time. Model parameters are

- the infection rate β

- and the recovery rate γ.

In this model, the average time of infection is γ^{-1} giving rise to a *basic reproduction number* of $R_0 = \beta\gamma^{-1}$.

SIR models and extensions are widely used in epidemic modeling. The have also been applied to the understand the dynamics of the ongoing Covid-19 pandemic [IV]. In particular, models including the possibility of unobserved cases or including a reporting delay have been developed. Within the SIR framework, both effects can be included in several ways, most easily by assuming that observed cumulative infections are simply a fraction $\alpha \in [0,1]$ of previous total infections $I_t + R_t$, i.e. $\alpha(I_{t-\tau} + R_{t-\tau})$. A more elaborate attempt instead considers more detailed dynamics of the form

$$\begin{aligned}
\frac{dS}{dt} &= -\beta_I \frac{S_t}{N} I_t - \beta_O \frac{S_t}{N} O_t - \beta_U \frac{S_t}{N} U_t \\
\frac{dI}{dt} &= \beta_I \frac{S_t}{N} I_t + \beta_O \frac{S_t}{N} O_t + \beta_U \frac{S_t}{N} U_t - \gamma_I I_t \\
\frac{dO}{dt} &= \alpha\gamma_I I_t - \gamma_R O_t \\
\frac{dU}{dt} &= (1-\alpha)\gamma_I I_t - \gamma_R U_t \\
\frac{dR}{dt} &= \gamma_R (O_t + U_t)
\end{aligned}$$

where a fraction α of infected individuals I_t is observed (O_t) after an initial delay $\frac{1}{\gamma_I}$. In any case, whether observed or not, individuals recover (or die) after an additional delay. In general, the infection rates $\beta_I, \beta_O, \beta_U$ could be different for initial infections and observed vs unobserved cases[V].

In addition, mitigation measures, e.g. social distancing, can be easily included by assuming that β's are functions of time. E.g. [VI] assumes one or several (soft) step functions where β drops after measures have been implemented. Unfortunately, as we show now a model including a time-varying β as well as unobserved cases is not identifiable. For simplicity, consider the above model with $\beta_I = \beta_O = \beta_U =: \beta$. Then, new infections arise with intensity $\beta\frac{S_t}{N}(I_t + O_t + U_t)$ which in turn translate into observed cases with intensity $\alpha\gamma_I I_t$. Now assume a second model with $\alpha' = 1 > \alpha$. By using a time varying $\beta'(t)$ such that

$$\beta'(t) = \beta\frac{S_t}{S'_t}$$

we obtain exactly the same number of observed cases O_t. Note that as $\alpha' > \alpha$, we have that $S_t < S'_t$ and S_t is a sigmoidal function of time due to the SIR dynamics. Furthermore, when the population is large, i.e. $N \gg 1$ and $S_0 \approx N$ the resulting $\beta'(t)$ is mostly driven by the drop in S_t as compared to the much smaller change in S'_t. Indeed, Fig. 3 shows the dynamics of the above model with $\beta = 0.3, \gamma_I = \gamma_R = \frac{2}{10}$ [VII] and $\alpha = 0.1$ starting from $(N = 10^8, 1, 0, 0, 0)$. In turn, the dynamics is approximated by the best-fitting logistic sigmoid scaling $\beta'(t)$ and assuming $\alpha' = 1$. Note that the number

[IV] B. F. Maier and D. Brockmann. Effective containment explains subexponential growth in confirmed cases of recent covid-19 outbreak in mainland china, 2020; J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann. Inferring covid-19 spreading rates and potential change points for case number forecasts, 2020; R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*, 2020; and S. Zhao and H. Chen. Modeling the epidemic dynamics and control of covid-19 outbreak in china. *medRxiv*, 2020

[V] An effective quarantine would be modeled via $\beta_O \equiv 0$.

[VI] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann. Inferring covid-19 spreading rates and potential change points for case number forecasts, 2020

[VII] Giving rise to an R_0 of 3.

of observed cases is almost identical whereas the final fraction of susceptible individuals is vastly different. Indeed, in the first case the epidemic is stopped by group immunity whereas in the second case effective mitigation measures are imposed. Correspondingly, police implications would be vastly different in the two situations even though they are observationally indistinguishable.

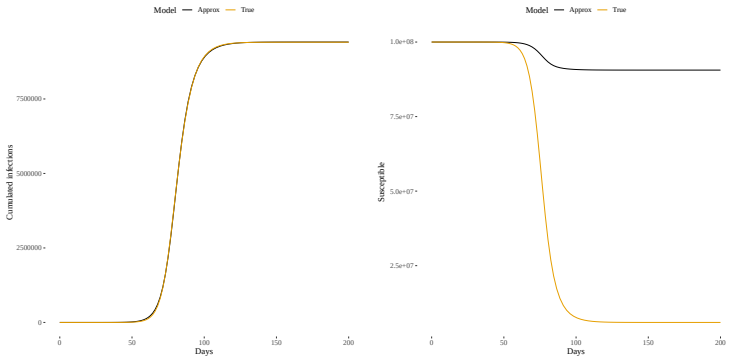


Figure 3: Total cumulative observed infections and number of susceptible individuals in two simulated model with observation fraction $\alpha = 0.1$ (true) and observation fraction $\alpha' = 1$ (approx). In the second model, the epidemic is stopped due to mitigation measures which are modeled via $\beta'(t)$ as explained in the main text.

Phenomenological growth model In order to build an identifiable model, we turn to a simpler phenomenological description (compare [VIII] for a similar approach). In particular, as sigmoidal growth provides a good approximation to the SIR model dynamics we now directly model the observed growth of reported cases and deaths. Furthermore, we use a strong assumption to identify the model.

Overall, I believe it unlikely that the death rate is very different across different countries[IX]. Thus, my phenomenological model builds on the assumption that the *death rate is constant across all countries*[X] and differences purely arise from delays in reporting positively tested cases and deaths. The model assumes the following:

- The probability of death is the same for all countries whereas the testing prevalence is country specific.
- Observed counts are negative binomial distributed – as an over-dispersed Poisson – and delayed wrt the actual cases.
- Actual case and death counts in each country grow according to a sigmoid function – as an approximation to an SIR type model – with country specific parameters[XI].

As shown in Fig. 4, this model[XII] indeed finds a consistent difference in the delay of death counts between Austria, Germany and France, Italy, Spain. Furthermore, the death rate – assumed constant across all countries – is estimated as $3.0 \pm \frac{0.8}{0.6}\%$. Estimates in the range of 3 to 5% appear to be rather robust for the present model, yet seem to be somewhat high[XIII] given that the death rate applies to the actual and not just the reported cases. It remains to be seen if my model estimate holds up over time and wrt estimates derived from more realistic models ... Yet, the assumption of a global death rate not just allows to compare observed counts across countries,

[VIII] R. Kubinec. A retrospective bayesian model for measuring covariate effects on observed covid-19 test and case counts. *SocArXiv*, April 2020

[IX] There are certainly demographic, medical or other aspects though.

[X] Note that this is not the case fatality rate which is commonly computed as the quotient of observed deaths and cases. Instead, the death rate applies to the actual, hidden number of cases and reveals itself in reported observations after the fact.

[XI] The basic model assumes

$$c(t) = \frac{a}{1 + e^{-\beta(t-\tau)}} ,$$

i.e. the logistic function. Extensions using Gompertz or generalized logistic functions are also explored.

[XII] Full code of this and other models can be found at my accompanying `https://github.com/bertschi/Covid` repository.

[XIII] Note that a naive estimation of the case fatality rate, i.e. dividing contemporal case by death counts is biased downwards by the delay effect as actual deaths only realize about a week after subjects had been tested positive. Accordingly the fatality rate estimate should be based on the substantially lower case counts a week ago.

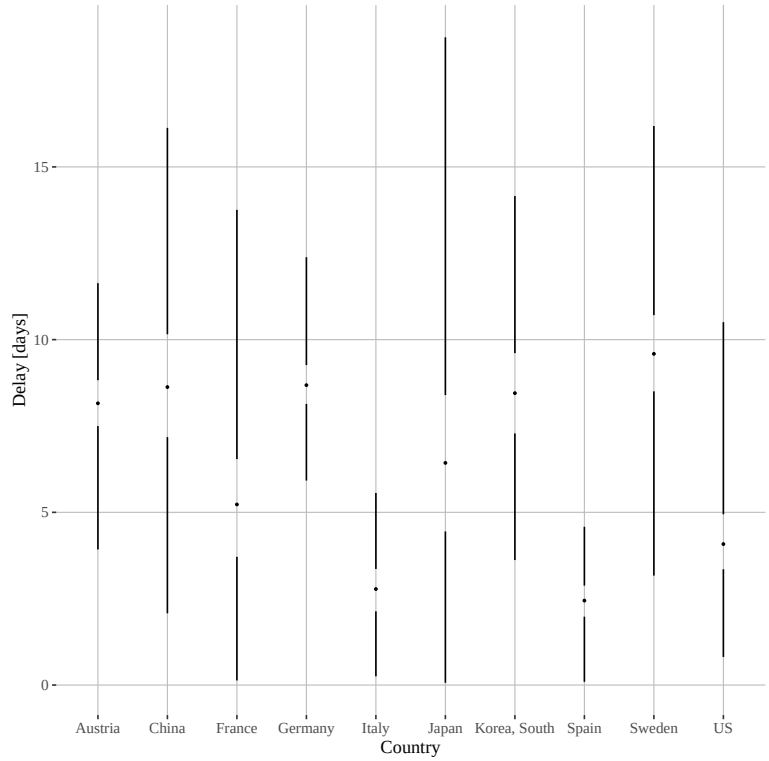
but also identifies the model as all curves are scaled in the same absolute fashion. Furthermore, jointly fitting the model on several countries thereby pools information from observed case and death counts explaining their differences purely as arising from different growth dynamics and delays.

Detailed model predictions for all considered countries are shown in Fig. 5. The predictions are mostly reasonable, but the model has difficulty of matching the rapid leveling off observed in China and South Korea – which has been shown to be rather well explained by SIR type dynamics with effective quarantine [XIV]. Interestingly, the model predicts that the curve has already slowed markedly in Germany and Italy even though this is barely visible in the raw numbers by now – another example of why the delay effect is important in understanding the dynamics of the COVID-19 pandemic. Yet, this model prediction relies heavily on the assumption of sigmoidal growth and I would not be too optimistic about it. Indeed, the predicted slow-down is less pronounced when assuming Gompertz growth which has also been observed in [XV]. Yet, in terms of leave-one-out likelihood the logistic model is preferred (not shown).

[XIV] B. F. Maier and D. Brockmann. Effective containment explains subexponential growth in confirmed cases of recent covid-19 outbreak in mainland china, 2020

[XV] W. Yang, D. Zhang, L. Peng, C. Zhuge, and L. Hong. Rational evaluation of various epidemic models based on the covid-19 data of china. *medRxiv*, 2020

Figure 4: Model estimated delay between reported case and death counts.



Discussion

Accurately understanding and modeling the ongoing Covid-19 pandemic is crucial in order to establish effective and timely counter

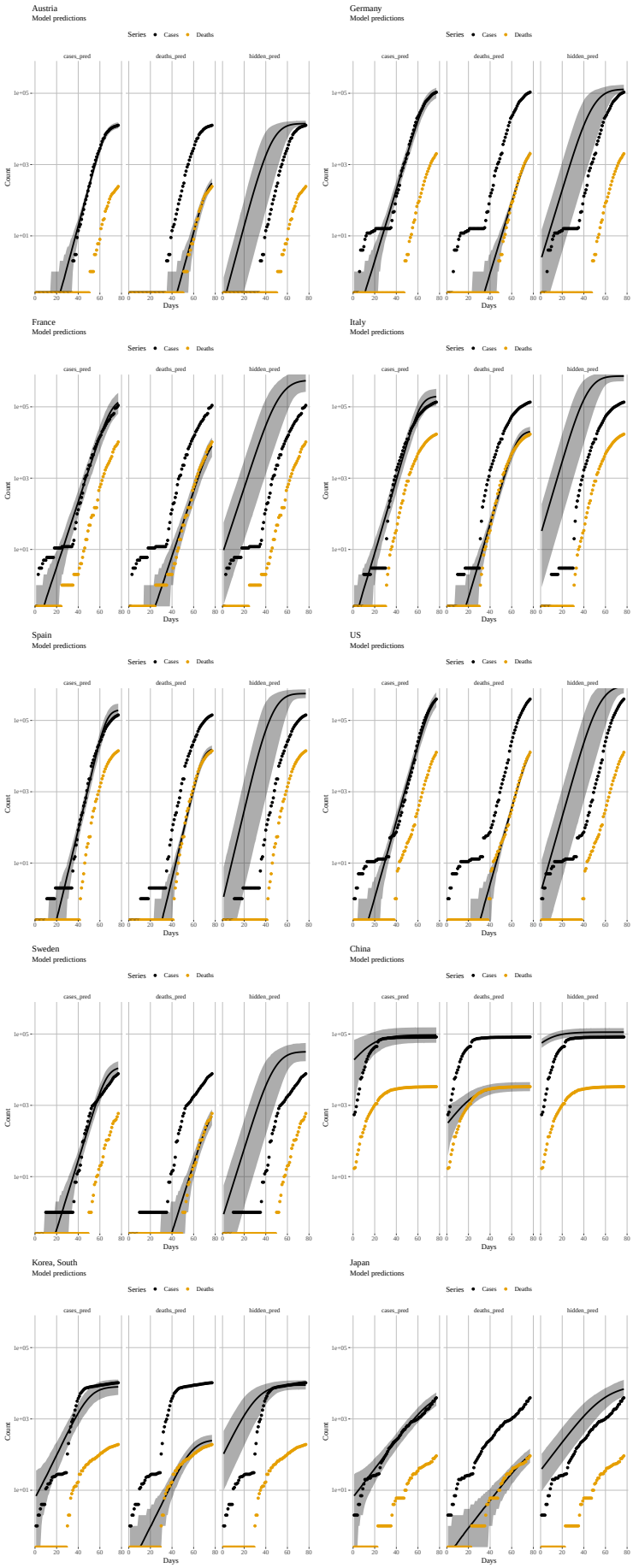


Figure 5: Model predictions (mean and 95% credible region) of actual (hidden) and observed case and death counts for different countries. Note the log scale on the vertical axis.

measures. Yet, major quantities of interest are not known precisely and especially delays of several weeks between first infection, eventual symptoms and death make it particularly challenging to quantify the effectiveness of social distancing. In turn, losening restrictions prematurely could reestablish exponential spreading with devstating consequences.

Here, I have shown that in SIR type models the fraction of reported cases and effectiveness of social distancing cannot be estimated jointly. Indeed, the model is strictly non-identifiable wrt these parameters. In turn, I have developed a phenomenological model based on sigmoidal growth dynamics. As the model does not strive to model precise epidemic dynamics, longer term predictions should be considered with care [XVI]. Here, I only provide nowcasts of actual case numbers which nevertheless provide a look into several weeks of the future due the delay effect build into the model. In particular, my main findings are that

[XVI] W. Yang, D. Zhang, L. Peng, C. Zhuge, and L. Hong. Rational evaluation of various epidemic models based on the covid-19 data of china. *medRxiv*, 2020

- the death rate might be higher than commonly assumed. This is mainly driven by the rather long delay of 5 to 10 days between reported cases and eventual death.
- the fraction of unobserved cases might be rather low and definitely does not exceed the reported case numbers tenfold as sometimes stated.

Both of these are bad news, as the epidemic could be worse than commonly assumed. Furthermore, containment is challenging due to week long incubation times [XVII] but a conbination of case tracing and isolation policies could be effective [XVIII]. On the other hand, if most cases are actually known effective policies can be implemented as evidenced in the currently successful containment in China and South Korea. In the end, data analysis alone only gets us only that far and more extensive testing is urgently needed.

[XVII] W. Bock, B. Adamik, M. Bawiec, V. Bezborodov, M. Bodych, J. P. Burgard, T. Goetz, T. Krueger, A. Migalska, B. Pabjan, T. Ozanski, E. Rafajlowicz, W. Rafajlowicz, E. Skubalska-Rafajlowicz, S. Ryfczynska, E. Szczurek, and P. Szymanski. Mitigation and herd immunity strategy for covid-19 is likely to fail. *medRxiv*, 2020

[XVIII] C. Fraser, S. Riley, R. Anderson, and N. Ferguson. Factors that make an infectious disease outbreak controllable. *Proc Natl Acad Sci USA*, 101(16):6146–6151, 2004; and R. Kubinec. A retrospective bayesian model for measuring covariate effects on observed covid-19 test and case counts. *SocArXiv*, April 2020

References

[1] W. Bock, B. Adamik, M. Bawiec, V. Bezborodov, M. Bodych, J. P. Burgard, T. Goetz, T. Krueger, A. Migalska, B. Pabjan, T. Ozanski, E. Rafajlowicz, W. Rafajlowicz, E. Skubalska-Rafajlowicz, S. Ryfczynska, E. Szczurek, and P. Szymanski. Mitigation and herd immunity strategy for covid-19 is likely to fail. *medRxiv*, 2020.

[2] J. Dehning, J. Zierenberg, F. P. Spitzner, M. Wibral, J. P. Neto, M. Wilczek, and V. Priesemann. Inferring covid-19 spreading rates and potential change points for case number forecasts, 2020.

[3] C. Fraser, S. Riley, R. Anderson, and N. Ferguson. Factors that make an infectious disease outbreak controllable. *Proc Natl Acad Sci USA*, 101(16):6146–6151, 2004.

[4] R. Kubinec. A retrospective bayesian model for measuring covariate effects on observed covid-19 test and case counts. *SocArXiv*, April 2020.

[5] R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*, 2020.

[6] B. F. Maier and D. Brockmann. Effective containment explains sub-exponential growth in confirmed cases of recent covid-19 outbreak in mainland china, 2020.

[7] M. Newman. *Networks*. Oxford University Press, 2nd edition, 2018.

[8] W. Yang, D. Zhang, L. Peng, C. Zhuge, and L. Hong. Rational evaluation of various epidemic models based on the covid-19 data of china. *medRxiv*, 2020.

[9] S. Zhao and H. Chen. Modeling the epidemic dynamics and control of covid-19 outbreak in china. *medRxiv*, 2020.

Additional figures

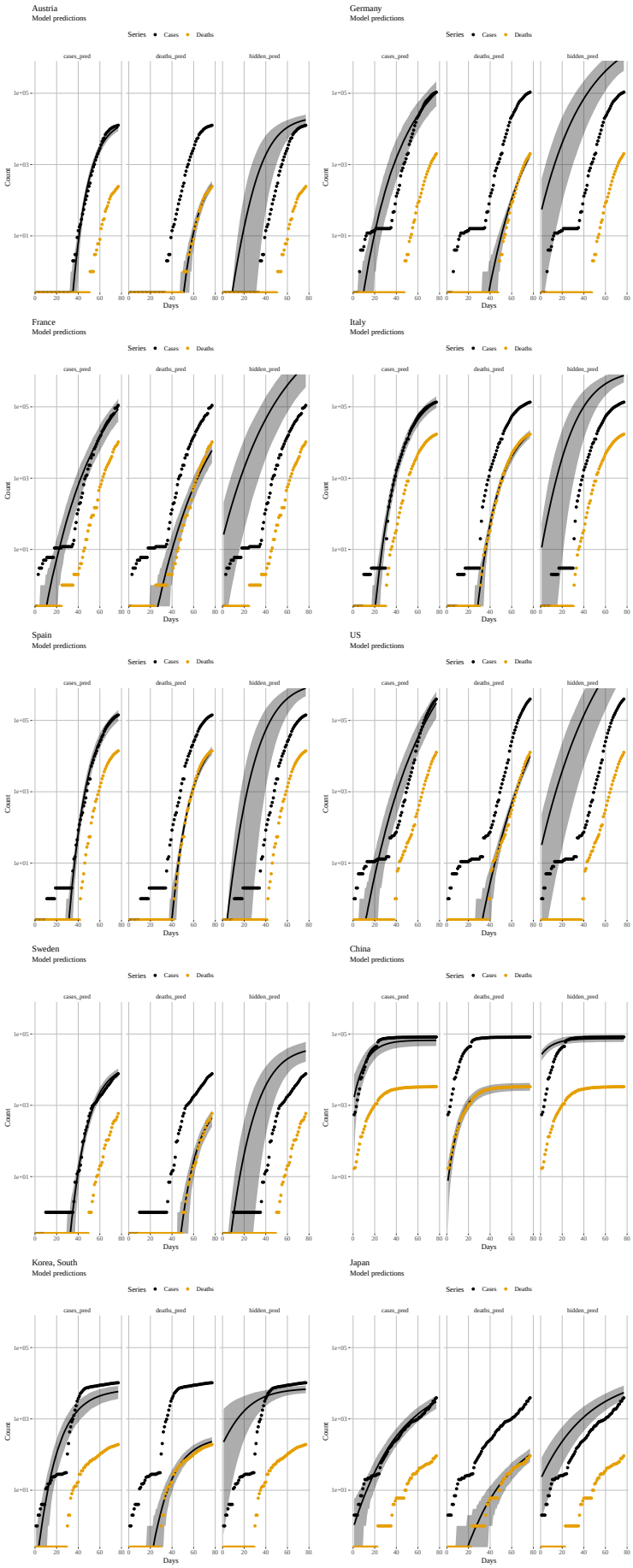


Figure 6: Model predictions using Gompertz growth dynamics (mean and 95% credible region) of actual (hidden) and observed case and death counts for different countries. Note the log scale on the vertical axis.