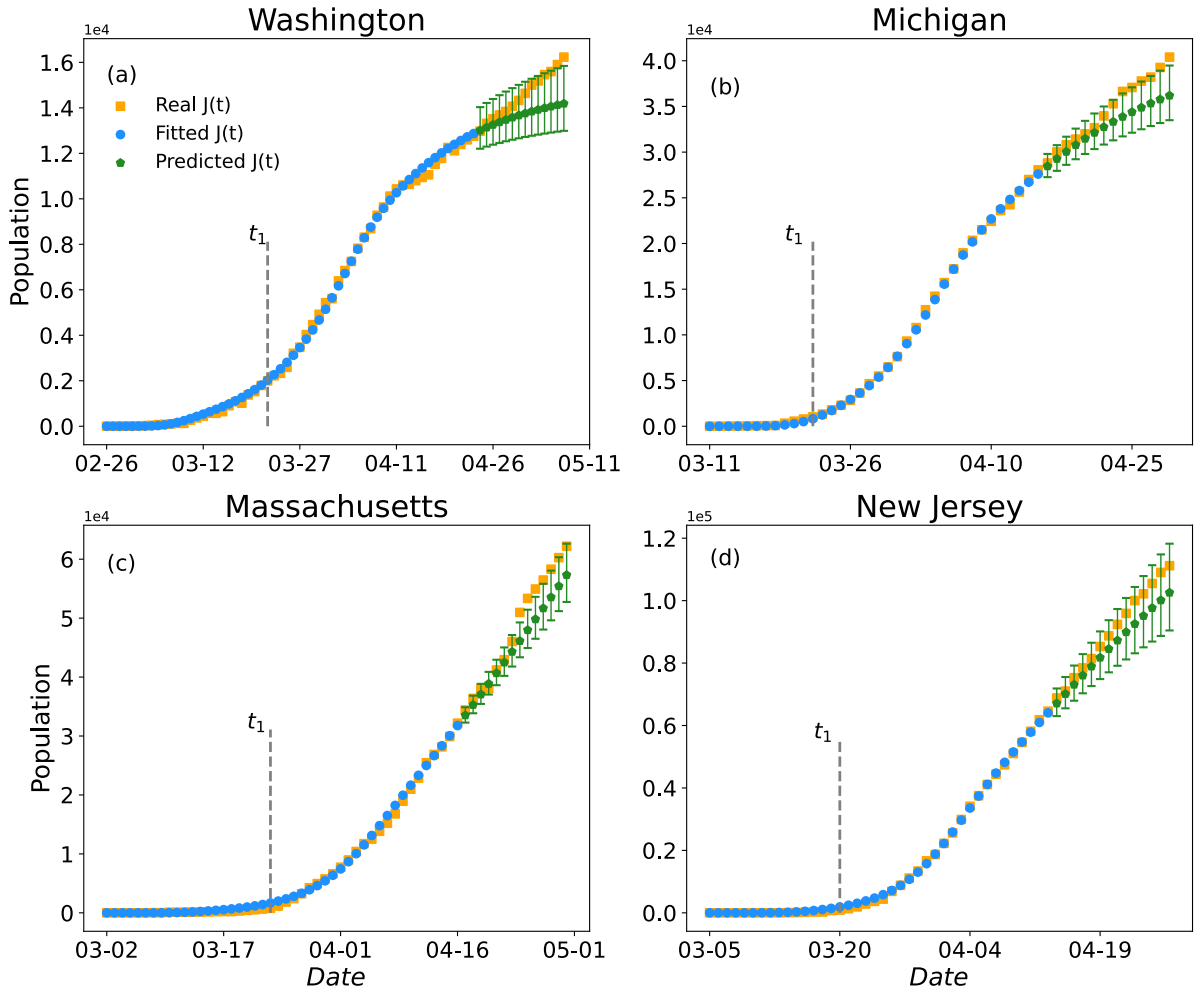
Supplementary Information for

When did COVID-19 start? - Optimal inference of time ZERO

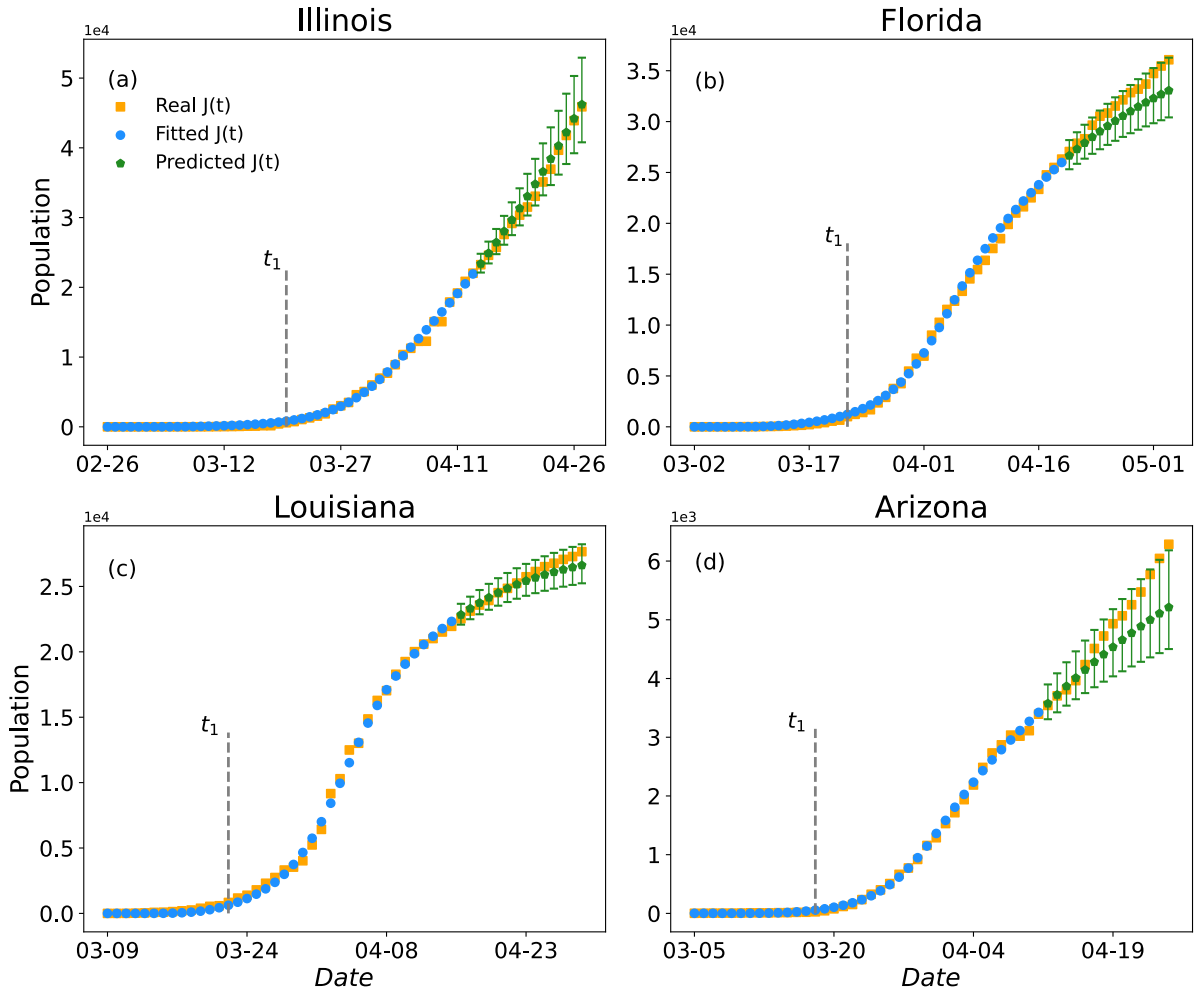
Zheng-Meng Zhai, Yong-Shang Long, Ming Tang*, Zonghua Liu, and Ying-Cheng Lai*

Contents

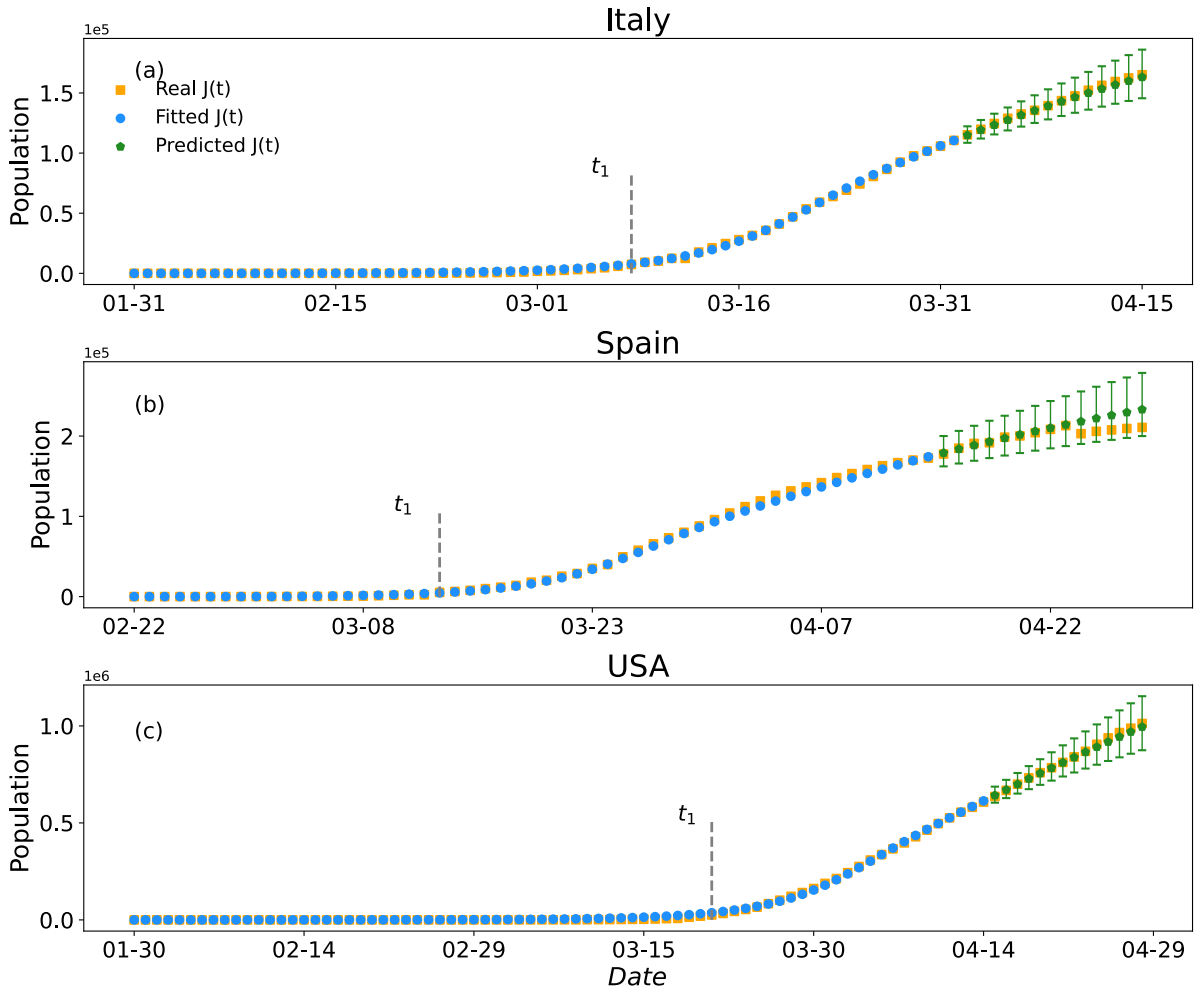
Supplementary Figures



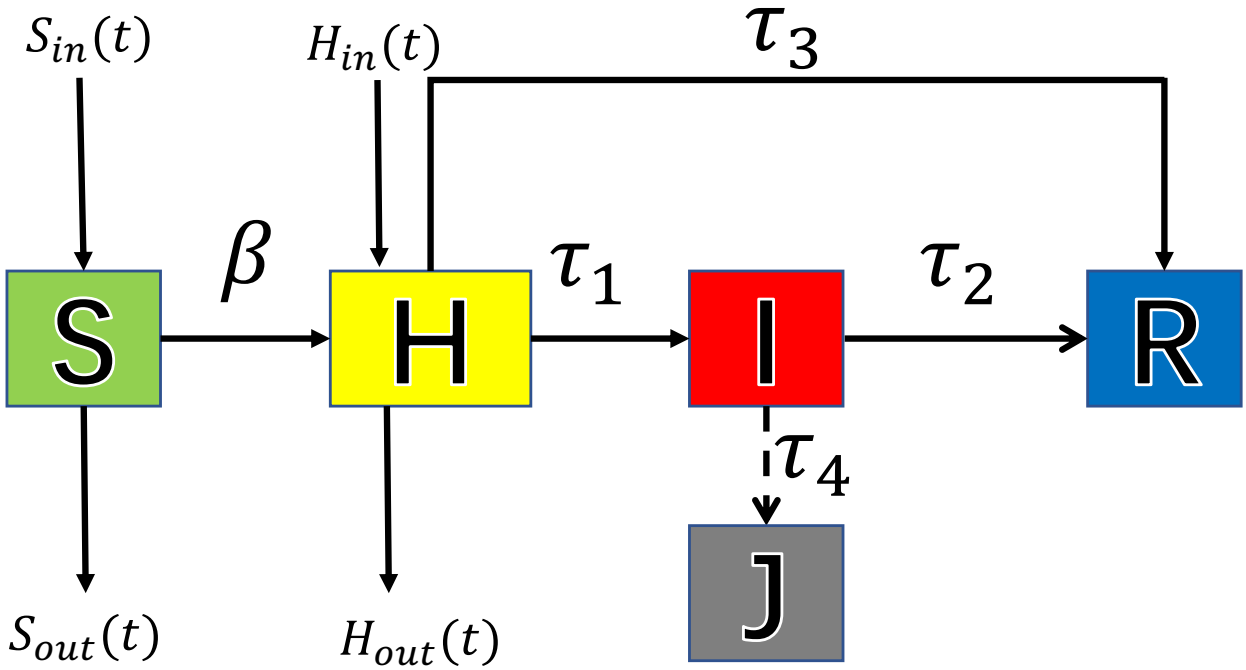
Supplementary Figure S1. Demonstration of the predictive power of the non-Markovian SHIJR model for COVID-19. Shown are the results of $J(t)$, the daily accumulative number of confirmed cases for four systems: (a) Washington State treated as an open system, (b) Michigan as an open system, (c) Massachusetts as an open system, and (d) New Jersey as an open system. The orange squares are the real time series $J(t)$. The blue dots in the first phase are the model fitted $J(t)$, in which the four key model parameters are estimated based on the real data in this sub-time interval. The green pentagons in the last 13 days of the whole time interval are the predicted $J(t)$, whose agreement with the real data attests to the predictive power of the model. The estimated parameters are: (a) $(\beta, H(t_0), \eta) = (0.16, 850, 0.6)$ and $\lambda \in [0.123, 0.157]$; (b) $(\beta, H(t_0), \eta) = (0.20, 7600, 0.8)$ and $\lambda \in [0.168, 0.193]$; (c) $(\beta, H(t_0), \eta) = (0.2, 1120, 0.75)$ and $\lambda \in [0.103, 0.117]$; (d) $(\beta, H(t_0), \eta) = (0.26, 1750, 0.55)$ and $\lambda \in [0.128, 0.153]$. The range of variations in the estimated value of λ is used to generate the green error bars in the predicted $J(t)$.



Supplementary Figure S2. Demonstration of the predictive power of the non-Markovian SHIJR model for COVID-19. Shown are the results of $J(t)$, the daily accumulative number of confirmed cases for four systems: (a) Illinois treated as an open system, (b) Florida as an open system, (c) Louisiana as an open system, and (d) Arizona as an open system. The estimated parameters are: (a) $(\beta, H(t_0), \eta) = (0.23, 180, 0.60)$ and $\lambda \in [0.092, 0.108]$; (b) $(\beta, H(t_0), \eta) = (0.23, 440, 0.55)$ and $\lambda \in [0.150, 0.170]$; (c) $(\beta, H(t_0), \eta) = (0.25, 1850, 0.75)$ and $\lambda \in [0.239, 0.261]$; (d) $(\beta, H(t_0), \eta) = (0.28, 120, 0.7)$ and $\lambda \in [0.174, 0.206]$. Legends are the same as those of Fig. S1.



Supplementary Figure S3. Demonstration of the predictive power of the non-Markovian SHIJR model for COVID-19. Shown are the results of $J(t)$, the daily accumulative number of confirmed cases for four systems: (a) Italy treated as a closed system, (b) Spain as a closed system, and (c) USA as a closed system. The estimated parameters are: (a) $(\beta, H(t_0), \eta) = (0.20, 330, 0.70)$ and $\lambda \in [0.127, 0.153]$; (b) $(\beta, H(t_0), \eta) = (0.25, 1400, 0.55)$ and $\lambda \in [0.138, 0.162]$; (c) $(\beta, H(t_0), \eta) = (0.22, 90, 0.55)$ and $\lambda \in [0.108, 0.132]$. Legends are the same as those of Fig. S1.



Supplementary Figure S4. Generalized COVID-19 model for an open system. Individuals in the S state are infected by H individuals at the rate β and switch to the H state. A fraction η of H-state individuals recover spontaneously or die, a process that requires τ_3 days. The parameter η thus represents the fraction of undocumented infections, and its value is determined by the testing capability of the country or State. The remaining $(1-\eta)$ fraction of H individuals go through a transition to the I state after an average incubation period of τ_1 days - a typical non-Markovian process. With medical treatment, individuals in the I state recover or die after τ_2 days. Finally, a time delay exists for the transition from I to J: on average the I individuals will need τ_4 days to be confirmed. The four quantities $S_{in}(t)$, $S_{out}(t)$, $H_{in}(t)$, and $H_{out}(t)$ represent the populations moving into and out of the S and H states of the open system.

Supplementary Tables

Supplementary Table 1. Additional information and results for ten States in the US, New York city, Italy, Spain, and the United Kingdom. The quantity t_{WHO} is the range of dates of first emergence of case(s) reported by the World Health Organization (WHO). Columns 3-6 list the confidence intervals of the four estimated parameters required of the non-Markovian SHIJR spreading model under government imposed control measures: β, $H(t_0)$, η, and λ. Abbreviations: It - Italy, WA - Washington State, NYC - New York City, NY - New York State, UK - United Kingdom, MI - Michigan, SP - Spain, MA - Massachusetts, CA - California, NJ - New Jersey, IL - Illinois, FL - Florida, LA - Louisiana, AZ - Arizona.

System	t_{WHO}	$(\beta_{low}, \beta_{upp})$	$(H_{low}(t_0), H_{upp}(t_0))$	(η_{low}, η_{upp})	$(\lambda_{low}, \lambda_{upp})$
It	1/31-2/20	(0.191,0.208)	(243,491)	(0.625,0.781)	(0.127,0.153)
USA	1/22-1/25	(0.213,0.227)	(72,115)	(0.497,0.607)	0.108,0.132)
WA	1/22-2/28	(0.156,0.164)	(749,966)	(0.560,0.639)	(0.123,0.157)
NY	3/2-3/3	(0.184,0.196)	(10220,14382)	(0.552,0.653)	(0.111,0.129)
NYC	3/2-3/4	(0.182,0.198)	(5015,7405)	(0.488,0.617)	(0.098,0.122)
UK	1/31-2/7	(0.193,0.207)	(727,1084)	(0.544,0.655)	(0.088,0.112)
MI	3/11-3/12	(0.194,0.206)	(5734,11486)	(0.742,0.865)	(0.168,0.193)
SP	2/1-2/24	(0.243,0.257)	(1192,1657)	(0.514,0.587)	(0.138,0.162)
MA	2/1-3/5	(0.190,0.207)	(820,2026)	(0.678,0.852)	(0.103,0.117)
CA	1/26-2/2	(0.210,0.229)	(265,500)	(0.574,0.735)	(0.100,0.120)
NJ	3/5-3/6	(0.248,0.272)	(1327,2443)	(0.472,0.643)	(0.128,0.153)
IL	1/25-3/1	(0.220,0.240)	(134,260)	(0.518,0.694)	(0.092,0.108)
FL	3/2-3/4	(0.224,0.237)	(369,523)	(0.503,0.597)	(0.150,0.170)
LA	3/9-3/10	(0.241,0.258)	(1381,3204)	(0.680,0.844)	(0.239,0.261)
AZ	2/1-3/6	(0.263,0.296)	(76,241)	(0.590,0.826)	(0.174,0.206)

Supplementary Table 2. The dates on which travel restriction orders were issued in the ten States studied. For each state, the date is effectively one on which an exponential decay in intrastate traffic begins.

State	Date
Arizona	March 19, 2020
Washington	March 23, 2020
New York	March 22, 2020
New Jersey	March 21, 2020
California	March 19, 2020
Michigan	March 23, 2020
Florida	March 23, 2020
Illinois	March 21, 2020
Massachusetts	March 24, 2020
Louisiana	March 23, 2020

Supplementary Table 3. Average daily outbound and inbound populations and the total population for the ten States studied

State	Outbound	Inbound	Population
Arizona	13,679	6,515	7,278,717
Washington	27,339	15,495	7,614,893
New York	55,012	133,095	19,453,561
New Jersey	133,507	71,056	8,882,190
California	21,171	19,979	39,512,223
Michigan	21,179	11,625	9,986,857
Florida	28,871	22,111	21,477,737
Illinois	48,160	45,451	12,671,821
Massachusetts	30,217	44,987	6,892,503
Louisiana	11,754	14,994	4,648,794

Supplementary Note: Construction of non-Markovian SHIJR epidemic model

Figure S4 illustrates the generalized model. An individual can be in one of the five states at each time step: susceptible (S), hidden (H), infected (I), confirmed and isolated (J), and removed (R). The states S, I, and R have the same meanings as in the classical SIR model for infectious disease, but states H and J are unique for COVID-19. In particular, an individual in H has had the virus and is infectious but is asymptomatic or only mildly symptomatic, in contrast to the I state in which individuals show symptoms. The H individuals are capable of infecting others. The J state contains individuals who are confirmed with COVID-19. Individuals in the R state, by definition, are not infectious.

We treat each State in the US as an open system, regarding the influences from all the other States as perturbations, mathematically represented by the populations moving into and out of the S and H states, denoted as $S_{in}(t)$, $S_{out}(t)$, $H_{in}(t)$, and $H_{out}(t)$, respectively, as shown in Fig. S4. These functions are determined by the travel intensity as a function of time. Two types of travel need to be distinguished: interstate and intrastate, with the corresponding intensity functions $l_{inter}(t)$ and $l_{intra}(t)$. Due to government imposed travel restrictions, these functions decay exponentially from an initial value to a final smaller constant value. For instance, a recent estimate [1] has given that, for several major US cities, the travel restrictions would reduce the outbound human movements by 50%. In general, we have $l_{inter}(t) = 1$, $e^{-\lambda_{inter}(t-t_c)}$, and $e^{-\lambda_{inter}(t_s-t_c)}$ for $t < t_c$, $t_c \le t \le t_s$, and $t > t_s$, respectively, and $l_{intra}(t) = 1$, $e^{-\lambda_{intra}(t-t_c^i)}$, and $e^{-\lambda_{intra}(t_s^i-t_c^i)}$ for $t < t_c^i$, $t_c^i \le t \le t_s^i$, and $t > t_s^i$, respectively. For simplicity, we set $\lambda_{inter} = \lambda_{intra}$. The quantities t_c and t_c^i are, respectively, the starting dates of interstate and intrastate travel restrictions and the exponential decay in the movement activities occurs between t_c and t_s or between t_c^i and t_s^i. The starting dates differ from State to State. The values of t_s and t_s^i are set as $t_c + 7$ and $t_c^i + 7$, respectively. Our generalized SHIJR model for COVID-19 epidemic for any given target State in the US can be described by the following set of delayed integro-differential equations:

$$\begin{aligned}
\frac{dS(t)}{dt} &= -H^n(t) - S_{out}(t) + S_{in}(t), && (1)\\
\Phi(t) &= H^n(t) - H_{out}(t) + H_{in}(t), && (2)\\
\frac{dH(t)}{dt} &= \Phi(t) - (1-\eta)\int_{t_0}^{t} f_1(\tau)\Phi(t-\tau)d\tau - \eta\int_{t_0}^{t} f_3(\tau)\Phi(t-\tau)d\tau \\
&\quad -(1-\eta)f_1(t)H(t_0) - \eta f_3(t)H(t_0), \\
\frac{dI(t)}{dt} &= (1-\eta)\int_{t_0}^{t} f_1(\tau)\Phi(t-\tau)d\tau && (3)\\
&\quad -(1-\eta)\int_{t_0}^{t} f_2(\tau')d\tau' \int_{t_0}^{t-\tau'} f_1(\tau)\Phi(t-\tau'-\tau)d\tau_1
\end{aligned}$$

$$+(1-\eta)f_1(t)H(t_0) - (1-\eta)\int_{t_0}^{t} f_2(\tau)f_1(t-\tau)H(t_0)d\tau - f_2(t)I(0), \quad (4)$$

$$\frac{dR(t)}{dt} = \eta\int_{t_0}^{t} f_3(\tau)\Phi(t-\tau)d\tau \quad (5)$$
$$+(1-\eta)\int_{t_0}^{t} f_2(\tau')d\tau' \int_{t_0}^{t-\tau'} f_1(\tau)\Phi(t-\tau'-\tau)d\tau_1$$
$$+\eta f_3(t)H(t_0) + (1-\eta)\int_{t_0}^{t} f_2(\tau)f_1(t-\tau)H(t_0)d\tau + f_2(t)I(0)$$
$$-R_{out}(t) + R_{in}(t)$$

$$\frac{dJ(t)}{dt} = (1-\eta)\int_{t_0}^{t} f_4(\tau')d\tau' \int_{t_0}^{t-\tau'} f_1(\tau)\Phi(t-\tau'-\tau)d\tau \quad (6)$$
$$+(1-\eta)\int_{t_0}^{t} f_4(\tau)f_1(t-\tau)H(t_0)d\tau,$$

$$\frac{dN}{dt} = (F_{in} - F_{out})l_{inter}(t), \quad (7)$$

where the quantity $H^n(t)$ in the first two equations is the rate of increase in the H-state population: $H^n(t) = \beta S(t) l_{intra} H(t)/N(t)$ with $l_{intra}(t)H(t)$ representing the active H-state population that has not been isolated, $f_1(\tau)$, $f_2(\tau)$, $f_3(\tau)$, and $f_4(\tau)$ are the normal probability distribution functions of the delay time τ_1, τ_2, τ_3, and τ_4, respectively, and $N(t)$ is the population of the State as a function of time. The quantity $\Phi(t)$ is the increment of the H-state population. The input and output functions are

$$S_{in}(t) = F_{in} l_{inter}(t) \frac{S_{total}(t)}{S_{total}(t) + H_{total}(t) + R_{total}(t)}, \quad (8)$$

$$S_{out}(t) = F_{out} l_{inter}(t) \frac{S(t)}{S(t) + H(t) + R(t)}, \quad (9)$$

$$H_{in}(t) = F_{in} l_{inter}(t) \frac{H_{total}(t)}{S_{total}(t) + H_{total}(t) + R_{total}(t)}, \quad (10)$$

$$H_{out}(t) = F_{out} l_{inter}(t) \frac{H(t)}{S(t) + H(t) + R(t)}, \quad (11)$$

$$R_{in}(t) = F_{in} l_{inter}(t) \frac{R_{total}(t)}{S_{total}(t) + H_{total}(t) + R_{total}(t)}, \quad (12)$$

$$R_{out}(t) = F_{out} l_{inter}(t) \frac{S(t)}{S(t) + H(t) + R(t)}, \quad (13)$$

where the quantities $S_{total}(t)$, $H_{total}(t)$ and $R_{total}(t)$ are the total S, H and R-state populations of the US excluding the target State, $F_{in}(t)$ and $F_{out}(t)$ are the fluxes into and out of the State, which can be extrapolated from empirical data. To numerically solve the whole set of equations, the values of the initial H-state population, $H(t_0)$, and of the infection rate β are needed, which can be estimated with a mathematical optimization procedure. In particular, we look for that optimal parameter

combination $\hat{\Theta} = (\beta,\, H(t_0),\, \eta,\, \lambda)$ that minimizes the weighted difference squared between the data points $J(t_i)$ and the predicted values of $f(t_i, \Theta)$ ($i = 0, 1, \ldots, n-1$), according to

$$\hat{\Theta} = \text{argmin} \sum_{i=1}^{n} w_{t_i} [f(t_i, \Theta) - J_{t_i}]^2. \tag{14}$$

Because of the necessity of assigning larger weights for more recent data, we set the weights to be $w_{t_{n-i}} = \alpha(1-\alpha)^{i-1}$ with $\alpha = 0.1$. The optimization problem can be solved to yield the optimal values β^*, $H^*(t_0)$, η^* and λ^* using, e.g., the Levenberg-Marquardt (LM) method [2–5].

To determine the daily fluxes $F_{in}(t)$ and $F_{out}(t)$, we will use the commuting data from the US Census Bureau (https://www.census.gov/topics/employment/commuting.html), which were obtained from sampling the home and work addresses of the working population in the five-year period (2011-2015). Our estimation method is as follows. Assume that the commuting population is distributed uniformly among the States. On average, each working individual commutes 0.11 time per day. Multiplying this number by the population of the State gives the daily average number of people who commute. Denoting this number by Q and letting the commuting populations from the Census Bureau's data base be $P_{in/out}$, we obtain the ratio $C_{in/out} = Q/P_{in/out}$. Let $D_{in/out}$ be populations in and out of the State from the data base. The daily fluxes can be obtained as $F_{in/out} = C_{in/out} \cdot D_{in/out}$.

For the ten States studied, the dates t_c^i are listed in Supplementary Table 2 and the flux values are listed in Supplementary Table 3

Supplementary References

[1] Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* (2020).

[2] Levenberg, K. A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.* **2**, 164–168 (1944).

[3] Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Indus. Appl. Math.* **11**, 431–441 (1963).

[4] Chowell, G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infect. Dise. Model.* **2**, 379–398 (2017).

[5] Kaltenbacher, B., Neubauer, A. & Scherzer, O. *Iterative Regularization Methods for Nonlinear Ill-Posed Problems*, vol. 6 (Walter de Gruyter, 2008).