Development of an interactive, agent-based local stochastic model of COVID-19 transmission and evaluation of mitigation strategies illustrated for the state of Massachusetts, USA

Alexander Kirpich\textsuperscript{1,2}, Vladimir Koniukhovskii\textsuperscript{3\*}, Vladimir Shvartc\textsuperscript{3}, Pavel Skums\textsuperscript{4}, Thomas A. Weppelmann\textsuperscript{5}, Evgeny Imyanitov\textsuperscript{6}, Semyon Semyonov\textsuperscript{7}, Konstantin Barsukov\textsuperscript{3}, Yuriy Gankin\textsuperscript{7}\*

\textsuperscript{1}Department of Population Health Sciences, Georgia State University, Atlanta, Georgia, United States
\textsuperscript{2}School of Public Health, Georgia State University, Atlanta, Georgia, United States
\textsuperscript{3}EPAM Systems, Saint Petersburg, Russia
\textsuperscript{4}Department of Computer Science, Georgia State University, Atlanta, Georgia, United States
\textsuperscript{5}Department of Internal Medicine, University of South Florida, Tampa, Florida, United States
\textsuperscript{6}N.N. Petrov Research Institute of Oncology, Saint Petersburg, Russia
\textsuperscript{7}Quantori, Cambridge, Massachusetts, United States

\*These authors contributed equally to this work.
* yuriy.gankin@quantori.com

\section*{S1 Appendix}

\textbf{The Model Structure}

The model stochastic simulations are generating the infected individuals at different times and stages. Those individuals are indexed by $k$ and have individual characteristics presented below:

\begin{equation}
Q_k = \left((x_k, y_k), tin_f_k, det_k, stg_k, age_k, rad_k, p_{cont(k)}, cont_k, R_0(k), sever_k, dur_k, st_k(t)\right).
\end{equation}

The characteristics of each individual $Q_k$ are determined and updated within the simulation process and have the following details:

- $(x_k, y_k)$ - the Cartesian coordinates (in pixels) of the individual $Q_k$ that do not change within the simulation process after they are introduced;
- $tin_f_k$ – the time of infection onset for the individual $Q_k$ which is measured in relation to the simulation baseline time denoted as 0;
- $det_k$ – the detection time variable measured in days that corresponds to a period from an infection acquisition until the proper diagnosis and reporting of individual $Q_k$;
- $stg_k = (stg_{1(k)}, stg_{2(k)}, stg_{3(k)})$ – the vector of durations of three disease infection stages measured in days that characterize the infectivity of the given individual during those stages. It is assumed that $stg_{1(k)} + stg_{2(k)} + stg_{3(k)} = det_k$;
• \( age_k \) – the age of the individual \( Q_k \) at the time of the infection onset;

• \( rad_k \) – the distance (in meters) up to which the individual \( Q_k \) is able to infect the nearby individuals;

• \( p_{\text{cont}(k)} \) – the probability that during each day the individual \( Q_k \) has any contacts which lead to new infections;

• \( cont_k = (\mu_{\text{cont}(k)}, \sigma_{\text{cont}(k)}^2) \) – the individual-specific parameters that define the distribution of the number of successful infection transmissions to other individuals within a given day. This number is generated randomly for each day \( t \), provided that the individual has any transmissions on the given day (according to the contact probability \( p_{\text{cont}(k)} \)).

• \( R_{0(k)} \) – the individual’s reproduction number. This variable stores the number of individuals that are infected by \( Q_k \) during the infection period \( det_k \). The average of those reproduction numbers across individuals and simulations are used to estimate the basic reproduction number \( R_0 \) \[1\] which is a characteristic of the entire epidemic;

• \( sever_k \) – the disease severity variable for the individual \( Q_k \) that takes three values, where 1 corresponds to lethal, 2 corresponds to severe, and 3 corresponds to mild; the disease severity does not change for a given individual after it is determined randomly from a trinomial distribution;

• \( dur_k \) – the disease duration from the infection onset to cure (or death) in days, which is generated randomly based on the \( sever_k \) parameter;

• \( st_k(t) \) – the status of the individual \( Q_k \) at a given day \( t \). The status of the individuals within the simulation is expected to change over time and is expected to take the following values:
  
  \[- st_k(t) = 0 \] the individual is detected based on the external information i.e. from the reported data that are used as the model input;

  \[- st_k(t) = 1 \] the individual is infected but has not been identified as such yet;

  \[- st_k(t) = 2 \] the individual has been infected and detected as such, which has also implied the individual’s isolation (quarantine).

  \[- st_k(t) = 3 \] the individual has recovered and is immune;

  \[- st_k(t) = 4 \] the individual has deceased.

The asymptomatic infections are incorporated inside the model via three severities that are generated for each infected individual. This incorporation of asymptomatic infections into those with a “mild” severity status is made to improve the model computational tractability. The first severity status is “mild” which incorporates asymptomatic individuals as well. Then after an individual within this undetected class with any severity status is getting detected (i.e. tested and confirmed to have an infection) this individual is assigned the “detected” status \( (st_k(t) = 1) \) and is immediately quarantined. This approach allows to incorporate all infected individuals with different disease severities into the infected status and to isolate the confirmed cases. Since the model parameters are calibrated based on the detected cases and the output of the model presents the detected cases as well the prediction abilities for detected and reported cases within the model are preserved.

In the beginning of the simulations the model utilizes multiple local epidemic epicenters \( E = \{E_1, E_2, \ldots, E_T\} \). Those epicenters serve as the model initial conditions.
and represent the geographic points for the index cases that are introduced into the susceptible population. The epicenters can either correspond to the actual address coordinates for those places where the initial outbreaks were detected or to the centers of the corresponding aggregated geographic units. The latter may be the case, if either the exact infection acquisition locations are not known, or the privacy concerns prevent the inclusion of such data into the model. In the latter case the centers of the aggregated geographic units are taken as epicenters $E_i$ for each $i = 1, 2, \ldots, \mathcal{I}$.

The local epicenters in the model are defined by a pair of geographic coordinates $(\text{Lat}, \text{Long})$ and by an epicenter-specific region radius $R_i$ which is defined in meters. Therefore, for $i = 1, 2, \ldots, \mathcal{I}$ the epicenter region is defined by a triplet:

$$E_i = ((\text{Lat}_i, \text{Long}_i), R_i).$$

(2)

The epicenter regions are defined from the surveillance epidemiological data. As the initial conditions in addition to the local epicenters the model incorporates the areas of high density $P = \{P_1, P_2, \ldots, P_J\}$ for $j = 1, 2, \ldots, \mathcal{J}$, where each $P_j$ represents a large city or a densely populated area and which is also defined by a triplet:

$$P_j = ((\text{Lat}_j, \text{Long}_j), R_j).$$

(3)

In the model the reporting times (days) for the initial index cases for each epicenter $i$ precede the modeled epidemic starting time which corresponds to the baseline time slot $t = 0$. Therefore, the reported time slot indexes across the epicenters $E_i$ are denoted as $s = 1, 2, \ldots, \mathcal{S}$ with the corresponding times $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_\mathcal{S}$. The earliest reported cases and their dates are used for the model input with indexes $s = 1, 2, \ldots, \mathcal{S}$ such that $\mathcal{S} < \mathcal{S}$ and the corresponding times $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_\mathcal{S}$. The corresponding number of confirmed and reported infections for each local epicenter $E_i$ up to and including the time $\hat{t}_s$ for $s = 1, 2, \ldots, \mathcal{S}$ is denoted as $n_i(\hat{t}_s)$. The corresponding set of infected and reported (i.e. with the status $s_{t_k}(t) = 1$) individuals across all epicenter is denoted as:

$$D = \{\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_{K_S}\},$$

(4)

where $k = 1, 2, \ldots, K_S$ is the global index for initial cases across all times $\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_\mathcal{S}$ and $K_S$ is the total number of the initial index cases simulated within the model based on the input data. The tilde notation for $Q_k$-s in $D$ emphasizes the link to the model input data.

The time index that corresponds to individual day within the model is denoted as $t$ and is equal to 0 at the model baseline. The simulation baseline time $t = 0$ corresponds to the latest reporting time $\hat{t}_\mathcal{S}$ of the earliest reported cases that are used for the model input. The actual infection times for those index cases precede the selected baseline simulation time $t = 0$ due to the infectivity periods generated for those index cases prior to their reporting. The actual simulation starting time that accounts for the infectivity periods is denoted as $t = T_{\text{min}}$ and is smaller than the baseline time $t = 0$. This simulation starting time $t = T_{\text{min}}$ is generated within the model, while the baseline time $t = 0$ is defined by the data and is defined by the largest index within the set of calibration indexes $s = 1, 2, \ldots, \mathcal{S}$. The largest simulation time $t = T_{\text{max}}$ is determined by the model user based on the desired length of prediction. The initial set of index cases $D$ from (4) defines the model initial cases that are allocated across the local epicenters (2) at times up to the baseline time $t = 0$.

Based on the model geographic characteristics (2) and (3) and the initial set of reported individuals $D$ from (4) the new lists $L(t)$ of modeled individuals are simulated for time slots $t \in [T_{\text{min}}, T_{\text{max}}]$ where $T_{\text{max}} - T_{\text{min}} + 1$ is the total number of the simulated time slots. The simulated lists $L(t)$ have the following format:

$$L(t) = \{Q_1, Q_2, \ldots, Q_{K(t)}\}$$

(5)
where the value of $K(t)$ is defined by the simulation at every simulation time step $t \in [T_{\text{min}}; T_{\text{max}}]$. During this procedure the input set individuals $D$ defined in (4) is allocated between the different epicenters and time slots within the lists $L(t)$ defined by (5). The allocations of the set $D$ is performed during the time slots $t \in [T_{\text{min}}; t_S]$ where $t_S < T_{\text{max}}$.

**Simulation of the Model Inputs Based on the Reported Data**

The initial reported time point $\tilde{t}_1$ corresponds to the indexing time $t = -\tilde{S}$ in relation to the selected baseline simulation time $t = 0$. For that time $\tilde{t}_1$ the cumulative number of infected and reported individuals at each epicenter $E_i$ for $i = 1, 2, \ldots, I$ is known from the model input and is denoted as $n_i(\tilde{t}_1) = n_i(-\tilde{S})$. Those numbers are also assumed to correspond to the initial increase in cases during that first reported time $\tilde{t}_1$ i.e. it is assumed $\Delta n_i(\tilde{t}_1) = n_i(\tilde{t}_1)$ for each epicenter $i$. The subsequent time steps are indexed by $t = \tilde{t}_1 + \tau = -\tilde{S} + \tau$ for the positive integer values of $\tau > 0$. Then the increase in the number of new cases that corresponds to index $t = -\tilde{S} + \tau$ is defined as the difference in cumulative counts:

$$
\Delta n_i(t) = \Delta n_i(\tilde{t}_1 + \tau) = \Delta n_i(-\tilde{S} + \tau) = n_i(-\tilde{S} + \tau) - n_i(-\tilde{S} + \tau - 1).
$$

In those notations for each epidemic epicenter $E_i \in E$ at time index $t \in [-\tilde{S}, \ldots, 0]$ the amount of the reported cases for the epicenter $E_i$ is denoted as $\Delta n_i(t)$ and corresponds to the list:

$$
D_i(t) = \{ \tilde{Q}_1, \tilde{Q}_2, \ldots, \tilde{Q}_{\Delta n_i(t)} \}. \quad (6)
$$

In the end, the initially reported individuals from (4) for each epicenter $E_i$ and timing index $t$ are simulated as the model input in the following way:

- The geographic coordinates $(\text{Lat}_i, \text{Long}_i)$ of the epicenter $E_i$ in the model are re-coded as Cartesian coordinates $C_i = (x_{i(c)}, y_{i(c)})$ which are defined in pixels. The character $c$ in the subscript emphasizes that they belong to the epicenter. Since the modeled geographic region is relatively small in comparison to the entire planet surface, the discrepancies between the geographic and the Cartesian coordinate systems such as the differences in latitude scales within the region are negligible. The radius $R_i$ (in peters) of each modeled circular epicenter in geographic coordinates is recalculated for the Cartesian coordinate system (in pixels) and is denoted as $r_i$.

The initial numbers of infected individuals within each epicenter $E_i$ is determined by the model input and those numbers are fixed. If the exact addresses and the corresponding geographic coordinates for individuals $Q_k$ are known, they can be imputed directly as points $M_k(x_k, y_k)$ in the Cartesian coordinate system within clusters $E_i$ that they belong to. More often, however, due to privacy concerns only the geographic unit level data are known, and the individuals are reported for the geographic center of the administrative unit which is treated in the model as the center of $E_i$ for some $i = 1, 2, \ldots, I$. In this case the geographic locations of the reported individuals $\tilde{Q}_k$ within the given cluster $E_i$ are randomly generated. The individual’s coordinates $M_k = M_k(x_k, y_k)$ in the Cartesian coordinate system are generated in the following way:

- The distance $\rho_k$ from the center $C_i$ to the generated point $M_k$ is generated randomly from the exponential distribution with the scale parameter $r_i$ i.e. $X \sim \mathcal{E}(1/r_i)$ with the density function:

$$
f(x) = \frac{1}{r_i} \exp \left( -\frac{x}{r_i} \right) \mathbb{I} [x \geq 0]
$$
where $r_i$ is the radius of the epicenter $i$.

- The polar angle $\varphi_k$ of the point $M_k$ in relation to the epicenter $C_i$ is generated randomly from the uniform distribution defined on the closed interval that represents the unit radius circle i.e. from the $\mathcal{U}[0; 2\pi]$.
- The coordinates of the point $M_k$ take the form:

$$
x_k = x_{i(c)} + \rho_k \cos(\varphi_k) \text{ and } y_k = y_{i(c)} + \rho_k \sin(\varphi_k),$$

where the angle $\varphi_k$ is counted counterclockwise in relation to the reference direction which corresponds to the $x$-axis direction in the Cartesian coordinate system.

- The values of the detection time $det_k$ are generated from a log-normal distribution ($LN$) with parameters $(\mu_{det}, \sigma^2_{det})$ and the generated outcomes are rounded to integer values. The log-normal random variable $X$ with parameters $(\mu_{det}, \sigma^2_{det})$ is assumed to be generated by taking the value $X = \exp(\sigma_{det} Z + \mu_{det})$ where $Z$ is generated from the standard normal distribution (i.e. $Z \sim \mathcal{N}(0, 1)$). In this parameterization:

$$E[X] = \exp \left[ \mu_{det} + \frac{\sigma^2_{det}}{2} \right] \text{ and } Var[X] = \left[ \exp \left[ \frac{\sigma^2_{det}}{2} \right] - 1 \right] \cdot \exp \left[ 2\mu_{det} + \sigma^2_{det} \right].$$

This implies the formulas for $\mu_{det}$ and $\sigma^2_{det}$ in terms of $Var[X]$ and $E[X]$:

$$\sigma^2_{det} = \ln \left[ 1 + \frac{Var[X]}{E[X]^2} \right] \text{ and } \mu_{det} = \ln \left[ E[X] \right] - \frac{1}{2} \ln \left[ 1 + \frac{Var[X]}{E[X]^2} \right],$$

where $\ln$ stands for a natural logarithm function.

- For each index case $Q_k$ the infection time $tinf_k$ is defined as $tinf_k = \hat{t}_s - det_k$ for some index $s \in \{1, 2, \ldots, S\}$ where $\hat{t}_s$ corresponds to the time index when the individual $Q_k$ has been reported. The infection time $tinf_k$ for some individuals may be earlier than the time $\hat{t}_v$ of the first reported case and the smallest of those times are taken as the actual beginning of the simulation $T_{min}$ which is described in the simulation details.

- The lengths of the disease stages $stg_{1(k)}$ and $stg_{2(k)}$ for the individual $Q_k$ are generated from normal distributions with empirically calibrated parameters via the following sequence of steps. The random variable $S_{1(k)}$ is generated from the normal distribution $\mathcal{N}(0, (0.08 \cdot det_k)^2)$ and the length of first stage $stg_{1(k)}$ is defined as

$$stg_{1(k)} = \max \left\{ 0.4 \cdot det_k + S_{1(k)} , 0 \right\},$$

which is subsequently rounded to an integer value. Then the random variable $S_{2(k)}$ is generated from the normal distribution $\mathcal{N}(0, (0.08 \cdot(det_k - stg_{1(k)}))^2)$ and the length of the second stage $stg_{2(k)}$ is defined as

$$stg_{2(k)} = \max \left\{ 0.4 \cdot (det_k - stg_{1(k)}) + S_{2(k)} , 0 \right\},$$

which is subsequently rounded to an integer value. Finally, the duration of the last stage $stg_{3(k)}$ is defined as

$$stg_{3(k)} = det_k - stg_{1(k)} - stg_{2(k)}.$$
• The age $age_k$ of the individual $Q_k$ is generated based on the empirical cumulative distribution function for the Massachusetts population constructed from the 2020 state of Massachusetts Population Pyramid \([2]\). More precisely, the empirical cumulative distribution function is constructed from the pyramid as $F_{age}(a) := P([\text{pyramid age}] < a)$ for a given age variable $a$. The age of the individual $Q_k$ is generated as $age_k := F_{age}^{-1}(U)$ where $U$ is a random variable generated from the standard uniform distribution (i.e. $U \sim \mathcal{U}(0; 1)$).

• The value of the radius $rad_k$ up to which the individual $Q_k$ is able to infect the nearby individuals is generated from a normal distribution with parameters $\mathcal{N}(\mu_{\text{dist}}, \sigma_{\text{dist}}^2)$. The value of the mean parameter $\mu_{\text{dist}}$ for the simulations of $rad_k$ is multiplied by 0.7 and 0.6 for those individuals who are younger than 10 and older than 60 years since they are assumed to have more limited social mobility.

• The amount of contacts for every individual $Q_k$ for each day $t$ is generated from a normal distribution with parameters from $\text{cont}_k$ i.e. from $\mathcal{N}(\mu_{\text{cont}}(k), \sigma_{\text{cont}}^2(k))$ with parameters that are specific to the individual $Q_k$. In particular, the mean distribution parameter $\mu_{\text{cont}}(k)$ is itself randomly generated once for each individual $Q_k$ and is adjusted in time based on the individual’s disease state. The parameter $\mu_{\text{cont}}(k)$ is changed to $0.2 \cdot \mu_{\text{cont}}(k)$ during the latent period $stg_l(k)$, is equal to $\mu_{\text{cont}}(k)$ during the early symptomatic period $stg_s(k)$, and is changed to $1.5 \cdot \mu_{\text{cont}}(k)$ during the severe infection period $stg_h(k)$. The value of $\mu_{\text{cont}}(k)$ is generated from a normal distribution with parameters $\mathcal{N}(\mu_{\text{dense}}, \sigma_{\text{dense}}^2)$ if the individual’s coordinates $M_k$ belong to any area of high density $P_j$, for some $j = 1, 2, \ldots, J$, and from a normal distribution with parameters $\mathcal{N}(\mu_{\text{usual}}, \sigma_{\text{usual}}^2)$ otherwise. In the same way the value of $\sigma_{\text{cont}}(k)$ is generated from a normal distribution with parameters $\mathcal{N}((0.5 \cdot \mu_{\text{dense}}), (0.25 \cdot \mu_{\text{dense}})^2)$ if the individual’s coordinates $M_k$ belong to any area of high density $P_j$, for some $j = 1, 2, \ldots, J$, and from a normal distribution with parameters $\mathcal{N}((0.5 \cdot \mu_{\text{usual}}), (0.25 \cdot \mu_{\text{usual}})^2)$ otherwise. The corresponding means $\mu_{\text{dense}}$ and $\mu_{\text{usual}}$ are multiplied by 0.7 and 0.5 for individuals who are younger than 10 and older than 60 years due to the assumed limited social mobility for those age groups. The resulting values are rounded to integers after they are generated.

• The parameter $p_{\text{cont}}(k)$ for individual $Q_k$ defines the probability of any infection-induced contacts for a single day when the quarantine measures are not implemented. The values of $p_{\text{cont}}(k)$ are generated once for each individual $Q_k$ from a log-normal distribution ($\mathcal{LN}$) with parameters $(\mu_{\text{ocp}}, \sigma_{\text{ocp}}^2)$. The log-normal random variable $X$ with parameters $(\mu_{\text{ocp}}, \sigma_{\text{ocp}}^2)$ is generated as $X = \exp(\sigma_{\text{ocp}} Z + \mu_{\text{ocp}})$ where $Z$ is generated from the standard normal distribution (i.e. $Z \sim \mathcal{N}(0, 1)$). The detailed formulas for the $\mu_{\text{ocp}}$ and $\sigma_{\text{ocp}}^2$ in terms of mean and variance of the distribution $X$ are described earlier in \([4]\). The abbreviation $\text{ocp}$ in the subscript stands for the “ordinary contact probability”. The parameter $\mu_{\text{ocp}}$ is equal to the population-level parameter denoted as $\text{Ordinary\_Contact\_Probability}$ which is fitted for the model. The value $\sigma_{\text{ocp}}$ is assumed to be small and fixed at 0.05. The values of the $\mu_{\text{ocp}}$ are multiplied by 0.8 and 0.6 for individuals who are younger than 10 and older than 60 years due to the assumed limited social mobility for those age groups. The generated probability values $p_{\text{cont}}(k)$ for each $Q_k$ are checked to be within $[0; 1]$ range and re-generated if they are larger to ensure the correct boundaries for the probability values. The value of the $p_{\text{cont}}(k)$ is used for the binomial distribution with a single trial to a $t$ every time step $t$ to generate the presence of absence of contacts at a given day i.e. from the distribution $\mathcal{BLN}(1, p_{\text{cont}}(k))$. 

January 27, 2021

6/13
• The individual’s reproduction number \( R_{0(k)} \) is saved from the simulations for every infected person and contains the number of the other individuals that this person infects during the simulations. The numbers \( R_{0(k)} \) across individuals and simulation runs are used to compute the averaged values and the corresponding confidence bands that are used as estimates of the population basic reproduction number \( R_0 \).

• The \( \text{sever}_k \) parameter is generated as an outcome of a trinomial distribution with a single trial with probabilities \( MN(1, (p_1(g), p_2(g), p_3(g))) \) which is a special case of a multinomial distribution. The first probability subscript contains the index of the disease severities described for (1). The second subscript \( g \) emphasizes that those probabilities are specific to age groups. There are three age groups within the model which are defined as:

\[
G = \{(< 20), (20 - 50), (> 50)\}.
\]

(11)

• The duration of treatment parameter \( \text{dur}_k \) is generated from the normal distribution \( N(\mu_d\text{ur(sever)}, \sigma^2_d\text{ur(sever)}) \), which depends on disease severity. The resulting duration is rounded to integer values to be represented in days. The generated values that are smaller than 3 are changed to 3 since three days are assumed to be the shortest disease duration period [3] [4] [5].

• The status \( (st_k(t)) \) of the individual \( \tilde{Q}_k \) is defined as \( st_k(t) = 0 \) when \( t \) is the reporting date for that individual and is updated for the other time steps \( t \) within the simulation. The initially reported individuals \( \tilde{Q}_k \) have the status \( st_k(t) = 1 \) before the reporting data which changes to \( st_k(t) = 0 \) on the day of reporting.

The entire set of originally reported cases that corresponds to the reporting times \( \tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_S \) is denoted as \( D \) and is presented with the global indexing in (4). The same set \( D \) can also be presented as a union of sets \( D_i(t) \) defined in (6) across all local epicenter \( i = 1, 2, \ldots, I \) at all time points up to and including the baseline time \( t = 0 \) and has the following representation:

\[
D = \bigcup_{t=-S}^{0} \bigcup_{i=1}^{I} D_i(t).
\]

(12)

In the formula (12) the number of reported cases within each epicenter \( D_i(t) \) during time \( t \) is different from epicenter to epicenter and is equal to \( \Delta n_i(t) \). In the simulation every reported individual \( \tilde{Q}_k \) from \( D \) is infectious with the status \( st_k(t) = 1 \) during the time interval \( [\inf f_k; \inf f_k + \det_k] \) and is isolated and is no longer able to infect the others after the time \( \inf f_k + \det_k \) with the status \( st_k(t) = 2 \). The actual simulation starting time that accounts for all infectivity periods that precede the individual’s reporting is denoted as \( T_{\min} \) and is defined as:

\[
T_{\min} = \min_{\tilde{Q}_k \in D} \inf f_k.
\]

(13)

**Generation of the New Infections Within the Model**

Every individual \( Q_k \) within the model is represented by the class (1). During the time indexes from interval \( t \in T_{\inf f(k)} = [\inf f_k; \inf f_k + \det_k] \) the individual \( Q_k \) has the status \( st_k(t) = 1 \) and can transmit infection to the others. Therefore, at each time step \( t \in T_{\inf f(k)} \) the individual \( Q_k \) infects \( \inf f_k(t) \) new individuals with probability \( P_{\text{cont}(k)} \)
and parameters \(\text{cont}_k\). The lists of individuals infected by \(Q_k\) during the time slot \(t\) are denoted as \(L_k(t)\) and have the format:

\[
L_k(t) = \{I_1, I_2, \ldots, I_{\text{inf}_k(t)}\},
\]

where the total number of those newly infected individuals \(\text{inf}_k(t)\) in the list \(L_k(t)\) depends on both on the infecting individual \(k\) and on the time slot \(t\). The individuals within the list \((14)\) have identical structure to the class described in \((1)\) and have the format:

\[
I_l = (x_l, y_l, \text{tinf}_l, \text{det}_l, \text{st}_l, \text{age}_l, \text{rad}, p_{\text{cont}(l)}, \text{cont}_l, R_l, \text{stg}_l, \text{sev}_l, \text{dur}_l, \text{st}_l(t))
\]

where \(l = 1, 2, \ldots, \text{inf}_k(t)\). The procedures for the \(L_k(t)\) simulation is outlined as follows:

- The amount of infected contacts \(\text{inf}_k(t)\) for every individual \(\tilde{Q}_k\) for each day \(t\) is generated based on the set of rules and assumptions outlined in details in the definition of \(\text{inf}_k(t)\) for classes \((1)\).

- For every infected contact index \(l = 1, 2, \ldots, \text{inf}_k(t)\) the individual class \((15)\) is simulated and denoted as \(I_l\). The new individuals defined by \((15)\) are simulated in the same way as the reported cases \(\tilde{Q}_k\) that are introduced into the model in the beginning. The only difference is the simulated locations of the new individuals which are generated in relation to the new center \(M_k(x_k, y_k)\) defined by \(\tilde{Q}_k\) rather than the center \(C_i\) of some epicenter \(E_i\) for \(i \in \{1, 2, \ldots, T\}\) to which the individual \(Q_k\) belongs to.

- The parameter \(p_{\text{cont}(l)}\) for individual \(I_l\) defines the probability of any infection-induced contacts for each day when the quarantine measures are or are not implemented. In the first case the values \(p_{\text{cont}(l)}\) are generated from a log-normal distribution \((\mathcal{LN})\) with parameters \((\mu_{\text{ocp}}, \sigma^2_{\text{ocp}})\) in the same was as for the index cases \(\tilde{Q}_k\) and the results are rounded to integer values. The subscript \(\text{ocp}\) stands for the \(\text{Ordinary Contact Probability}\) parameter. In the second case the values \(p_{\text{cont}(l)}\) are generated from a log-normal distribution \((\mathcal{LN})\) with parameters \((\mu_{\text{icp}}, \sigma^2_{\text{icp}})\) and the results are rounded to integer values. The subscript \(\text{icp}\) stands for the \(\text{Isolated Contact Probability}\) parameter. The shift from the \(\text{Ordinary Contact Probability}\) to the \(\text{Isolated Contact Probability}\) parameter is controlled by the \(\text{Quarantine In}\) time point. The shift in the distribution of parameters for \(p_{\text{cont}(l)}\) from the \(\text{ocp}\) to \(\text{icp}\) set happens at time index \(\text{Quarantine In}\) i.e. when \(t > \text{Quarantine In}\).

- For the newly simulated infected individuals \(I_l\) for all \(l = 1, 2, \ldots, \text{inf}_k(t)\) the infection time is assigned to \(t\) i.e. \(\text{tinf}_l = t\) and the status changes at time \(t + 1\) i.e. \(\text{st}_l(t + 1) = 1\).

The simulation of infected individuals defined in \((15)\) is implemented sequentially for every time slot \(t \in [T_{\text{min}}; T_{\text{max}}]\) and the modeled individuals at time \(t\) are combined in the overall list \(\mathcal{L}(t)\). The overall list \(\mathcal{L}(t)\) contains all of the detected individuals \(\tilde{Q}_k\) across all local epicenters \(i = 1, 2, \ldots, T\) that are incorporated into the model from the data as well as the lists of newly infected individuals \(L_k(t)\) that appear during the simulation process steps. For the smallest times slot \(t = T_{\text{min}}\) it is assumed that \(\mathcal{L}(t) = \mathcal{L}(T_{\text{min}})\) contains only the earliest model input cases from the list \(\mathcal{L}(t) = \mathcal{L}(T_{\text{min}})\). The overall lists for subsequent time slots \(\mathcal{L}(t + 1)\) are formed sequentially from the lists \(\mathcal{L}(t)\) for \(t \in [T_{\text{min}}; (T_{\text{max}} - 1)]\) via the following two scenarios:
• The first scenario corresponds to the time slots $t \leq 0$ i.e. before the selected baseline simulation time. In this case both the reported individuals $D(t)$ and the lists $L_k(t)$ of individuals infected during time $t$ by all infectious individual indexed by all $k$-s are incorporated into the simulation for the next time slot $t + 1$.

  - If the time index $t \leq 0$ i.e. before the selected baseline simulation time then the newly generated list $L(t + 1)$ is initially formed from the previous list $L(t)$ and expanded by adding the detected individuals from the reported data that correspond to time $D(t + 1)$ across all epicenter i.e. with the list

$$D(t + 1) = \bigcup_{i=1}^{I} D_i(t + 1)$$

where the number of individuals in each epicenter $i$ is equal to $\Delta n_i(t + 1)$.

  - All individuals within the list $L(t)$ that are indexed by $k$ are checked for the infectiousness condition $t \in T_{inf}(k)$. If the condition holds then the list of individual $L_k(t)$ that is infected by individual $k$ at time $t$ is added to the list $L(t + 1)$.

  - If for the individual indexed by $k$ from $L(t)$ and for the current time $t$ the condition $tinf_k + det_k < t$ holds, the individual is still moved to the list $L(t + 1)$. The individual's status, however, is changing to quarantined (i.e. $st_k(t + 1) = 2$) and the given individual stops infecting the others.

  - If for the individual indexed by $k$ from $L(t)$ and for the current time $t$ the condition $tinf_k + det_k + dur_k < t$ holds, the individual is still moved to the list $L(t + 1)$. The individual's status, however, is changing to either recovered (i.e. $st_k(t + 1) = 3$) if the individual's severity variable $sever_k = 2$ or 3, or deceased (i.e. $st_k(t + 1) = 4$) if the individuals severity variable $sever_k = 1$.

• The second scenario corresponds to the time slots $t > 0$ i.e. after the selected baseline simulation time. In this case only the lists $L_k(t)$ of individuals infected during time $t$ by the infectious individual with index $k$ are incorporated into the simulation for the next time slot $t + 1$.

The Overall Model Flow

The entire modeling process can be summarized via the following steps:

• The model input time interval is determined by fixing the first $\tilde{S}$ reporting indexes out of the total $S$ where $\tilde{S} < S$. Those indexes correspond to the reporting time slots $\tilde{t}_1, \tilde{t}_2, \ldots, \tilde{t}_{\tilde{S}}$. This completely defines the list of reported index cases $D$ from (4) that are used as the model initial conditions. The baseline time of the model $t = 0$ is assumed to correspond to $\tilde{t}_{\tilde{S}}$.

• The individuals from the reported set $D$ that are defined in (4) are assigned to the local epicenters of the future epidemic $E_i$ for $i = 1, 2, \ldots, I$ based on the available (from the input data) geographic distribution.

• The geographic data about the areas of high density $P_j$ for $j = 1, 2, \ldots, J$ are incorporated into the model.

• The model is initialized with the index cases from $D$. Based on those index cases that are defined in (4) the initial infection time $T_{min}$ is determined. This step is necessary to incorporate the infection times that have been present before the first reporting time $\tilde{t}_1$ into the model.
• The final time point of the stochastic simulations $T_{max}$ is defined by the user based on the desired study and prediction goals.

• The initial list of infected individuals $L(T_{min})$ is initialized at time $T_{min}$ only with the earliest model input cases from the list $D(T_{min})$.

• The infected list of individuals $L(t + 1)$ for the time slot $t + 1$ is generated sequentially for all $t \in [T_{min}; T_{max} - 1]$ based on the list of individuals $L(t)$ from previous time slot $t$ as described in the new infection generation process and the individual’s characteristics within the list $L(t + 1)$ are updated at this time step $t + 1$.

Based on the lists $L(t)$ at every time slot $t \in [T_{min}; T_{max}]$ the infected modeled population summaries can be computed and summarized. In particular, the total number of currently infected but not identified individuals (i.e. those with the status $st(t) = 1$) is saved into $Inf(t)$ variable for every $t$. The total number of treated or quarantined individuals (i.e. with the status $st(t) = 2$) is saved into $Treat(t)$ variable for every $t$. The total number of recovered individuals (i.e. with the status $st(t) = 3$) is saved into $Recov(t)$ variable for every $t$. The total number of deceased individuals (i.e. with the status $st(t) = 4$) is saved into $Dead(t)$ variable for every $t$. Those numbers are used in the model calibration procedures, epidemiological summaries and in the model predictions. The model input utilizes only the first $S$ reported indexes with the corresponding reported times $\tilde{t}_s$ for $s = 1, 2, \ldots, S$ with the total number of reported indexes equal to $S$ and $\tilde{S} < S$. The remaining reported indexes $\tilde{S} + \tau, \tilde{S} + \tau + 1, \ldots, \tilde{S}$ for some integer $\tau \geq 0$ are divided into the two groups:

$$\left\{ \tilde{S} + \tau, \tilde{S} + \tau + 1, \ldots, \tilde{S} \right\}$$

$$\left\{ \tilde{S} + 1, \tilde{S} + 2, \ldots, \tilde{S} \right\}.$$  \hspace{1cm} (17)

The first group of the reported indexes from (17) is used for the model calibration and estimation of the unknown parameters. The second group of the reported indexes from (17) is used to evaluate the quality of the model predictions. The Massachusetts surveillance data that are used for the model calibration, validation and predictions are freely available at the Massachusetts Department of Public Health web site [6]. The first reported date which corresponds to the time index $\tilde{t}_1$ in the model is March 13, 2020. The latest reported date that is used for the model input is March 26, 2020 which corresponds to the time index $\tilde{t}_S$ in the model. The time indexes that correspond to $\tilde{t}_{S+\tau}$ and $\tilde{t}_S$ are April 14, 2020 and April 22, 2020 respectively. The parameters optimization is performed by minimizing the sum of squared differences between the model-produced outputs and the calibration data by using the Nelder–Mead numerical minimization method [7].

The Model Calibration Details

The set of global parameters denoted as $HP$ that the model is optimized over is presented below:

$$HP = \left\{ \text{Detect, Ordinary, Contact, Prob, Isolated, Contact, Prob, Usual, cont, Dense, cont} \right\}.$$  

Based on the model-produced lists $L(t)$ at every time slot $t$ the model-produced summaries of the generated infected population can be computed and summarized for a given set of parameters from $HP$ which includes the time series of $Inf(t)$, $Treat(t)$, $Recov(t)$ and $Dead(t)$. In the proposed model formulation the set of global parameters $HP$ is estimated based on the $Total(t) = Treat(t) + Recov(t) + Dead(t)$ information about the infected individuals (i.e. those with the statuses $st(t) = 2, 3, 4$) time series produced by the model for time slot $t \in [T_{init}; T_{term}]$. The time slots between $T_{init}$ and
The objective function $\text{ObjFunc}(\text{HP})$ for minimization across the set of parameters from $\text{HP}$ is the standardized sum of squared differences between the model produced outputs and the reported data time series:

$$\text{ObjFunc}(\text{HP}) = \sqrt{\frac{1}{T_{\text{term}} - T_{\text{init}}} \sum_{t=T_{\text{init}}}^{T_{\text{term}}} \left[ \text{Total}(t) - \text{TotalAct}(t) \right]^2}.$$  (18)

The Nelder–Mead [7] numerical minimization method for multiple dimensions has been used to minimize $\text{ObjFunc}(\text{HP})$ function from (18) with respect to $\text{HP}$. The initial values for the algorithm $\text{HP}_0$ are selected from the plausible ranges of parameters of interest which provide a good visual alignment between the collected data and the model produced output.

### The model availability

The model has been implemented in multiple environments which include the application tool for Microsoft Windows [8] and the web prediction tool [9] [10] (summaries only). The model application tool for Microsoft Windows is freely available under the terms of the MIT license [11]. The model parameters can either be changed manually by the user within the parameters table that is used by the tool or, more conveniently, within the tool graphic user interface (GUI). The example of the parameter’s table is presented in Fig. A and the tool source code, the application, and the relevant tool documentation are available on GitHub [8].

The current model tool has been calibrated based on the state of Massachusetts (United States) incidence data [6]. In addition to that the user has an option to adjust interactively the tool parameters which include, in particular, the transmission parameters and the quarantine implementation dates(s). Overall, the proposed framework and the code are fairly general and can be adopted for other areas and territories where the demographics of the incidence cases and population characteristics are known with at least some geographic precision, and where the rapid evaluations of social distancing measures have to be quantified.
Fig A. The illustration of the tool configuration file. The configuration file parameter values are loaded by the tool when the tool is initially launched. Then the user has an option to use the loaded parameter's values, to adjusted them manually or to use the tool estimation procedures.
References


2. worldpopulationreview.com;
   https://worldpopulationreview.com/states/maasachusetts-population/


10. tableau.com; https://public.tableau.com/profile/quantori#!/vizhome/ModelingCovidMassachusetts_Map/InteractiveMap

11. MIT Licence; https://opensource.org/licenses/MIT