An Exploratory Evaluation into the Effect of Data Availability and Indictator Coverage on Parameter Estimates using Hamiltonian Monte Carlo

Jim Duggan[1]

1 School of Computer Science, University of Galway, Galway, Ireland.

Abstract

There are typically two important questions addressed via the model calibration process: (1) does the time series of the fitted model match to the historical data; and (2) can reliable parameter estimates be inferred that are bounded within credible intervals. The evolution of Markov Chain Monte Carlo (MCMC) methods provide powerful methodological and computational frameworks for parameter estimation, and recent studies confirm the value of the Hamiltonian Monte Carlo approach for system dynamics models. This paper addresses an important research question for the calibration process, namely: what is the impact of data availability and indicator coverage on parameter estimation. It presents a 3 x 7 factorial study based on an SEIR model with hospitalisations and deaths. An extensive exploratory analysis is presented, and the highlights differences for a number of posterior distributions calculated, depending on the data availablity, and the set of indicators.

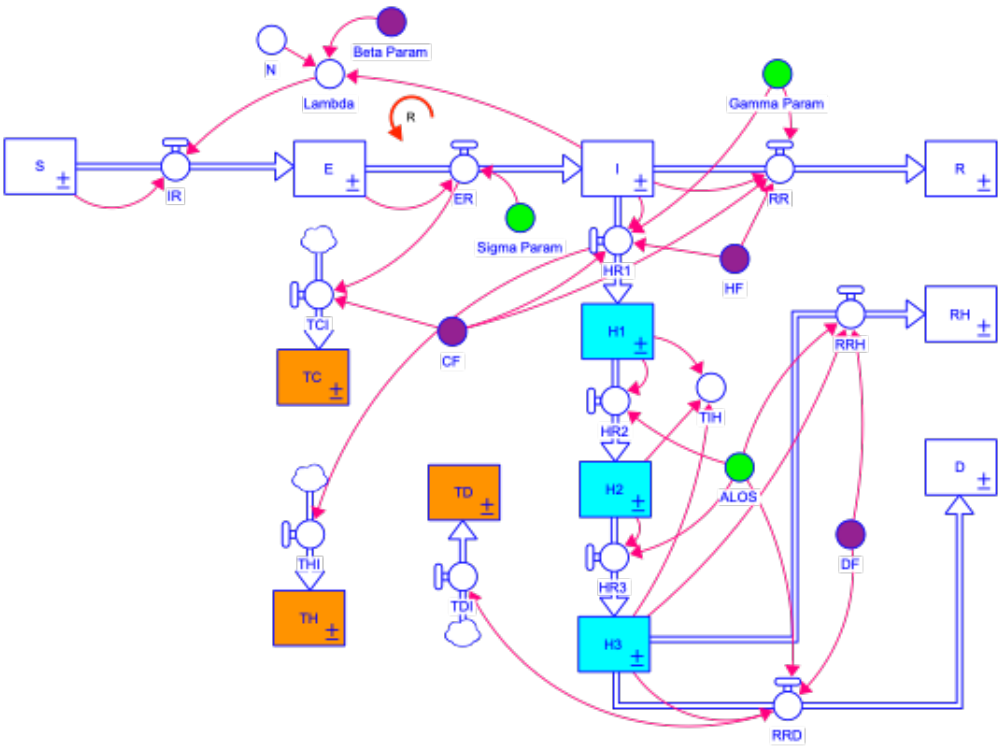

Fig 1. An SEIR Influenza Model with Hospitalisations and Deaths

1 Introduction

Contributions

Three indicator model SEIR simple model

Overall pipeline for efficiently processing inference results

Use of a new method (overlapping analysis) to measure differences in parameter estimates

2 Model Structure and Experimental Design

The model used is an extension the well-known deterministic SEIR structure [1], and is shown in Figure 1. People start out as being susceptible to a novel pathogen, and with the introduction of *patient zero* to the infectious (I) stock, the contagion loop is activated. People then move to an exposed (E) stock, where they do not contribute to the force of infection (λ), before entering an infectious state. Once in an infectious state people contribute to the force of infection (λ), before exiting via a first order exponential delay structure. A certain proportion of those infected have symptoms (clinical fraction), and a proportion of infectious are hospitalised, while the remainder recover. Hospitalisation is modelled as a third order delay structure, and 90% of those hospitalised recover, while 10% do not and move to the deaths stock (D). Three stocks (TC, TH, and TD) are used to record the cumulative numbers of cases, hospitalisations, and deaths.

Equations (1-4) for the main SEIR structure are shown below. The parameters σ and γ represent the inverse of the latent and infectious delays, c is the clinical fraction, and h is the hospitalisation fraction. The model assumes that only those who show symptoms can end up in the hospital stream. The force of infection λ is calculated based

on product of the effective contact rate (β) with the number of infectious (I), divided by the total population (N).

$$\dot{S} = -\lambda S \tag{1}$$
$$\dot{E} = \lambda S - \sigma E \tag{2}$$
$$\dot{I} = \sigma E - (1 - ch)\gamma I - ch\gamma I \tag{3}$$
$$\dot{R} = (1 - ch)\gamma I \tag{4}$$
$$\lambda = \beta \frac{I}{N} \tag{5}$$
$$\beta = 1.0 \tag{6}$$

The hospitalisation stream is modelled via equations 7-13. It involves a straightforward sequence of stocks that model people staying in hospital, with the average length of stay (L) set to 10 days. The hospitalisation rate is governed by the fraction ch exiting the infectious stock, and therefore for this model, that will evaluate to $0.6 \times 0.1 = 0.06$. Upon exiting hospital, $(1 - d)$ move to the stock R_H, while the fraction d move to the stock D.

$$\dot{H}_1 = ch\gamma I - \frac{H_1}{L_1} \tag{7}$$
$$\dot{H}_2 = \frac{H_1}{L_1} - \frac{H_2}{L_2} \tag{8}$$
$$\dot{H}_3 = \frac{H_2}{L_2} - d\frac{H_3}{L_3} - (1 - d)\frac{H_3}{L_3} \tag{9}$$
$$\dot{R_H} = (1 - d)\frac{H_3}{L_3} \tag{10}$$
$$\dot{D} = d\frac{H_3}{L_3} \tag{11}$$
$$L_1 = L_2 = L_3 = \frac{L}{3.0} \tag{12}$$
$$L = 10 \tag{13}$$

As part of the inference process, we need to generate rates for cases (14), hospitalisations (15), and deaths (16), and these numbers are recorded as cumulative (stock) values. The Stan file generated (via the package `readsdr`) will difference these values in order to compare the simulated model values with the (synthetic) data values as part of the likelihood calculations for the posterior distributions.

$$\dot{T_C} = c\sigma E \tag{14}$$
$$\dot{T_H} = ch\gamma I \tag{15}$$
$$\dot{T_D} = d\frac{H_3}{L_3} \tag{16}$$

Finally, we present the relevant model parameters for the experiments. Two of these are based on the literature relating to pandemic influenza [2], and the remaining values

are arbitrary choices used as part of the experimentation process. For example, four of these parameters (β, c, d and h) will be estimated as part of the inference process.

Name	Symbol	Value	Units	Source
Latent Duration	σ^{-1}	2.0	$Days$	Vynnycky et al. [3]
Infectious Duration	γ^{-1}	2.0	$Days$	Vynnycky et al. [3]
Effective Contact Rate	β	1.0	$Days^{-1}$	Model estimate
Clinical Fraction (CF)	c	0.60	$Dimn$	Model estimate
Death Fraction (DF)	d	0.10	$Dimn$	Model estimate
Hospitalisation Fraction (HF)	h	0.10	$Diml$	Model estimate
Average Length of Stay	L	10.0	$Days$	Model estimate

3 Computational Framework

Figure 2 captures the overall computational framework used to configure the experiments, generate the results, and produce the analysis. In order to execute this workflow, which is open-source and designed using R [3], a number of additional components were required:

- Stan [4], a statistical modelling platform, which provides an interface to perform Bayesian inference via the No-U-Turn-Sampler (NUTS). Stan models can contain ordinary differential equations, and is used to perform inference for deterministic models of infectious disease [5, 6, 7]
- `cmdstanr` [8], which is a lightweight interface to Stan for R users.
- `readsdr` [9], a package that automatically converts XMILE files from Stella and Vensim to Stan code.
- R's tidyverse packages, specifically `dplyr` for data manipulation, `ggplot2` for visualisation and `tidyr` for nesting data frames, a feature that allows the results to be conveniently organised into a 21 by 12 table.

The overall framework is divided into three main functions.

- **Generate Synthetic Data**, which runs the SEIR model using the package `readsdr` for one instance, and is based on the differencing of the three indicator stocks (see equations 13-15). The negative binomial distribution is used to generate random count variables based on the model outputs. For this experiment, the dispersion parameters selected are (10, 20, 40) for (Cases, Hospitalisations, and Deaths). Sample output from this process is shown in Figure 3.
- **Estimate Parameters**, which performs the fitting process. For each experiment, this will (1) use `readsdr` to generate a stan file, (2) run the stan file using the `cmdstanr` interface, and (3) prepare and store all of the results in a single multidimensional data structure. This structure will contain one row for each experiment, and encapsulate variables such as the number of indictors, the measurement model, the data used for the calibration, the posterior samples for each parameter (4,000 each), the time series output for each model run, and the duration of the run. Given the computational resources needed to run the fitting, an advantage of storing all the results in a database (RDS file used) is that the analysis stage can be conducted at a later stage.
- **Analyse Results**, which provides a number of scripts to generate a range of results to support analysis. These include: trace plots to explore convergence; time series showing the fits; boxplots highlighting the parameter distributions

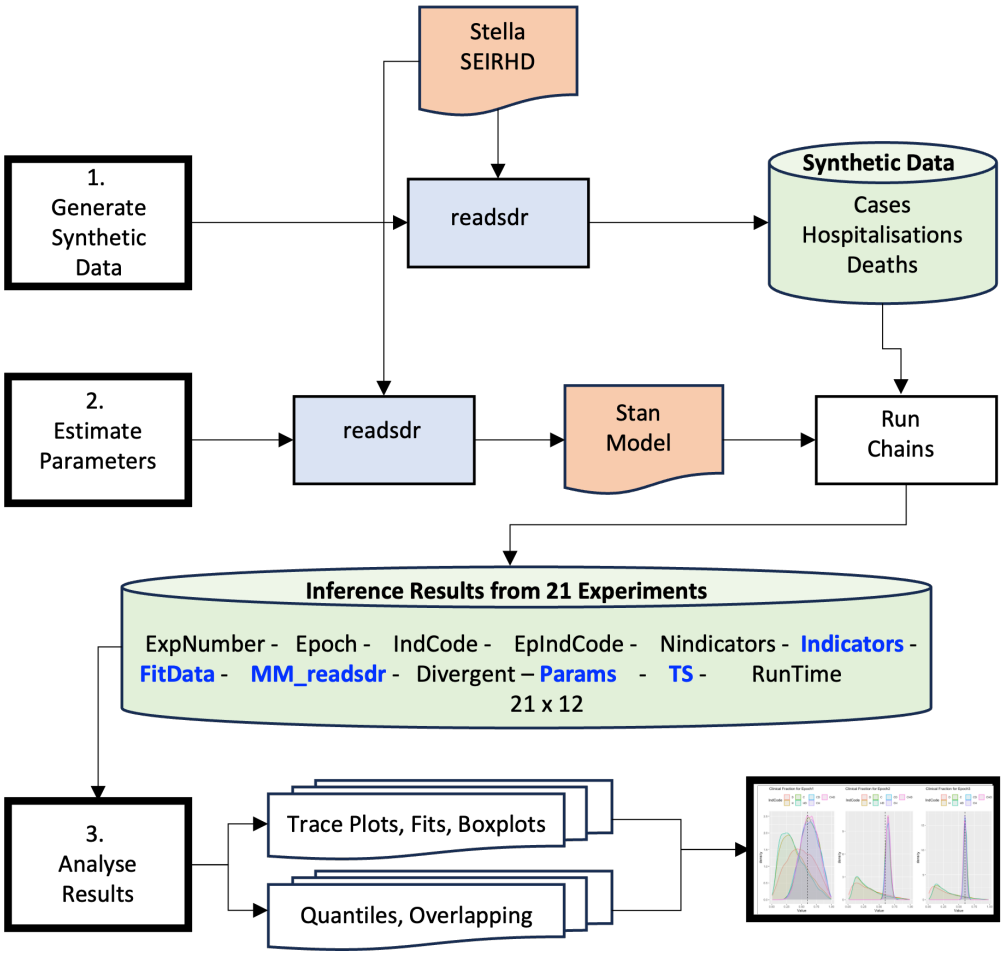

Fig 2. Overall data analytics framework for experiments

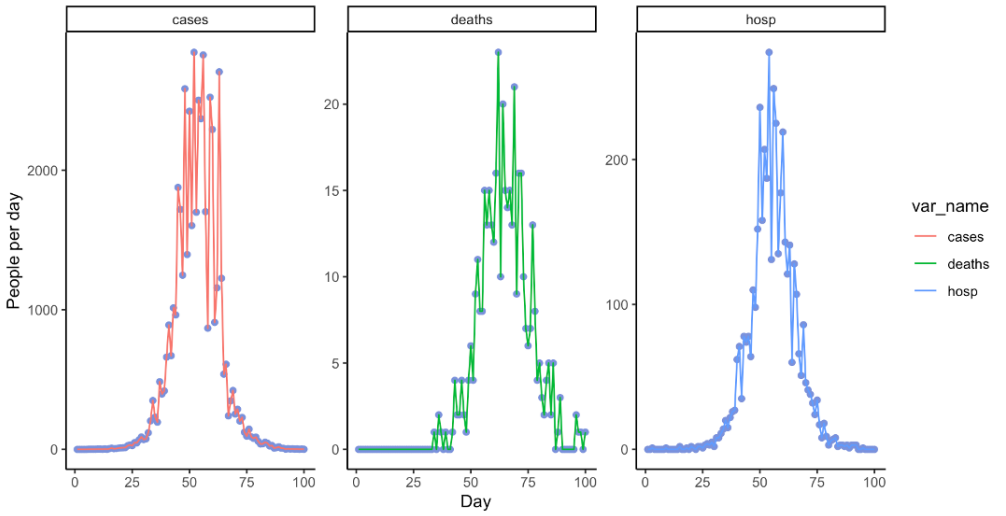

Fig 3. Sample synthetic data generated for the experiments

across all 21 experiments; quantile analysis to show the 95% credible intervals from the inference process; and overlap analysis to provide insights into how close the posterior distributions are for each combination of epoch and indicator.

The results are now presented in more detail.

4 Experimental Results

Overall, the inference process for the 21 experiments generates over 273MB of data, so there is a wide range of analysis that can be performed. Given that the overall goal is to explore possible differences in parameter estimates, our analysis comprises the following stages:

- Presentation of convergence tests for each parameter over the four MCMC chains, for each of the 21 data configurations.
- Exploring of the model fits, to confirm that the parameter estimates generate plausible dynamic behaviour for each model.
- Boxplots to highlight the distribution of inferred parameters, and to see how the estimates line up with the original values used to construct the synthetic data sets.
- Overlapping analysis, defined as the area intersected by two or more probability density functions [10, 11], to provide a measure of how close the estimates are across each of the 21 experiments.

4.1 Convergence Tests

Our first analysis, capture in Figure 4, displays the trace plots showing how the estimates of the parameter values change over the MCMC process. Four MCMC chains are run, with 1000 iterations for the warm-up phase, and 1000 for the sampling process. The goal is to ensure that a stationary distribution is achieved for all parameters [5], and this is achieved, and confirmed by analysing the output from stan, which showed no divergences within all the samples. This chain convergence is an important property of

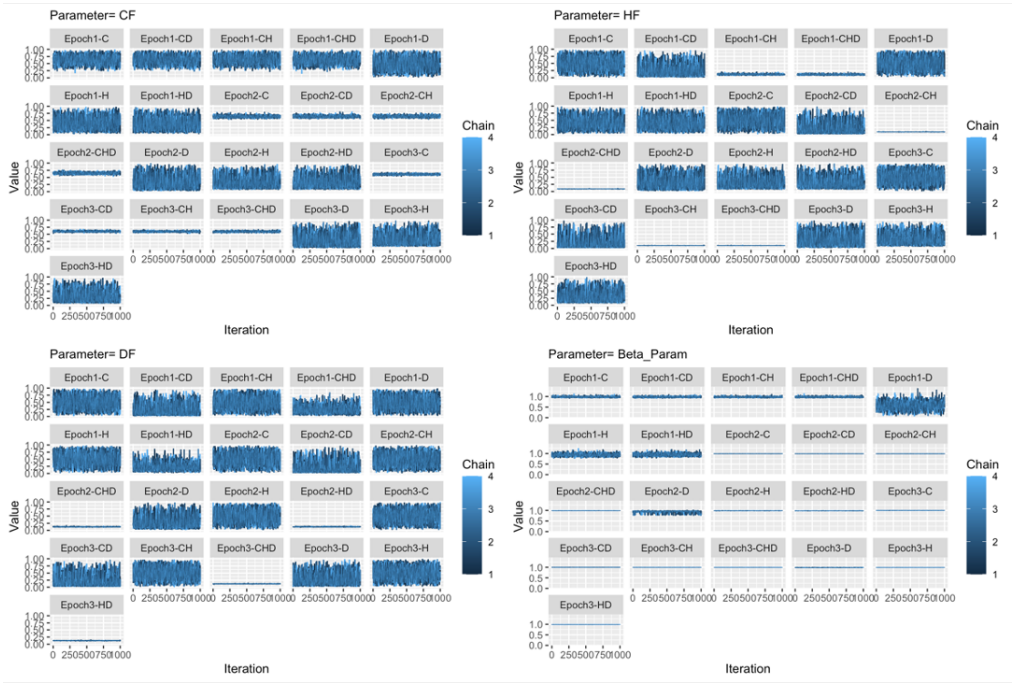

Fig 4. Exploring parameter convergence

the inference method. However, the plots can also illustrate interesting properties of the fitted parameters. Through observation, we can see the different ranges of the estimates, for example:

- Experiments involving epoch one (the opening phase of the epidemic) typically have wider ranges, which is as expected, as there is less data available to inform the fitting process. For example, for the parameter `HF` (true value 0.1), the values for `Epoch1-C` seem to cover the interval from 0 to 1.
- Experiments where the three indicators were used tend to have narrower bands, which again is not surprising given that more information is available for the calibration process. Returning to parameter `HF`, we can see that its values for `Epoch1-CHD` remain much closer to the true value of 0.1.

A more detailed analysis of these parameter values will be explored through boxplots, quantile analysis and overallping analysis.

4.2 Model fits

An important requirement for calibration is that the model provides a plausible representation of historical data, and in MCMC fitting we can also explore the time series quantiles returned as part of the posterior distribution. From a process perspective, having fits that align with historical data is needed to confirm that the model structure can generate the behaviour of interest, and these outputs also can be deployed as a confidence-building measure for modellers and clients. A sample of the fits are displayed in Figure 5, for the indicators *Cases* and *Deaths*, across three of the epochs. The mean value from the MCMC samples is also shown, and overall, on visual inspection, these look like good fits to the data.

An interesting aspect is what the fits do not show, which is the range of parameter values underling the generated time series. For example, if you take the set of plots in

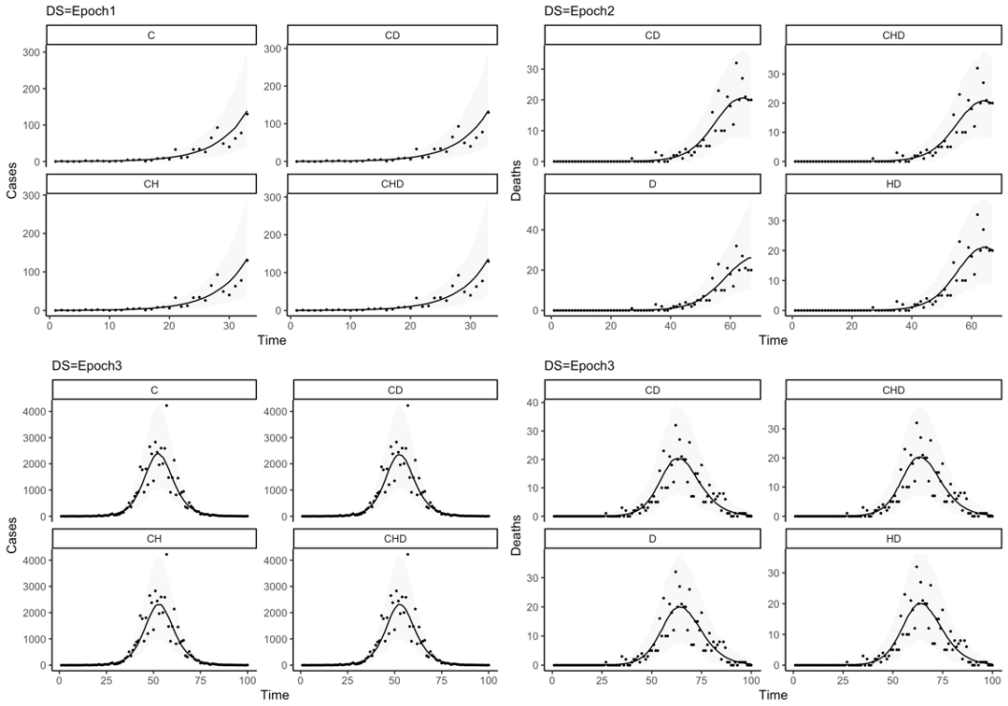

Fig 5. A sample of fits from the inference process (95% CI)

row 1 column 1 (epoch 1), and compare the fits for indicator C and indicators CHD, there is not an observable distinction between both, in that both fits look plausible. However, when we compare the parameters estimated for these two samples, differences emerge. This can first be explored using boxplots.

4.3 Boxplots of parameter estimates

The boxplot is a valuable method to summarise data, as it shows the median, the interquartile range (IQR), the location of values 1.5 times above and below the 75th and 25th percentiles, and outliers. We use the boxplot as a means to compare parameter values from all 21 experiments, and estimated across (1) each of the three epochs and (2) all of the seven combinations of data indicators. The 84 boxplots are presented in Figure 6.

We can observe a number of patterns from these descriptive statistics:

Why might this be the case? NArrow to expansive... tipping point?

4.4 Quantile analyis of fitted parameters

The plots displayed in the preceding section are useful to gain an overall appreciation of the posterior distributions, and to supplement that analysis, it is important to show the 95% credible intervals. For the four parameters, this information is presented in Figure 7, and arranged in descending order by the difference between the upper and lower quantiles.

A number of points can be made when reflecting on these results, and focusing on the first three records for each parameter (sorted in ascending order based on the narrowness of the credible interval):

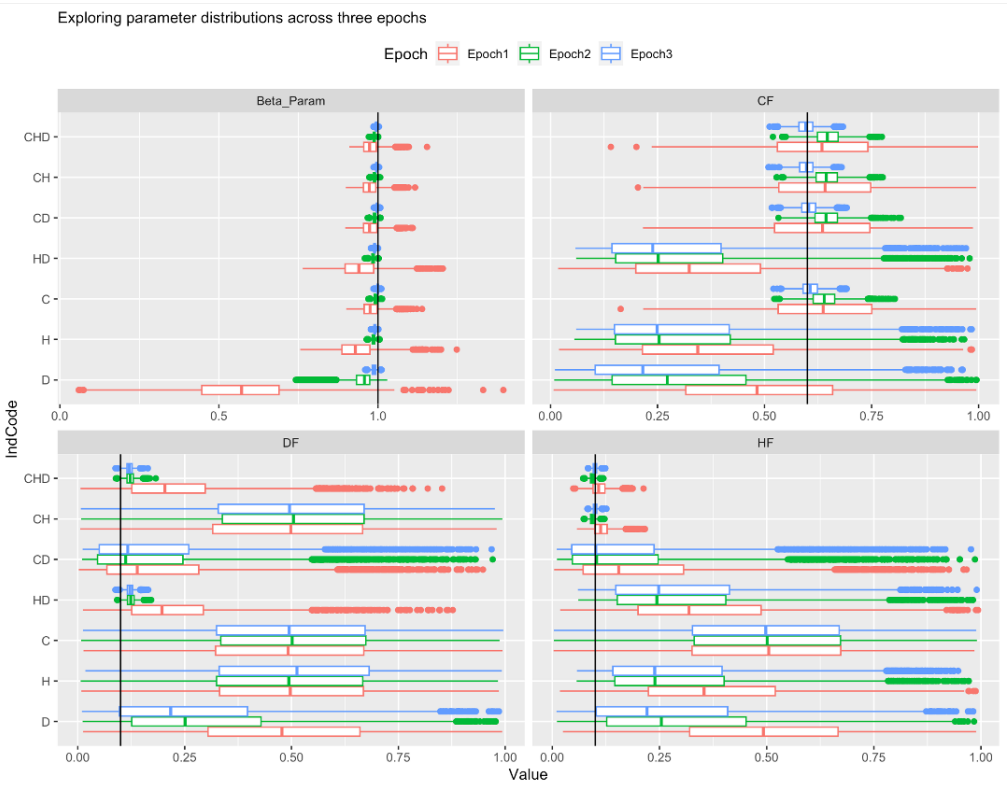

Fig 6. Box plot analysis for parameters by epoch and indicator source(s)

4.5 Overlap analysis of parameter densities

Disadvantage (scale)

5 Discussion

References

1. Vynnycky E, White R. An introduction to infectious disease modelling. OUP oxford; 2010.
2. Vynnycky E, Edmunds W. Analyses of the 1957 (Asian) influenza pandemic in the United Kingdom and the impact of school closures. Epidemiology & Infection. 2008;136(2):166–179.
3. Duggan J. System dynamics modeling with R. Springer; 2016.
4. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. Stan: A probabilistic programming language. Journal of statistical software. 2017;76.
5. Andrade J, Duggan J. A Bayesian approach to calibrate system dynamics models using Hamiltonian Monte Carlo. System Dynamics Review. 2021;37(4):283–309. doi:https://doi.org/10.1002/sdr.1693.

	EpIndCode	Parameter	Q_0.025	Median	Mean	Q_0.975	Q95Rang
1	Epoch3-CH	Beta_Param	0.991	0.995	0.995	1.000	0.009
2	Epoch3-CHD	Beta_Param	0.990	0.995	0.995	0.999	0.009
3	Epoch3-CD	Beta_Param	0.992	0.998	0.998	1.003	0.011
4	Epoch3-C	Beta_Param	0.993	0.999	0.999	1.005	0.012
5	Epoch3-HD	Beta_Param	0.984	0.990	0.990	0.996	0.012
6	Epoch3-H	Beta_Param	0.983	0.990	0.990	0.997	0.014
7	Epoch2-CH	Beta_Param	0.981	0.989	0.989	0.997	0.016
8	Epoch2-CHD	Beta_Param	0.980	0.988	0.988	0.996	0.016
9	Epoch2-H	Beta_Param	0.975	0.986	0.986	0.997	0.022
10	Epoch2-CD	Beta_Param	0.978	0.990	0.990	1.001	0.023
11	Epoch2-HD	Beta_Param	0.973	0.985	0.985	0.996	0.023
12	Epoch2-C	Beta_Param	0.980	0.992	0.992	1.004	0.024
13	Epoch3-D	Beta_Param	0.973	0.987	0.987	1.003	0.030
14	Epoch1-CH	Beta_Param	0.928	0.973	0.976	1.039	0.111
15	Epoch1-CHD	Beta_Param	0.930	0.974	0.977	1.041	0.111
16	Epoch1-C	Beta_Param	0.930	0.976	0.979	1.046	0.116
17	Epoch1-CD	Beta_Param	0.927	0.974	0.977	1.046	0.119
18	Epoch2-D	Beta_Param	0.783	0.958	0.941	0.999	0.216
19	Epoch1-HD	Beta_Param	0.827	0.940	0.944	1.078	0.251
20	Epoch1-H	Beta_Param	0.815	0.929	0.933	1.073	0.258
21	Epoch1-D	Beta_Param	0.196	0.571	0.570	0.948	0.752

	EpIndCode	Parameter	Q_0.025	Median	Mean	Q_0.975	Q95Range
1	Epoch3-CH	CF	0.552	0.598	0.598	0.645	0.093
2	Epoch3-CD	CF	0.559	0.603	0.604	0.653	0.094
3	Epoch3-CHD	CF	0.550	0.596	0.597	0.645	0.095
4	Epoch3-C	CF	0.560	0.606	0.607	0.657	0.097
5	Epoch2-CHD	CF	0.581	0.647	0.648	0.721	0.140
6	Epoch2-CH	CF	0.578	0.644	0.646	0.723	0.145
7	Epoch2-C	CF	0.570	0.639	0.640	0.719	0.149
8	Epoch2-CD	CF	0.574	0.644	0.645	0.727	0.153
9	Epoch1-CHD	CF	0.360	0.634	0.634	0.908	0.548
10	Epoch1-CH	CF	0.358	0.641	0.641	0.921	0.563
11	Epoch1-C	CF	0.346	0.637	0.639	0.915	0.569
12	Epoch1-CD	CF	0.342	0.635	0.635	0.920	0.578
13	Epoch3-HD	CF	0.077	0.239	0.294	0.773	0.696
14	Epoch2-HD	CF	0.078	0.252	0.301	0.785	0.707
15	Epoch3-H	CF	0.076	0.249	0.304	0.787	0.711
16	Epoch2-H	CF	0.077	0.254	0.307	0.790	0.713
17	Epoch1-HD	CF	0.072	0.324	0.359	0.819	0.747
18	Epoch3-D	CF	0.026	0.216	0.271	0.775	0.749
19	Epoch1-H	CF	0.076	0.344	0.380	0.842	0.766
20	Epoch2-D	CF	0.040	0.273	0.320	0.829	0.789
21	Epoch1-D	CF	0.091	0.482	0.487	0.898	0.807

	EpIndCode	Parameter	Q_0.025	Median	Mean	Q_0.975	Q95Range
1	Epoch3-CH	HF	0.089	0.100	0.100	0.111	0.022
2	Epoch3-CHD	HF	0.089	0.099	0.100	0.112	0.023
3	Epoch2-CH	HF	0.081	0.093	0.094	0.107	0.026
4	Epoch2-CHD	HF	0.081	0.093	0.093	0.107	0.026
5	Epoch1-CHD	HF	0.073	0.108	0.110	0.157	0.084
6	Epoch1-CH	HF	0.076	0.113	0.114	0.163	0.087
7	Epoch2-CD	HF	0.015	0.103	0.178	0.699	0.684
8	Epoch3-HD	HF	0.076	0.248	0.300	0.760	0.684
9	Epoch2-HD	HF	0.079	0.244	0.298	0.769	0.690
10	Epoch3-CD	HF	0.016	0.101	0.176	0.713	0.697
11	Epoch2-H	HF	0.078	0.239	0.295	0.784	0.706
12	Epoch3-H	HF	0.076	0.239	0.294	0.785	0.709
13	Epoch1-CD	HF	0.018	0.155	0.218	0.728	0.710
14	Epoch1-HD	HF	0.076	0.319	0.359	0.811	0.735
15	Epoch1-H	HF	0.080	0.354	0.384	0.836	0.756
16	Epoch3-D	HF	0.025	0.221	0.277	0.782	0.757
17	Epoch2-D	HF	0.033	0.254	0.308	0.826	0.793
18	Epoch3-C	HF	0.102	0.498	0.499	0.895	0.793
19	Epoch2-C	HF	0.102	0.501	0.501	0.898	0.796
20	Epoch1-D	HF	0.099	0.492	0.495	0.900	0.801
21	Epoch1-C	HF	0.084	0.505	0.501	0.907	0.823

	EpIndCode	Parameter	Q_0.025	Median	Mean	Q_0.975	Q95Range
1	Epoch3-CHD	DF	0.104	0.121	0.121	0.140	0.036
2	Epoch3-HD	DF	0.104	0.122	0.122	0.141	0.037
3	Epoch2-CHD	DF	0.102	0.122	0.123	0.146	0.044
4	Epoch2-HD	DF	0.103	0.124	0.124	0.149	0.046
5	Epoch1-HD	DF	0.043	0.197	0.222	0.541	0.498
6	Epoch1-CHD	DF	0.045	0.204	0.227	0.562	0.517
7	Epoch2-CD	DF	0.016	0.112	0.178	0.676	0.660
8	Epoch1-CD	DF	0.018	0.139	0.204	0.701	0.683
9	Epoch3-CD	DF	0.017	0.117	0.188	0.718	0.701
10	Epoch2-D	DF	0.036	0.251	0.298	0.786	0.750
11	Epoch3-D	DF	0.027	0.217	0.273	0.779	0.752
12	Epoch2-CH	DF	0.104	0.505	0.502	0.897	0.793
13	Epoch2-C	DF	0.104	0.502	0.503	0.905	0.801
14	Epoch3-C	DF	0.097	0.495	0.498	0.900	0.803
15	Epoch1-D	DF	0.093	0.478	0.483	0.897	0.804
16	Epoch3-CH	DF	0.096	0.495	0.499	0.901	0.805
17	Epoch1-H	DF	0.096	0.497	0.500	0.906	0.810
18	Epoch1-C	DF	0.094	0.492	0.495	0.907	0.813
19	Epoch3-H	DF	0.095	0.513	0.506	0.912	0.817
20	Epoch1-CH	DF	0.090	0.498	0.494	0.908	0.818
21	Epoch2-H	DF	0.088	0.494	0.495	0.908	0.820

Fig 7. Summary of quantiles for parameters across the 21 experiments.

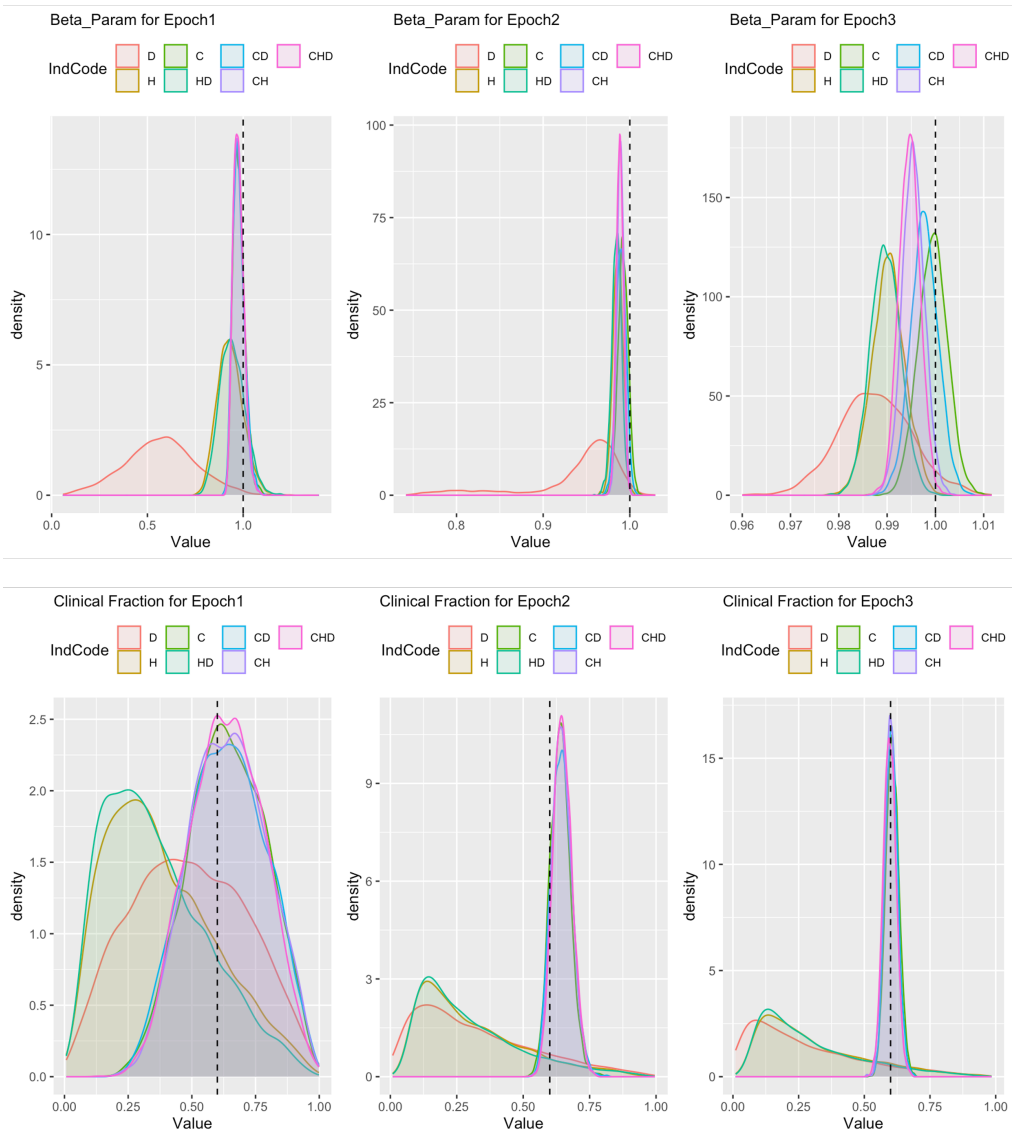

Fig 8. Visualizing a list of three elements

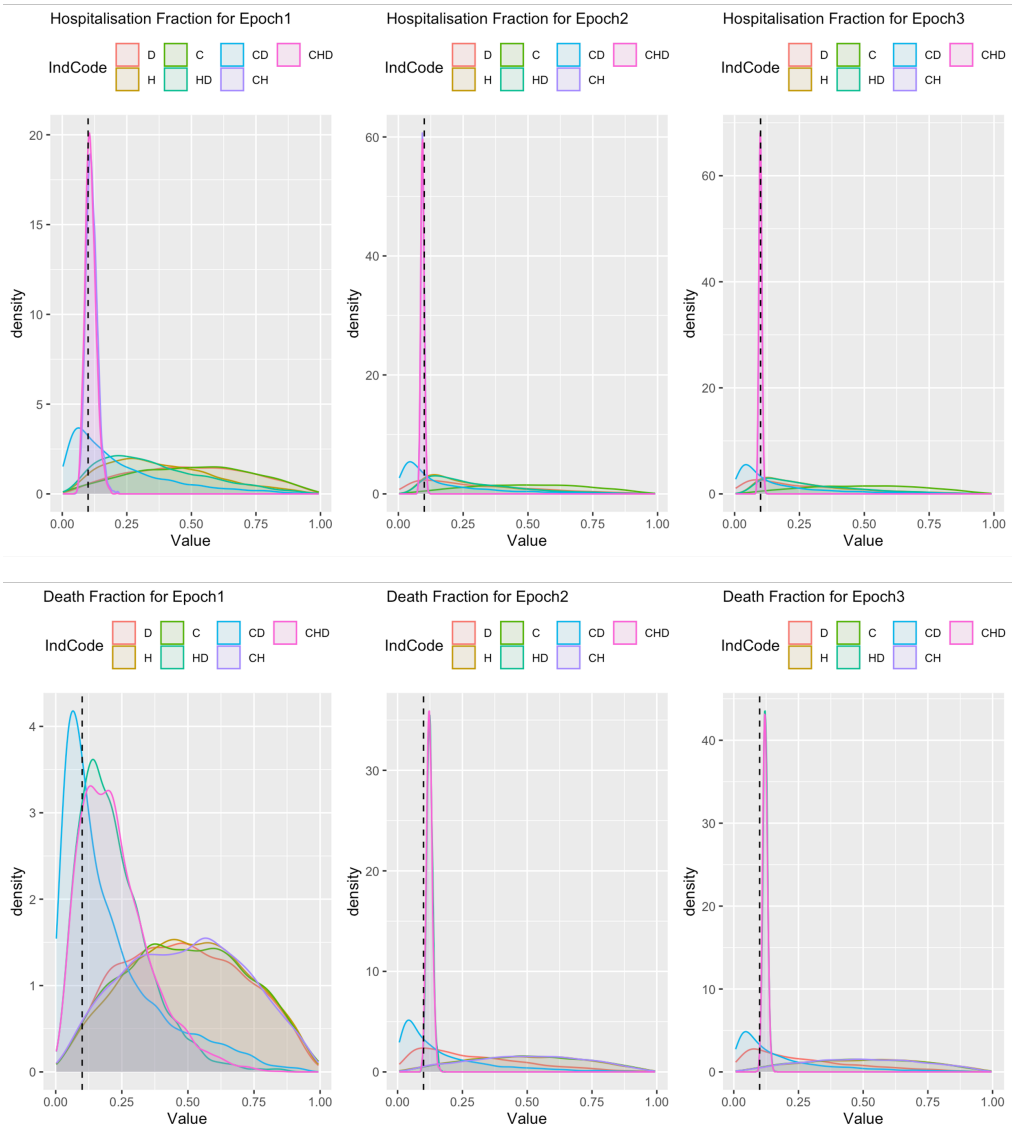

Fig 9. Visualizing a list of three elements

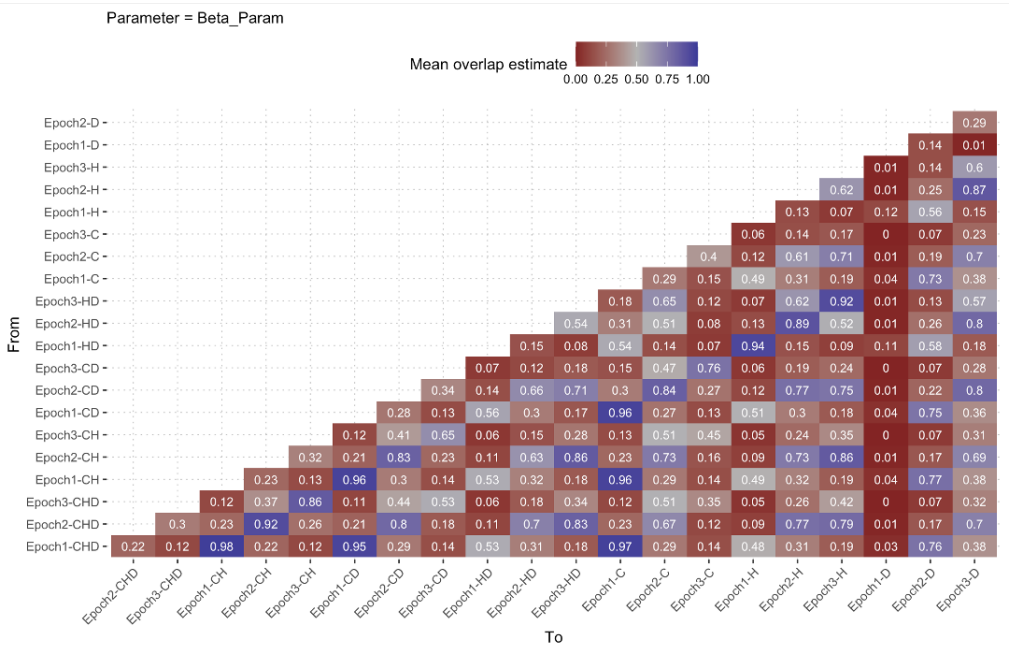

Fig 10. Visualizing a list of three elements

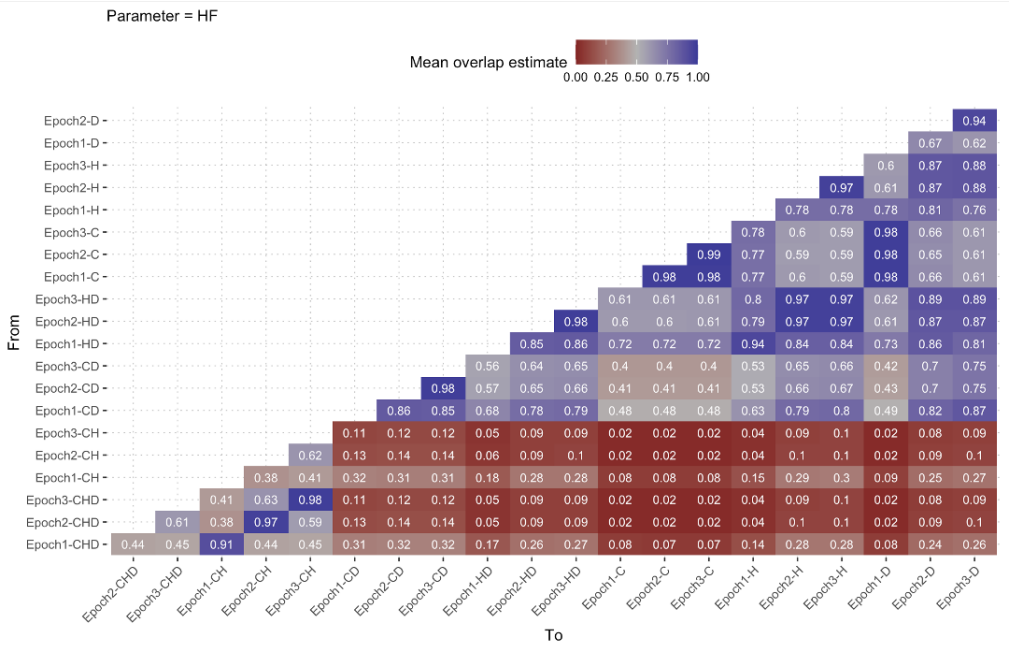

Fig 11. Overlapping calculations for hospitalisation fraction (HF)

6. Andrade J, Duggan J. An evaluation of Hamiltonian Monte Carlo performance to calibrate age-structured compartmental SEIR models to incidence data. Epidemics. 2020;33:100415. doi:https://doi.org/10.1016/j.epidem.2020.100415.

7. Andrade J, Duggan J. Anchoring the mean generation time in the SEIR to mitigate biases in R0 estimates due to uncertainty in the distribution of the epidemiological delays. Royal Society Open Science. 2023;10(8):230515. doi:10.1098/rsos.230515.

8. Gabry J. cmdstanr: R Interface to'CmdStan'. (No Title). 2021;.

9. Andrade J. Readsdr: translate models from system dynamics software into R; 2021. Available from: `https://github.com/jandraor/readsdr`.

10. Pastore M. Overlapping: a R package for Estimating Overlapping in Empirical Distributions. Journal of Open Source Software. 2018;3(32):1023. doi:10.21105/joss.01023.

11. Pastore M, Calcagnì A. Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index. Frontiers in Psychology. 2019;10. doi:10.3389/fpsyg.2019.01089.