Commuter Mobility and the Spread of Infectious Diseases: Application to Influenza in France

Segolene Charaudeau[1,2,3,4]*, **Khashayar Pakdaman**[3,4], **Pierre-Yves Boëlle**[1,2]

[1] INSERM, UMR S 707, Paris, France, [2] Université Pierre et Marie Curie - Paris 6, Paris, France, [3] Institut Jacques Monod, Paris, France, [4] Université Denis Diderot, Paris, France

Abstract

Commuting data is increasingly used to describe population mobility in epidemic models. However, there is little evidence that the spatial spread of observed epidemics agrees with commuting. Here, using data from 25 epidemics for influenza-like illness in France (ILI) as seen by the Sentinelles network, we show that commuting volume is highly correlated with the spread of ILI. Next, we provide a systematic analysis of the spread of epidemics using commuting data in a mathematical model. We extract typical paths in the initial spread, related to the organization of the commuting network. These findings suggest that an alternative geographic distribution of GP accross France to the current one could be proposed. Finally, we show that change in commuting according to age (school or work commuting) impacts epidemic spread, and should be taken into account in realistic models.

Citation: Charaudeau S, Pakdaman K, Boëlle P-Y (2014) Commuter Mobility and the Spread of Infectious Diseases: Application to Influenza in France. PLoS ONE 9(1): e83002. doi:10.1371/journal.pone.0083002

Editor: Alessandro Vespignani, Northeastern University, United States of America

Received April 30, 2013; **Accepted** November 7, 2013; **Published** January 9, 2014

Funding: This work has been funded by the Fondation pour la recherche medicale (http://www.frm.org/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: segolene.charaudeau@u707.jussieu.fr

Introduction

The multi-scale network of social interactions [1,2] makes rapid dissemination of transmissible diseases possible, as illustrated recently by pandemic A/H1N1 2009 influenza and SARS [3,4]. In this context, predicting the efficacy of public health interventions requires the identification of the most relevant factors for dissemination [4]–[5]. For instance, international air travel was found to provide good prediction for the worldwide spread of SARS and influenza A/H1N1 2009 [3,4]; it was however shown that intervention on the global air traffic would be of limited efficacy [6]. At a more local scale, air travel is less relevant and other types of movement must be taken into account. Commuting, i.e. daily movements from residence to work or school, has been widely used to describe spatial mobility in models, using exhaustive datasets [7,8] or gravity models [9,10].

Except for a report on the correlation between influenza epidemic peak timing and inter-states commuting in the USA [9], whether commuting may explain the spatial spread of epidemics has been little studied. Influenza like illness (ILI) incidence time series, as monitored by the Sentinelles network since 1984 in France, provide data at a high spatial resolution (NUTS3) that can be used in this respect (http://www.sentiweb.org). These data, unique in duration and spatial resolution, helped elucidate long sought questions like the impact of school closure during epidemics [11] and to validate model predictions for pandemic flu [12]. Commuting data based on the census of the population is also available at an even finer scale.

Using these two databases we first analyzed how commuting data relates to disease spread at a local level. We then examind the underlying mechanisms of propagation using an epidemic model derived from commuting networks An indicator based on the similarity of epidemic courses in excess of random movements was developed. Finally, we investigated how age differences in commuting networks, i.e. to school or to work, led to changes in the spatial spread of diseases.

Materials and Methods

Data

Sentinelles data. The Sentinelles network [13] is comprised of over thirteen hundred general physicians (GPs), accounting for approximately 2% of the total number of French GPs. They report the number of observed influenza-like illness cases on a regular basis, using a standardized case definition (more than 39C fever with myalgia and respiratory syndromes). We used the data of 26 consecutive seasonal influenza epidemics, from 1985 to 2010 (Figure 1). The data was obtained on a weekly basis at the NUTS3 ('department') level. There are 95 NUTS3 areas in France. To jointly analyse multi-year epidemics, we defined each year week 0 as the national epidemic peak, and considered 15 weeks of data before and after this date.

Demography and commuting. We used the data collected in the 1999 census data in France. All data were obtained at the LAU1 level, that we refer to as 'district' afterwards. There are 3704 districts in France. In each district, the population was split into 5 age classes : less than 3 years old; 3 to 10; 11 to 18; 18 to 65 and more than 65. These categories were retained to capture large changes in mixing groups due to schooling (3–10 and 11–18) and work (18–65). The frequency of each age class was obtained from census data in each district, as well as the percentage of population with a professional occupation. We also computed the average

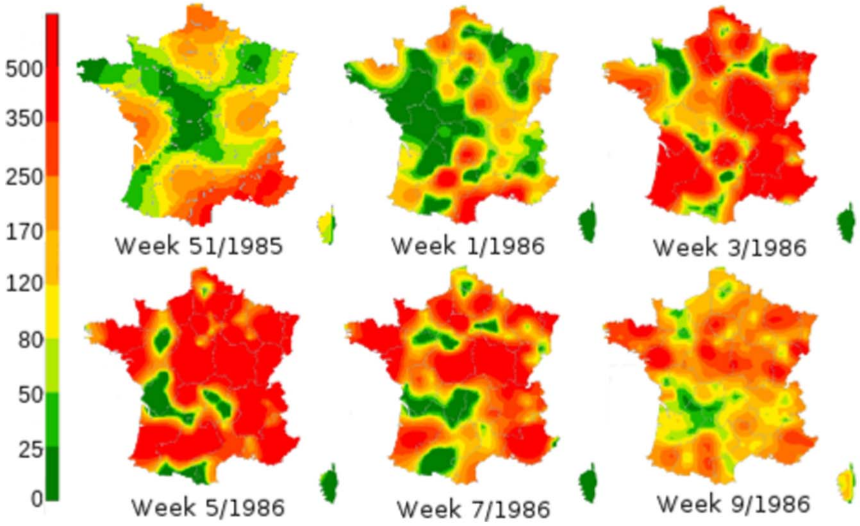

Figure 1. Spatial spread of influenza like illness in France. Incidence for 100000 inhabitants as monitored by the Sentinelles network during season 1985–1986. Maps are 2 weeks apart.
doi:10.1371/journal.pone.0083002.g001

number of contacts of an individual of age a with members of the same household of age a' in each district, denoted by $M_H^D(a,a')$ in district D.

The commuting dataset, derived from census data, contains the movements of more than 25 millions of adults and 9 millions of children. Commuting frequencies between districts were computed as a matrix $M_S(D,D')$ for school-based commuting and $M_W(D,D')$ for work-based commuting, where D stands for the district of residence and D' for the district of destination. The matrices were normalized by rows, yielding the percentage of the population of the district of residence commuting to the district of destination; for example $M_X(D,D)$ was the percentage of people remaining in their district of residence for school or work.

We identified communities using the weighted 'Louvain' algorithm [14]. This algorithm clusters nodes by maximizing the weight of links within each cluster while minimizing that between clusters. The communities identified with the school commuting network and the work commuting network were compared with the Jaccard index, which compares 2 clusterings by measuring the number of district pairs that are gathered together in both clusterings over the number of comparable district pairs (a pair of districts is considered comparable if the 2 units belong to the same community in at least one clustering).

Disease transmission model

Natural history of influenza infection. The natural history of influenza infection was described as a 4 stage SEIR process: individuals were first susceptible to the disease (stage S), then latent (infected but not infectious yet; stage E), infectious (stage I) and finally recovered and removed from transmission (stage R). We simulated transmission using the generation time distribution, i.e. the time from infection in a primary case to infection in a secondary case, as in Mills et al. [15]. For all asymptomatic cases and symptomatic cases within households, the generation time distribution was modelled by a gamma distribution with mean 3.7 days and standard deviation 3.1 days. For symptomatic cases in the community, the generation time was gamma distributed with mean 1.1 days and standard deviation 0.4 day [16]. These differences account for the reduced time spent in the community, school or workplace by symptomatic cases. We assumed an initial percentage of susceptibility of 80%, irrespective of age.

Transmission. A discrete time (time step 0.2 days) deterministic transmission model was implemented. We assumed that only professionally active individuals in age class 18–65 would commute to work, and that all children aged 3 to 18 attended and commuted to school. School-based commuting matrices were the same in age classes 3–10 and 11–18. No births and deaths were considered during the time of simulation, nor any change in place of residence or of destination.

At each time step, the number of incident cases $\Delta I_{a,D}(t)$ in age class a and district D was computed as $S_{a,D}(t)\times P_{a,D}(t)$ where $S_{a,D}(t)$ was the number of susceptible individuals and $P_{a,D}(t)$ the probability of infection. The probability of infection was calculated according to the following equation:

(1).

$$P_{a,D}(t)=1-e^{-(\lambda_{a,D}^{H}(t)+\lambda_{a,D}^{S}(t)+\lambda_{a,D}^{W}(t)+\lambda_{a,D}^{Co}(t))\Delta t} \tag{1}$$

where $\lambda_{a,D}^{X}(t)$ was the force of infection exerted on an individual of age a in district D from place X.

Household based force of infection was computed using the age-specific average number of contacts in the household. More precisely, the force of infection was proportional to the density of infected contacts among household members as follows (2) :

$$\lambda_{a,D}^{H}(t)=\beta_H\frac{\sum_{a'} M_H^D(a,a')\times(I_{a',D}^{A}(t)+I_{a',D}^{S}(t))}{\sum_{a'} M_H^D(a,a')\times N_{a',D}(t)} \tag{2}$$

where β_H was the pairwise rate of contact leading to transmission in the household. $I_{a',D}^{A}(t)$ and $I_{a',D}^{S}(t)$ were respectively the number of asymptomatic and symptomatic incident cases, which were considered equally able to transmit the infection.

For school-based (X = S) and workplace-based (X = W) force of infections, we used a similar approach, computing the expected density of infection among contacts as (3):

$$\lambda^{X}_{a,D}(t)= \beta_X \frac{\sum_{D'} M_X(D,D') \times \sum_{D''} M_X(D'',D') \times (I^{A}_{a',D}(t)+I^{S}_{a',D}(t))}{\sum_{D'} M_X(D,D') \times \sum_{D''} M_X(D'',D') \times N_{a,D''}} \tag{3}$$

here β_X was the pairwise rate of contact leading to transmission. Using this formulation, the contacts in place D' are counted with all people effectively commuting to this place, from place D as well as from all places D'' directly connected to D'.

For community based transmission, the force of infection was computed using the same principle as above by (4).

$$\lambda^{Co}_{a,D}(t)=\beta_{Co} \times \frac{\sum_{D'} M_{Co}(D,D') \times \sum_{D''} M_{Co}(D'',D') \sum_{a'} (I^{A}_{a',D}(t)+I^{S}_{a',D}(t))}{\sum_{D'} M_{Co}(D,D') \times \sum_{D''} M_{Co}(D'',D') \sum_{a'} N_{a',D''}} \tag{4}$$

where the sum was on all districts D' sharing a border with district D. To take into account the different behavior of people during day and night, we considered that individuals were only commuting during the day, and staying at home during the night. Therefore, we considered that individuals could only interact within their households at night.

We calibrated transmission parameters β_S, β_W, β_C and β_H so that simulated epidemics had durations and attack rates consistent with observed epidemics (see http://www.sentiweb.org). More precisely, in the Sentinelles network, a typical epidemic starts when incidence increases over 150 cases/100000 per week, and remains above this threshold for approximately 10 weeks; the cumulated excess cases during this period ranges between 2 and 8 percent of the population. We selected parameters with which the duration with an incidence larger than 150/100000 was 10 weeks, and the excess cumulated cases was 5.5% of the population. Several sets of β values were still possible, and we finally selected values so that one half of the cases were due to school or work transmission (respectively $35.6\% \pm 0.008$ and $10.1\% \pm 0.001$), and the other half to local transmission (household and community, respectively $29.9\% \pm 0.005$ and $24.2\% \pm 0.006$ of transmission). This repartition compared with other choices reported in [17] and [7], although we put a little more weight on school/work transmission. Using these parameters, the initial exponential growth coefficient of the epidemic was 0.75 log(person)/week, in the same range as those observed during the last 25 epidemic seasons in France (0.5 to 1.0).

Statistical analysis of data and results

Spatial auto-correlation analysis. Moran's I statistic [18] was used to evaluate the spatial auto-correlation of ILI incidence data. Moran's I was calculated by:

$$I=\frac{N}{\sum w_{ij}} \times \frac{\sum_i \sum_j w_{ij}(x_i - x)(x_j - x)}{\sum_i (x_i - x)^2} \tag{5}$$

where N is the number of spatial units, x_i the incidence observed in unit i and w_{ij} the spatial weight of the link between i and j. Moran's I ranges between -1 and 1, with negative values indicating negative correlation among neighbors, while positive values indicate positive correlation. To assess whether commuting agreed with spatial incidence, we computed the w_{ij} as the size of the population commuting between i and j [19].

Moran's I was computed for each week before and after epidemic peaks, and averaged, week-wise. The same procedure was repeated 1000 times using random permutations to calculate p-values. To test for the specific role of the commuting network as opposed to commuting distance only, we compared these indices with those obtained using random commuting networks, where the distribution of distance travelled was kept the same as in the original data, but commuting trips were chosen at random in any direction. We repeated the above calculation for 100 such random networks.

We also used Mantel's test as described in [9]. The correlation between incidence time series was first calculated for all pairs of departments, then compared with the flows (ingoing and outgoing) between departments.

In all cases, permutation tests were used to calculate P-values.

Overlap between epidemics. We used the overlap measure introduced in Colizza [20], that takes into account the similarity in spatial spread, as well as in total incidence. Values close to 1 indicate similar incidence in all places at a given time, while values of 0 correspond with little overlap. In all cases, epidemics were started with one infected children in a single district. The overlap between two epidemics, started in districts I and II, was calculated as

$$\Theta(t)=\left(\sqrt{\frac{\sum_j I^{I}_{j}(t)}{N}\frac{\sum_j I^{II}_{j}(t)}{N}} + \sqrt{(1-\frac{\sum_j I^{I}_{j}(t)}{N}) \times (1-\frac{\sum_j I^{II}_{j}(t)}{N})} \right) \times \sum_j \sqrt{\Pi^{I}(t) \times \Pi^{II}(t)} \tag{6}$$

where $\Pi^{I}(t)$ described the geographical distribution of incidence among districts at time step t in epidemic I, and $i^{I}(t)$ was the incidence per population at time t. The overlap measurement is for a given time t. Irrespective of the starting places, the overlap measure always grew to 1 with time.

For each pair of districts in France, we aimed to identify up to what date after first introduction epidemics grew more similarly than expected if commuting was at random. This is measured by criterion, C_1 that we computed as follows. First, the commuting networks were reshuffled, by permuting, at random, the destinations in the original network. This procedure retained the distribution of degrees in incoming and outgoing links, but randomized the destinations all over France, implementing a random commuting network. Then, epidemics were simulated starting from the same pair of districts using the reshuffled networks. The "above randomness" part was computed as the time during which the overlap of the epidemics simulated using the original networks was larger than that with the reshuffled networks (Figure 2). Large values of C_1 indicated that the two epidemics looked alike for a long time.

Sensitivity analysis. To test the sensitivity of the model to the proportion of infections occuring in each context, we performed 100 simulations with a set of parameters, for which $32.0\% \pm -0.005$ of transmission occured at home, $36.5\% \pm 0.0056$ at school or work and $31.3\% \pm -0.0009$ in the community, starting from randomly selected districts. Overlap was used to compare these simulations to the former ones.

An analysis of sensitivity was also performed to test the impact of the hypothesis that adults asymptomatic individuals had a reduced generation time, by simulating 100 outbreaks with a random initial case where only children would have it. As before, overlap was used to compare the simulations to the former ones.

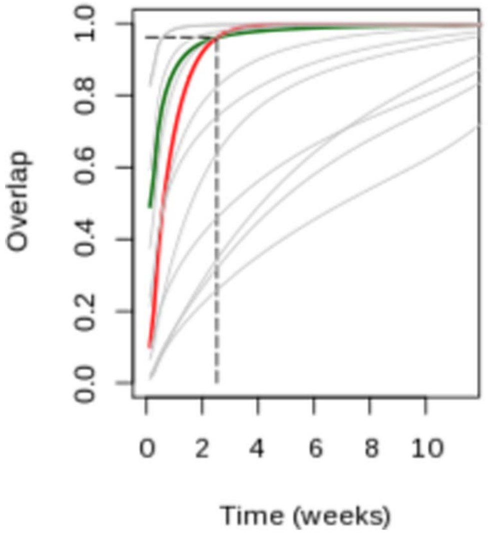

Figure 2. Measuring similarity in spread above randomness C_1. Lines correspond with overlap measures for a given pair of district at different times after introduction of a single infected. For a particular pair (green line), we also present the overlap measure obtained using reshuffled networks for the same pair (red line). Criterion C_1 was defined as the time when the green line crossed the red line.
doi:10.1371/journal.pone.0083002.g002

The sensitivity of the results to the proportion of adults initially immunized was also tested, simulating 100 outbreaks intitialized in randomly chosen districts with different rate of immunity (0, 10, 20, 30, 40, 50, 60 and 70%). Simulations were compared to outbreaks generated with a 80% rate of immunity for adults using overlap.

Results

Commuting networks

Workers from one district commuted on average to 133 other districts, and school aged children to an average 75 destinations (Figure 3-a,b). The average commuting distance was 14.8 km and 12.4 km for work and school, with 15% of workers commuting outside their department, but only 6.7% for children (Figure 3-c). Long distance travel (>100 km away) was however as common for work and school (1.5% of the cases).

The diameter (i.e. the longest minimal path from one place to the other) of the commuting network was 3 for work and 4 for school.

The importance of short-distance commuting also showed in the communities found by clustering (Figure 3-d,e). Indeed, all communities were constituted of adjacent districts, although this is not a constraint of the method. The Jaccard index for the work and school communities was 0.519, showing that approximately half the districts belonged to the same community in both the work and school networks. The differences arose for the most part from places along the borders between clusters. The work network produced less communities than the school network, especially in

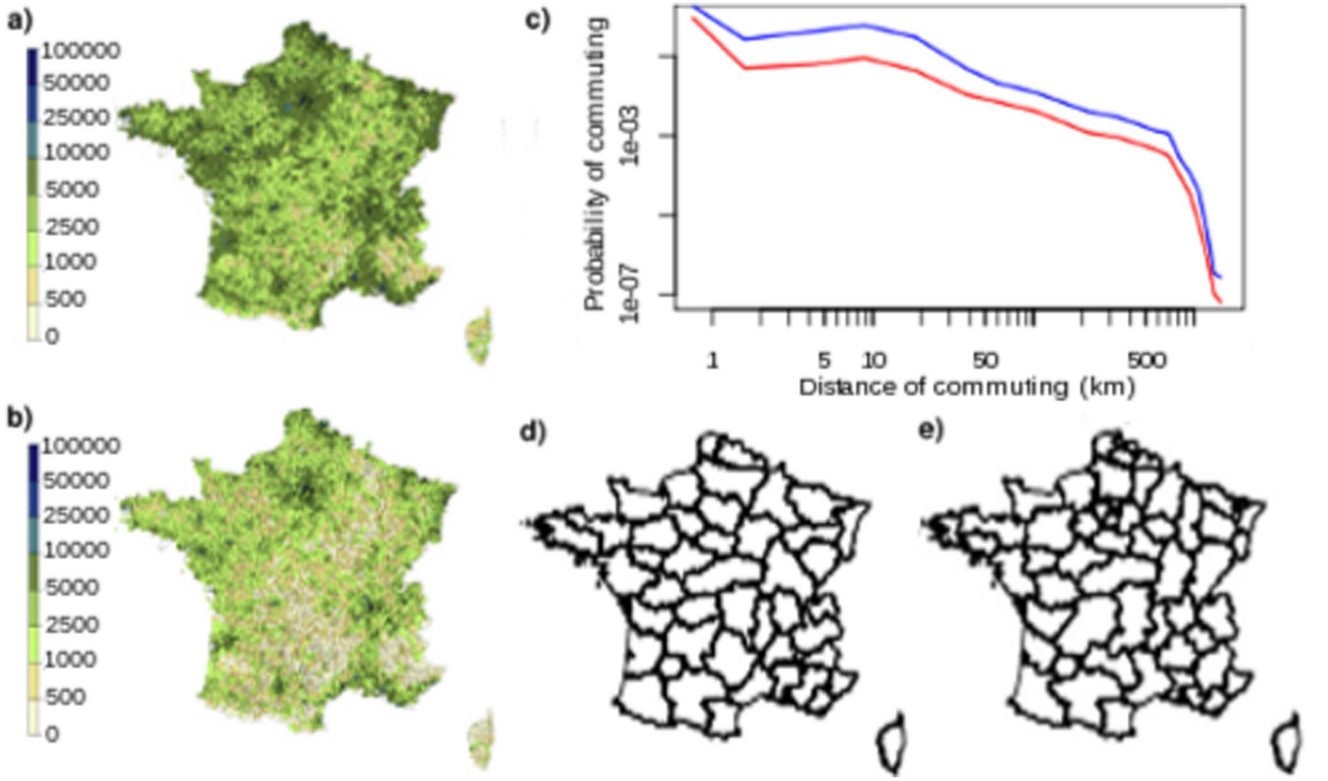

Figure 3. Commuter mobility in France. (a,b)Total number of individuals leaving each district via work commuting (a) and school commuting (b). (c) Proportion of commuters and travelled distance in the school network (red) and the work network (green). (d,e) Clusters identified in the work (d) and schoool (e) commuting networks.
doi:10.1371/journal.pone.0083002.g003

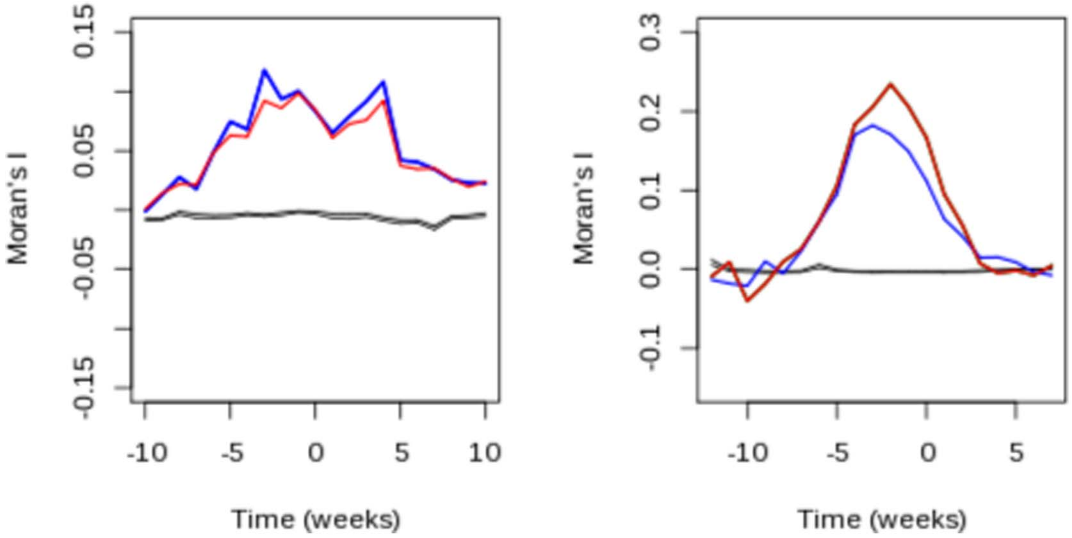

Figure 4. Autocorrelation in incidence for observed and simulated epidemics. (a) Mean value of Moran's Index computed on the 26 epidemics from the Sentinelles network, and (b) on 100 simulated epidemics. In each case, the blue line uses work commuting based weights or school (red line). Gray areas corresponds to the 95% expected values when no autocorrelation is present.
doi:10.1371/journal.pone.0083002.g004

the Paris region, highlighting the more local structure of school commuting.

Commuting and observed epidemics in France

In the 26 epidemics observed in the Sentinelles network, the spatial autocorrelation computed with weights derived from school and work commuting was significantly greater than 0. In other words, incidence increased synchronously in strongly linked areas. Moran's I was significantly greater than 0 ($P<0.001$) as soon as 8 weeks before the national peak and remained greater than 0 up to 9 weeks afterwards(Figure 4-a), with maximum value 1 to 3 weeks before the date of the national peak. The magnitude of Moran's I was approximately the same with all spatial weights.

Likewise, Mantel's test performed with weights matrix derived from school and work commuting was positive (Mantel's correlation being equal to 0.069 for work commuting and 0.060 for school commuting), confirming the existence of a spatial autocorrelation linked to commuting movements ($P<0.001$).

Commuting and simulated epidemics

Simulated epidemics started from different places were all similar in timing and incidence at the national level. Moran's I analysis exhibited the same behavior as in the observed epidemics (Figure 4-b) and was significantly positive using all weight matrices. Here again, the index increased as the epidemic spread and was the largest shortly before the date of national peak.

As for observed epidemics, Mantel's test was found to be positive for simulated epidemics (mantel correlation was equal to 0.106 with work commuting and 0.121 with school commuting).

Overlap in initial epidemic spread. Irrespective of the starting district, national incidence was very similar over the course of the epidemic. Even if the national incidence were similar, overlap changed depending on the pair of districts considered. Initial overlap was very variable using the observed commuting network, but always increased to 1 with time. Remarkably, the overlap in epidemics using reshuffled networks was also large, and quickly increased to 1 as well.

The excess in overlap, as measured by criterion C_1, ranged from 0 to more than 180. The first case arose for epidemics started from distant places, with C_1 increasing in neighboring districts. There was a large negative correlation between C_1 and distance ($r=-0.916\pm0.040$, Spearman correlation). Almost all district pairs more than $d_{lim}=100$ km away had $C_1=0$, in other words epidemics started from districts more than d_{lim} km away showed little resemblance in initial spread.

On the contrary, C_1 increased when the two starting districts were closer, indicating spread on common paths. However, the variance of C_1 was large, even at small distances, indicating that

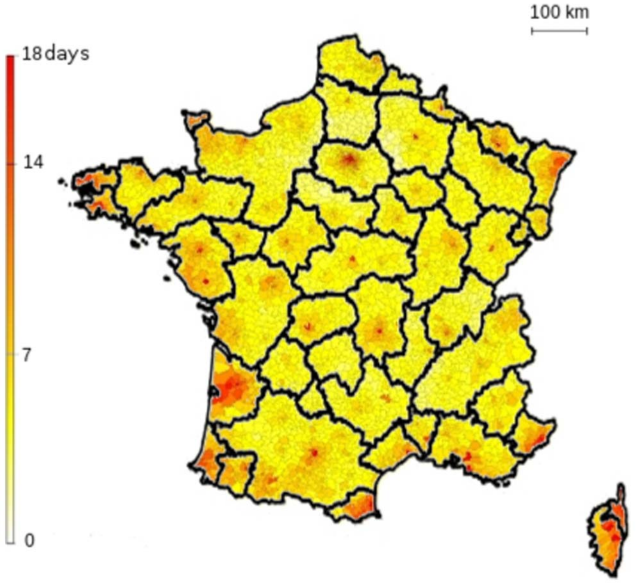

Figure 5. Typical pathways according to initial infective location. For each district, C_1 values were averaged over all neighbors less than 100 km away. Basins of attraction were identified by clustering.
doi:10.1371/journal.pone.0083002.g005

distance was not the only condition for similar spread. For example, 2 epidemics started in districts less than 10 km away could be less similar than 2 epidemics started more than 50 km away; and epidemics started from less than 10 km away could have a very similar spread or quickly diverge depending on the pair of districts considered.

We found that the correlation between C_1 and the proportion of commuters between districts was also large ($r = 0.854 \pm 0.038$), and that both distance and volume contributed to the value of C_1: The partial correlation between C_1 and the proportion of commuters, conditional on distance, was 0.415. The coefficient of determination of distance and proportion of commuters on C_1 was large: $r^2 = 0.852 \pm 0.022$.

To get a picture of initial common paths of spread, we averaged the value of C_1, in each district, over all neighbors less than 100 km away. A large value indicated common initial paths in all epidemics started in close neighbors. Figure 5 illustrates these preferential paths, as evidenced by large values of average C_1 in several places. Among the districts having the largest values of C_1, many were large French cities, like Paris, Toulouse or Marseille: 30 of the 50 largest French cities were among those with the largest C_1 values. Other districts with large average C_1 were found as suburban cities close to large cities; and some in coastal or border districts. Overall, there was a large correlation between average C_1 and the number of inhabitants in each district ($r = 0.654 \pm 0.019$).

Based on the average C_1 value, we obtained 49 communities based on Louvain clustering (Figure 5). Most of these clusters included one or two very populated French cities, for which the average value of C_1 was the highest of the community. 33 clusters included one of the 50 largest French cities and 5 other included a city less important in size, but large relative to its neighboring districts. Other large French cities were included in previous clusters, as they were strongly connected to a large city (Aix, for example, 22nd biggest city in France, was aggregated with Marseille, 2nd most populated city, which is both close and well connected to it). 6 of the remaining clusters did not include major French cities and corresponded with sparsely populated areas. Finally, coastal or border districts tended to cluster together on a geographical basis.

Age dependent commuting networks

Commuting for work and school created two layers of mixing that could lead to differences in the spatial spread. Indeed, the distance traveled to work was larger, suggesting increased dissemination, but transmission in children is typically larger and could take precedence on transmission by adults. We therefore simulated the spread of epidemics in models where either commuters for school or work remained in their place of residence, with the same number of contacts.

Epidemics were started from 100 random districts with the 3 possibilities : commuting to work and school, only to school or only to work (Figure 6-a,b,c). Epidemics simulated with the two commuting reached a national peak in a narrow time window, the time of peak slightly depending on the size of district of departure

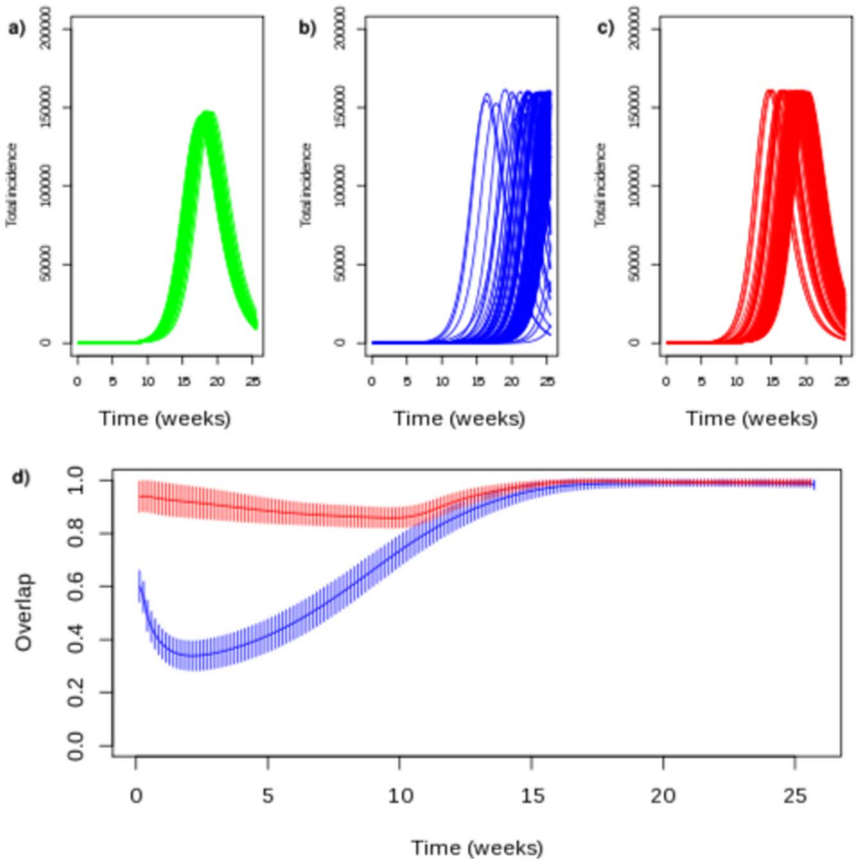

Figure 6. School and work commuting networks and the spatial spread of epidemics. (a,b,c) ILI epidemic curves using all commuting networks (a), only work commuting (b) and only school commuting (c). Epidemics were started form 1000 randomly chosen districts. (d) Overlap between epidemics using work (blue curve) or school commuting (red curve).
doi:10.1371/journal.pone.0083002.g006

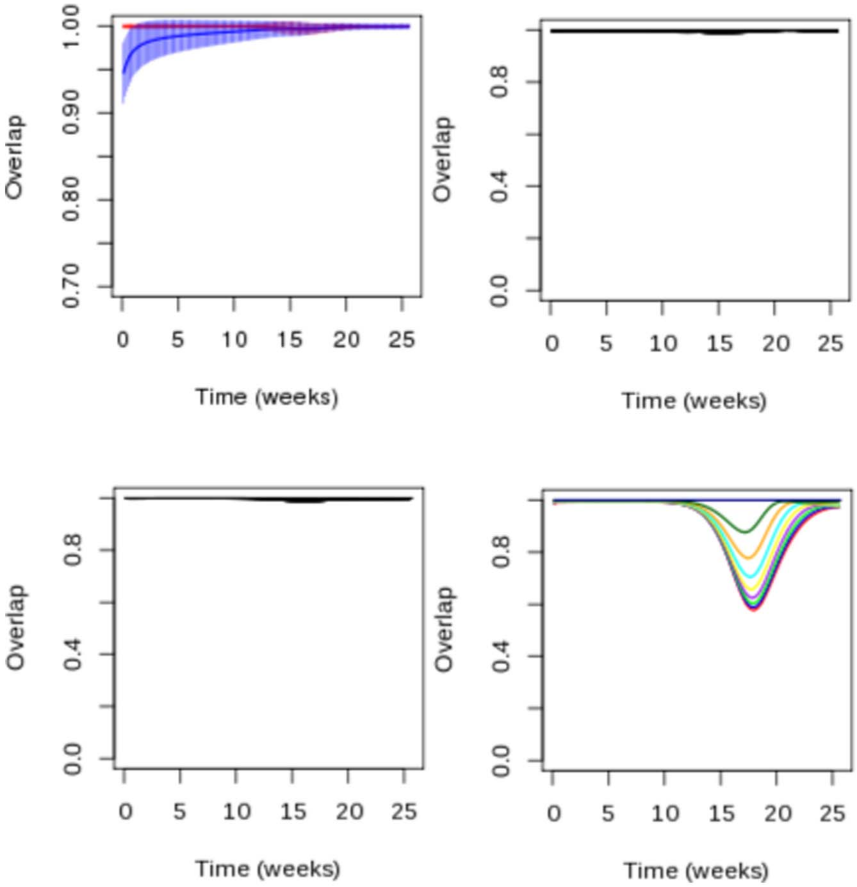

Figure 7. Sensitivity analysis. Overlap between epidemics simulated with first model and epidemics propagating only by school (red) or work (blue) commuting (a), with epidemics for which asymptomatic adults do not have a reuced genration time (b), with epidemics simulated with different parameters of transmission (c). (d) Overlap between epidemics in which 80% of adults are susceptible with epidemics with different rates of susceptibility.

doi:10.1371/journal.pone.0083002.g007

population (correlation -0.087 ± 0.031) or on the number of commuters sent by the district of departure in the school and the work network (correlation were respectively -0.106 ± 0.032 and -0.133 ± 0.032). The final attack rate was not influenced by the district of departure. The spread of epidemics simulated with only one type of commuting was more variable, with an increased range of time to the national peak.

Not unexpectedly, ignoring one commuting network led to epidemics that spread less rapidly. The peak of epidemics simulated with school commuting were on average delayed by 2 weeks, although with large variability. For some simulations, the propagation was faster when only school commuting was present, but this was independent of the district of departure (correlation of delay with district population : 0.037 ± 0.198; correlation with the number of children commuting from the district : 0.011 ± 0.197). The impact was more important for epidemics simulated with work commuting, which were more delayed, and with highest variability.

Finally, simulated outbreaks where all commuters followed the same commuting pattern, either school or work, were much in line with the results above. Overlap with original simulations was almost perfect when using only the school network but differed markedly from the start when using only work commuting (w Figure 7 -a).

Sensitivity analysis. The overlap between simulations with different rates of contacts and the original simulations started in the same district was very large (Figure 7 -b) as 95% of overlap values ranged between 0.9929 and 0.9998 through the entire course of the epidemic. This indicates that the spread of the epidemic was very similar in both cases and that our results regarding to how networks shape the initial spread were robust to this modification.

Similarly, the overlap between epidemics with a reduced generation time for symptomatic adults and without was very large (Figure 7-c) with 95% of overlap values ranging between 0.9931 and 0.9999 during the whole course of the epidemics. This showed that the results regarding initial spread of the disease was robust to this assumption.

The overlap between simulations with 80% of susceptible adults and other percentages of immunization decreased with the rate of susceptibility of adults (Figure 7-d).

Discussion

Our analysis showed that commuting data determines the spread of influenza in modern populations, as evidenced by the large autocorrelation in observed ILI incidence in regions connected by commuting. Building on this observation, we provided an in depth study of the consequences of mobility as described by commuting in the initial spread of epidemics, showing how to identify preferential paths in a densely connected territory. Last, we showed that age specific heterogeneity in commuting leads to different patterns of spread, depending on the age category the most involved in transmission.

The spatial structure of epidemics in France was manifest according to the change in Moran's index over time. The index increased up to a maximum just before the national epidemic peak, and decreased afterwards. This spatial structure was hinted at by the non random structure of spatial incidence pointed out by Bonabeau et al. [21] and the decreasing correlation with distance found by Crepey et al. [22]. However, neither of these studies linked these observations with human mobility. Here, we showed that these properties could be explained by commuting, strengthening the case for using commuting data to model the spatial spread of diseases at a regional scale. We measured the correlation between incidence and commuting using Moran's I and Mantel's test. These provide complementary information regarding the association of commuting with spatial disease spread. Indeed, Moran's I compares magnitudes in connected regions, while Mantel's test is more sensitive to the timing of the peaks between epidemics. As in Viboud [9], Mantel's test supported the hypothesis of correlation between epidemic spread and commuting volume. Our conclusions are further supported by the fact that in the simulated epidemics, Moran's I and Mantel's test displayed the same pattern as for observed epidemics.

In our systematic exploration of the model dynamics, a three stages scenario for the spread of epidemics emerged. The first stage followed introduction of an infected individual in the population. The lack of large C_1 value for districts more than 100 km apart reflected the spatial scale of this first phase, and the large variance in C_1 values evidenced the strong dependence on the initial location for initial spread. During this stage, transmission occurred in the initial community and its proximal districts over a few weeks. It ended when infection reached an amplifier district. This was illustrated by the existence of districts with a large average C_1 value, showing that these places produced epidemics that were very similar to those started around. The second stage saw the spread from the first amplifier district to other districts at a longer range, via long distance links. In this second stage, it was mostly large cities that were attained all over the territory. The last stage started with the spread around large cities, but quickly led to transportation of cases both locally and globally, yielding the national epidemics. Importantly, this structure arose from the features of observed commuting data. One of the challenges was to be able to identify the amplifier nodes and their basins of attraction, and the downstream propagation paths directly from such data. This is where the methods introduced in our paper are of broader interest.

We used the raw commuting data from the census, instead of a smoothed version based on a gravity model [9,23,24]. As our data was exhaustive, it was not necessary to use modelling in the first place. Using raw data leads to more heterogeneity in commuting links, given different districts at the same distance and with the same population may not receive the same number of commuters. It may also lead to results that are very dependent on the reported mobility, which captures only a part of human mobility. Allowing individuals to mix in a local community (district and close neighbors) was a way to keep the particular features of the commuting data, while allowing for inaccuracies or random moves not measured in commuting. We also chose to differentiate school and work commuting, when most metapopulation models either ignore school commuting [9,23] or assume the same rate of contact between individuals in the 2 contexts [24]. In our simulations, we found that the interactions of the two networks tended to homogenize epidemic curves, irrespective of the starting location. Indeed, the timing of the peak was in a very limited range, irrespective of the starting place. With our choice of parameters, the spatial spread of the disease was driven more strongly by school commuting than by work commuting: removing the work network affected less overall transmission than the converse. The prominence of the school network is likely a consequence of our assumption that over 40% of all transmissions occurred in school. However, this analysis shows that differences in commuting networks could lead to changes in spatial spread. For example, it was reported that school holidays mostly affected how quick a disease would spread [25,26], but this result did not take into account differences between work and school commuting. Our results show that closing schools may also affect preferential paths of spread.

Seeding epidemics with only one case, as we did in the systematic analysis, is presumably not very realistic. Indeed, real epidemics may be seeded by repeated introductions from abroad over a few weeks. We however selected this simple seeding pattern to study systematically the influence of the initial place of introduction, as it allowed a rather simple way to compare epidemic courses through their overlap. This type of seeding likely reduces noise and leads to increased spatial autocorrelation, as noted in Figure 4.

Thanks to the systematic search for locations having large similarity with others, we identified preferential paths for epidemic spread due to human mobility. Clustering districts according to the average C_1 measure allowed to define clusters showing the 'basin of attraction' for these preferential paths, as shown in Figure 5. Most clusters were centered around an important city of the area, which may not be highly populated compared to other cities, but was relatively important compared to neighboring places. The role of such places must be studied further in the context of epidemiologic surveillance. Indeed, it suggests that to capture a new epidemic, it would be interesting to have at least a GP in each cluster. It must be studied whether this would be more effective than allocating surveillance based on population coverage [27]. Moreover, as the behavior of epidemics from any district in a cluster tends to resemble the behavior from a central city, focusing on the main cities identified in the study could lead to the optimal use of GPs for surveillance.

Author Contributions

Conceived and designed the experiments: SC KP PYB. Performed the experiments: SC. Analyzed the data: SC. Contributed reagents/materials/analysis tools: SC. Wrote the paper: SC KP PYB.

References

1. Brockmann D, Hufnagel L, Geisel T (2006) The scaling laws of human travel. Nature 439: 462–465.
2. Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453: 779–782.
3. Khan K, Arino J, Hu W, Raposo P, Sears J, et al. (2009) Spread of a Novel Inuenza A (H1N1) Virus via Global Airline Transportation. New England journal of medicine 361: 212–214.
4. Colizza V, Barrat A, Barthelemy M, Vespignani A (2007) Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. BMC Medicine 5.
5. Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, et al. (2004) Modelling disease outbreaks in realistic urban social networks. Nature 429: 180–184.

6. Hollingsworth TD, Ferguson NM, Anderson RM (2006) Will travel restrictions control the international spread of pandemic inuenza? Nature Medicine 12: 497–499.
7. Merler S, Ajelli M (2010) The role of population heterogeneity and human mobility in the spread of pandemic inuenza. Proceedings of the Royal Society B - Biological Sciences 277: 557–565.
8. Ajelli M, Goncalves B, Balcan D, Colizza V, Hu H, et al. (2010) Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. BMC Infectious Diseases 10.
9. Viboud C, Bjornstad ON, Smith DL, Simonsen L, Miller MA, et al. (2006) Synchrony, waves, and spatial hierarchies in the spread of inuenza. Science 312: 447–451.
10. Truscott J, Ferguson NM (2012) Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. PLoS Computational Biology 8.
11. Cauchemez S, Valleron AJ, Boelle PY, Flahault A, Ferguson NM (2008) Estimating the impact of school closure on inuenza transmission from Sentinel data. Nature 452: 750–U6.
12. Balcan D, Gonçalves B, Hu H, Ramasco JJ, Colizza V, et al. (2010) Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. Journal of computational science 1: 132–145.
13. Flahault A, Vergu E, Coudeville L, Grais RF (2006) Strategies for containing a global inuenza pandemic. Vaccine 24: 6751–6755.
14. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. Journal of statistical Mechanics -Theory and experiment.
15. Mills CE, Robins JM, Lipsitch M (2004) Transmissibility of 1918 pandemic inuenza. Nature 432: 904–906.
16. Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, et al. (2011) Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic inuenza. Proceedings of the National Academy of Sciences of the United States of America 108: 2825–2830.
17. Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, et al. (2006) Strategies for mitigating an inuenza pandemic. Nature 442: 448–452.
18. Moran PAP (1950) Notes on Continuous Stochastic Phenomena. Biometrika 37: 17–23.
19. Bavaud F (1998) Models for spatial weights: A systematic look. Geographical Analysis 30: 153–171.
20. Colizza V, Barrat A, Barthelemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. Proceedings of the National Academy of Sciences of the United States of America 103: 2015–2020.
21. Bonabeau E, Toubiana L, Flahault A (1998) The geographical spread of inuenza. Proceedings of the Royal Society B - Biological Sciences 265: 2421–2425.
22. Crepey P, Barthelemy M (2007) Detecting robust patterns in the spread of epidemics: A case study of inuenza in the united states and France. American journal of epidemiology 166: 1244–1251.
23. Balcan D, Colizza V, Goncalves B, Hu H, Ramasco JJ, et al. (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences of the United States of America 106: 21484–21489.
24. Lunelli A, Pugliese A, Rizzo C (2009) Epidemic patch models applied to pandemic inuenza: Contact matrix, stochasticity, robustness of predictions. Mathematical Biosciences 220: 24–33.
25. Merler S, Ajelli M, Pugliese A, Ferguson NM (2011) Determinants of the spatiotemporal dynamics of the 2009 H1N1 pandemic in Europe: implications for real-time modelling. PLoS Computational Biology 7: e1002205.
26. Eames KTD, Tilston NL, Brooks-Pollock E, Edmunds WJ (2012) Measured dynamic social contact patterns explain the spread of H1N1v inuenza. PLoS computational biology 8: e1002425.
27. Polgreen PM, Chen Z, Segre AM, Harris ML, Pentella MA, et al. (2009) Optimizing Inuenza Sentinel Surveillance at the State Level. American journal of epidemiology 170: 1300–1306.