Part 1: Generation of (pseudo-)random numbers (30 marks)

Given the distribution:

$$f(x) = \begin{cases} \dfrac{\alpha\left(\dfrac{k-\mu+x}{k}\right)^{-\alpha-1}}{k}, & x \geq \mu \\ 0 & \text{otherwise} \end{cases} \qquad \alpha, k \in \mathbb{R}^+,\ \mu \in \mathbb{R}$$

the inversion method can generate pseudorandom values distributed as $f(x)$. This involves generating $U \sim U[0,1]$ and applying the quantile function of $f(x)$ to U (i.e. $x = F^{-1}(U)$) to produce pseudorandom values distributed as $f(x)$. Given that $f(x)$ is a probability density function (PDF), we require the CDF $F(x)$. Integrating our PDF, w.r.t. x, between μ and x – the bounds specified by the PDF:

$$F(x) = \int_\mu^x \frac{\alpha}{k}\left(\frac{k-\mu+x}{k}\right)^{-\alpha-1}\, dx$$

$$\text{Let } u = \frac{k-\mu+x}{k} \implies \frac{du}{dx} = \frac{1}{k} \ \therefore\ du = \frac{dx}{k} \ \therefore\ dx = k \cdot du$$

$$F(x) = \int_\mu^x \frac{\alpha}{k} u^{-\alpha-1} \cdot k\, du = \int_\mu^x \alpha u^{-\alpha-1}\, du = \alpha \int_\mu^x u^{-\alpha-1}\, du = \alpha \cdot \left[\frac{u^{-\alpha}}{-\alpha}\right]$$

$$= -u^{-\alpha}\Big|_\mu^x = -\left[\left(\frac{k-\mu+x}{k}\right)^{-\alpha} - \left(\frac{k-\mu+\mu}{k}\right)^{-\alpha}\right] = -\left[\left(\frac{k-\mu+x}{k}\right)^{-\alpha} - 1\right]$$

$$\boxed{\therefore\ F(x) = 1 - \left(\frac{k-\mu+x}{k}\right)^{-\alpha}} \tag{1}$$

With $F(x)$, we can find the quantile function $F^{-1}(U)$. Then, by setting $F(x) = U$, we can find $x \sim f(x)$:

$$U = 1 - \left(\frac{k-\mu+x}{k}\right)^{-\alpha}$$

$$1 - U = \left(\frac{k-\mu+x}{k}\right)^{-\alpha}$$

$$\frac{1}{1-U} = \left(\frac{k-\mu+x}{k}\right)^{\alpha}$$

$$\left(\frac{1}{1-U}\right)^{1/\alpha} = \frac{k-\mu+x}{k}$$

$$\boxed{\therefore\ x = F^{-1}(U) = k\left(\frac{1}{1-U}\right)^{1/\alpha} - k + \mu} \tag{2}$$

Figure 1 shows the distribution of 30,000 pseudorandom variates, generated according to $f(x)$ using the quantile function $F^{-1}(U)$, with parameters $\alpha = 3, k = 1, \mu = 0$:

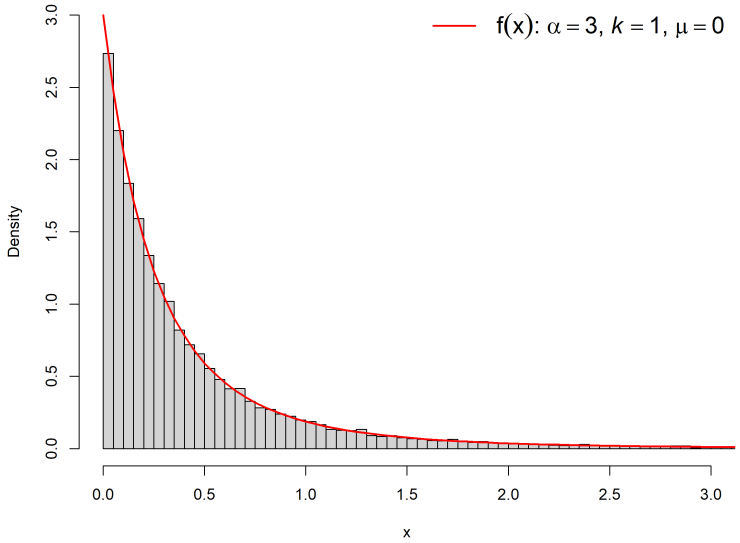

Figure 1: Histogram of 30,000 generated pseudorandom variates, distributed as $f(x)$

Goodness-of-fit testing. To quantify the goodness-of-fit of our sample to the pdf, as a function of N, we can apply the Kolmogorov-Smirnov (KS) test (Kolmogorov, 1933) on independent samples, from size $1 \to N$, drawn from $f(x)$ using $F^{-1}(U)$.

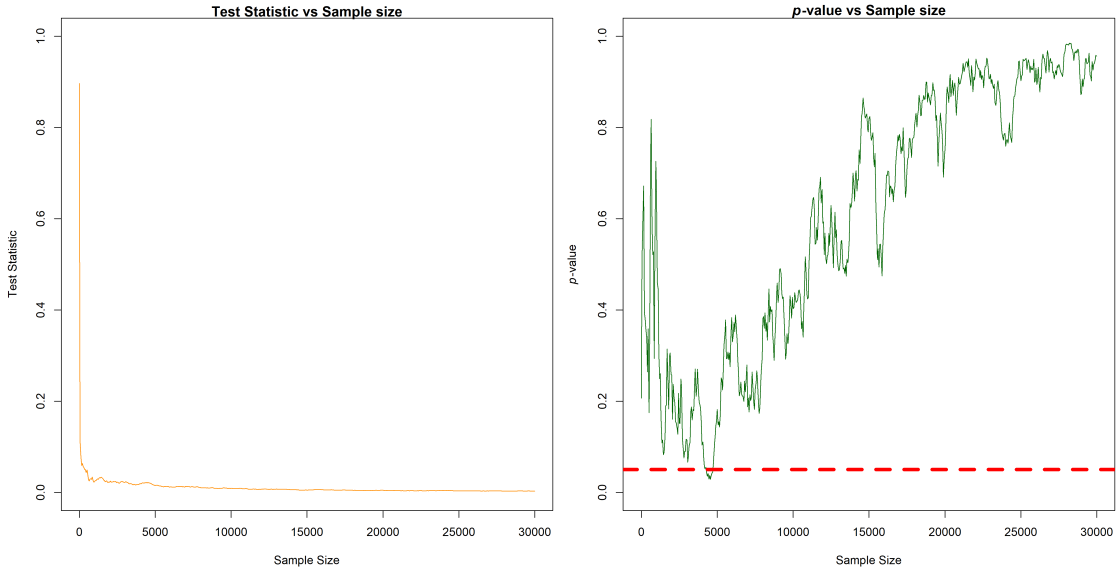

Figure 2: Test Statistics and p-values from ks-test, performed on $F^{-1}(U)$ samples of size 1-30,000

Test statistics $T_i \to 0.009$ as N increases, indicating the maximum difference between the sample CDF and the theoretical CDF. Meanwhile, p-values remain above the 0.05 significance threshold, for all value of N, except between 4,000 and 5,000, indicating that we fail to reject the null hypothesis for almost all values of N.

Find $\mathbb{E}(|\mathbf{X}|)$ ***using importance sampling (IS).*** Rather than estimating $\mathbb{E}(|X|)$ directly from $F^{-1}(U)$ variates, IS involves using a proposal distribution $g(x)$ and applying weightings $w(x) = \frac{f(x)}{g(x)}$ to each generated value from $g(x)$. This enables the variance on $\mathbb{E}(|X|)$ to be reduced as values which contribute more to the expectation can be sampled more frequently, improving sampling efficiency. Given that we can sample from $f(x)$, we can use a $g(x)$ which has the same form as $f(x)$ but with different parameter values. $\mathbb{E}(|X|)$ and $\text{Var}(X)$ can be estimated via IS from the following:

$$\mathbb{E}(|X|) = \mathbb{E}(h(X)) = \int h(x)f(x)\ dx = \int h(x)\frac{f(x)}{g(x)}g(x)\ dx = \mathbb{E}(w(X)h(X))$$

$$\because\ \frac{1}{N}\sum_{i=1}^{N} w(X_i)h(X_i) \xrightarrow{\text{a.s}} \mathbb{E}(w(X)h(X)) \quad \therefore\ \frac{1}{N}\sum_{i=1}^{N} w(X_i)h(X_i) \xrightarrow{\text{a.s}} \mathbb{E}(h(X))$$

$$\boxed{\mathbb{E}(|X|) = \frac{1}{N}\sum_{i=1}^{N} w(X_i)h(X_i)} \tag{3}$$

$$\text{Var}(|X|) = \text{Var}\left(\frac{1}{N}\sum_{i=1}^{N} w(X_i)h(X_i)\right) = \frac{1}{N^2}\sum_{i=1}^{N} \text{Var}\left(w(X_i)h(X_i)\right) = \frac{\text{Var}(w(X)h(X))}{N}$$

$$\boxed{\text{Var}(|X|) = \frac{\text{Var}(w(X)h(X))}{N}} \tag{4}$$

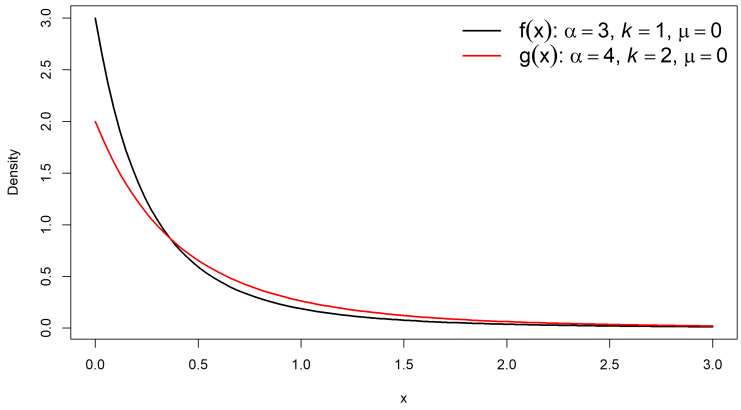

Figure 3: Comparision of Proposal Distribution $g(x)$ with Target Distribution $f(x)$

Table 1 shows how IS results in reduced variance, relative to simple sampling.

	$\mathbb{E}(\|X\|)$	$Var(\|X\|)$
Simple Sampling	0.508	0.778
IS	**0.506**	**0.0000178**

Table 1: Estimates of $\mathbb{E}(|X|)$ and $Var(|X|)$, via Simple Sampling and IS (N = 30,000)

Part 2: Markov Chains and Markov Chain Monte Carlo (30 marks)

Why MCMC works. Markov chains are a type of stochastic process where each state $X_t \in \mathcal{S}$ (where $\mathcal{S}$ is the state space) depends only upon the previous state X_{t-1}, and not any prior states. The probability of moving from some state X_i to another X_j in k steps (i.e. the transition probability) can be expressed as:

$$p_{i,j}^{(k)} = \mathbb{P}(X_{t+k} = j | X_t = i)$$

A transition matrix P will describe each of these transition probabilities. If a Markov chain possesses the properties of **irreducibility**, **aperiodicity** and **positive recurrence**, it will converge towards a unique **invariant distribution** $\pi(X)$ as the number of steps tends to infinity. An **irreducible** Markov chain means that there is a non-zero probability of reaching every other state from any starting state. Irreducibility is required, otherwise the chain will not converge towards a unique invariant distribution. Irreducibility can be expressed as:

$$\exists\ k \in \mathbb{N} : P_{i,j}^{(k)} > 0, \qquad \forall i, j \in \mathcal{S}$$

An **aperiodic** Markov chain has a period $D = 1$ iff 1 is the greatest common divisor of $\left\{k \geq 1 : p_{i,i}^{(k)} > 0\right\}\ i \in \mathcal{S}$. Given that we require an irreducible Markov chain, we can determine that the Markov chain is aperiodic if:

$$\exists\ i \in \mathcal{S} : P_{i,i} > 0$$

A Markov chain is **positive recurrent** when there is a finite expected time for the chain to return to its starting state. Let T_i be the 1^{st} return time to X_i:

$$\mathbb{E}[T_i | X_0 = i] < \infty$$

In the case of an irreducible Markov chain, positive recurrence is assured, provided the state space $\mathcal{S} < \infty$. With only finitely many states to move to, and the ability to move to any of them from any other state (irreducibility), the chain will eventually return to its starting point in finitely many steps. These properties in combination enable us to sample from a unique invariant distribution. We can calculate this invariant distribution from our transition matrix P. Provided P is irreducibility and aperiodic, the unique invariant distribution can be expressed as:

$$\pi(X_k) = \sum_{j=i}^{M} \pi(X_j) p_{j,k} \qquad \forall j, k \in \mathcal{S}$$

π is a row-vector $(\pi(X_1), \ldots, \pi(X_M))$ representing each state's probability mass in the invariant distribution. π can be found by identifying the left-eigenvector of P with eigenvalue = 1. Computing this provides the invariant distribution. In summary, given some target distribution π, Markov chain Monte Carlo aims to find an irreducible and aperiodic Markov chain with transition matrix P which has π as its invariant distribution. The Ergodic Theorem then allows us to estimate the distribution's expectation value:

$$\forall f : \mathcal{S} \to \mathbb{R} \qquad \mathbb{E}(f(X)_n) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \xrightarrow{\text{a.s.}} \mathbb{E}(f(X))$$

MCMC Simulation. Suppose we have a Markov chain, with $\mathcal{S} = \{S_1, S_2, S_3, S_4, S_5\}$, described by the following transition matrix P:

$$P = \begin{pmatrix} 0.31604254 & 0.25869199 & 0.1489175 & 0.01231435 & 0.26403364 \\ 0.03150148 & 0.36524669 & 0.2841283 & 0.24752339 & 0.07160011 \\ 0.30609973 & 0.14349155 & 0.1403304 & 0.12099784 & 0.28908044 \\ 0.27962029 & 0.28624100 & 0.1593571 & 0.19573749 & 0.07904413 \\ 0.24392730 & 0.08553483 & 0.2373811 & 0.18115231 & 0.25200447 \end{pmatrix}$$

We can note that since each transition probability $p_{i,j} > 0$, P is irreducible. Furthermore, given its irreducibility, P is also aperiodic, due to each $p_{i,i} > 0$. Moreover, with irreducibility and the fact that $|\mathcal{S}| < \infty$, the Markov chain is also positive recurrent. As such, we can compute the unique invariant distribution of the Markov chain using P by finding the left-eigenvector of P with eigenvalue = 1:

$$\pi(X) = \{0.2289303, 0.2309293, 0.1972263, 0.1481236, 0.1947905\}$$

Figure 4 shows the difference between empirical realisations of this invariant distribution $\widehat{\pi}_n(X)$ and the theoretical invariant distribution $\pi(X)$, as a function of sample size. Evidently, estimates converge to $\pi(X)$ as $n \to \infty$

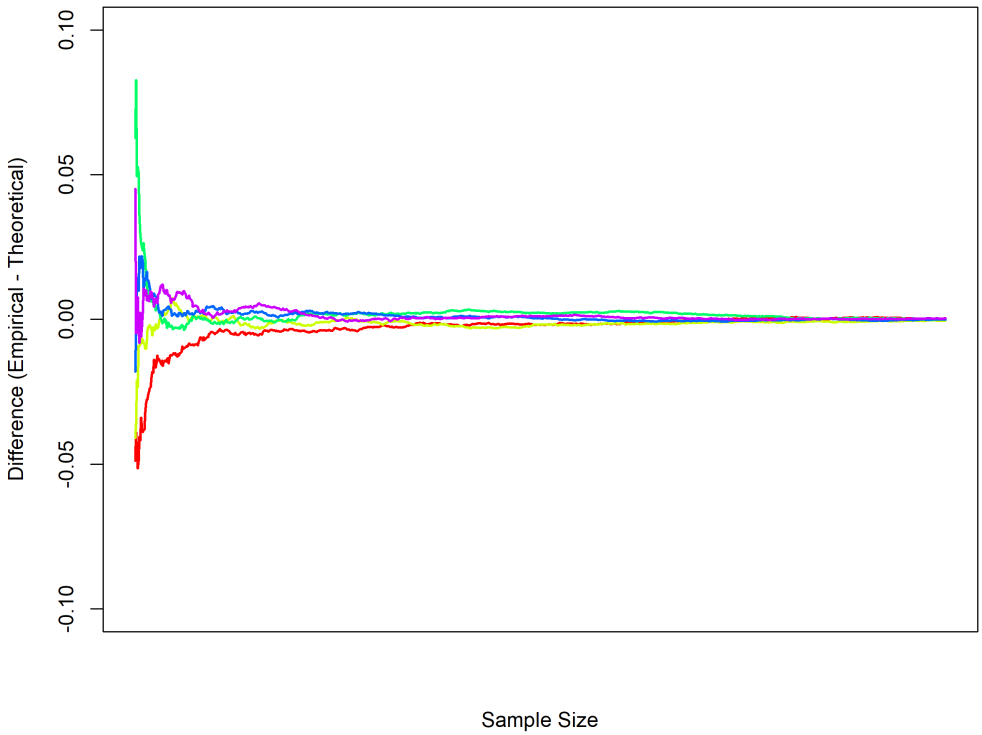

Figure 4: Difference Between Empirical and Theoretical Invariant Distributions, by Sample Size

Part 3: MCS in current research topics (40 marks)

Overview of De-Leon and Aran (2023). De-Leon and Aran (2023) aimed to compare two methods for predicting the spread of COVID-19 in Israel: the susceptible-infectious-removed (SIR) model and the authors' Monte Carlo Agent-based Model (MAM).

Firstly, the SIR model places individuals in one of three states: susceptible, where the individual can develop the disease but has yet to catch it; infectious, where the individual is currently infected and capable of spreading the disease, and removed wherein the individual can no longer develop the disease, either through developing immunity (natural or vaccinated) or death. Each individual's transition between these three states is determined by a set of ordinary differential equations (ODEs) to determine the number of individuals in each state at each time point t.

In the case of De-Leon and Aran (2023), these ODEs take the form:

$$\frac{dI^i(t)}{dt} = -\mu \cdot I^i(t) + S^i(t) \cdot \beta^i \cdot \sum_{j=1} I^j(t)$$

$$\frac{dS^i(t)}{dt} = -S^i(t) \cdot \beta^i \cdot \sum_{j=1} I^j(t)$$

$$\frac{dR^i(t)}{dt} = \mu \cdot I^i(t)$$

$\frac{dI^i(t)}{dt}$, $\frac{dS^i(t)}{dt}$ & $\frac{dR^i(t)}{dt}$ represent the change in the infected I, susceptible S and recovered R populations of each i^{th} group at each time point t. A key issue with the SIR model approach is that such equations fail to represent some important behavioural differences between subgroups. For instance, the authors point out how "since nonpharmaceutical interventions (NPIs) are essential in controlling COVID-19 and they tend to fluctuate across regions and countries, the model must be able to distinguish regionality" (De-Leon and Aran, 2023, p. 2). This SIR model approach thus operates at the population level and so assumes homogeneity in behaviour amongst individuals within the population. An alternative approach might instead seek to operate at the level of each individual. For instance, an individual's proximity to infectious persons influences their probability of infection; or an individual's vaccination status might protect or expose them to infection different to other members of their subgroup. To account for such indivudla variation, an SIR model approach would require a unique ODE for each combination.

To address this limitation, the authors propose their MAM approach. This approach is based on agent-based modelling, wherein each individual in the population is represented as a particle within a space. Particles can then move around in this space at each time step and interact with other particles based on their characteristics. MAM assigns an initial position to each particle within an $L \times L = A$ simulation area, wherein the positions are assigned randomly by sampling uniformly between 0 and L for the x and y co-ordinates of each particle. The authors then allow these particles to move stochastically by updating their x and y by sampling from a normal distribution with $\mu = 0$ and $\sigma = L_0/2$, where L_0 is the starting side length of the simulation area A.

Particle movement. To simulate particle movement at each time step, the Box-muller method can be utilised. This method generates pairs of variates distributed $\mathcal{N}(0,1)$, which can then be transformed to any desired $\mathcal{N}(\mu,\sigma)$.

We can begin by representing two random standard normal variates as polar co-ordinates:

$$X_1, X_2 \sim \mathcal{N}(0,1) \implies R_1 = R\cos(\theta),\ R_2 = R\sin(\theta)$$

where $\theta \sim U(0, 2\pi)$ and R, being the straight-line distance to the origin $R^2 = x^2 + y^2$. If x are y are distributed $\mathcal{N}(0,1)$, then x^2 and y^2 are each distributed χ^2_1. In combination, $x^2 + y^2$ is distributed χ^2_2

We can express χ^2_2 as Gamma$(1, 1/2)$, since:

$$Z_1, ..., Z_k \sim \mathcal{N}(0,1) \implies \sim \text{Gamma}(k/2, 1/2)$$

In our scenario, $k = 2$, so:

$$R^2 = x^2 + y^2 \sim \text{Gamma}(2/2, 1/2) \implies \sim \text{Gamma}(1, 1/2)$$

Next, we can then note that Gamma$(1, 1/2) \implies$ Exp$(\lambda = 1/2)$:

$$\text{Gamma}(\alpha = 1, \beta = 1/2) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{\beta x} = \frac{(1/2)^1}{(1-1)!} x^{1-1} e^{(1/2)x} = 1/2 e^{(1/2)x} = \text{Exp}(\lambda = 1/2)$$

Exp$(1/2)$ can then be expressed as:

$$R^2 \sim \text{Exp}(1/2) \implies \sim -2\log(U_1), \quad U_1 \sim U(0,1)$$

Thus, by using:

$$R = \sqrt{-2\log(U_1)}, \quad \theta = 2\pi U_2$$

We can generate $X_1, X_2 \sim \mathcal{N}(0,1)$:

$$X_1 = \sqrt{-2\log(U_1)} \cdot \cos(2\pi U_2), \qquad X_2 = \sqrt{-2\log(U_1)} \cdot \sin(2\pi U_2)$$

Finally, to have X_1 and X_2 distributed as $\mathcal{N}(\mu,\sigma)$, we can observe that since X_1 and X_2 are distribured according to a standard normal distribution, we can re-arrange the standard score formula to convert these standard normal variates to arbitrary normal variates:

$$z = \frac{x-\mu}{\sigma} \implies x = \mu + \sigma z$$

Let's call these arbitrary normal variates Y_1 and Y_2:

$$Y_1 = \mu + \sigma X_1, \qquad Y_2 = \mu + \sigma X_2$$

With $\mu = 0$ and $\sigma = L_0/2$, each time step in the simulation will move the position of each particle by $Y_1 = \Delta x$ and $Y_2 = \Delta y$.

Infection mechanism. With the particle movement defined, we can now outline particle infection. At $t = 0$, some number of particles N_I are initialised as infectious. For each particle which is susceptible $n_{SI} \in N_{SI}$, the probability of infection at time step t is given by:

$$P_{infected} = \left\lfloor \sum_{N_I} \text{Infectious}(t_i, t) \times (1 - VE) \times \exp\left(\frac{(X_{N_I} - X_{N_{SI}})^2 + (Y_{N_I} - Y_{N_{SI}})^2}{2\sigma_r^2}\right) + \epsilon \right\rfloor$$

If $P_{infected} > 0$, the particle is then infected.

Infectious(t_i, t) is a function to define when an infected particle is actively infectious. For the alpha and delta variants, this is 4-7 days (time steps) after infection, and for omicron it is 2-5 days.

VE is the vaccine effectiveness (thus $1 - VE$ is the risk of infection, given the particle's vaccination effectiveness) of a given particle at a given time step t. For each age group, an initial number of particles are vaccinated at time step $t = 0$, propotional to their actual vaccination levels in Israel as of January 8^{th}, 2021. "Vaccination rates from this date forward were estimated using an exponential function" (De-Leon and Aran, 2023, p. 4). Whilst unclear what function is actually used, I am assuming it to take the form: $1 - e^{\lambda x}$. As for vaccine effectiveness, it is initially 0 for the first week then linearly rises to 0.9 by day 28. VE then begin linearly decreasing after day 150, down to 0.6 after 180 days.

The exponential expression is the distance that a given particle is to every infectious particle at time step t. Lastly, ϵ adds a uniform variate $U(0, 1)$ to each probability. Combined with the floor function, this means the ϵ will convert the probability into a binary value (i.e. infected or not infected).

Summary of De-Leon and Aran (2023)'s findings. The authors ran the MAM using 11,000 particles and simulated it 800 times to scale up to the population size of Israel (9,200,000). To assess model performance, the authors looked at the mean absolute percentage error (MAPE). During the first outbreak, where the alpha variant was dominant, a naive SIR model overestimated infections whilst the multiage-SIR model and MAM successfully modelled when infections peaked and declined. However, MAM outperformed SIR in predicting the decline of infections and was overally superior in terms of accuracy (MAPE = 89% SIR, MAPE = 95% MAM). The second outbreak is marked by the emergence of the delta variant and the waning immunity from the first vaccination. This waning immunity was counter-acted by a booster dose. The authors again found MAM to outperform the multiage-SIR model by more successfully predicting the peak of infections and having greater overall accuracy (MAPE = 84% SIR, MAPE = 88% MAM). However, MAM failed to predict the observed sharp decline in infections during this period, with the authors suggesting that this may be due to MAM being unable to model the Jewish high-holidays when schools close. Lastly, the third outbreak introduced the omicron variant, with its shorter latent period (the time between initial infection and becoming infectious). The multiage-SIR and MAM were both successful at predicting overall infections, but MAM was again superior in predicting the peak and had a higher MAPE (MAPE = 84% SIR, MAPE = 90% MAM). That being said, multiage-SIR and MAM were unable to capture the sharp decline in infections amongst 60+ year

olds, likely resulting from a fourth dose only given to those over 60. Overall, MAM showed superior accuracy to the multiage-SIR model, and MAM was better able to prepresent phenomena such as the transition between COVID-19 variants and variation in vaccination effectiveness between subgroups.

Reproduction of specific MC application from De-Leon and Aran (2023). The application from De-Leon and Aran (2023) chosen for replication is modelling of the first outbreak, marked by the initial dose of the vaccine in response to the alpha variant.

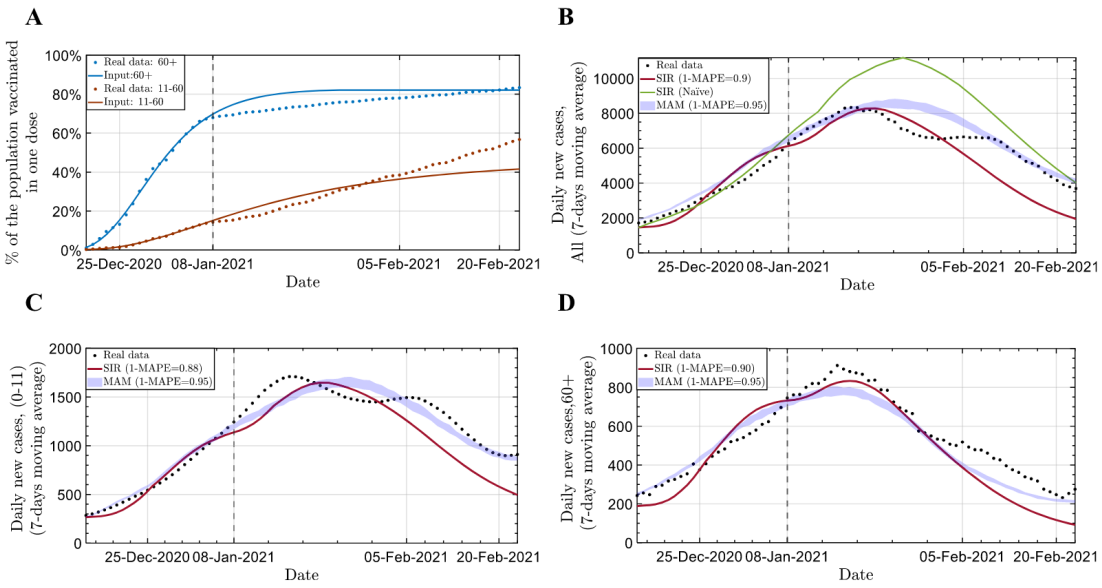

Figure 5: Figure 1 from De-Leon and Aran (2023)

Modelling from 25/12/2020 to 20/02/2021, using a simulation with particle size 1.1×10^4:

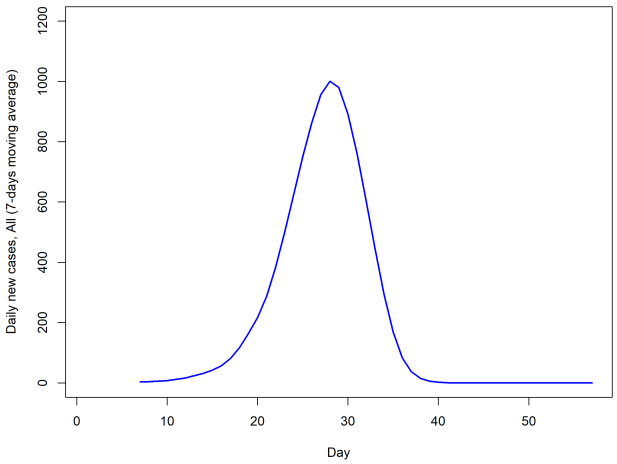

Figure 6: Reproduction of Figure 1B from De-Leon and Aran (2023)

This reproduction varies significantly from De-Leon and Aran (2023). The reproduction does match the overall expected pattern where infections initially increase, reach a maximum and then decline, but the increase and decrease are much sharper than in De-Leon and Aran (2023) and the peak occurs around 10 days eariler. I also found running the simulation to be too computationally

expensive to run 800 times, although this would not have changed the overall pattern observed in **Figure 6**.

One potentialy reason for why the replication failed is that some parameters in the model were unclear. Firstly, it was unclear how vaccination rates should change between age groups. As mentioned, it was assumed that the increase followed the form: $1 - e^{\lambda x}$. but no value for λ was provided, nor was it clear that this was how De-Leon and Aran (2023) acutally modelled vaccination increase. Furthermore, in the calculation of infection probability, it was unclear what the $2\sigma_r^2$ term referred to, and so the value was chosen arbitrarily.

In conclusion, De-Leon and Aran (2023) successfully demonstrated how a Monte Carlo Agent-based Model approach can yield superior prediction accuracy compared to the traditional SIR model approach. MAM consistently outperformed the multiage-SIR model on mean absolute percentage error. Interestingly, some challenges remained in predicting COVID-19's spread in Israel as MAM was unable to capture some specific events which occurred such as public holidays and aspects of the vaccination campaign. The reproduction attempt of the first outbreak failed to find similar results, potentially due to limited information about model parameters.

References

De-Leon, H. and Aran, D. (2023), 'MAM: Flexible Monte-Carlo Agent based model for modelling COVID-19 spread', *Journal of Biomedical Informatics* **141**, 104364.

Kolmogorov, A. (1933), 'Sulla determinazione empirica di una legge didistribuzione', *Giorn Dell'inst Ital Degli Att* **4**, 89–91.