Covid-19: Heterogeneity in R_0 Approximations using Bayesian Statistics

Susan Holmes

joint with Claire Donnat

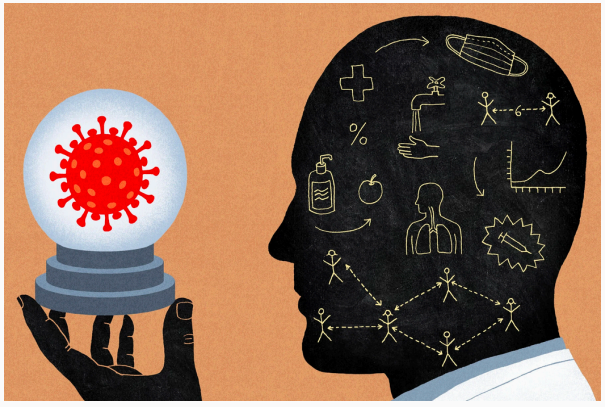
MSRI, August, 13th 2020

Statistics, Stanford
webpage

@SherlockpHolmes

Part I

Bayesian Statistics for mathematicians



`link to NYTimes`

Bayesian statistics for mathematicians/probabilists

A probability illustration (Diaconis and Holmes, 2002 [3]).

- Vanilla Birthday problem (we fix $\mathbf{p} = \frac{1}{365}$): How many people, $\mathbf{k}$, in the room for a 50-50 chance of a birthday?
- The Birthday problem for students all the same year with heterogeneous $\mathbf{p}$'s.
- Replace $\mathbf{p} = \frac{1}{365}$ by a distribution of possible probabilities that take into account for instance the difference between probabilities of a birth on a weekend and that on a weekday.

Birthday Problem with Dirichlet$(c, c, \ldots, cs)$

What k required for a $50-50$ chance of a match when $n = 365$:

- Uniform Prior, c=1 $k \doteq .83\sqrt{n}$, for $n = 365$, $k \doteq 16$
- Dirichlet$(a_1, a_2, \ldots, a_{365})$ Symmetric Prior, $a_i = c$

c	.5	1	2	5	20	∞
k_c	13.2	16.2	18.7	20.9	21.9	22.9

Honest Priors: Dirichlet$(a_1, a_2, \ldots, a_{365})$

Construct a 2 "hyper"parameter family of Dirichlet priors writing $a_i = A\pi_i$, with $\pi_1 + \pi_2 \cdots + \pi_n = 1$. Assign weekdays parameter $\pi_i = a$, weekends $\pi_i = \gamma a$, with $260a + 104\gamma a = 1$. Here γ is the parameter 'ratio of weekends to weekdays', (roughly we said $\gamma \doteq .7$) and A measures the strength of prior conviction. The table below shows how k varies as a function of A and γ. We have assumed the year has $7 \times 52 = 364$ days.

A \ γ	.5	.7	1
1	2.2	2.2	2.2
364	16.1	16.3	16.4
728	18.4	18.6	18.8
∞	22.2	22.4	22.6

The Coupon Collector's Problem

In its classical version (Laplace (1812), Feller (1968)) k balls are dropped uniformly and independently into n boxes, all boxes are **covered** if each contains at least one ball. The classical approximations (Feller (1968),page 105) show that for

$$k = n \log n + \theta n, \quad -\infty < \theta < \infty$$

$$P(cover) = \sum_{k=0}^{n} (-1)^k \binom{n}{k} (1 - \frac{k}{n})^k \doteq e^{-e^{-\theta}}$$

For example, when $n = 365$, $P(cover) \doteq \frac{1}{2}$ for $k = 2287$ or as Feller (1968) puts it : in a village of 2300 inhabitants it is about even odds that every day is someone's birthday.

Bayesian version

- With a uniform prior

$$P(\mathrm{cover}) = \frac{\binom{k-1}{n}}{\binom{n+k-1}{n}} \doteq e^{-\frac{1}{\theta}}, \text{ for } k = \theta n^2, 0 < \theta < \infty \qquad (1)$$

For example, $P_u(\mathrm{cover}) = \frac{1}{2}$ for $\theta = 1.44$.
When $n = 365$ this will need $k = 191,844$.

Thus a Bayesian with a uniform prior would think it takes a town-sized village to have even odds that every day is someone's birthday.

Bayesian version

- With a general Dirichlet prior $D_{\tilde{a}}$

$$\text{Let } A = a_1 + a_2 + \cdots + a_n$$

$$\text{and } \lambda_i = \frac{(A - a_i)(A - a_i + 1)\cdots(A - a_i + (k-1))}{A(A+1)\cdots(A+k-1)}$$

and if $\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$, then

$$P\{cover\} \sim e^{-\lambda}$$

As an example, if $a_i \equiv c$,

$$P_c(cover) \sim e^{-\left(\frac{c}{\theta}\right)^c} \quad \text{for } k = \theta n^{\frac{c+1}{c}}$$

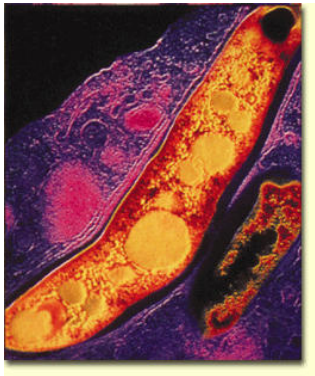
when $n = 365$, to have $P_c(cover) \doteq \frac{1}{2}$ requires k_c of:

c	1	2	5	10	20	∞
k_c	191,844	16,000	4,555	3176	2685	2297

Mathematical remarks

1. In the Birthday problem, use of a prior makes a mild difference. For the coupon collectors problem the prior makes a huge difference.
   Here, the same prior on the same underlying space seems sensible for one event (birthdays) and strange for other events (coverings).
2. Maths: We used Stein's method is to prove a Poisson approximation for the number of empty cells. This is proved with error bounds which shows that the numbers in the table above are accurate to at least 5%.

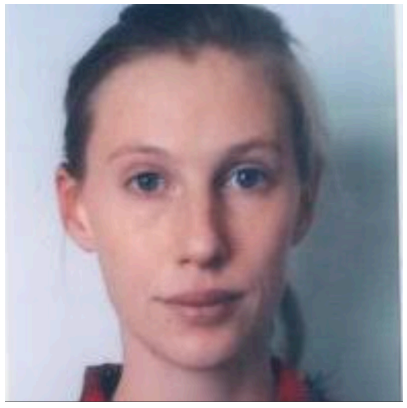
Applications: When does heterogeneity matter?



Strains of *Myobacterium tuberculosis*.
How many samples need to be collected if the strains are very unequally distributed if we want a coverage of 90% of all possible strains?
How many people need to be infected to attain a coverage of 75%?

Part II

Heterogeneity of R_0 and R (Joint work with Claire Donnat)



Asst Prof., U Chicago, Dept of Statistics.

Basic parameter and decomposition

The reproductive number R characterizes the expected number of secondary cases produced by one single typical infectious case.

This quantity can be further broken down into different categories.

- R_0 (basic reproductive number) — assumes that the population is completely susceptible for modeling a completely novel virus
- "effective" R_{eff}, assumes a mixed population of susceptible and immune hosts.

We focus on discussing how to deal with the heterogeneity of these parameters, that we replace by random variables.

$$R = \tau \times \bar{c} \times D_I \tag{2}$$

where τ is the transmissibility (i.e., probability of infection given contact between a susceptible and infected individual), $\bar{c}$ is the average number of contact per day between susceptible and infected individuals, and D_I is the duration of infectiousness.

τ and c can be considered random variables.

Basic decomposition: many sources of heterogeneity.

R thus serves as an epidemiological metric to describe the contagiousness or transmissibility of infectious agents: the outbreak is expected to continue if R is greater than 1, or to naturally subside if R is strictly less than 1.

This coefficient inherently depends on some local characteristics of the population and of the virus.

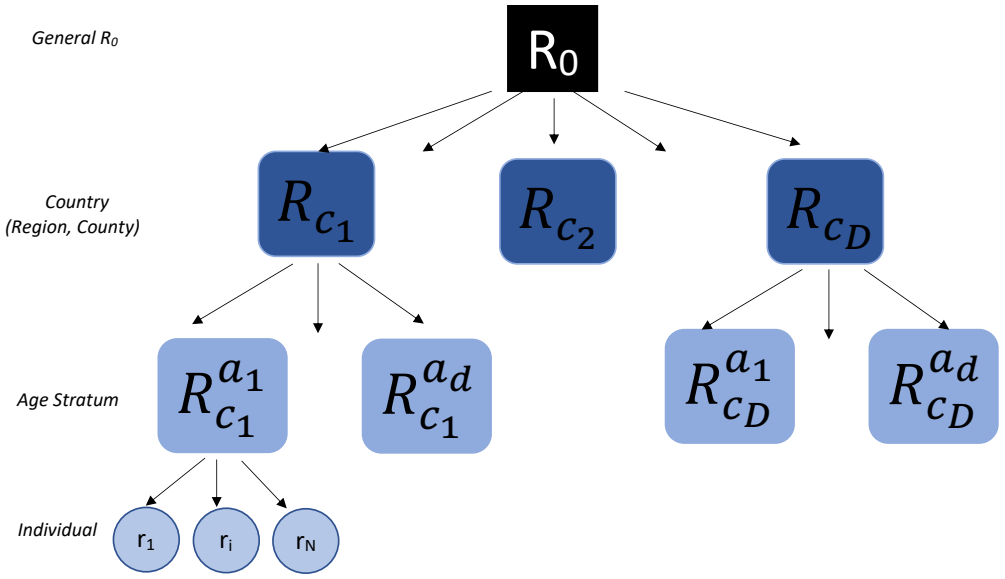
It is tied to temporal and spatially-varying factors, such as population age demographics, political or environmental variables, cultural or social dynamics, or the density of the population — all favoring or diminishing the rate of contacts $\overline{c}$ between individuals.

It can also vary according to differences in climate or infectiousness that impact $\mathrm{D_I}$ (heterogeneity of strains is also a possibility).

Hierarchical model for R

The expected number of secondary cases is contingent on each primary cases' socio-economic status, age, etc., and perhaps even time — as one could imagine the contact rates varying between weekends and weekdays.

A very fine-grain analysis of the R's heterogeneity would thus model R as a distribution over cases and time in a given population.



R could be hierarchically broken down according to regions, age groups, and, at the most granular level, across subjects.

Model the heterogeneity of the reproductive number R

Bayesian hierarchical extension to standard models of R_0.
Let G be the number of groups that we want to analyze (these could either be localized virus outbreak clusters , regions or countries).
Let N_g denote the population of each of these groups, initially assumed to be completely susceptible.

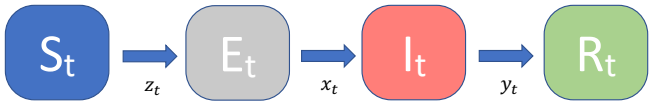


Figure 1: Standard compartmental SEIR model

Estimate the number of new cases per day using the model.

$$
\begin{aligned}
\frac{dS_k(t)}{dt} &= -\frac{S_k(t)}{N_k}\frac{R_0^{(k)}}{D_I}I_k(t) \\
\frac{dE_k(t)}{dt} &= -\frac{S_k(t)}{N_k}\frac{R_0^{(k)}}{D_I}I_k(t) - \frac{E_k(t)}{D_E} \\
\frac{dI_k(t)}{dt} &= \frac{E_k(t)}{D_E} - \frac{I_k(t)}{D_I}
\end{aligned}
\tag{3}
$$

where:

- $S_k(t), E_k(t), I_k(t)$, and $R_k(t)$ are the number of susceptible, latent, infectious, and removed individuals at time t in group k;
- D_E and D_I are the mean latent (assumed to be the same as incubation) and infectious period (equal to the serial interval minus the mean latent period);
- $R_0^{(k)}$ is the basic reproductive number is population k.

Deterministic equations do not provide natural uncertainty quantification for estimates of R_0, nor account for heterogeneities.

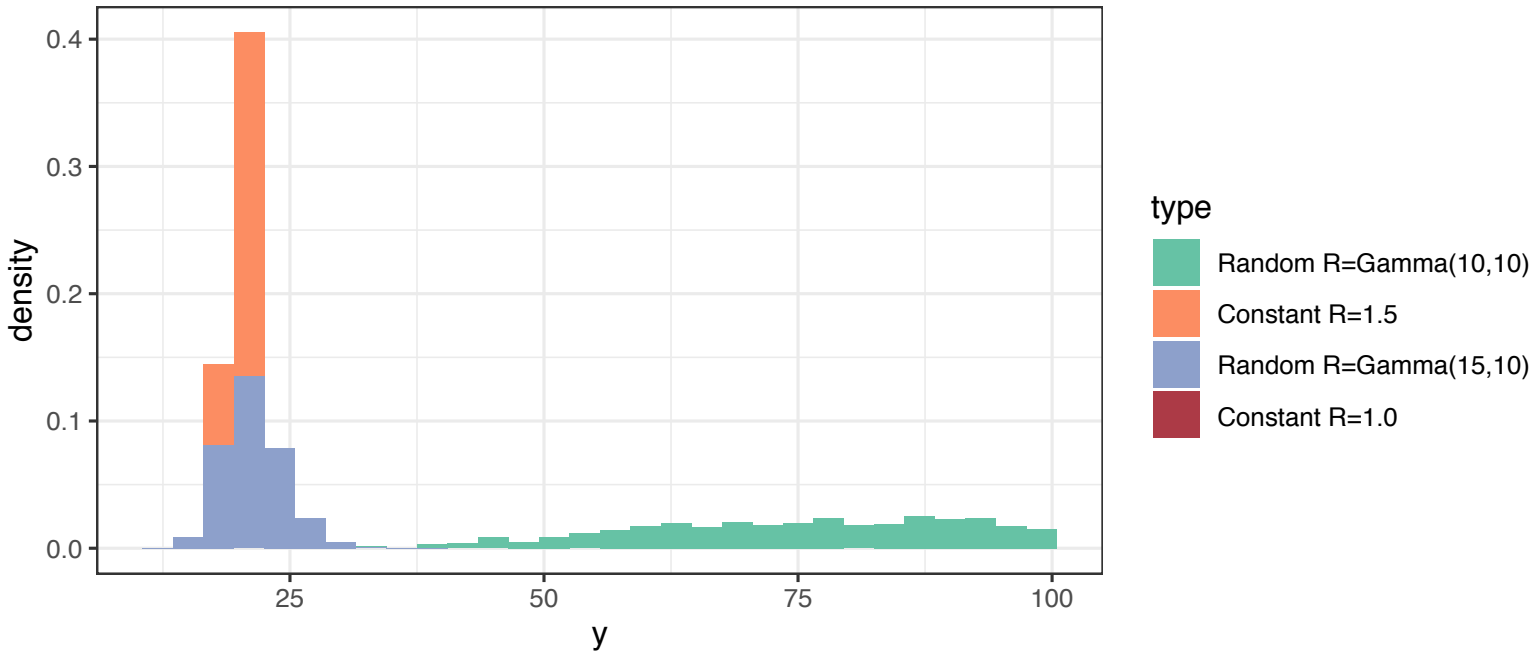
Stochastic components in SEIR models as in the study of Ebola [11].

Doing this using Bayesian methods avoids the identifiability issues associated to simply adding more parameters to account for the heterogeneity of the basic reproductive numbers R_0.

Similar to a non-parametric model by Fraser [7], also used for estimating R_0 in Cori et al [2].

A version of this model was implemented in the R-package `EarlyR` [12], which has been used in recent studies[13].

Simplest case: time until 5,000 deaths



Comparison of the distributions of the stopping time (number of deaths reaching 5,000) for varying Rs vs constant R

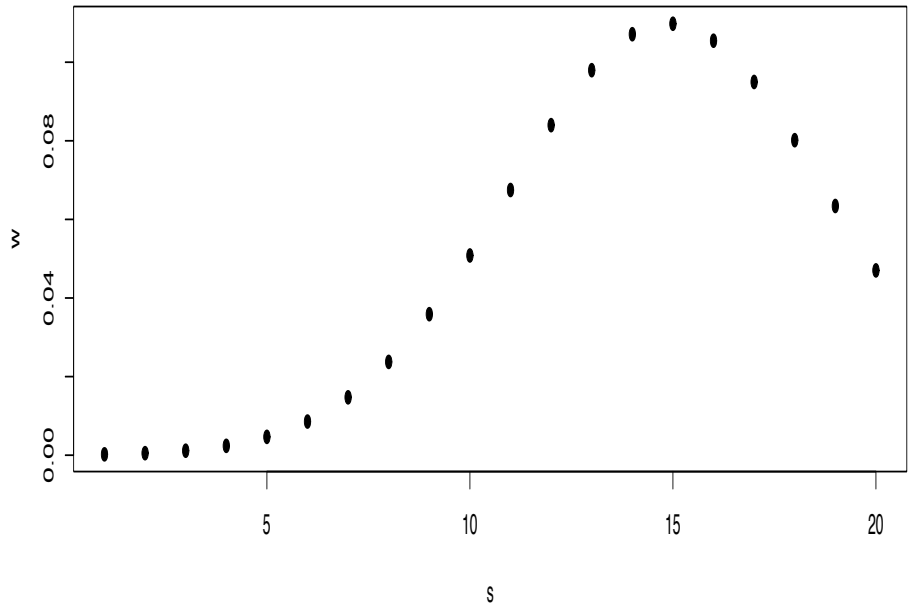
Compression of the exposed and infected periods– the model relies solely on inferring the number of new cases from previous observations using an “infectivity profile” [2].

In this setting, each infected case is expected to contaminate on average of R_0 patients (by definition) — but the distribution of this number of new infections is given by a probability distribution which only depends on the time s elapsed since infection: one could indeed imagine a patient becoming increasingly contagious over the first few days of the infection as the viral load builds up, and decreasingly so after the peak of the illness.

This infectious profile is typically modeled as a Gamma distribution.

Infectious profile

Cori et al [2] propose the use of the parameters of the serial interval (more substantial observational data and better estimation) as a good proxy.



We focus solely on R, which we assume to have a distribution over space and time.

We assume the parameters of the serial interval to have been correctly estimated and thus, the coefficients w_s to be known.

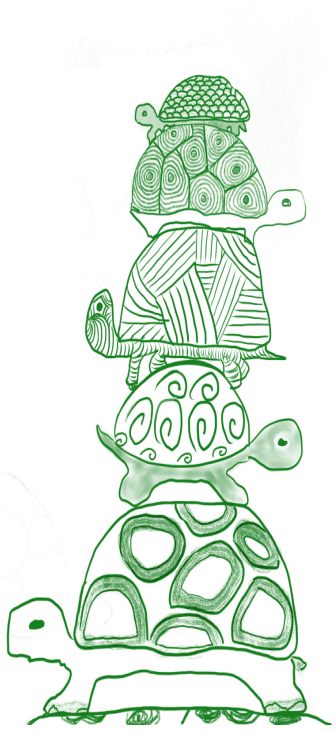
We call X the number new infectious cases each day.

The incidence on day t conditioned on the previous incidences can be modeled by a Poisson distribution of the form:

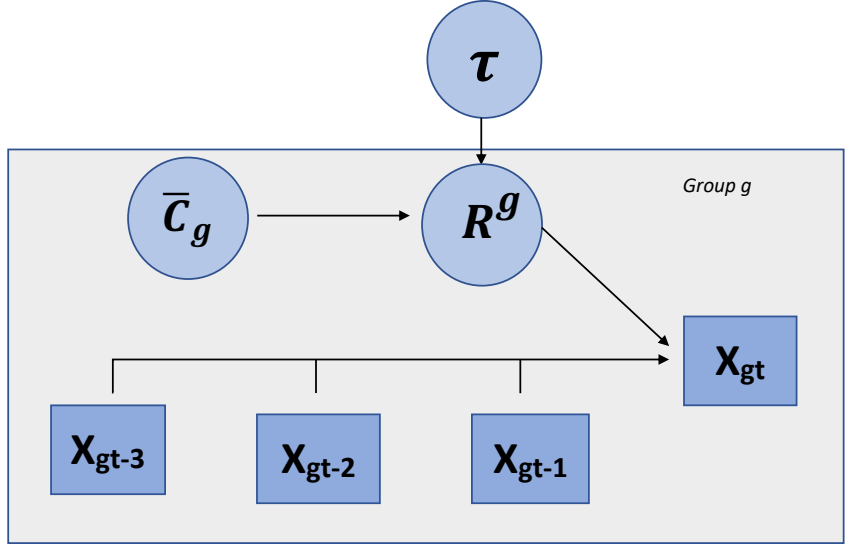
$$\forall t \leqslant T, \quad X_t \sim \text{Poisson}(R \sum_{s=1}^{t-1} w_s X_{t-s}) \quad (*)$$

where $w_s = \mathbb{P}\big[\Gamma_{\alpha,\beta} \in (s, s+1)\big]$.

Hierarchical Models



Here, we assume a hierarchical structure for R, decomposes it as the product of the transmissibility τ, the daily contact $\overline{c}_g$ and the duration of individual infections D_I — which we assume to be known.
Rate of contact $\overline{c}_g$ is group-specific.



The model is summarized as:

$$\forall t \leqslant T, \forall g \leqslant G, \quad X_{t,g} \sim \text{Poisson}(R^{(g)} \sum_{s=1}^{t-1} w_s X_{g,t-s})$$

$$\forall g = 2 \cdots G, \quad \bar{c}_g \sim \Gamma(2,1) \tag{4}$$

$$\tau \sim \beta(1,39)$$

$$R^{(g)} = \bar{c}_g \tau D_I$$

It is extended to a full random-effect version by taking the effective reproductive rate $R^{(g)}$, at each time step, to be sampled from a gamma distribution: $R_t^{(g)} \sim \Gamma(R^{(g)} * 10, 10)$.

Mathematics/Statistics issue: Non-identifiability

If we could have several values for the parameters that give the same observables, we say there is non-identifiability.

$\bar{c}\tau$ is invariant by rescaling of the two factors.

Fix the first group's daily contact rate $\bar{c}_1$ to a fixed value — we pick it here to be 1.

All other values of $\bar{c}$ can be understood as relative measures with respect to this benchmark group — thus a $\bar{c}_2$ with value 2 would indicate that, on a daily basis, an infected individual in population 2 has twice as many contacts in the susceptible group than in population 1.

This benchmark value could be either an arbitrary benchmark value (which should allow the potential R to vary within reasonable ranges), or an informed measure of social interactions — for instance, a daily contact of one person per day might seem like an appropriate value for a population in complete lockdown, such as seen in Wuhan as of January 22^{nd}.

Computational Details

- `RStan` programming suite[1].
- This uses Hamiltonian Monte Carlo to generate samples and estimate the different parameters of the model.
- We use a total of 8 chains, with 5,000 warmup iterations and 1,000 sampling steps.
- All the associated code and data are provided on Github[1].
- There is no theory that says that these Hamiltonian Monte Carlo methods have converged, we set up synthetic experiments that we use to benchmark the accuracy of our method.
- Computations were done on a HPC cluster and took a few hours.

[1] `https://github.com/donnate/heterogeneity_R0`

Heterogeneity of R_0's

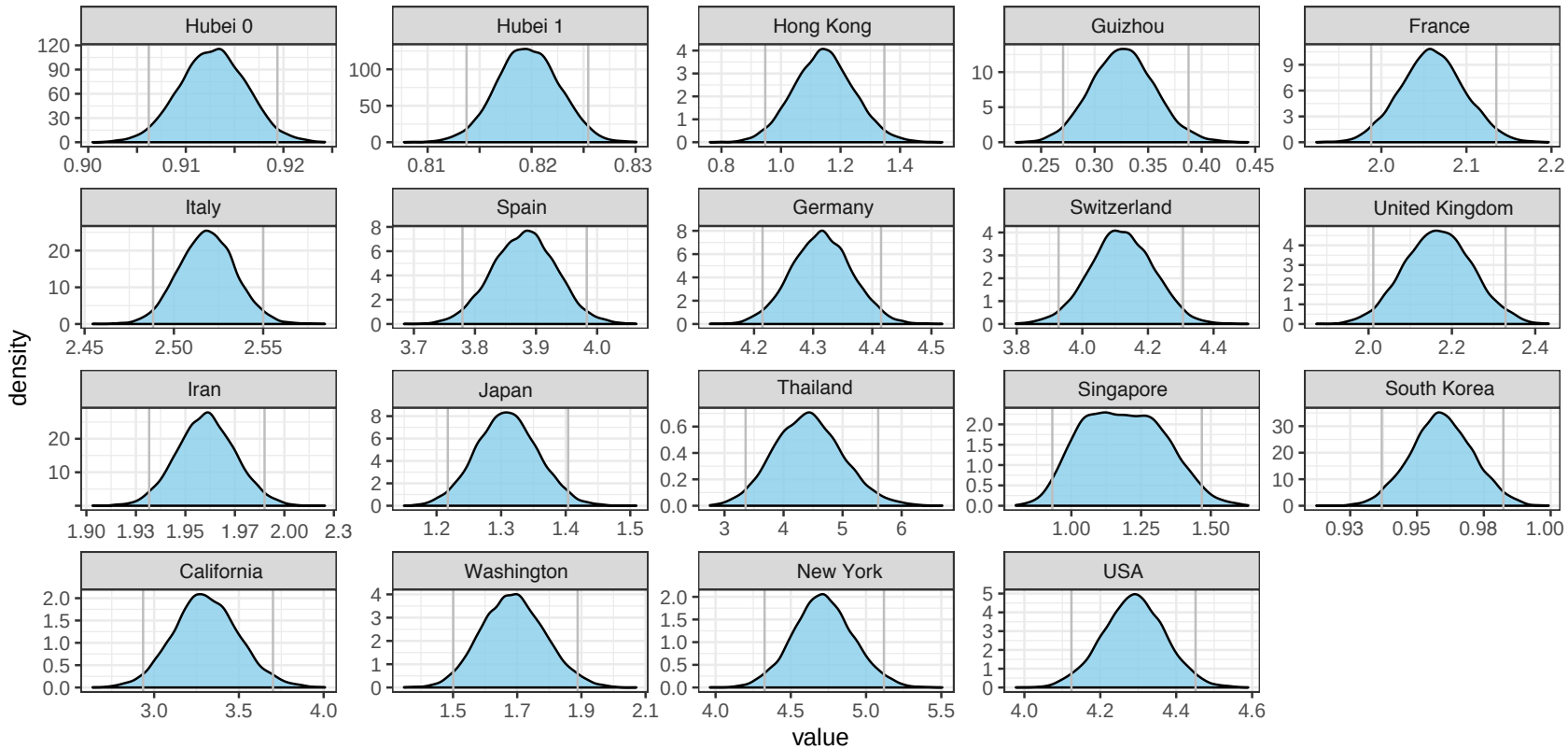


Figure 2: Distribution of the recovered spatial reproductive numbers R for the spatial Random-Effects Model. (**Beware the different scales**)

Evaluating the impact heterogeneity in predictive scenarios.

Our goal is to assess the toll of hospital load that a rapidly propagating pandemic can induce. As such, we emphasize that we focus on **short-term estimation**, and the study of the uncertainty for time frames of a few weeks.

Quantify the effect of governmental measures on the “flattening of the curve”.

Assess how informative our exponential growth model truly is when used in the context of drawing predictive scenarios under such huge uncertainty.

We use our fitted reproductive number to generate new predictions for the next 200 days.

We do not assume here that a given policy can manage to bring the R_0 to a given value (e.g. 1) – in other words, that the effect of the policy is absolute.

Spectrum of social distancing measures and their effect in relative terms.

We thus consider policies that divide the daily contact rate by a certain factor, rather than in absolute value.

The model that we adopt here is the following. For each day:

1. We generate the number of new incident cases based on the Bayesian model detailed and fitted in the previous section.
2. We then generate the number of people among these incidence cases that will require hospitalization. This number is generated by a binomial distribution, with a hospitalization rate that is contingent on the geographic localization and takes into account the age demographic layout of each cluster:

$$\pi_{\text{Hosp}}^{(g)} \sim 0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{Hosp}}, 1)$$

   where α_g is the proportion of each age group in location g (divided in 4 groups: from "0-19" years-old, "20-54", "54-65", and "65+"), and π_α is the hospitalization rate per group (expressed in percentages, and assumed to be universal across all contagion groups).
3. Once the number of newly hospitalized people has been selected, we choose among them using a binomial distribution

the people directly admitted into an Intensive Care Unit (ICU). The parameter for the binomial is also contingent on the demographics:

$$\pi^{(g)}_{\text{ICU}|\text{Hosp}} \sim \frac{0.01 * \Gamma(\alpha_g^T \pi^{\text{ICU}}_{\alpha}, 1)}{\pi^{(g)}_{\text{Hosp}}}$$

where $\pi^{\text{ICU}}_{\alpha}$ is the ICU rate per group (also expressed in percentages, and assumed to be universal across all contagion groups).

4. Finally, the fatalities are chosen among the people placed in the ICU, and sampled from a binomial distribution with probability:

$$\pi^{(g)}_{\text{death}|\text{ICU}} \sim \frac{0.01 * \Gamma(\alpha_g^T \pi^{\text{death}}_{\alpha}, 1)}{\pi^{(g)}_{\text{ICU}}}$$

5. For the hospitalizations, ICU and number of deaths selected, we assign time of death and of departure from the hospital/ICU by sampling from a matched normal distribution.

The scenarios are thus sampled as follows:

$$\tau \sim \text{Posterior}(\tau)$$

$$\forall g, \quad \bar{c}_g \sim \text{Posterior}(\bar{c}_g)$$

$$X_{t,g} \sim \frac{1}{2}\left(\mathcal{N}\left(2\sqrt{R\sum_{s=1}^{K} w_s X_{t-s} + \frac{3}{8}}, 1\right)\right)^2 - \frac{3}{8}$$

$$\pi^{(g)}_{\text{Hosp}} \sim 0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{Hosp}}, 1)$$

$$\pi^{(g)}_{\text{ICU|Hosp}} \sim \frac{0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{ICU}}, 1)}{\pi^{(g)}_{\text{Hosp}}}$$

$$\pi^{(g)}_{\text{death|ICU}} \sim \frac{0.01 * \Gamma(\alpha_g^T \pi_\alpha^{\text{death}}, 1)}{\pi^{(g)}_{\text{ICU}}}$$

$$\text{Hosp}_{t,g} \sim \text{Binomial}(X_{t,g}, \pi^{(g)}_{\text{Hosp}})$$

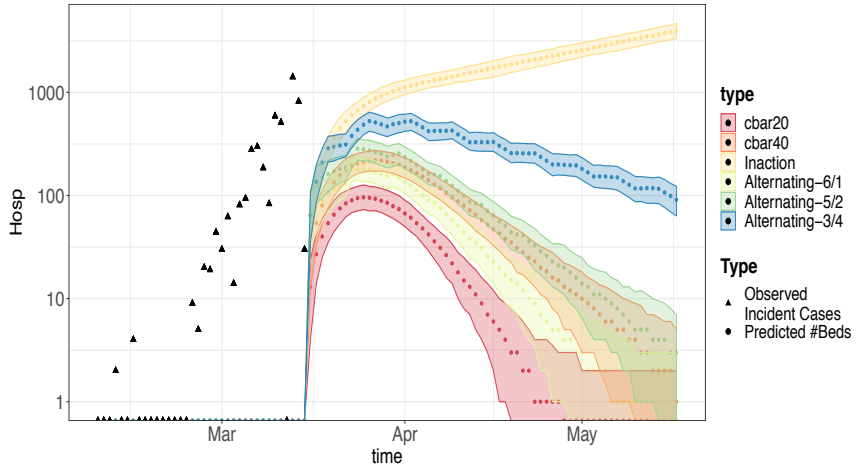
$$\text{ICU}_{t,g} \sim \text{Binomial}(\text{Hosp}_{t,g}, \pi^{(g)}_{\text{ICU|Hosp}})$$

$$\text{Deaths}_{t,g} \sim \text{Binomial}(\text{ICU}_{t,g}, \pi^{(g)}_{\text{death|ICU}})$$

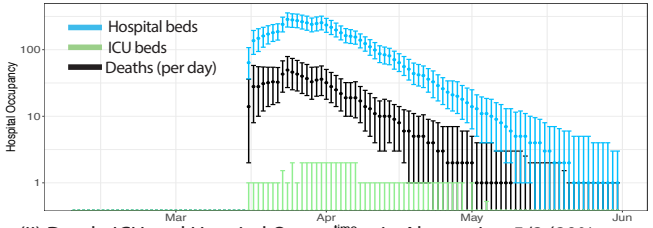
$$\forall i \in [1...\text{Deaths}_{t,g}], \quad T_i^{\text{Deaths}_{t,g}} \sim N(\mu_d, \sigma_d)$$

$$\forall i \in [1...\text{ICU}_{t,g}] \quad T_i^{\text{ICU}_{t,g}} \sim N(\mu_{\text{ICU}}, \sigma_{\text{ICU}})$$

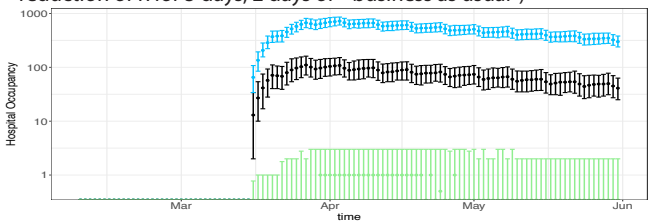
(A) Hospital Occupancy using a general R (world R_0)



(i) Hospital Occupancy

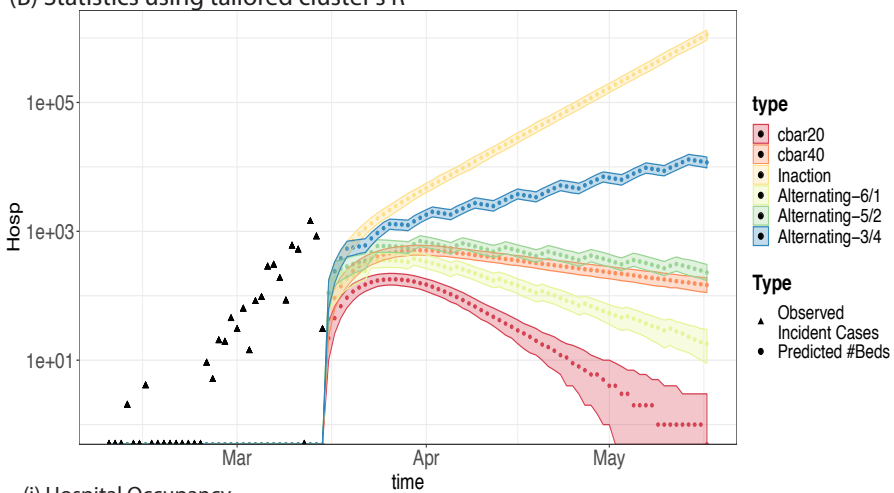


(ii) Death, ICU and Hospital Occupancy in Alternating 5/2 (80% reduction of R for 5 days, 2 days of ``business as usual")

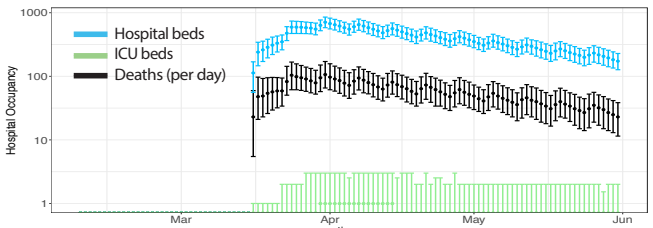


(iii) Death, ICU and Hospital Occupancy in Alternating 2/5 (80% reduction of R for 2 days, 5 days of ``business as usual")

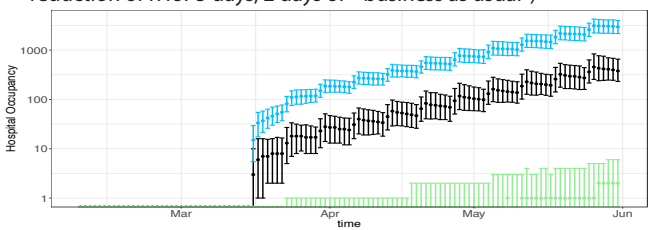
(B) Statistics using tailored cluster's R



(i) Hospital Occupancy



(ii) Death, ICU and Hospital Occupancy in Alternating 5/2 (80% reduction of R for 5 days, 2 days of ``business as usual")



(iii) Death, ICU and Hospital Occupancy in Alternating 2/5 (80% reduction of R for 2 days, 5 days of ``business as usual")

Spatial

Random-Effects Model: Comparisons of the outcomes of the different strategies. Estimated trajectories in terms of occupied hospital beds using various R: the group's tailored Bayesian R, as well as a general R estimated from the aggregated data. We note the difference in the impact on the healthcare systems that the aggregation vs the spatially heterogeneous R

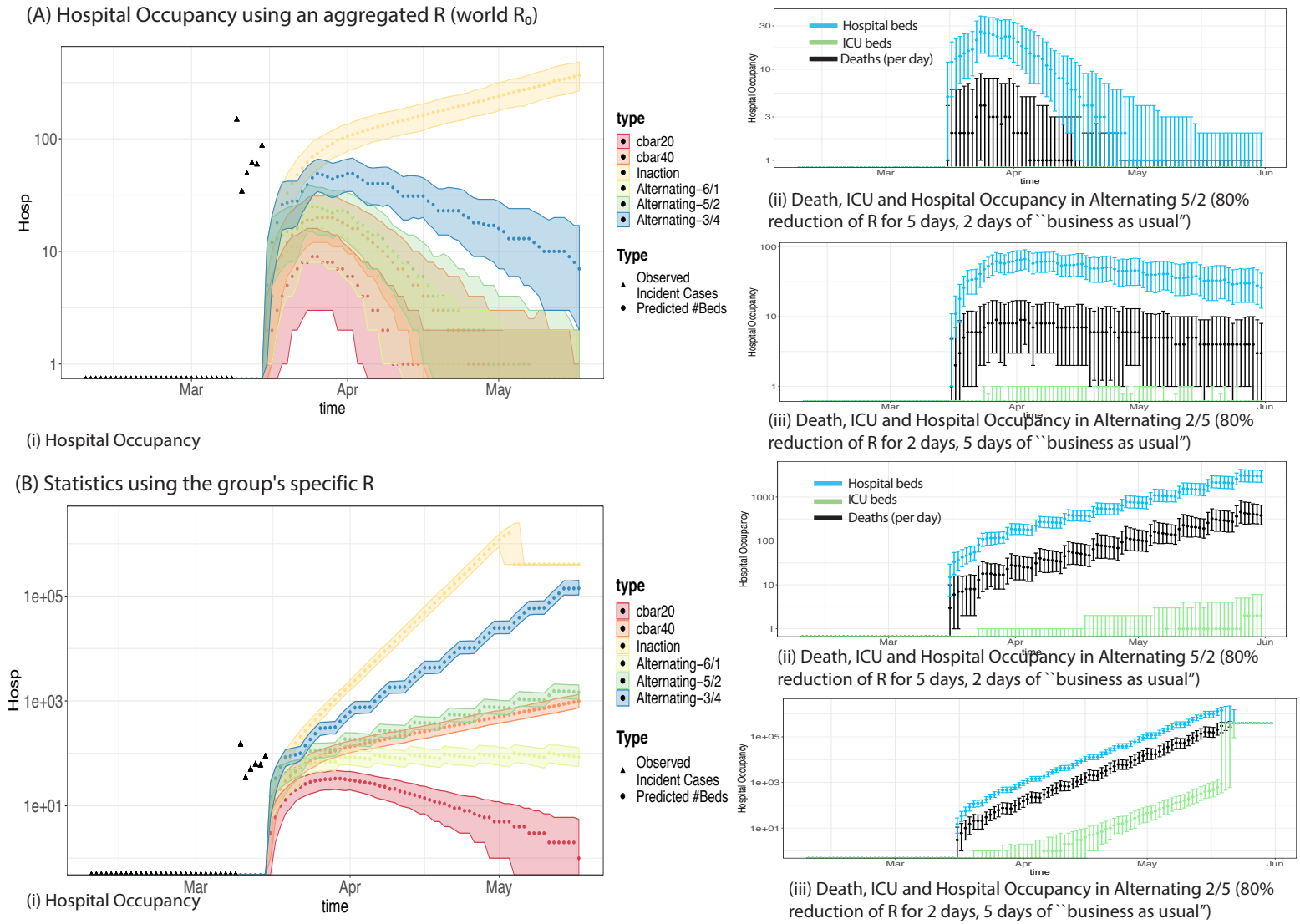
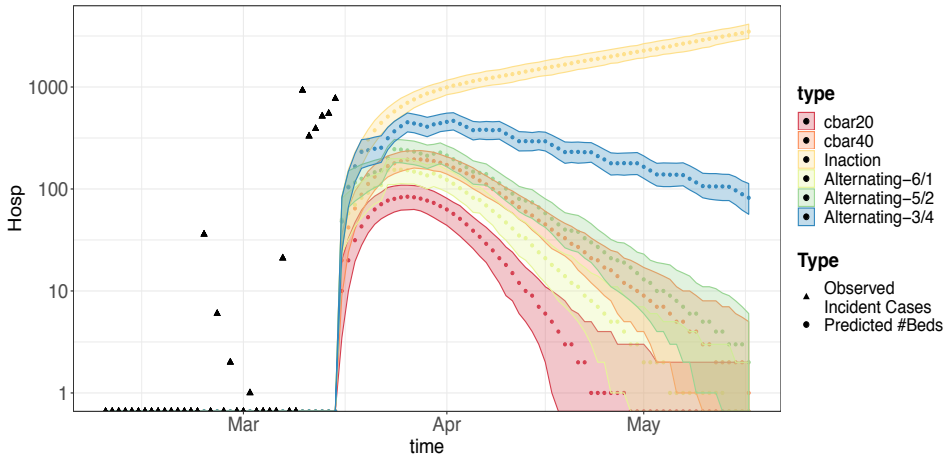
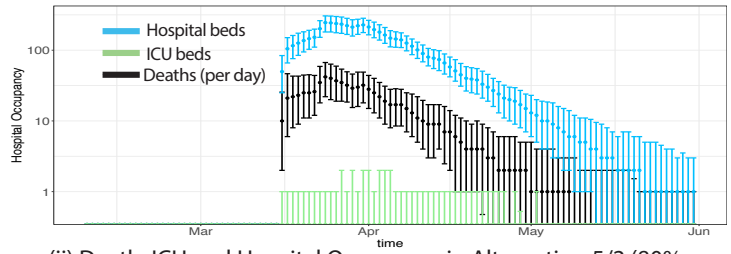


Figure 3: California. Spatial random effects trajectories in terms of occupied hospital beds using various R: the group's specific Bayesian R, as well as a general R estimated from the aggregated data. Impact on the healthcare projections that the aggregation vs the spatially heterogeneous R yield.

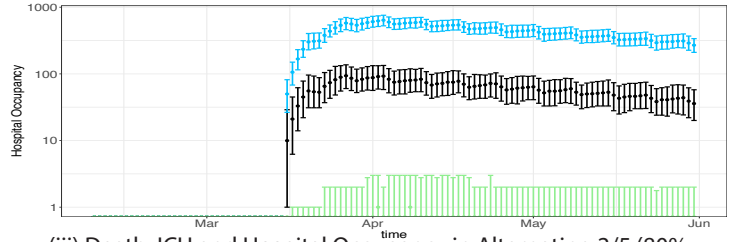
(A) Hospital Occupancy using an aggregated R (world R_0)



(i) Hospital Occupancy

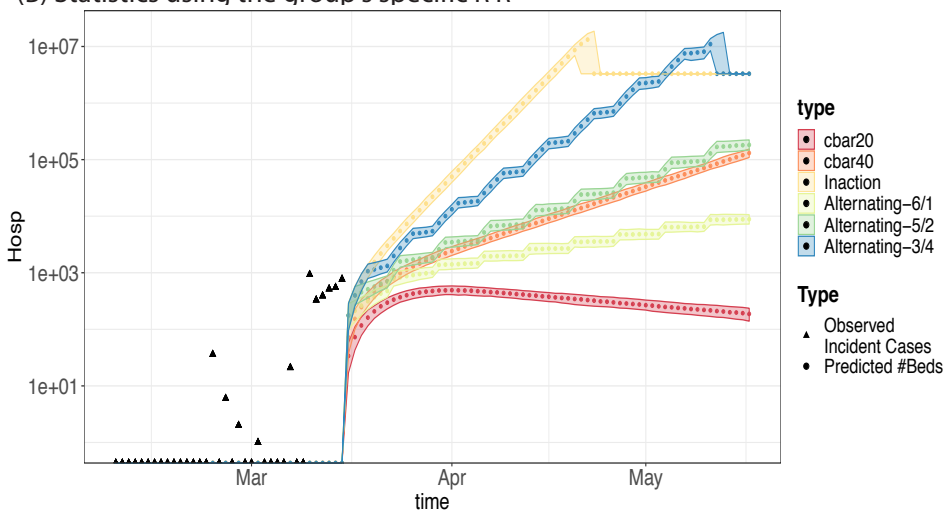


(ii) Death, ICU and Hospital Occupancy in Alternating 5/2 (80% reduction of R for 5 days, 2 days of ``business as usual")

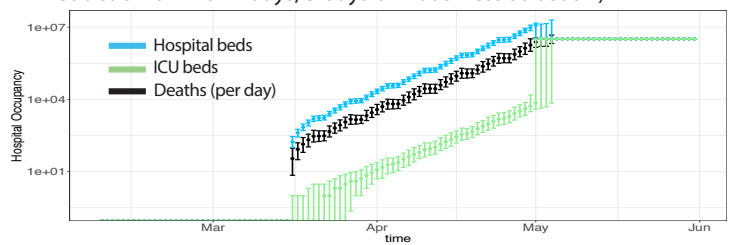


(iii) Death, ICU and Hospital Occupancy in Alternating 2/5 (80% reduction of R for 2 days, 5 days of ``business as usual")

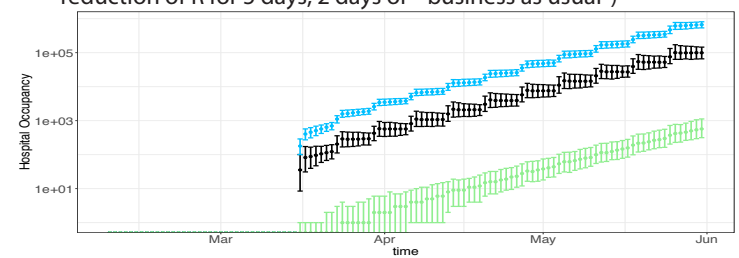
(B) Statistics using the group's specific R R



(i) Hospital Occupancy



(ii) Death, ICU and Hospital Occupancy in Alternating 5/2 (80% reduction of R for 5 days, 2 days of ``business as usual")



(iii) Death, ICU and Hospital Occupancy in Alternating 2/5 (80% reduction of R for 2 days, 5 days of ``business as usual")

Figure 4: Spatial Random-Effects. United States of America

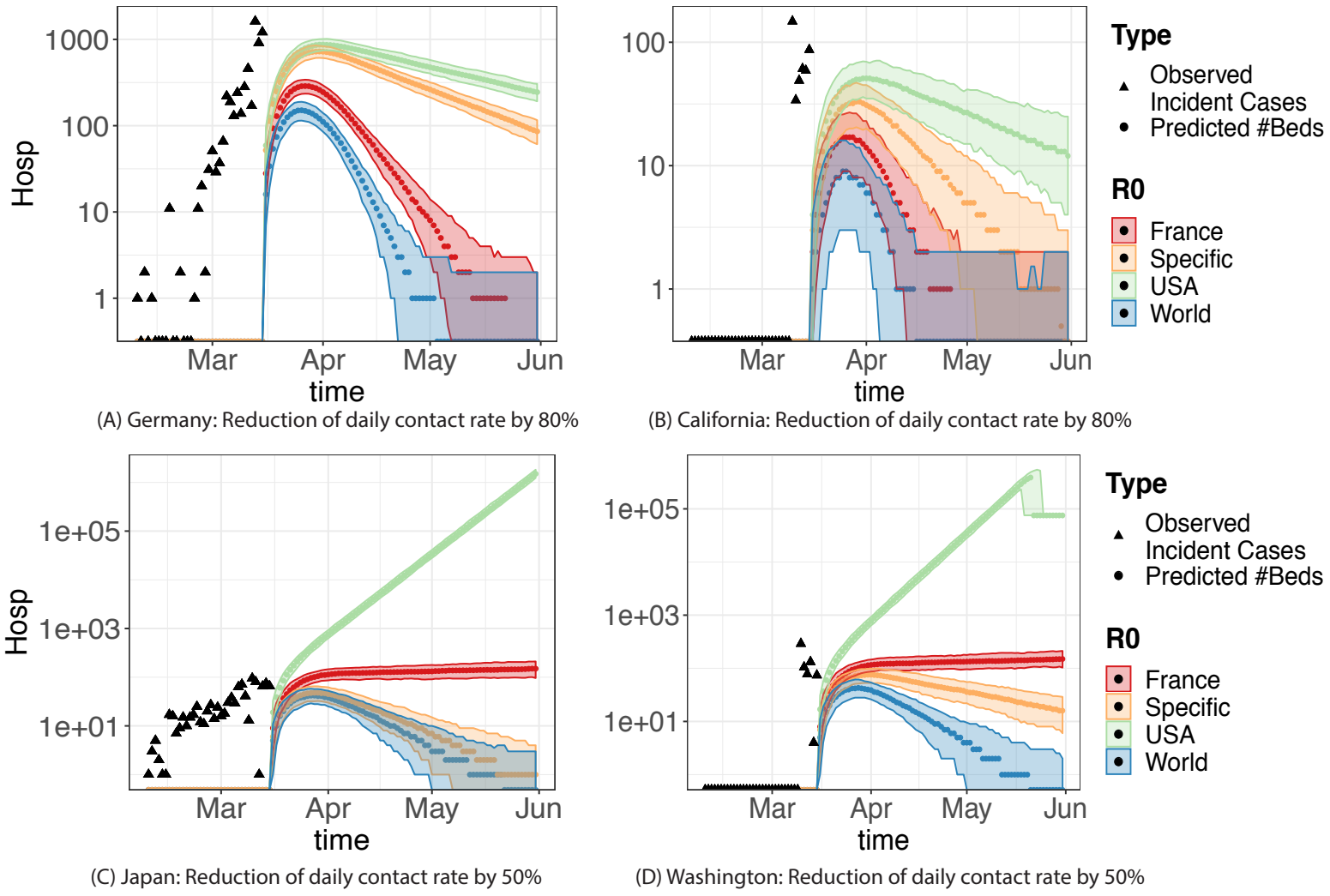
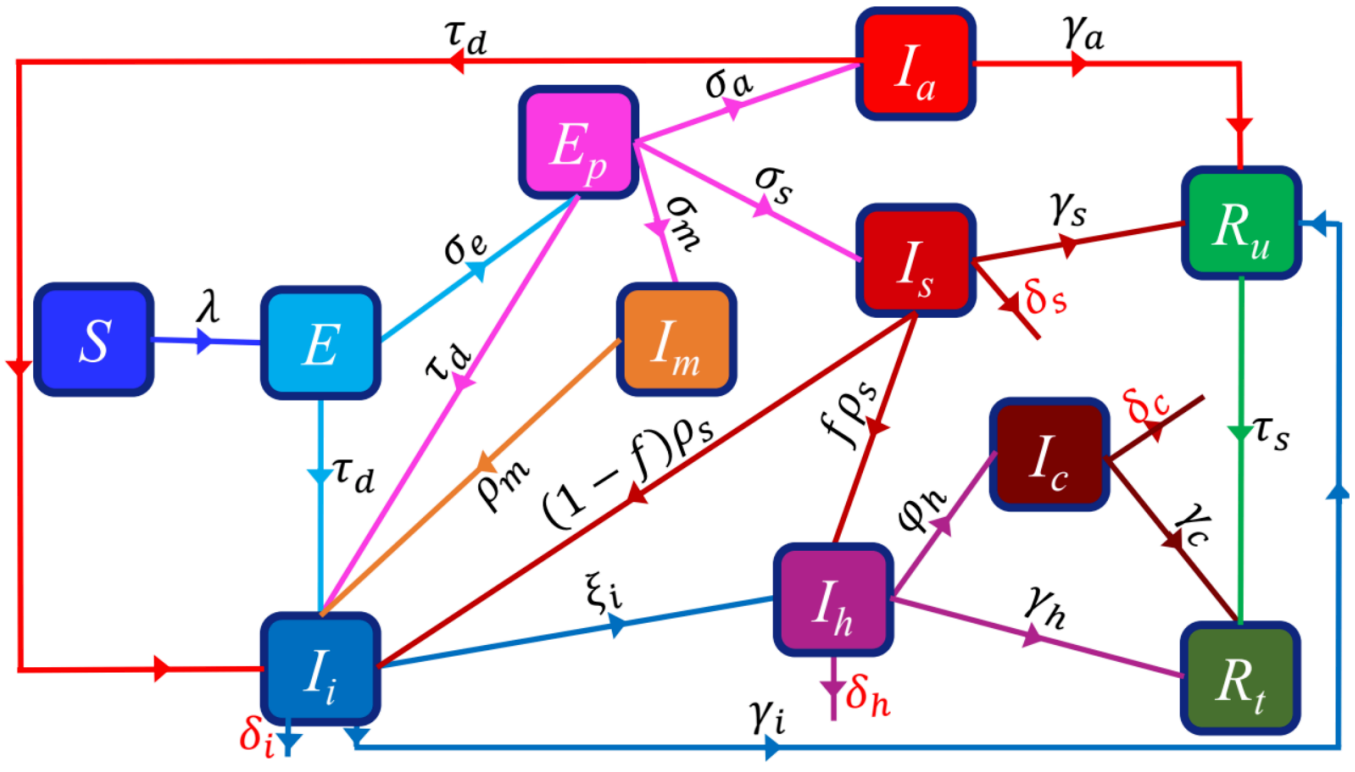


Figure 5: Spatial Random-Effects. Four different groups (impact of a given policy), using different R_0s. This shows the importance of correctly accounting for group-wise heterogeneity in the model.

Part III

Next hurdle: Identifiability.

Deconvolution: Identifiability



source: Abba Gumel, MSRI Talk, Aug 12, 2020.

$$N(t) = S(t) + E(t) + I(t) + A(t) + H(t) + Q(t) + R(t)$$

Identifiability

Identifiability of infection model parameters early in an epidemic

Timothy Sauer[a,*], Tyrus Berry[a], Donald Ebeigbe[b], Michael M. Norton[b], Andrew Whalen[b,e,f], Steven J. Schiff[b,c,d]

[a] *Department of Mathematical Sciences, George Mason University, Fairfax, VA, USA*
[b] *Centers for Neural Engineering and Infectious Disease Dynamics, Department of Engineering Science and Mechanics, The Pennsylvania State University, University Park, PA, USA*
[c] *Department of Neurosurgery, Penn State College of Medicine, Hershey, PA, USA*
[d] *Department of Physics, The Pennsylvania State University, University Park, PA, USA*
[e] *Department of Neurosurgery, Massachusetts General Hospital, Boston, MA, USA*
[f] *Department of Neurosurgery, Harvard Medical School, Boston, MA, USA*

Abstract

It is known that the parameters in the deterministic and stochastic SEIR epidemic models are structurally identifiable. For example, from knowledge of the infected population time series $I(t)$ during the entire epidemic, the parameters can be successfully estimated. In this article we observe that estimation will fail in practice if only infected case data during the early part of the epidemic (pre-peak) is available. This fact can be explained using a long-known phenomenon called dynamical compensation. We use this concept to derive an unidentifiability manifold in the parameter space of SEIR that consists of parameters indistinguishable to $I(t)$ early in the epidemic. Thus, identifiability depends on the extent of the system trajectory that is available for observation. Although the existence of the unidentifiability manifold obstructs the ability to exactly determine the parameters, we suggest that it may be useful for uncertainty quantification purposes. A variant of SEIR recently proposed for COVID-19 modeling is also analyzed, and an analogous unidentifiability surface is derived.

Identifiability: problem solved in Bayesian framework

With proper prior and posterior distributions, the posterior distributions enable the estimation of parameters and solutions to ”deconvolution problems” and the identifiability problems go away.

Other complete Bayesian analyses (but with fixed R_0)

Flaxman et., 2020[6] also used a Bayesian model with Stan to look for changepoints.

Article

Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe

https://doi.org/10.1038/s41586-020-2405-7

Received: 30 March 2020

Accepted: 22 May 2020

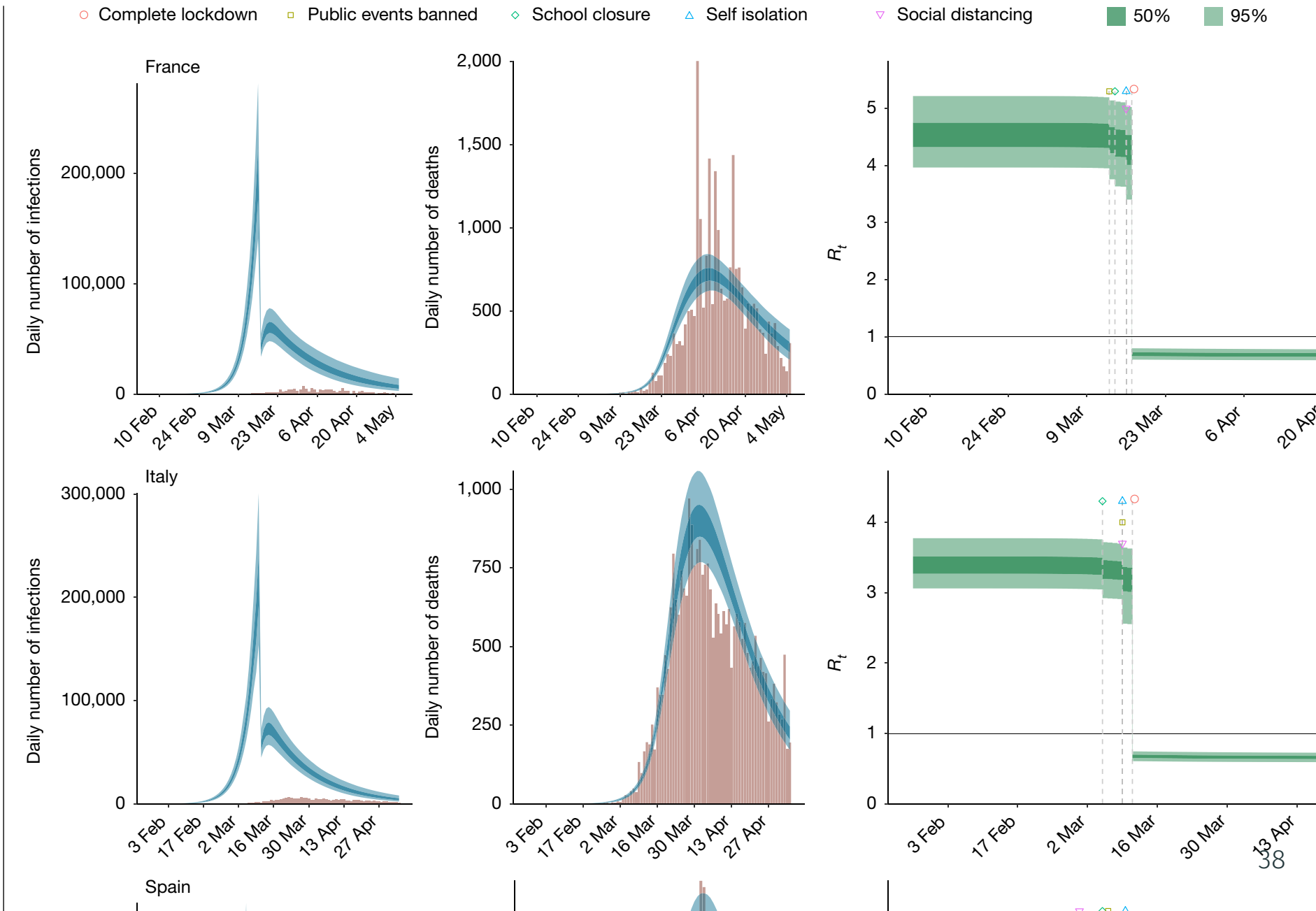
Published online: 8 June 2020

Check for updates

Seth Flaxman[1,7], Swapnil Mishra[2,7], Axel Gandy[1,7], H. Juliette T. Unwin[2], Thomas A. Mellan[2], Helen Coupland[2], Charles Whittaker[2], Harrison Zhu[1], Tresnia Berah[1], Jeffrey W. Eaton[2], Mélodie Monod[1], Imperial College COVID-19 Response Team*, Azra C. Ghani[2], Christl A. Donnelly[2,3], Steven Riley[2], Michaela A. C. Vollmer[2], Neil M. Ferguson[2], Lucy C. Okell[2] & Samir Bhatt[2,7]✉

Following the detection of the new coronavirus[1] severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and its spread outside of China, Europe has experienced large epidemics of coronavirus disease 2019 (COVID-19). In response, many European countries have implemented non-pharmaceutical interventions, such as the closure of schools and national lockdowns. Here we study the effect of major interventions across 11 European countries for the period from the start of the COVID-19 epidemics in February 2020 until 4 May 2020, when lockdowns started to be lifted. Our model calculates backwards from observed deaths to estimate transmission that occurred several weeks previously, allowing for the time lag between infection and death. We use partial pooling of information between countries, with both individual and shared effects on the time-varying reproduction number (R_t). Pooling allows for more

Effect of lockdown (Flaxman et al, 2020)



A simple and appropriate Bayesian computation for specificity/sen

Carpenter and Gelman, 2020[8] provide an elegant solution to accounting for the uncertainties in testing by using priors on the false positive and false negative rates.

	(a) *Posterior inference with weak prior*		(b) *Posterior inference with stronger prior*	
Parameter	median	(95% interval)	median	(95% interval)
Prevalence, π	0.016	(0.000, 0.160)	0.016	(0.001, 0.021)
Specificity, γ_1	0.997	(0.987, 1.000)	0.995	(0.987, 0.999)
Sensitivity, δ_1	0.797	(0.065, 1.000)	0.821	(0.622, 0.959)
μ_γ	5.54	(4.43, 6.72)	5.234	(4.60, 5.91)
μ_δ	1.54	(0.24, 2.89)	1.54	(0.90, 2.22)
σ_γ	1.62	(0.82, 2.61)	0.72	(0.26, 1.15)
σ_δ	0.87	(0.11, 2.16)	0.39	(0.00, 0.73)

Bayesian approaches

- Allow explicit inclusion of sources of heterogeneity through hierarchical models.
- Are intermediary between full resolution of ABM and completely summarized parametric methods.
- Simulations leading to uncertainty quantification which can serve for experimental design and followup.
- Overcomes problems with non-identifiability.

More research needed on

- Model selection and sensitivity analyses.
- Age/ stratification affects network structure and $\bar{\mathbf{c}}$.
- Using a more flexible model for $\mathrm{D_I}$, infection times (do different virus strains vary).
- Superspreader and extreme value tail events.
- Communication of orders of approximation and uncertainty.

R code and resources

Mathematicians show all their work and in statistics and epidemiology, reproducible research is the only research. (Data and Code).

Github: `https://github.com/donnate/heterogeneity_R0`

RECON: `https://www.repidemicsconsortium.org`

epidemia: `https://github.com/ImperialCollegeLondon/epidemia`

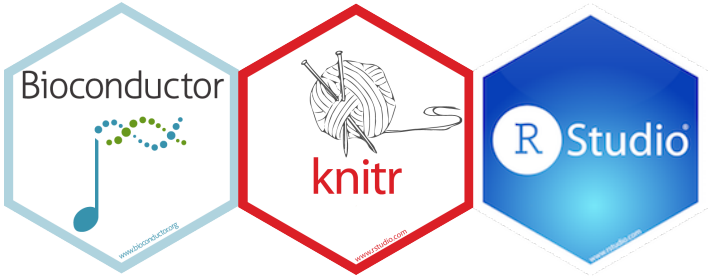
socialmixr: `https://github.com/sbfnk/socialmixr`

RStan: `https://mc-stan.org/rstan/`

book: Modern Statistics for Modern Biology `http://bios221.stanford.edu/book/`

Benefitting from the tools and schools of Statisticians.....

Thanks to the R, Stan and Bioconductor community and to co-authors.



Persi Diaconis, Wolfgang Huber, JJ Allaire and Rob Gentleman.

Thank you to the organizers for inviting me.

References


Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell.
Stan: A probabilistic programming language.
Journal of statistical software, 76(1), 2017.


Anne Cori, Neil M Ferguson, Christophe Fraser, and Simon Cauchemez.
A new framework and software to estimate time-varying reproduction numbers during epidemics.
American journal of epidemiology, 178(9):1505–1512, 2013.


Persi Diaconis and Susan Holmes.
A bayesian peek into feller volume I.
Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 64(3):820–841, 2002.


Persi Diaconis, Susan Holmes, and Richard Montgomery.

Dynamical bias in the coin toss.
SIAM review, 49(2):211–235, 2007.


Claire Donnat, Susan Holmes, et al.
Tracking network dynamics: A survey using graph distances.
The Annals of Applied Statistics, 12(2):971–1012, 2018.


Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, Mélodie Monod, Imperial College COVID-19 Response Team, Azra C Ghani, Christl A Donnelly, Steven Riley, Michaela A C Vollmer, Neil M Ferguson, Lucy C Okell, and Samir Bhatt.
Estimating the effects of non-pharmaceutical interventions on COVID-19 in europe.
Nature, June 2020.


Christophe Fraser.
Estimating individual and household reproduction numbers in an emerging epidemic.

PloS one, 2(8), 2007.


Andrew Gelman and Bob Carpenter.
Bayesian analysis of tests with unknown specificity and sensitivity.
medRxiv, 2020.


Susan Holmes.
Successful strategies for human microbiome data generation, storage and analyses.
Journal of Biosciences, 44(5):111, 2019.


Susan Holmes and Wolfgang Huber.
***Modern Statistics for Modern Biology*.**
Cambridge University Press, IBSN: 1108705294, 2019.


Phenyo E Lekone and Bärbel F Finkenstädt.
Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study.
Biometrics, 62(4):1170–1177, 2006.


Pierre Nouvellet Thibaut Jombart, Anne Cori.
earlyR: Estimation of Transmissibility in the Early Stages of a Disease Outbreak.
https://CRAN.R-project.org/package=earlyR, 2017.


Shi Zhao, Qianyin Lin, Jinjun Ran, Salihu S Musa, Guangpu Yang, Weiming Wang, Yijun Lou, Daozhou Gao, Lin Yang, Daihai He, et al.
Preliminary estimation of the basic reproduction number of novel coronavirus (2019-ncov) in china, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak.
International Journal of Infectious Diseases, 2020.