Quantifying the over-estimation of predicted pandemic curves caused by population clustering

Mathias L. Heltberg,[1,2,3,†,*] Christian Michelsen,[1,†]
Emil S. Martiny,[1] Lasse Engbo Christensen,[4]
Mogens H. Jensen,[1] Tariq Halasa,[5] and Troels C. Petersen[1]

[1]Niels Bohr Institute, University of Copenhagen, Blegdamsvej 17, 2100 Copenhagen E

[2]Laboratoire de Physique, Ecole Normale Superieure, Rue Lhomond 15, Paris 07505

[3]Infektionsberedskab, Statens Serum Institute, Artillerivej, 2300 Copenhagen S

[4] DTU Compute, Section for Dynamical Systems,
Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Anker Engelunds Vej 101A, 2800 Kongens Lyngby

[5] Animal Welfare and Disease Control, University of Copenhagen,
Gronnegårdsvej 8, 1870 Frederiksberg C

[†]These authors contributed equally.

[*]To whom correspondence should be addressed; heltberg@nbi.ku.dk.

The modeling of Sars-CoV-2 has become a critical aspect of present life. A standard procedure has become the implementation of more or less detailed differential equations which estimate the time development in the number of Susceptible, (Exposed,) Infected, and Recovered (SIR/SEIR) individuals. However, these models are based on very simple assumptions which constitute obvious approximations. In this paper, we introduce an agent based model including spatial clustering. Based on Danish population data, we estimate how this impacts the early prediction of a pandemic and the long term developments. Our results show that early phase SEIR model predictions overestimate the peak number of infected and the equilibrium level by a at least a factor two. These results are robust to variation of parameters influencing connection distances within a realistic range and independent of distribution of infection rates.

Since the emergence of reports of a new disease from Wuhan, China, the pathogen now known as SARS-CoV-2 has spread dramatically, paralyzing societies, resulting in a large number of deaths and severe economic damage worldwide (*1, 2*). Mathematical models were developed to estimate the reproduction number and guide the authorities in an attempt to minimize the damage caused by this novel virus (*3, 4, 5, 6*). Generally, the models have been variants of the SIR/SEIR model and vary in complexity including simple deterministic compartmental models (*6, 7*), meta-population compartmental models (*8, 9, 10*), individual based models without including spatial specifications (*4, 11, 12*), and finally spatio-temporal agent based models (*13*).

One important aspect in the modelling is the ability to predict the infection peak height and the steady state level of individuals having been infected based on the early

rise in the number of infected before governmental interference.

Earlier work has pointed out the importance of including heterogeneity when modeling the spread of infectious disease such as contact patterns between individuals (*14*), population mixing assumptions (*15*), heterogeneities caused by super-spreaders (*12*), and the spatial dependency of COVID-19 (*16, 17*). However, these mathematical models have not combined heterogeneous elements nor quantified how much the early SIR/SEIR predictions for Sars-CoV-2 might be biased.

We address these deficiencies by constructing a stochastic agent based model (ABM) on a realistic, spatially distributed population where each individual (agent) could be in one of four states and switch according to the rates schematized in Fig. 1A. We used this ABM to investigate how heterogeneities can bias our prediction in the early phases of a pandemic. In order to reduce the space of parameters, we do not include political countermeasures during the pandemic, but rather test the levels of herd immunity and endemic steady state. Before including heterogeneity, we compared the ABM to the corresponding SEIR model, and found them to agree within 5% for all parameter configurations tested (see Supplementary Information).

We constructed a network of spatially distributed contacts using data based on:

- The geographic location of people in Denmark (from Boligsiden (*18*))
- The average number of contacts per individual per day of 11 (from HOPE (*19*)). Given an average infectious period of 4 days, we approximate the average number of effective contacts to be $\mu = 40$.
- The average commuting distance $\rho = 0.1\,\mathrm{km}^{-1}$ and the fraction of long distance commutes $\epsilon_\rho = 4\,\%$ (from Statistics Denmark (*20*))

This is schematized in Fig. 1B (see also Supplementary Information) and all ten parameters in this model are explained in Table 1. Having introduced heterogeneity, the

distribution of connections in this network are created automatically through the population clustering, see Fig. 2A. This naturally leads to individuals living in densely populated areas having higher number of connections.

In an example simulation with 100 initially infected individuals, $N_{\text{init}} = 100$, we observed a spatial difference in areas affected by the disease (Fig. 2B), as expected. One region reached local endemic steady state (green arrow, Fig. 2B) while other regions of similar density were highly infected (red arrow, Fig. 2B) and yet other districts were almost unaffected (grey arrow, Fig. 2B).

To quantify the effect of population clustering, we compared the ABM result to the reference SEIR model of similar parameters. Generally, we observed that the epidemic developed faster with a higher infection peak I_{peak}, but also subsided quicker, leading to a lower number of infected once reaching endemic steady state, R_∞ (Fig. 2C and Fig. 2D).

In order to explore how population clustering affects the epidemic, we chose a reference value of infection rates, $\beta = 0.01$, and an alternative value of $\beta = 0.007$. In the absence of spatial dependence ($\rho = 0\,\text{km}^{-1}$) these correspond to initial reproduction numbers $\mathcal{R}_0 \approx 1.7$ and 1.1, respectively. Increasing the spatial dependence (i.e. increasing ρ) lead to a significant rise in the infection peak for the ABM, $I_{\text{peak}}^{\text{ABM}}$, compared to the (unaffected) SEIR model, $I_{\text{peak}}^{\text{SEIR}}$ for both the reference value and the alternative lower value of β (black and blue points, Fig. 2E). We introduced heterogeneity in infection strengths ($\sigma_\beta = 1$, see Fig. 1B), thus making some individuals much more infectious than others (i.e. including '*super shedders*'). We found no significant impact from this effect (red points in Fig. 2E). Similarly, we introduced heterogeneity in connection weights ($\sigma_\mu = 1$, see Fig. 1B), thus making some individuals much more likely to form contacts than others (i.e. including '*super connecters*'). This leads to a significant effect for $\rho = 0\,\text{km}^{-1}$, which converges towards the other curves for $\rho > 0.1\,\text{km}^{-1}$ (orange (only super connecters) and green

(super connecters and super shedders) points in Fig. 2E). The total number of infected individuals when the epidemic is over, R_∞, converged towards half of the SEIR model prediction as a function of ρ except for $\beta = 0.007$ where the endemic steady state level is larger than the one obtained by the SEIR model (Fig. 2F). Fixing $\rho = 0.1\,\mathrm{km}^{-1}$ and increasing the fraction of distance-independent contacts, ϵ_ρ, we found that $I_{\mathrm{peak}}^{\mathrm{ABM}}$ is almost unaffected for $\epsilon_\rho < 0.5$ (Fig. 2G), while R_∞^{ABM} increases linearly towards the SIER model R_∞^{SEIR}, as expected (Fig. 2H).

Next we consider how these heterogeneities bias the traditional SEIR model predictions, especially the predictions based on fits to the number of infected (i.e. the curve to be flattened) in the beginning of the epidemic (see Supplementary Information). Without spatial dependence, the predicted curves fitted the number of infected individuals very well (Fig. 3A). Introducing spatial dependence ($\rho = 0.1\,\mathrm{km}^{-1}$) leads to a severe overestimation of the epidemic based on the number of early infection cases (Fig. 3B). This result can be interpreted by the fact that in societies where population density and thus individual contact number varies significantly, the early phase will be driven by people with many contacts (super connecters). This typically happens in cities where the population density is high. Increasing the spatial dependence ρ we found that the SEIR model predictions overestimated the infection peak height I_{peak} and the total number of infected R_∞ significantly even for very small spatial heterogeneities (Fig. 3C and Fig. 3D). We observed this general trend for all tested combinations of parameters and heterogeneities. In particular we found that if long-distance connections ϵ_ρ are below $10\,\%$, the bias in the estimated infection peak height, I_{peak}, was constant within statistical uncertainty (Fig. 3E). For the total number of infected, R_∞, we observed an almost linear regression to the SEIR model as ϵ_ρ approaches one. However, even when $\epsilon_\rho = 0.25$, the prediction bias was still a factor of two (Fig. 3F). We concluded from these curves a general trend;

if one fits a SEIR model to infection numbers during the beginning of an epidemic, and use these estimates to predict the characteristics of the epidemic at a national level, one overestimates the number of infected by at least a factor of two.

In summary, our research reveals that the degree of population clustering in Denmark creates a discrepancy between the early predictions made by the SEIR models and the underlying agent based interactions. It results in a significant overestimation of the impact of the disease, both in terms of maximal number of simultaneously infected (by a factor of 3) and the endemic steady state level (by a factor of 2.5). Such discrepancies have been observed for earlier pandemics, for instance the 1918 Spanish flu, where the predicted herd immunity level was severely overestimated (*21*). These results can be an important element in explaining these mismatches, even though other elements, as for instance social distancing and mutations to the viral strain, also play a part. During 2020, numerous countries have been faced with the task of laying out strategies to minimize the consequences of SARS-CoV-2, including the importance of '*flattening the curve*'. While this is truly crucial to avoid overpopulated hospitals, it should be taken seriously enough that we might specify to a higher degree of certainty which curve to be flattened. The mean reproduction number, $\mathcal{R}_0$, was around $2.5 - 3$ across countries in the early phase (*22*). This led to predictions estimating that 60% of the populations would be infected at the end of the pandemic. Our results estimate this number to be closer to 30%. This has two important implications which are largely of a positive character. First, we do not have to fear as many infected as predicted by SEIR models and other models not including spatial clustering. This could already have significant impact on the outbreaks that will occur during the rest of the pandemic, which are likely to be smaller than expected, especially in countries that have already measured a high fraction of infected in the population. Secondly, our study emphasize the great benefits by making lock-downs early in the pan-

demic, when it is driven by '*super connectors*'. Since people in living in city-clusters are more likely to catch the infection, they are, on average, more likely to be affected in the beginning and by removing contacts from these individuals, one can avoid the worst peak while affecting the fewest number of people.

During this pandemic, mathematical predictions have been heavily criticized (*23, 24*) and it is now important to improve the accuracy of them, in order to increase the confidence in the predictions. Our work seriously questions the validity of predictions based on SEIR models, quantifies their biases from not including spatial clustering, and suggests that the precise effect of population clustering should be addressed more seriously. While our work has absolutely no political agenda, it should serve as an input in the current debate of how to handle the severe consequences of a crisis like SARS-CoV-2.

Figures and Tables

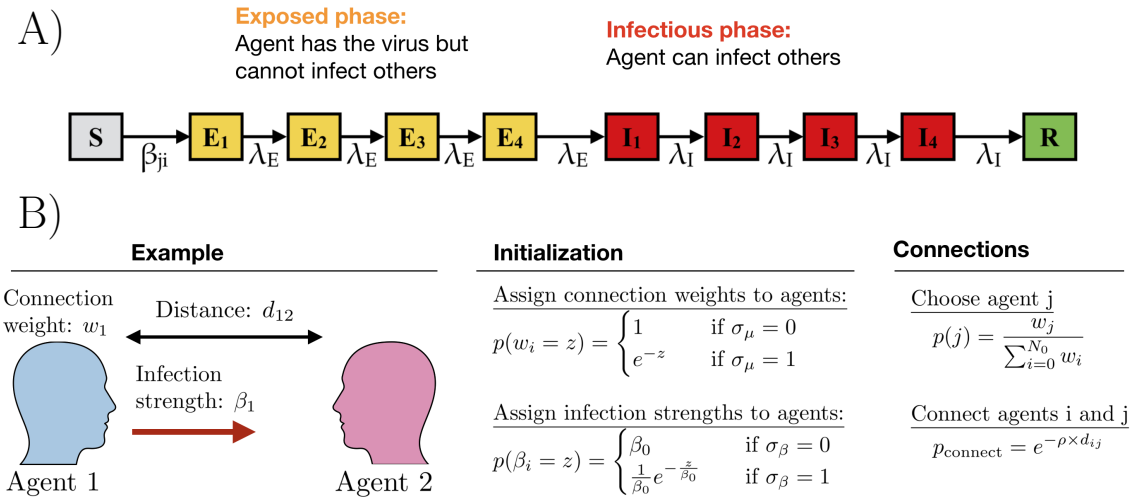

Figure 1: **A)** Illustration of the modified susceptible-exposed-infected-removed (SEIR) model used. It consists of ten consecutive states (S, E_{1-4}, I_{1-4}, and R), with transition rates governed by β, λ_E, and λ_I, respectively. **B)** Illustration of how the spatial network is generated and heterogeneities in individuals included.

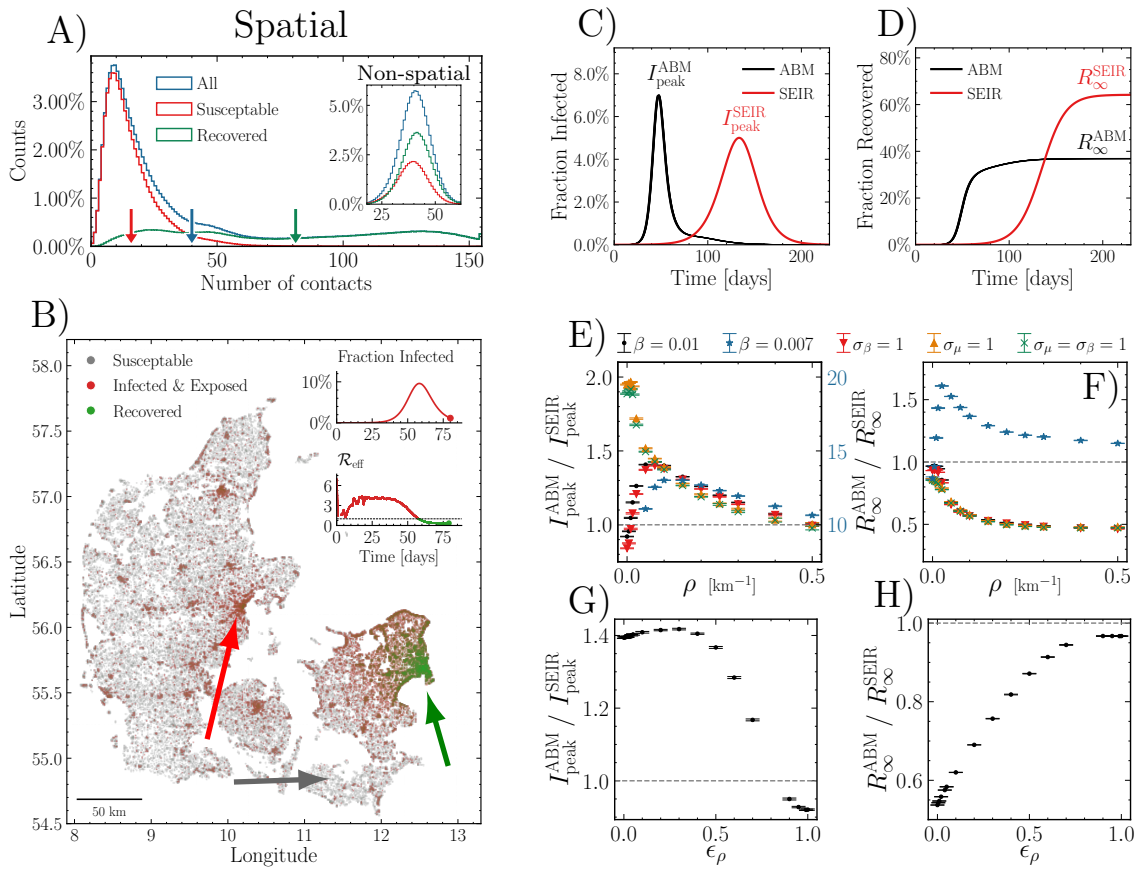

Figure 2: **A)** Histograms showing the number of susceptible (red) and recovered (green) individuals at the end of an epidemic with $\rho = 0.1\,\mathrm{km}^{-1}$. The distribution before the epidemic is shown in blue. The arrows show the mean of each distribution. The inset shows the same for $\rho = 0\,\mathrm{km}^{-1}$. **B)** Visualisation of the spatial position of individuals during the infection and which state they are in. Green arrow: Largest city in Denmark (Copenhagen): mostly recovered. Red arrow: Second largest city in Denmark (Aarhus): mostly infected. Grey arrow: low-population area: mostly susceptible (i.e. have not been infected). **C)** Number of infected individuals as a function of time. Data shown for the spatially distributed network ($\rho = 0.1\,\mathrm{km}^{-1}$). Simulation was repeated 10 times. **D)** cumulative sum of individuals who have had the disease as a function of time (with $\rho = 0.1\,\mathrm{km}^{-1}$). **E)** Relative difference in maximal number of infected, I_{peak}, between deterministic (SEIR) and ABM as a function of ρ, and shown for different parameters. Note the data for $\beta = 0.007$ are shown in blue with a factor 10 scaling (right y-axis). **F)** Relative difference in total number of infected at the end of the epidemic, R_∞, between deterministic (SEIR) and ABM as a function of ρ. Colors similar to **E**. **G)** Same as **E**, but as a function of ϵ_ρ. **H)** Same as **F**, but as a function of ϵ_ρ.

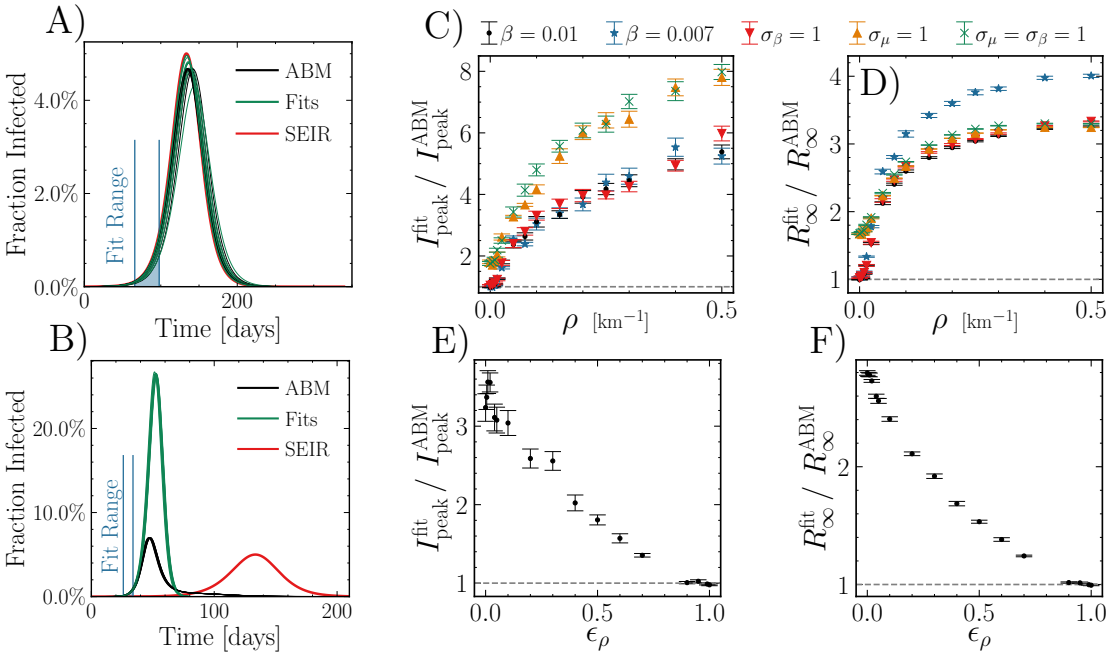

Figure 3: **A)** Number of infected individuals for the ABM in black, the SEIR model in red and the SEIR fits to the ABM data in green. Blue lines show the interval where parameters are fitted (also shown below the curves). Here $\rho = 0\,\text{km}^{-1}$. **B)** Same as **A** but with population clustering ($\rho = 0.1\,\text{km}^{-1}$). **C)** Relative difference in maximal number of infected, I_{peak}, between the fit and the ABM for different values of ρ. Simulations repeated 10 times for each data-point. **D)** Relative difference in total number of infected at the end of the epidemic, R_∞, between the fit and the ABM for different values of ρ. **E)** Same as **C**, but as a function of ϵ_ρ. **F)** Same as **D**, but as a function of ϵ_ρ.

Variable	Description	Value	Range	Units
N_0 :	Population size	$5.8 \cdot 10^6$	$10^5 - 10^7$	–
N_{init} :	Number of individuals initially infected	100	$1 - 10^4$	–
μ :	Average number of network contacts	40	$10 - 100$	–
β :	Typical infection strength	0.01	$0.001 - 0.1$	day^{-1}
λ_E :	Rate to move through ¼ of latency period	1	$0.5 - 4$	day^{-1}
λ_I :	Rate to move through ¼ of infectious period	1	$0.5 - 4$	day^{-1}
σ_μ :	Population clustering spread	0	$0 - 1$	–
σ_β :	Interaction strength spread	0	$0 - 1$	–
ρ :	Typical acceptance distance	0.1	$0 - 0.5$	km^{-1}
ϵ_ρ :	Fraction of distance-independent contacts	0.04	$0 - 1$	–

Table 1: Overview of the ten parameters applied in this study, their typical value, and the ranges we have considered. The first six parameters are standard SEIR parameters, whereas the last four parameters define the heterogeneity in the model. These four parameters do not affect the SEIR model.

References

1. M. Chinazzi, *et al.*, *Science* **368**, 395 (2020).

2. W.H.O., Listings of WHO's response to COVID-19, `www.who.int/news/item/29-06-2020-covidtimeline`.

3. R. M. Anderson, H. Heesterbeek, D. Klinkenberg, T. D. Hollingsworth, *The Lancet* **395**, 931 (2020).

4. J. Hellewell, *et al.*, *The Lancet Global Health* **8**, e488 (2020).

5. M. J. Keeling, T. D. Hollingsworth, J. M. Read, *medRxiv* p. 2020.02.14.20023036 (2020).

6. T. Kuniya, *Journal of Clinical Medicine* **9**, 789 (2020).

7. R. Li, *et al.*, *Science* **368**, 489 (2020).

8. K. Prem, *et al.*, *The Lancet Public Health* **5**, e261 (2020).

9. B. A. D. van Bunnik, *et al.*, *medRxiv* p. 2020.05.04.20090597 (2020).

10. L. Danon, E. Brooks-Pollock, M. Bailey, M. J. Keeling, *medRxiv* p. 2020.02.12.20022566 (2020).

11. S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, M. Prokopenko, *arXiv:2003.10218 [cs, q-bio]* (2020).

12. K. Sneppen, R. J. Taylor, L. Simonsen, *medRxiv* p. 2020.05.17.20104745 (2020).

13. G. J. Milne, S. Xie, *medRxiv* p. 2020.03.20.20040055 (2020).

14. S. Bansal, B. T. Grenfell, L. A. Meyers, *Journal of The Royal Society Interface* **4**, 879 (2007).

15. L. Kong, J. Wang, W. Han, Z. Cao, *International Journal of Environmental Research and Public Health* **13** (2016).

16. D. Kang, H. Choi, J.-H. Kim, J. Choi, *International Journal of Infectious Diseases* **94**, 96 (2020).

17. D. Giuliani, M. M. Dickson, G. Espa, F. Santi, *arXiv:2003.06664 [stat]* (2020).

18. Boligsiden.dk, `www.boligsiden.dk`.

19. HOPE Project, `www.hope-project.dk`.

20. Statistics Denmark, Statistikbanken, `www.statistikbanken.dk`.

21. V. Andreasen, C. Viboud, L. Simonsen, *The Journal of infectious diseases* **197**, 270 (2008).

22. P. Boldog, *et al.*, *Journal of Clinical Medicine* **9**, 571 (2020).

23. I. Holmdahl, C. Buckee, *New England Journal of Medicine* (2020).

24. L. Wynants, *et al.*, *BMJ* **369** (2020).