

Supplementary Information

Multiscale mobility networks and the large scale spreading of infectious diseases

Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J. Ramasco & Alessandro Vespignani

1 Databases

1.1 World Airport Network

The World Airport Network (WAN) is composed of 3362 commercial airports indexed by the International Air Transport Association (IATA) that are located in 220 different countries. The database contains the number of available seats per year for each direct connection between two of these airports. The coverage of the dataset is estimated to be 99% of the global commercial traffic. The WAN can be seen as a weighted graph comprising 16 846 edges whose weight, $\omega_{j\ell}$, represents the passenger flow between airports j and ℓ . The network shows a high degree of heterogeneity both in the number of destinations per airport and in the number of passengers per connection [1, 2, 3, 4].

1.2 Global Population and its Allocation

The population dataset was obtained from the Web sites of the "Gridded Population of the World" and the "Global Urban-Rural Mapping" projects [5, 6], which are run by the Socioeconomic Data and Application Center (SEDAC) of Columbia University. The surface of the world is divided into a grid of cells that can have different resolution levels. Each of these cells has been assigned an estimated population value.

Out of the possible resolutions, we have opted for cells of 15×15 minutes of arc to constitute the basis of our model. This corresponds to an area of each cell approximately equivalent to a rectangle of 25×25 kms along the Equator. The dataset comprises 823 680 cells, of which 250 206 are populated. Since the coordinates of each cell center and those of the WAN airports are known, the distance between the cells and the airports can be calculated. We have performed a Voronoi-like tessellation of the Earth surface assigning each cell to the closest airport that satisfies the following two conditions: (i) Each cell is assigned to the closest airport within the same country. And (ii), the distance between the airport and the cell cannot be longer than 200 kms. This cutoff naturally emerges from the distribution of distances between cells and closest airports, and it is introduced to avoid that in barely populated areas such as Siberia we can generate geographical census areas thousands of kilometer wide but with almost no population. It also corresponds to a reasonable

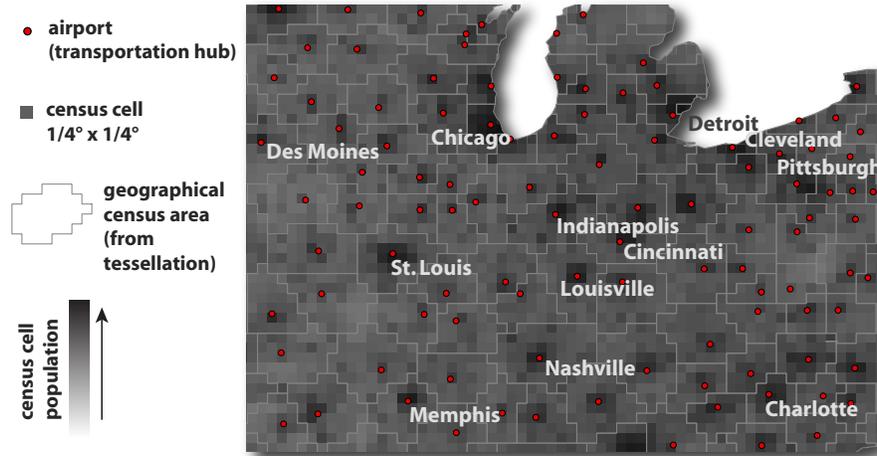


Figure 1: Population database and Voronoi tessellation around main transportation hubs. The world surface is represented in a grid-like partition where each cell - corresponding to a population values - is assigned to the closest airport. Geographical census areas emerge that constitute the sub-populations of the meta-population model.

upper cutoff for the ground traveling distance expected to be covered to reach an airport before traveling by plane.

Before proceeding with the tessellation, we need to take into account that some urban areas include more than one airport. For instance, London has up to six airports, Paris has two, and New York City has three. Our aim is to build a metapopulation model whose subpopulations correspond to the geographical census areas obtained from tessellation. Inside these geographical census areas a homogeneous mixing is assumed. The groups of airports that serve the same urban area need therefore to be aggregated since the mixing within the given urban area is expected to be high and cannot be represented in terms of separated subpopulations for each of the airports serving the same city. We have searched for groups of airports located close to each other and we manually processed the identified groups of airports to select those belonging to the same urban area. The airports of the same group are then aggregated in a single "super-hub". An example with the final result of the Voronoi tessellation procedure with cells and airports can be seen in Figure 1. The geographical census areas become thus the basic subpopulations of our metapopulation model. Their connections will determine the geographical spreading of an hypothetical epidemic. The air transportation is already integrated in the model, but a further step must be taken in order to also include ground transportation in a realistic way.

1.3 Commuting Networks

Our commuting databases have been collected from the Offices of Statistics of 29 countries in 5 continents (out of the 6 continents – Europe, North America, Latin America, Asia, Oceania,

Table 1: Commuting networks in each continent. Number of countries (N_c), number of administrative units (V) and inter-links between them (E) are summarized.

Continent	N_c	V	E
Europe	17	65880	4490650
North America	2	6986	182255
Latin America	5	4301	102117
Asia	3	2732	323815
Oceania	2	746	30679
Total	29	80645	5129516

Africa). The full dataset comprehends more than 80 000 administrative regions and over five million commuting flow connections between them (see Tables 1 and 2 for details). The definition of administrative unit and the granularity level at which the commuting data are provided enormously vary from country to country. For example, most European countries adhere to a practice that ranks administrative divisions in terms of geocoding for statistical purposes, the so called Nomenclature of Territorial Units for Statistics (NUTS). Most countries in the European Union are partitioned into three NUTS levels which usually range from states to provinces. The commuting data at this level of resolution is therefore strongly coarse-grained. In order to have a higher geographical resolution of the commuting datasets that could match the resolution scale of our geographical census areas, we looked for smaller local administrative units (LAU) in Europe. The US or Canada report commuting at the level of counties. However, even within a single country the actual extension, shape, and population of the administrative divisions are usually a consequence of historical reasons and can be strongly heterogeneous.

Such heterogeneity renders the efforts to define a universal law describing commuting flows likely to fail. The mobility behavior might indeed result different across countries simply due to the country specific partition of the population into administrative boundaries. In order to overcome this problem, and in particular to define a data/driven short range commuting for GLEaM, we used the geographical census areas obtained from the Voronoi tessellation as the elementary units to define the centers of gravity for the process of commuting. This allows to deal with self-similar units across the world with respect to mobility as emerged from a tessellation around main hubs of mobility and not country specific administrative boundaries.

We have mapped the different levels of commuting data into the geographical census areas formed by the Voronoi-like tessellation procedure described above. The mapped commuting flows can be seen as a second transport network connecting subpopulations that are geographically close. This second network can be overlaid to the WAN in a multi-scale fashion to simulate realistic scenarios for disease spreading. The statistical properties of the commuting network at the level of the geographical census areas are reported in Figure 2. The network exhibits important variability in the number of commuters on each connection as well as in the total number of commuters per geographical census area. Since the census areas are relatively homogeneous and

Table 2: Commuting networks in each country. Number of administrative units (V) and inter-links between them (E) are summarized.

Continent	Country	V	E
Europe	Austria	99	1886
	Belgium	589	71528
	Denmark	248	20990
	Finland	348	22484
	France	36602	1984825
	Germany	439	46465
	Greece	1034	26525
	Hungary	3140	45403
	Italy	8101	446056
	Netherlands	504	15120
	Norway	430	29285
	Portugal	308	27694
	Slovenia	192	3690
	Spain	52	826
	Sweden	290	31438
	Switzerland	2896	185172
	UK	10608	1531263
North America	Canada	3845	19202
	US	3141	163053
Latin America	Mexico	2443	63678
	Chile	342	29410
	Colombia	1101	18044
	El Salvador	262	11438
	Nicaragua	153	4786
Asia	Hong Kong	18	306
	Japan	2364	302339
	Taiwan	350	21170
Oceania	Australia	674	27688
	New Zealand	72	2991

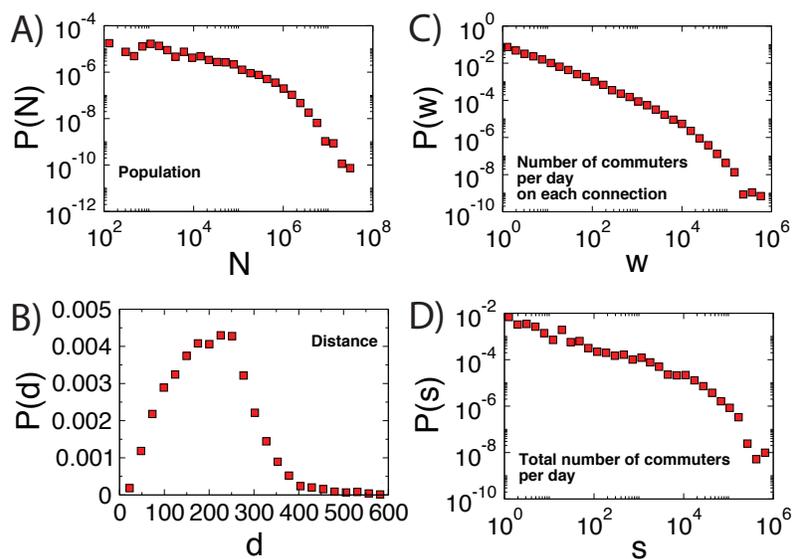


Figure 2: Properties of the commuting network after the mapping to the level of the geographical census areas. We display the distributions of A) population, B) distance between connected census areas, C) daily number of people commuting on each connection, and D) number of daily commuters traveling outside each census area. The distribution of distances between connected geographical census areas peaks around 250kms and decays exponentially afterwards, showing how people tend to avoid daily travels taking roughly more than two hours on each way. All the other properties are distributed in a broad range, varying as much as 6 orders of magnitude in some cases.

self-similar this allows us to estimate a gravity law that successfully reproduces the commuting data obtained across different continents, and provide us with estimations for the possible commuting levels in the countries for which such data is not available.

The layer of commuting network that we consider in GLEaM does not include the inter-countries commuting, as detailed data is usually missing. Commuting data is indeed available at the national level, and the mobility out of the country is usually provided as a single figure measuring international commuting, with no specification of the destination. We studied the magnitude of these fluxes in the few countries for which this information is available. We found that within the European Union – where we expect the intra-country mobility to be higher – France reports a total of 250 000 daily international commuters out of a total of 23 million. That is, 1% of the overall French commuting. Similarly, international commuters amount up to a 1.1% of the commuting in Austria and less than 1% in the UK (0.3%) or in Italy (0.2%). Other regions, such as South America, report similar or smaller values. Given the very small contribution of international commuting with respect to national commuting or international air travel, and the lack of extensive data, we did not consider the commuting between neighboring countries. A straightforward extension of

Table 3: Exponents of gravity law as obtained by applying a multivariate analysis to global commuting data.

d (km)	Parameter	Estimate	Standard Error	p-value	R ²
≤ 300	α	0.46	0.01	< 2E − 16	0.7972
	γ	0.64	0.01	< 2E − 16	
	β	0.0122	0.0002	< 2E − 16	
> 300	α	0.35	0.06	6.91E − 09	0.5369
	γ	0.37	0.06	2.12E − 09	

the model can be developed as additional data becomes available.

1.3.1 Gravity Law and commuting data statistical analysis

We use the following expression as a model for the number of commuters ω_{ij} traveling between two nearby geographical census areas i and j :

$$\omega_{ij} = C \frac{N_i^\alpha N_j^\gamma}{e^{\beta d_{ij}}}, \quad (1)$$

where d_{ij} is the distance between the two airports in kms and N_i and N_j are the populations of the census areas i and j , respectively. We have tested other expressions with a power-law dependence on d_{ij} , finding that the exponential behavior better describes the data.

The gravity law of equation (1) has 4 free parameters: the exponents, α and γ , the inverse characteristic distance β and the proportionality constant, C . A multivariate regression analysis is applied to obtain the values of the parameters that better fit our data as well as an estimation of their statistical significance. By applying a logarithmic transformation to both sides of Eq. (1) we get the expression

$$\log(\omega_{ij}) = \alpha \log N_i + \gamma \log N_j - \beta d_{ij} + \log C, \quad (2)$$

in which all the fit parameters enter linearly. The values estimated for α , γ , β and C are reported in the Table 3 along with their p -values and the regression coefficients.

The division that we have done of the fit in two regions corresponding to $d_{ij} \leq 300$ km and $d_{ij} > 300$ km is a result of the existence of two different regimes in ω_{ij} that emerge during the minimization of the residual sum of squares. The transition between these two different trends in ω_{ij} can be observed in Fig. 1 of the main paper where at around 250km a flattening of the commuter flows with respect to d_{ij} is seen.

The fit of the model parameters is done considering all the empirical commuter data, as mapped into our geographical census areas, and therefore aggregating data from different countries and regions of the world. One of such regions, like the US or Europe, could be dominant and therefore bias the multivariate regression. To make sure that this is not the case, we test *a posteriori* the agreement of the actual commuting flows with those generated synthetically with the obtained gravity

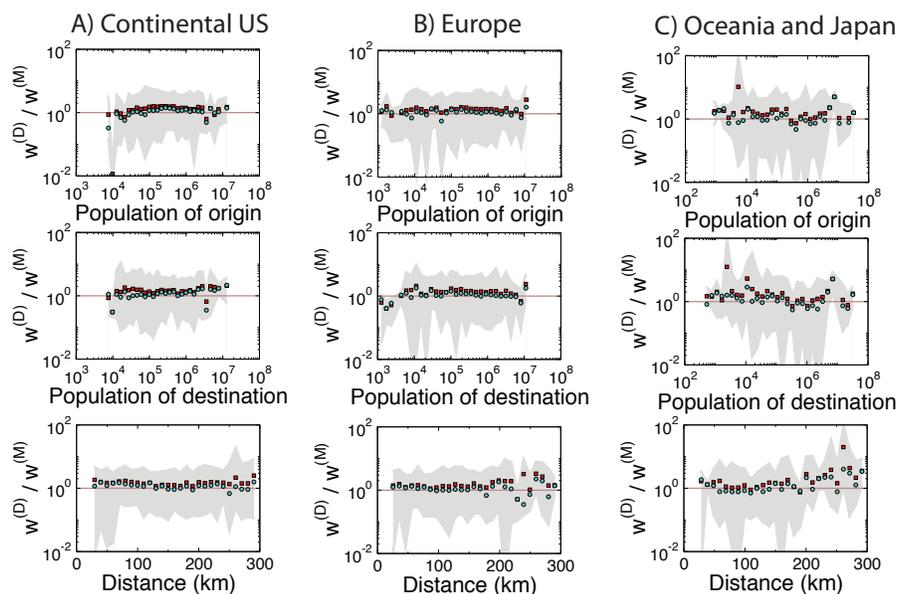


Figure 3: In panels A), B) and C) we display the mean (square), median (circle) and 95% CI (shaded area) for the ratio of actual commuting fluxes to the model values as a function of three variables (population of origin, population of destination, distance between airports) in the Continental US, Europe and Oceania together with Japan, respectively. We demonstrate that the functional form as well as the fitted exponents of gravity law enable us to successfully reproduce the actual commuting fluxes at a global scale.

law in the different regions and countries. We find that the synthetic commuting networks are statistical good representations of the actual data all over the world (see Figure 3 and 4), further supporting the use of the gravity law at a global scale.

It is important to stress that the obtained gravity law is working at the level of our geographical census areas, but in general cannot be extrapolated to different granularity. As we discuss in the main paper, the tessellation defines geographical areas centered around major transportation hubs. This construction is the same in all countries of the world, thus providing tassels which have a unique granularity. This granularity has statistical properties much more homogeneous with respect to the many different administrative boundaries and partitions used in different countries, and defines a framework compatible with a gravity law that is general enough to be applied in different parts of the world. Finally we must mention that we have analyzed the gravity law also by using a progressive decomposition allowing univariate regression and a bootstrapping procedure. These methodologies produce very similar results to those reported here.

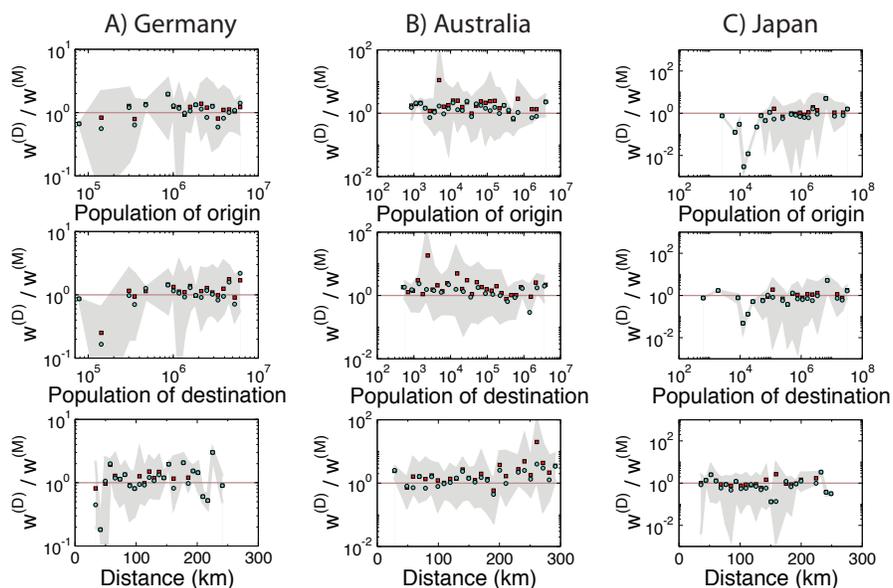


Figure 4: In panels A), B) and C) we display the mean (square), median (circle) and 95% CI (shaded area) for the ratio of actual commuting fluxes to the model values as a function of three variables (population of origin, population of destination, distance between airports) in Germany, Australia and Japan, respectively. We demonstrate that the functional form as well as the fitted exponents of gravity law enable us to successfully reproduce the actual commuting fluxes at a global scale.

1.3.2 Synthetic generation of commuting networks

By using the airport tassels we first determined the geographical neighbors of each subpopulation. This task is quite straightforward: If two cells with a common boundary are assigned to different geographical census areas, then the census areas are nearest neighbors and will have a link in the synthetic network. In this construction we have only considered those geographical census areas which are located in the same country. However, we have to make an exception for the airport of Basel because it is operated jointly by France and Switzerland. This process provides us with the connections of the commuting networks in each country. Then we have assigned the weight of each link, the flux of daily commuters, by applying the gravity law of Eq. (1) whose parameters have been fitted to the entire empirical commuting database.

In Figure 5, we show the distributions of population sizes and distances between commuting neighbors, as well as the distribution of weights and out-commuters of the census areas. The synthetic commuting network exhibits very similar properties to the actual commuting data (see Figure 2 for a comparison).

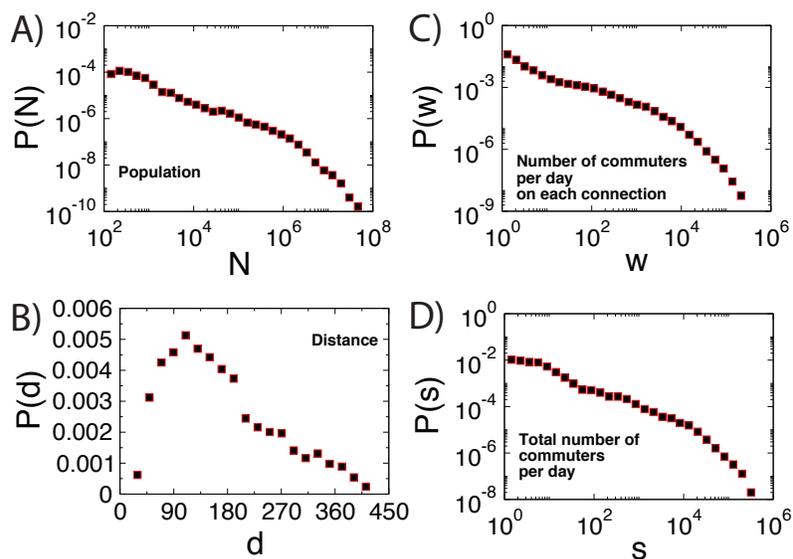


Figure 5: Properties of the synthetic commuting networks. We display the distributions of A) population, B) distance between connected census areas, C) daily number of people commuting on each connection, and D) number of residents in each geographical census area commuting outside per day. The distribution of distances peaks around 100 kms. However all the other quantities are distributed in a broad range varying over 6 orders of magnitude in intensity as also occurs for the empirical networks.

2 Epidemic dynamic model

Each geographical census area corresponds to a subpopulation in the metapopulation model, inside which we consider a Susceptible-Latent-Infectious-Recovered (SLIR) compartmental scheme, typical of influenza-like illnesses (ILIs), where each individual has a discrete disease state assigned at each moment in time. In Fig. 6, a diagram of the compartmental structure with transitions between compartments is shown. The contagion process, i.e. generation of new infections, is the only transition mechanism which is altered by short-range mobility, whereas all the other transitions between compartments are spontaneous and remain unaffected by the commuting. The rate at which a susceptible individual in subpopulation j acquires the infection, the so called force of infection λ_j , is determined by interactions with infectious persons either in the home subpopulation j or in its neighboring subpopulations on the commuting network.

Given the force of infection λ_j in subpopulation j , each person in the susceptible compartment (S_j) contracts the infection with probability $\lambda_j \Delta t$ and enters the latent compartment (L_j), where Δt is the time interval considered. Latent individuals exit the compartment with probability $\epsilon \Delta t$, and transit to asymptomatic infectious compartment (I_j^a) with probability p_a or, with the complementary probability $1 - p_a$, become symptomatic infectious. Infectious persons with

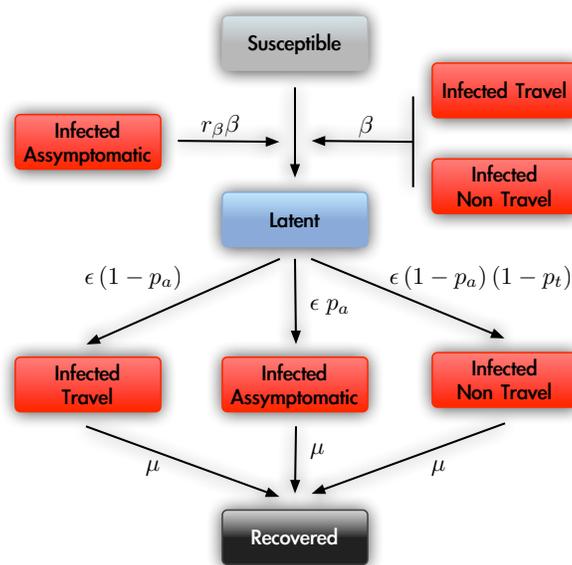


Figure 6: Compartmental structure of our epidemic model within each subpopulation.

symptoms are further divided between those who can travel (I_j^t), probability p_t , and those who are travel-restricted (I_j^{nt}) with probability $1 - p_t$. All the infectious persons permanently recover with probability $\mu\Delta t$, entering the recovered compartment (R_j) in the next time step. All transitions and corresponding rates are summarized in Table 4 and in Figure 6. In each subpopulation the variation of the number of individuals in each compartment $[m]$ can be written at any given time step as

$$X_j^{[m]}(t + \Delta t) - X_j^{[m]}(t) = \Delta X_j^{[m]} + \Omega_j([m]) \quad , \quad (3)$$

where the term $\Delta X_j^{[m]}$ represents the change due to the compartment transitions induced by the disease dynamics and the transport operator $\Omega_j([m])$ represents the variations due to the traveling and mobility of individuals. The latter operator takes into account the long-range airline mobility and defines the minimal time scale of integration as 1 day. The mobility due to the commuting flows is taken into account by defining effective force of infections by using a time scale separation approximations as detailed in the following sections.

2.1 Stochastic and discrete integration of the disease dynamics

In each subpopulation j , we define an operator acting on a compartment $[m]$ to account for all the transitions out of the compartment in the time interval Δt . Each element $\mathcal{D}_j([m], [n])$ of this operator is a random variable extracted from a multinomial distribution and determines the number of transitions from compartment $[m]$ to $[n]$ occurring in Δt . The change $\Delta X_j^{[m]}$ of a compartment $[m]$

Table 4: Transitions between compartments and their rates.

Transition	Type	Rate
$S_j \rightarrow L_j$	Contagion	λ_j
$L_j \rightarrow I_j^a$	Spontaneous	ϵp_a
$L_j \rightarrow I_j^t$	"	$\epsilon(1 - p_a)p_t$
$L_j \rightarrow I_j^{nt}$	"	$\epsilon(1 - p_a)(1 - p_t)$
$I_j^a \rightarrow R_j$	"	μ
$I_j^t \rightarrow R_j$	"	μ
$I_j^{nt} \rightarrow R_j$	"	μ

in this time interval is given by a sum over all random variables $\{\mathcal{D}_j([m], [n])\}$ as follows

$$\Delta X_j^{[m]} = \sum_{[n]} \{-\mathcal{D}_j([m], [n]) + \mathcal{D}_j([n], [m])\} . \quad (4)$$

As a concrete example let us consider the evolution of the latent compartment. There are three possible transitions from the compartment: transitions to the asymptomatic infectious, the symptomatic traveling and the non-traveling infectious compartments. The elements of the operator acting on L_j are extracted from the multinomial distribution

$$\text{Pr}^{\text{Multin}}(L_j(t), p_{L_j \rightarrow I_j^a}, p_{L_j \rightarrow I_j^t}, p_{L \rightarrow I_j^{nt}}) , \quad (5)$$

determined by the transition probabilities

$$\begin{aligned} p_{L_j \rightarrow I_j^a} &= \epsilon p_a \Delta t , \\ p_{L_j \rightarrow I_j^t} &= \epsilon(1 - p_a)p_t \Delta t , \\ p_{L \rightarrow I_j^{nt}} &= \epsilon(1 - p_a)(1 - p_t) \Delta t , \end{aligned} \quad (6)$$

and by the number of individuals in the compartment $L_j(t)$ (its size). All these transitions cause a reduction in the size of the compartment. The increase in the compartment population is due to the transitions from susceptibles into latents. This is also a random number extracted from a binomial distribution

$$\text{Pr}^{\text{Bin}}(S_j(t), p_{S_j \rightarrow L_j}) , \quad (7)$$

given by the chance of contagion

$$p_{S_j \rightarrow L_j} = \lambda_j \Delta t , \quad (8)$$

with a number of attempts given by the number of susceptibles $S_j(t)$. After extracting these numbers from the appropriate multinomial distributions, we can calculate the change $\Delta L_j(t)$ as

$$\Delta L_j(t) = -[\mathcal{D}_j(L, I^a) + \mathcal{D}_j(L, I^t) + \mathcal{D}_j(L, I^{nt})] + \mathcal{D}_j(S, L) . \quad (9)$$

2.2 The integration of the transport operator

The transport operator is defined by the airline transportation data and sets the integration time scale to 1 day. The number of individuals in the compartment $[m]$ traveling from the subpopulation j to the subpopulation ℓ is an integer random variable, in that each of the $X_j^{[m]}$ potential travellers has a probability $p_{j\ell} = w_{j\ell}/N_j$ to go from j to ℓ . In each subpopulation j the numbers of individuals $\xi_{j\ell}$ traveling on each connection $j \rightarrow \ell$ at time t define a set of stochastic variables which follows the multinomial distribution

$$P(\{\xi_{j\ell}\}) = \frac{X_j^{[m]}!}{(X_j^{[m]} - \sum_{\ell} \xi_{j\ell})! \prod_{\ell} \xi_{j\ell}!} (1 - \sum_{\ell} p_{j\ell})^{(X_j^{[m]} - \sum_{\ell} \xi_{j\ell})} \prod_{\ell} p_{j\ell}^{\xi_{j\ell}}, \quad (10)$$

where $(1 - \sum_{\ell} p_{j\ell})$ is the probability of not traveling, and $(X_j^{[m]} - \sum_{\ell} \xi_{j\ell})$ identifies the number of non traveling individuals of the compartment $[m]$. We use standard numerical subroutines to generate random numbers of travellers following these distributions. The transport operator in each subpopulation j is therefore written as

$$\Omega_j([m]) = \sum_{\ell} (\xi_{\ell j}(X_{\ell}^{[m]}) - \xi_{j\ell}(X_j^{[m]})), \quad (11)$$

where the mean and variance of the stochastic variables are $\langle \xi_{j\ell}(X_j^{[m]}) \rangle = p_{j\ell} X_j^{[m]}$ and $\text{Var}(\xi_{j\ell}(X_j^{[m]})) = p_{j\ell}(1 - p_{j\ell}) X_j^{[m]}$. Direct flights as well as connecting flights up to two-legs can be considered. It is worth remarking that on average the airline network flows are balanced so that the subpopulation N_j are constant in time, e.g. $\sum_{[m]} \Omega_j([m]) = 0$.

2.3 Time-scale separation and the integration of the commuting flows

The Global Epidemic and Mobility (GLEaM) modeler combines the infection dynamics with long- and short-range human mobility. Each of these dynamical processes operates at a different time scale. For ILI there are two important intrinsic time scales, given by the latency period ε^{-1} and the duration of infectiousness μ^{-1} , both larger than 1 day. The long-range mobility given by the airline network has a time scale of the order of 1 day, while the commuting takes place in a time scale of approx. $\tau^{-1} \sim 1/3$ day. The explicit implementation of the commuting in the model thus requires a time interval shorter than the minimal time of airline transportation. To overcome this problem, we use a time-scale separation technique, in which the short-time dynamics is integrated into an effective force of infection in each subpopulation.

We start by considering the temporal evolution of subpopulations linked only by commuting flows and evaluate the relaxation time to an equilibrium configuration. Consider the subpopulation j coupled by commuting to other n subpopulations. The commuting rate between the subpopulation j and each of its neighbors i will be given by σ_{ji} . The return rate of commuting individuals is set to be τ . Following the work of Sattenspiel and Dietz [7], we can divide the individuals original from the subpopulation j , N_j , between $N_{jj}(t)$ who are from j are located in j at time t and those, $N_{ji}(t)$, that are from j are located in a neighboring subpopulation i at time t . Note that by consistency

$$N_j = N_{jj}(t) + \sum_i N_{ji}(t). \quad (12)$$

The rate equations for the subpopulation size evolution are then

$$\begin{aligned}\partial_t N_{jj} &= -\sum_i \sigma_{ji} N_{jj}(t) + \tau \sum_i N_{ji}(t) \quad , \\ \partial_t N_{ji} &= \sigma_{ji} N_{jj}(t) - \tau N_{ji}(t) \quad .\end{aligned}\tag{13}$$

By using condition (12), we can derive the closed expression

$$\partial_t N_{jj} + (\tau + \sigma_j) N_{jj}(t) = N_j \tau \quad ,\tag{14}$$

where σ_j denotes the total commuting rate of population j , $\sigma_j = \sum_i \sigma_{ji}$. $N_{jj}(t)$ can be expressed as

$$N_{jj}(t) = e^{-(\tau+\sigma_j)t} \left(C_{jj} + N_j \tau \int_0^t e^{(\tau+\sigma_j)s} ds \right) \quad ,\tag{15}$$

where the constant C_{jj} is determined from the initial conditions, $N_{jj}(0)$. The solution for $N_{jj}(t)$ is then

$$N_{jj}(t) = \frac{N_j}{(1 + \sigma_j/\tau)} + \left(N_{jj}(0) - \frac{N_j}{(1 + \sigma_j/\tau)} \right) e^{-\tau(1+\sigma_j/\tau)t} \quad .\tag{16}$$

We can similarly solve the differential equation for the time evolution of $N_{ji}(t)$,

$$\begin{aligned}N_{ji}(t) &= \frac{N_j \sigma_{ji}/\tau}{(1 + \sigma_j/\tau)} - \frac{\sigma_{ij}}{\sigma_j} \left(N_{jj}(0) - \frac{N_j}{(1 + \sigma_j/\tau)} \right) e^{-\tau(1+\sigma_j/\tau)t} \\ &+ \left[N_{ji}(0) - \frac{N_j \sigma_{ji}/\tau}{(1 + \sigma_j/\tau)} + \frac{\sigma_{ij}}{\sigma_j} \left(N_{jj}(0) - \frac{N_j}{(1 + \sigma_j/\tau)} \right) \right] e^{-\tau t} \quad .\end{aligned}\tag{17}$$

The relaxation to equilibrium of N_{jj} and N_{ji} is thus controlled by the characteristic time $[\tau(1 + \sigma_j/\tau)]^{-1}$ in the exponentials. Such term is dominated by $1/\tau$ if the relation $\tau \gg \sigma_j$ holds. In our case, $\sigma_j = \sum_i \omega_{ji}/N_j$, that equals the daily total rate of commuting for the population j . Such rate is always smaller than one since only a fraction of the local population is commuting, and it is typically much smaller than $\tau \simeq 3 - 10 \text{ day}^{-1}$. Therefore the relaxation characteristic time can be safely approximated by $1/\tau$. This time is considerably smaller than the typical time for the air connections of one day and then consider the subpopulations $N_{jj}(t)$ and $N_{ji}(t)$ as relaxed to their equilibrium values,

$$N_{jj} = \frac{N_j}{1 + \sigma_j/\tau} \quad \text{and} \quad N_{ji} = \frac{N_j \sigma_{ji}/\tau}{1 + \sigma_j/\tau} \quad .\tag{18}$$

This approximation, originally introduced by Keeling and Rohani [8], allows us to consider each subpopulation j as having an effective number of individuals N_{ji} in contact with the individuals of the neighboring subpopulation i . In practice, this is similar to separate the commuting time scale from the other time scales in the problem (disease dynamics, traveling dynamics, etc.). While the approximation holds exactly only in the limit $\tau \rightarrow \infty$, it is good enough as long as τ is much larger than the typical transition rates of the disease dynamics. In the case of ILIs, the typical time scale separation between τ and the compartments transition rates is close to one order of magnitude or even larger.

Eqs. (19) can be generalized in the time scale separation regime to all compartments $[m]$. The number of individuals $X_j^{[m]}(t)$ in each compartment $[m]$ at time t in city j can be expressed as

the sum of individuals $X_{jj}^{[m]}(t)$ who are actually present in their home subpopulation and those $X_{ji}^{[m]}(t)$ who are visiting a neighboring city i [7]. By definition it follows that $X_j^{[m]}(t) = X_{jj}^{[m]}(t) + \sum_{i \in \nu(j)} X_{ji}^{[m]}(t)$, where $\nu(j)$ denotes the set of neighbors of j . All individuals in each traveling compartment visit a neighboring subpopulation at a rate of σ_{ji} for an average duration of τ^{-1} . At the equilibrium the populations $X_{jj}^{[m]}$ and $X_{ji}^{[m]}$ can be therefore expressed as:

$$X_{jj}^{[m]} = \frac{X_j^{[m]}}{(1 + \sigma_j/\tau)} \text{ and } X_{ji}^{[m]} = \frac{X_j^{[m]}}{(1 + \sigma_j/\tau)} \sigma_{ji}/\tau, \quad (19)$$

where $\sigma_j = \sum_{i \in \nu(j)} \sigma_{ji}$ denotes the total commuting rate of j and $X_{jj}^{[m]} = X_j^{[m]}$ and $X_{ji}^{[m]} = 0$ for all the other compartments which are restricted from traveling. These expressions will be used to obtain the effective force of infection taking into account the interactions generated by the commuting flows.

2.4 Effective force of infection

The force of infection λ_j that a susceptible population of a subpopulation j sees can be decomposed into two terms: λ_{jj} and λ_{ji} . The component λ_{jj} refers to the part of the force of infection whose origin is local in j . While λ_{ji} indicates the force of infection acting on susceptibles of j during their commuting travels to a neighboring subpopulation i . The effective force of infection can be estimated by summing these two terms weighted by the probabilities of finding a susceptible from j in the different locations, S_{jj}/S_j and S_{ji}/S_j , respectively. Using the time-scale separation approximation that establishes the equilibrium populations in Eq. (19), we can write

$$\lambda_j = \frac{\lambda_{jj}}{1 + \sigma_j/\tau} + \sum_i \frac{\lambda_{ji} \sigma_{ji}/\tau}{1 + \sigma_j/\tau}. \quad (20)$$

We will focus now on the calculation of each term of the previous expression. The force of infection occurring in a subpopulation j is due to the local infectious persons staying at j or to infectious individuals from a neighboring subpopulation i visiting j and so we can write

$$\lambda_{jj} = \frac{\beta_j}{N_j^*} \left[I_{jj}^{nt} + I_{jj}^t + r_\beta I_{jj}^a + \sum_i (I_{ij}^{nt} + I_{ij}^t + r_\beta I_{ij}^a) \right], \quad (21)$$

where β_j is introduced to account for the seasonality in the infection transmission rate (if the seasonality is not considered, it is a constant), and N_j^* stands for the total effective population in the subpopulation j . By definition, $I_{jj}^{nt} = I_j^{nt}$ and $I_{ji}^{nt} = 0$ for $j \neq i$. If we use the equilibrium values of the other infectious compartments (see Eq. (19)) we obtain

$$\lambda_{jj} = \frac{\beta_j}{N_j^*} \left[I_j^{nt} + \frac{I_j^t + r_\beta I_j^a}{1 + \sigma_j/\tau} + \sum_i \frac{I_i^t + r_\beta I_i^a}{1 + \sigma_i/\tau} \sigma_{ij}/\tau \right]. \quad (22)$$

The derivation of λ_{ji} follows from a similar argument yielding:

$$\lambda_{ji} = \frac{\beta_i}{N_i^*} \left[I_{ii}^{nt} + I_{ii}^t + r_\beta I_{ii}^a + \sum_{\ell \in \nu(i)} (I_{\ell i}^{nt} + I_{\ell i}^t + r_\beta I_{\ell i}^a) \right], \quad (23)$$

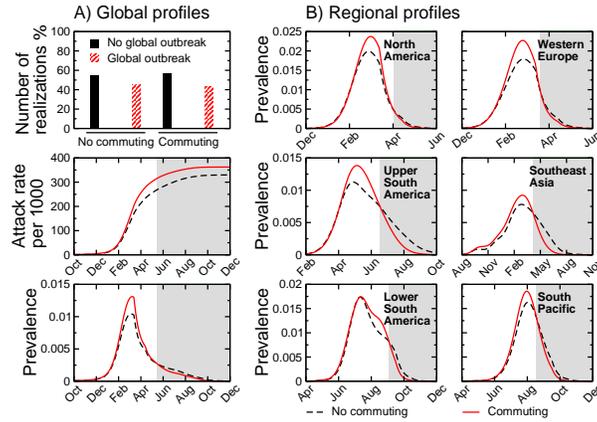


Figure 7: Initial time: April 1, $R_0 = 1.5$. In A) we compare simulation results for the global epidemic profiles with and without inclusion of commuting networks and $\tau = 3 \text{ day}^{-1}$. In B) prevalence profiles in 6 regions corresponding to 3 different climatic zones, Northern, Tropical and Southern, from top to bottom, respectively, are shown. The faster decay in the prevalence profiles are highlighted by shaded areas. The profiles of Lower South America and South Pacific refer to the epidemic activities in the following winter.

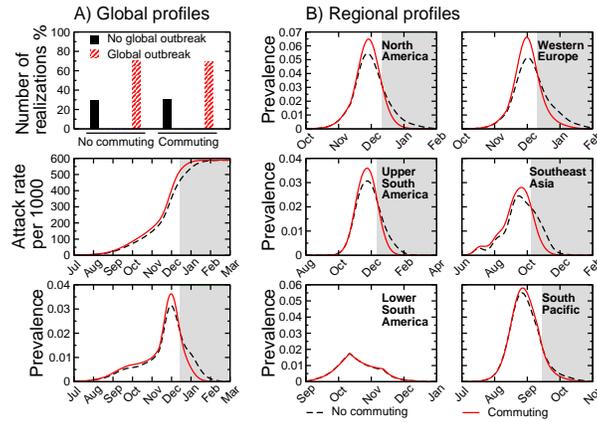


Figure 8: Initial time: April 1, $R_0 = 2.3$. In A) we compare simulation results for the global epidemic profiles with and without inclusion of commuting networks and $\tau = 3 \text{ day}^{-1}$. In B) prevalence profiles in 6 regions corresponding to 3 different climatic zones, Northern, Tropical and Southern, from top to bottom, respectively, are shown. The faster decay in the prevalence profiles are highlighted by shaded areas.

where $v(i)$ represents the set of neighbors of i , and therefore the terms under the sum are due to the visits of infectious individuals from the subpopulations ℓ , neighbors of i , to i . By plugging the

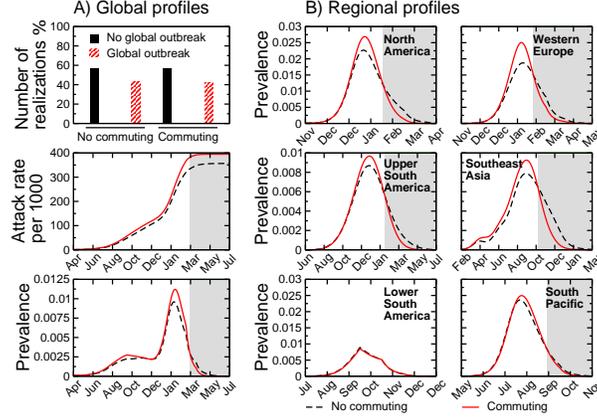


Figure 9: Initial time: October 1, $R_0 = 1.5$. In A) we compare simulation results for the global epidemic profiles with and without inclusion of commuting networks and $\tau = 3 \text{ day}^{-1}$. In B) prevalence profiles in 6 regions corresponding to 3 different climatic zones, Northern, Tropical and Southern, from top to bottom, respectively, are shown. The faster decays in the prevalence profiles are highlighted by shaded areas. The profiles of North America and Western Europe refer to the epidemic activities in the following winter.

equilibrium values of the compartment into the above expression, we obtain

$$\lambda_{ji} = \frac{\beta_i}{N_i^*} \left[I_i^{nt} + \frac{I_i^t + r_\beta I_i^a}{1 + \sigma_i/\tau} + \sum_{\ell \in v(i)} \frac{I_\ell^t + r_\beta I_\ell^a}{1 + \sigma_\ell/\tau} \sigma_{\ell i}/\tau \right]. \quad (24)$$

Finally, in order to have an explicit form of the force of infection we need to evaluate the effective population size N_j^* in each subpopulation j , i.e., the actual number of people staying at the location j . The effective population is $N_j^* = N_{jj} + \sum_i N_{ij}$, that in the time-scale separation approximation reads

$$N_j^* = I_j^{nt} + \frac{N_j - I_j^{nt}}{1 + \sigma_j/\tau} + \sum_i \frac{N_i - I_i^{nt}}{1 + \sigma_i/\tau} \sigma_{ij}/\tau. \quad (25)$$

Note that in these equations all the terms with compartments have an implicit time dependence. By inserting λ_{jj} and λ_{ji} into Eq. (20), it can be seen that the expression for the force of infection includes terms of zeroth, first and second order on the commuting ratios (i.e., σ_{ij}/τ). These three term types have a straightforward interpretation: The zeroth order terms represent the usual force of infection of the compartmental model with a single subpopulation. The first order terms account for the effective contribution generated by neighboring subpopulations with two different sources: Either susceptible individuals of subpopulation j having contacts with infectious individuals of neighboring subpopulations i , or infectious individuals of subpopulations i visiting subpopulation j . The second order terms correspond to an effective force of infection generated by the contacts of susceptible individuals of subpopulation j meeting infectious individuals of subpopulation ℓ (neighbors of i) when both are visiting subpopulation i . This last term is very small

in comparison with the zeroth and first order terms, typically around two order of magnitudes smaller, and in general can be neglected.

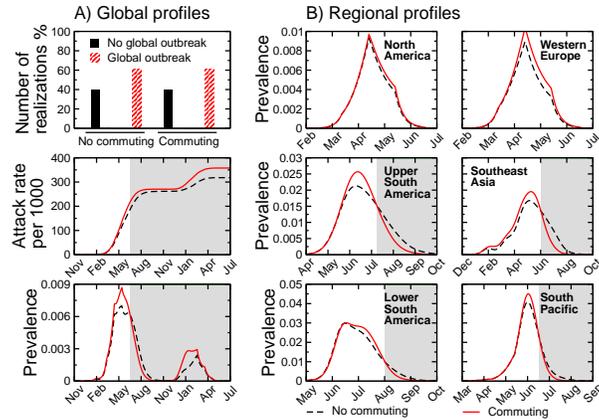


Figure 10: Initial time: October 1, $R_0 = 1.9$. In A) we compare simulation results for the global epidemic profiles without and with inclusion of commuting networks and $\tau = 3 \text{ day}^{-1}$. In B) prevalence profiles in 6 regions corresponding to 3 different climatic zones, Northern, Tropical and Southern, from top to bottom, respectively, are shown. The faster decays in the prevalence profiles are highlighted by shaded areas.

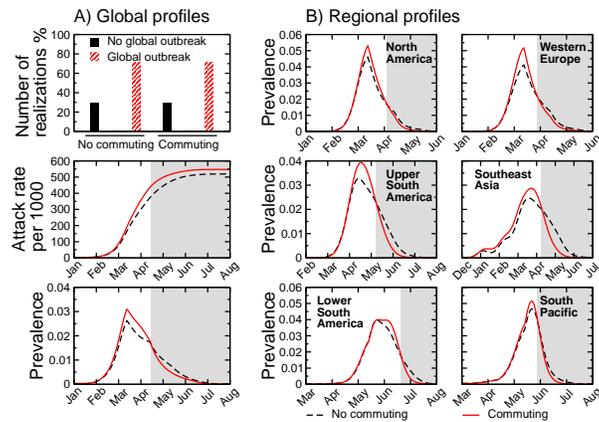


Figure 11: Initial time: October 1, $R_0 = 2.3$. In A) we compare simulation results for the global epidemic profiles without and with inclusion of commuting networks and $\tau = 3 \text{ day}^{-1}$. In B) prevalence profiles in 6 regions corresponding to 3 different climatic zones, Northern, Tropical and Southern, from top to bottom, respectively, are shown. The faster decays in the prevalence profiles are highlighted by shaded areas.

2.5 Construction of invasion trees

In the main text, we present a geographic invasion tree inside the continental United States, showing both the infection hierarchy and the most probable transmission routes between subpopulations. This is a directed, weighted minimum spanning tree among all the possible infection transmission paths.

For every subpopulation j , we keep track of the cumulative number of imported infections from all other neighboring subpopulations. This is done until time t_0 , when the first local generation of infection occurs in the subpopulation. For each subpopulation pair lj , we define p_{lj} as the probability of infection transmission from l to j . This probability shows the likelihood that subpopulation j 's infection is seeded by subpopulation l . Since infected cases are imported either by air traffic or commuting, p_{lj} is calculated by considering the transmission probability through air traffic p_{lj}^a , and commuting p_{lj}^c , depending on the connection type.

In the following, for subpopulation j , we use A for the set of subpopulations which are connected by air traffic, and C for the set of neighboring subpopulations having commuting flows. For p_{lj}^a , the imported seeds include latent (L), symptomatic infected traveling (I^t) and asymptomatic infected people (I^a). Assume $l(l \in A)$ is a subpopulation having air traffic with j , and $\xi_{lj}(t)$ ($\xi \in (L, I^t, I^a)$) is the number of people in each compartment traveling from l to j at time t . Then the probability is defined as:

$$p_{lj}^a = \frac{\sum_{t < t_0} (L_{lj}(t) + I_{lj}^t(t) + I_{lj}^a(t))}{\sum_{l \in A} \sum_{t < t_0} (L_{lj}(t) + I_{lj}^t(t) + I_{lj}^a(t))} .$$

For p_{lj}^c , because of the commuting short-range coupling between subpopulation j and every neighboring subpopulation $l(l \in C)$, all neighboring subpopulations which have infected cases (including infected people who are not traveling) should be considered as potential seeds for this subpopulation. Hence this probability is:

$$p_{lj}^c = \frac{\sum_{t < t_0} (I_l^t(t) + I_l^{nt}(t) + I_l^a(t))}{\sum_{l \in C} \sum_{t < t_0} (I_l^t(t) + I_l^{nt}(t) + I_l^a(t))} .$$

Therefore, for any connected subpopulation pair lj , the transmission probability is

$$p_{lj} = \frac{p_{lj}^a + p_{lj}^c}{\sum_{k \in A} p_{kj}^a + \sum_{k \in C} p_{kj}^c} .$$

Finally, similar to the distance based on correlation coefficients $d_{ij} = \sqrt{2(1 - r_{ij})}$ where r_{ij} is the correlation coefficient, we define a distance metric $d_{ij} = \sqrt{1 - p_{ij}}$ to measure dissimilarities for the infection probability. The minimum spanning tree is then calculated using Chu-Liu-Edmunds Algorithm.

3 Simulations of an hypothetical epidemic

We performed two different sets of simulations with the new ingredient, i.e. commuting, in which only the synthetic commuting network or the combination of synthetic commuting network with

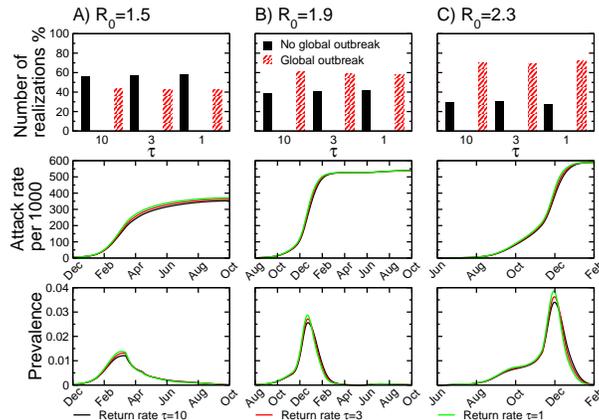


Figure 12: Initial time: April 1. We compare simulation results for the global epidemic profiles by varying commuting return rate τ one order of magnitude. Even though there are small variations in the profiles of the prevalence curves, the results are stable and illustrate that the intensity of the commuting fluxes do not significantly alter the basic evolution of the epidemic.

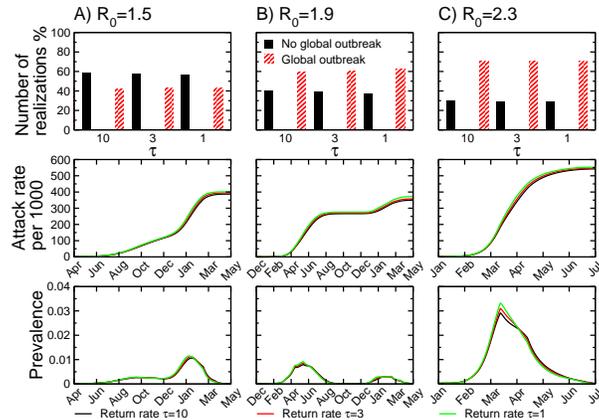


Figure 13: Initial time: October 1. We compare simulation results for the global epidemic profiles by varying commuting return rate τ one order of magnitude. Even though there are small variations in the profiles of the prevalence curves, the results are stable and illustrate that the intensity of the commuting fluxes do not significantly alter the basic evolution of the epidemic.

real data has been considered. These results have been compared with each other as well as with the simulations where only the airline traffic is taken into account.

Model parameters are same as of the main paper if not stated otherwise. All simulations were initiated by a single symptomatic infectious person and let evolve for a duration of two years. Only the runs with a global outbreak, defined as generation of new symptomatic cases in more

than one country, were considered for the analysis.

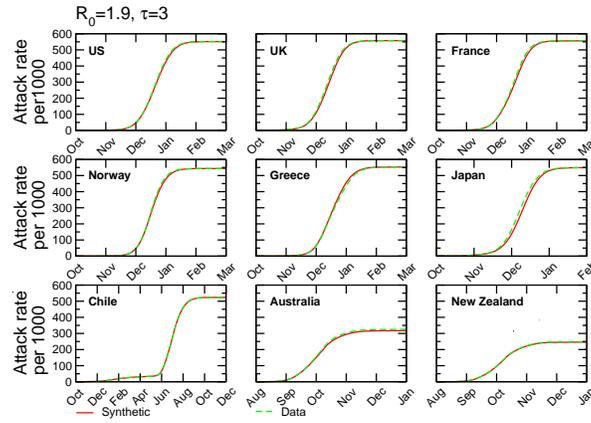


Figure 14: Initial time: April 1, $R_0 = 1.9$, $\tau = 3 \text{ day}^{-1}$. We compare attack rate curves of representative countries from 5 continents where we have commuting data. The profiles have been obtained by using the synthetic commuting network alone or the actual commuting data combined with the synthetic network in the simulations. The results are in good agreement, additionally confirming the reliability of our estimation for the gravity law.

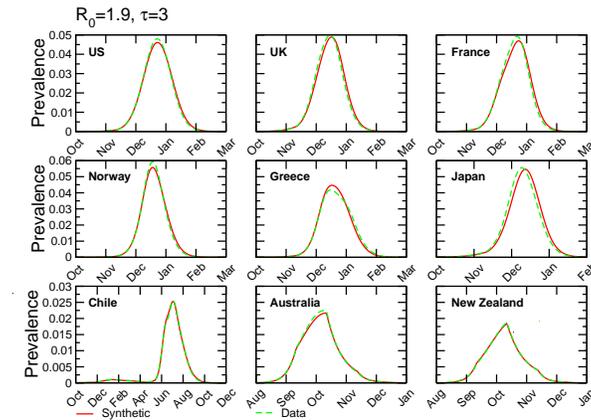


Figure 15: Initial time: April 1, $R_0 = 1.9$, $\tau = 3 \text{ day}^{-1}$. We compare prevalence curves of representative countries from 5 continents where we have commuting data. The profiles have been obtained by using the synthetic commuting network alone or the actual commuting data combined with the synthetic network in the simulations. The results are in good agreement, additionally confirming the reliability of our estimation for the gravity law.

3.1 Initial conditions and R_0

Hanoi has been chosen as the origin of the infection with two different initial conditions: April 1st and October 1st. Simulations with three different values of $R_0 = 1.5, 1.9$ and 2.3 have been carried out. In the Figures 7-11, we compare the results with and without ground transportation (commuting) for a fixed return rate of $\tau = 3 \text{ day}^{-1}$. The results are shown at a global scale and also at the level of regions in different climatic zones.

In the panel A of each figure, we display the chance of global outbreak at the end of two years from the initial date. Depending on the value of R_0 , between 57% and 29% of the realizations show no epidemic propagation out from the source country. If the infection does get out of the source country, it is almost certain that the epidemic will circulate in all the countries of our database given the values of R_0 simulated. We see that the initial date does not have any effect on the outbreak probability, but this should be taken with caution because we consider a very long temporal range in our simulations (2 years). Since the commuting networks are confined to the interior of each country and no commuting is allowed across country boundaries, we observe that the outbreak probability does not change by the inclusion of short-range mobility. The only way for an infection to cross borders in our model is via air-transportation.

In the same panel, we also show the mean global attack rate and the prevalence as a function of time. The shape of the epidemic curves strongly depends on the value of R_0 and on the initial time since the seasonality has been implemented in the simulations. The inclusion of commuting increases the global attack rate up to 4% from the baseline case with air-travel only. The major differences between the simulations with and without commuting concentrate in the areas where the fluctuations are more relevant. The addition of commuting does not have a high impact on the global magnitudes characterizing the epidemic spreading, but plays an important role in the local distribution of the disease. It increases the local synchrony between nearby subpopulations, an effect that is visible in the faster decays of the tails of the prevalence curves.

In the panel B of Figures 7-11, the average prevalence of 6 regions in different world hemispheres are shown (Northern, Tropical and Southern). To obtain the regional epidemic curves, we aggregate the data of the geographical census areas falling within each region but always discriminating by climatic zones to avoid the wash out of seasonality. As explained above, the effect of considering commuting becomes more clear when the geographical focus is localized in more restricted areas.

3.2 Commuting return rate τ

Our baseline value for the time spent away from the home subpopulation during commuting is $\tau^{-1} = 1/3 \text{ day}$, roughly corresponding to the average working hours in a weekday. In Figures 12 and 13, we compare the global epidemic profiles under one order of magnitude variation in the commuting return rate, $\tau = 1, 3$ and 10 day^{-1} . Once the commuting is considered, the value of τ does not cause a qualitative difference in the epidemic evolution (see the tails of the prevalence curves). Remembering that the relevant quantity in the calculation of the force of infection is the ratio of the commuting rate to the return rate, these results show that the commuting-flux intensities do not significantly alter the basic evolution of the epidemic. The reason for this is that the higher order terms in the effective force of infection that couple neighboring subpopulations are

relevant only when the number of infected individuals in the subpopulation of interest is virtually zero. Once the local outbreak is started, the main contribution to the evolution of the epidemics is given by the local force of infection. The intensity of the coupling is therefore not extremely relevant as it affects only marginally the evolution of the internal subpopulation epidemic evolution. This argument explains why varying over one order of magnitude the coupling induced by the commuting flows does not have an appreciable impact on the epidemic profiles. On the other hand, the presence or not of the coupling makes a difference by inducing synchronization effects in the early seeding of epidemics in neighboring subpopulations.

3.3 Synthetic commuting network versus real commuting data

In the Figures 14 and 15, we compare the attack rate and the prevalence profiles generated by using the commuting network obtained from real data with those generated by using the synthetic commuting network. The synthetic commuting network is quite successful in reproducing the same qualitative epidemic curves in specific countries and worldwide.

4 Model comparison with seasonal influenza data

In this section, we compare results of our model with the seasonal influenza activities in the year 2001-2002. During this season the predominant influenza virus type was A/H3N2 in most world wide locations [10], such as United States [13, 10], Europe [17], Australia [20], Africa and the Americas [10], while in the previous influenza season (2000-2001) the A/H1N1 type was the most common worldwide [9]. This fact makes it possible to distinguish the current epidemic from the remnants of the previous year epidemic by simply looking at the virus type. Influenza activity is then obtained from the Global Influenza Program (FluNet) database by WHO [15] and Flu Activity & Surveillance by the United States Centers for Disease Control and Prevention [12].

From the epidemiological records, Hong Kong is the only city in the SE Asia region having A/H3 influenza activity during 2001 summer [16]. This choice is supported by recent work on the geographical origin and the global spreading of A/H3N2 influenza [18]. We further assume that 10^{-5} of the initial city population is latent in accordance with [14, 19].

For the model parameters, we have considered a latent period of $\epsilon = 1.1$ days, and infectious period of $\mu = 2.95$ days. This choice results in an average generation interval of around 4 days, in accordance with published estimates [11]. Moreover, only a fraction $\alpha = 60\%$ [14] of the world population is susceptible to seasonal influenza. The effective reproductive number R_{eff} is defined as $R_{eff} = \alpha R_0$ and we have considered an R_{eff} value of 1.4 that is in the range of those observed for seasonal influenza. This value is obtained by a best fit of the peak time in the US surveillance regions.

The initial date of the simulation is obtained by calculating the best χ^2 value of the results produced by the simulation when compared with the empirical influenza activity in Australia. For tolerance, we use a range of dates for which the χ^2 value is less than 1.1 times of minimum, namely July 21 [July 9 - July 31] for the model including commuting and August 1 [July 22 - August 15] for the model without this component. In all cases we perform 1000 realizations for each set of parameters and initial dates.

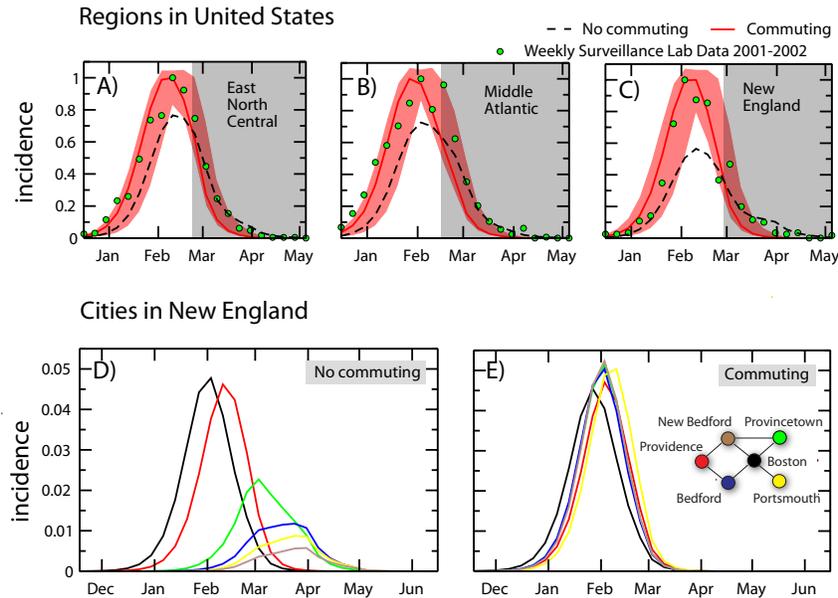


Figure 16: A), B) and C). Comparison between 2001-2002 seasonal influenza surveillance data and the simulation results in three separate US regions. Circles represent weekly surveillance data, while the median of predictions with (without) commuting is represented by a solid (dashed) line. The red shaded area represents the 95% CI for the range of best initial dates [July 9 – July 31] of the model with commuting networks. We rescale the empirical and commuting datasets to one, and divide the curves corresponding to the no commuting case by the maximum value of the diffusive case. This guarantees that all curves can be shown on the same scale, while illustrating the different model predictions for the cases with and without commuting as enhanced by the gray shaded area. D) and E). Median weekly incidence profiles for Boston area and surrounding cities with no commuting (D) and with commuting (E). A schematic network representation of the short-range connections is shown for guidance. The synchronization among the various incidence profiles is considerably enhanced when commuting is considered, with a reduction of over one month in the time interval between peaks in neighboring cities. Model profiles were calculated using 10^3 independent outbreak realizations.

In Figure 16 we compare our predictions with 2001-2002 weekly surveillance lab data in three different US regions. Along with a good agreement with the empirical data, Figure 16 captures the same features and behavior exhibited by the synthetic scenario used in the main text. Both models (without or with commuting networks) capture the temporal evolution of empirical influenza activities within the confidence intervals for selected range of best initial dates. The figure however shows the increased synchronization among the various incidence profiles of neighboring cities when commuting is considered analogously to what shown in Figure 3 of the main text. In addition it is evident that at the regional level, the model including commuting flows is providing a

better fit against the real data.

References

- [1] Barrat A, Pastor-Satorras R, Vespignani A (2004), The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. (USA)* 101:3747–3752.
- [2] Colizza V, Barrat A, Barthelemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci U S A* 103: 2015–2020.
- [3] Colizza V, Barrat A, Barthelemy M, Vespignani A (2006) The modeling of global epidemics: Stochastic dynamics and predictability. *Bull Math Biol* 68: 1893–1921.
- [4] Colizza V, Barrat A, Barthelemy M, Valleron A-J, Vespignani A (2007) Modeling the World-wide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *PLoS Med* 4(1): e13. doi:10.1371/journal.pmed.0040013.
- [5] Center for International Earth Science Information Network (CIESIN), Columbia University; and Centro Internacional de Agricultura Tropical (CIAT). The Gridded Population of the World Version 3 (GPWv3): Population Grids. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. <http://sedac.ciesin.columbia.edu/gpw>
- [6] Center for International Earth Science Information Network (CIESIN), Columbia University; International Food Policy Research Institute (IFPRI); The World Bank; and Centro Internacional de Agricultura Tropical (CIAT). Global Rural-Urban Mapping Project (GRUMP), Alpha Version: Population Grids. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. <http://sedac.ciesin.columbia.edu/gpw>
- [7] Sattenspiel L, Dietz K (1995) A Structured Epidemic Model Incorporating Geographic Mobility Among Regions. *Mathematical Biosciences* 128: 71–91.
- [8] Keeling MJ, Rohani P (2002) Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters* 5: 20–29.
- [9] Update: Influenza activity – United States and worldwide, 2000–01 season, and composition of the 2001–02 influenza vaccine. *MMWR*, 50:466–470, 2001.
- [10] Update: Influenza activity – United States and worldwide, 2001–02 season, and composition of the 2002–03 influenza vaccine. *MMWR*, 51:503–506, 2002.
- [11] Carrat F, Vergu E, Ferguson NM, Lemaître M, Cauchemez S, Leach S, Valleron AJ (2008). Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *American Journal of Epidemiology* 167(7):775.
- [12] CDC flu activity & surveillance, Centers for Disease Control and Prevention, United States. <http://www.cdc.gov/flu/weekly/fluactivity.htm>

-
- [13] Centers for Disease Control and Prevention, United States. 2001-02 influenza season summary, 2002.
- [14] Grais RF, Ellis JH, Kress A, Glass GE (2004). Modeling the spread of annual influenza epidemics in the US: The potential role of air travel. *Health Care Management Science*, 7(2):127–134.
- [15] Global influenza surveillance (FluNet), World Health Organization.
<http://www.who.int/csr/disease/influenza/influenzane트워크/en/>
- [16] World Health Organization. Influenza in the world. *Weekly Epidemiological Record*, 46:357–364, 2001.
- [17] Paget WJ, Meerhoff TJ, Goddard NL, et al (2002). Mild to moderate influenza activity in europe and the detection of novel A (H1N2) and B viruses during the winter of 2001-02. *Euro Surveill* 7(11):147–57.
- [18] Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, Gust ID, Hampson AW, Hay AJ, Hurt AC, et al (2008). The global circulation of seasonal influenza A (H3N2) viruses. *Science* 320(5874):340.
- [19] Rvachev LA, Longini IMJR (1985). A mathematical model for the global spread of influenza. *Mathematical Biosciences* 75(1):3–22.
- [20] Yohannes K, Roche P, Spencer J, Hampson A (2003). Annual report of the national influenza surveillance scheme, 2002. *Commun Dis Intell.* 27(2):162–172.