Movements networks in epidemiological models: integration, analysis and application to commuting movements in France

Ségolène Charaudeau

UNIVERSITE PARIS DIDEROT (PARIS 7)
ECOLE DOCTORALE: FRONTIERE DU VIVANT

DOCTORAT

BIOMATHEMATIQUES

MME SEGOLENE CHARAUDEAU

MOVEMENTS NETWORKS IN EPIDEMIOLOGICAL MODELS: INTEGRATION, ANALYSIS AND APPLICATION TO COMMUTING MOVEMENTS IN FRANCE

Thèse dirigée par:

Dr Pierre-Yves BOËLLE

Soutenue le 26 juillet 2013

JURY

M. le professeur Pierre-Yves Boëlle (Université Pierre et Marie Curie) Directeur

Mme. la docteure Vittoria Colizza (INSERM) Examinatrice

M. le docteur Jean-Claude Desenclos (Institut National de Veille Sanitaire) Rapporteur

M. le professeur Khashayar Pakdaman (Université Paris Diderot) Directeur

Mme. la docteure Laura Temime (Ecole Pasteur-CNAM) Rapporteure

M. le professeur Jean-Christophe Thalabard (Université Paris Descartes) Examinateur

Funding

This work has been funded by the "Fondation pour la recherche médicale"

Website: http://www.frm.org/

Acknowledgements

First of all, I would like to express my deep appreciation and gratitude to my advisors, Dr Pierre-Yves Boelle and Dr Khashayar Pakdaman, for the patient guidance and mentorship they provided to me.

Thanks to my Thesis Advisory committee members, Dr Bertran Auver and Dr Simon Cauchemez, who followed the progresses of this thesis and gave me useful guidance and suggestions.
Thanks also to Dr Laura Temime and Dr Jean-Claude Desenclos, who accepted to review this work and to Dr Vittoria Colizza and Dr Jean-Christophe Thalabard, who accepted to be in my jury.

I would also like to thank my colleagues, from Institut Jacques Monod and from UMR S 707 , the interesting discussions we had together helped me many times to find solutions to some issues in my work.

On an unrelated note, I would like to thank Anne Roos-Weil and all the team of Pesinet, who gave me the wonderful opportunity of discovering public health field work in Bamako, and gave me the will of pursuing my work in this field.
Thanks also to all the teachers of the "Business and poverty" chair at HEC, whose courses helped me to understand how I could interface my interest in research and my desire to work for international development.

More generally, thanks to all the people I have met and discussed with in MakeSense network and in my doctoral school "Frontieres du Vivant": a lot of amazing people, with plenty of wonderful ideas and unlimited energy, who I always enjoyed interacting with.

I would like to take this opportunity to thank my parents, my sister Bénédicte

and my brother in law Nicolas, who have always been here for me during these four years, listening and encouraging me in difficult times. This thesis would not exist if they had not been here to give me strength.
In a similar vein, I would like to thank my friend Veronique, who has been a permanent support and who helped me a lot across difficulties and doubts. I would also like to add a very special thanks to Baptiste, who kindly proposed to re-read this thesis and has been a very important support during the last moments of the writing.

Finally, I would like to dedicate this thesis to my grandfather, Jean-Pierre Michel, his curiosity and his unlimited scientific knowledge have deeply inspired me and given me the desire to seek for answers. One of my main regrets is that he is not here anymore to see the accomplishment of this work.

Résumé en français

Les infrastructures de transport, liées à la géographie du territoire, ont une influence importante sur les interactions entre individus. La progression récente de la rapidité et de la fréquence des déplacements [144] ont notamment aidé à la diffusion à l'échelle mondiale de plusieurs maladies: HIV [54], SARS [30], H1N1 [56, 91], MRSA [40]. Dans ce contexte, il est d'une importance cruciale de développer des outils théoriques permettant de comprendre comment la mobilité humaine influence la propagation de maladies infectieuses.

Au cours de cette thèse, nous avons tout d'abord réalisé une revue systématique de l'utilisation des mouvements de population dans la propagation épidémique. Nous nous sommes ensuite focalisés sur l'étude des mouvements de navette et avons développé plusieurs outils pour comprendre leur influence: l'exemple de la diffusion de la grippe en France a été utilisé pour cette étude.

Revue systématique de la mobilité humaine dans l'épidémiologie des maladies transmissibles

La première partie de cette thèse a été consacrée à une revue systématique des travaux portant sur des modèles de propagation épidémique et intégrant des données de mobilité. Si le rôle des transports aériens dans la diffusion des pathogènes est aujourd'hui bien documenté, à l'échelle nationale/régionale, les mouvements de navette en revanche sont souvent intégrés sans que cela soit justifié. Dans la suite de cette thèse, nous avons donc cherché à comprendre l'influence que ces mouvements pouvaient avoir, en se focalisant sur le cas spécifique de la propagation de la grippe en France.

Analyse des chemins de propagation de la grippe

Tout d'abord, nous avons effectué un calcul de l'auto-corrélation spatiale des données de suivi de propagation grippale enregistrées par le réseau Sentinelles basée sur les mouvements de navette, afin de confirmer leur influence sur la propagation de la grippe. Grâce à des outils de comparaison d'épidémies basés sur un modèle métapopulation, nous avons ensuite montré qu'il était important

d'intégrer les mouvements des enfants et des adultes, qui sont différents. Cette étude nous a également permis de montrer l'existence de chemins de diffusion typiques, empruntés selon le point de départ de l'épidémie.

Analyse spectrale

Dans un second temps, nous avons effectué une étude analytique de l'influence des mouvements de navette en développant un modèle linéaire basé sur le précédent. Nous proposons dans cette thèse deux outils basés sur l'analyse spectrale de la matrice de ce modèle et du sous-espace vectoriel généré par ses vecteurs propres dominants: ces outils ont permis d'étudier le rôle de chaque canton dans la diffusion de la grippe d'une part et la dynamique globale du modèle d'autre part, ainsi que de confirmer les résultats précédents.

Summary of the thesis in English

Transportation networks have a major influence on the interactions between people. The worldwide diffusion of some disease, like HIV [54], SARS [30], H1N1 [56, 91] or MRSA [40] has for instance been eased by the recent increases in the velocity of transportation means and the frequency of movements [144]. In this context, it is of crucial importance to develop theoretical tools to understand the influence of human mobility on infectious disease propagation.
During this thesis, we have performed at first a systematic review of the use of population movements in models of epidemic propagation. Then, we focused on the national/regional scale, using commuting movements, and we developed several tools to analyze their influence on the propagation of influenza in France.

Systematic review of human mobility in epidemiology of infectious diseases

The first part of this thesis has been dedicated to the systematic review of works on epidemic propagation integrating mobility data. Nowadays, the influence of air travel on disease diffusion has been extensively studied: on national/regional scale however, commuting movements are often integrated in models without justification. In the following of this thesis, we investigated the influence of these movements: the example of influenza propagation in France was used to make this study.

Propagation pathways analysis

We started this work with the calculation of spatial auto-correlation of surveillance data from the Sentinelles network, based on commuting data, in order to confirm they had an influence on influenza propagation. Then, we developed tools based on a metapopulation model, to compare epidemics propagation, thanks to which we showed that both children and adults movements, which are different, should be integrated in models. In this study, we also evidenced the existence of preferential pathways for diffusion, that are followed depending on the seeding district of the epidemic.

Spectral analysis

We then performed an analytical analysis of commuting movements influence by developing a linear model based on the previous one. In this thesis, we propose two tools based on the spectral analysis of the matrix of this model and the study of the vectorial sub-space generated by its dominant eigenvectors: with these tools, we studied the role of each district in the propagation on one hand and the global dynamics of the model on the other hand. Previous results were confirmed by this second analysis.

Contents

1 Preface **12**

2 Background & definitions **16**

2.1 Why do we use models ? 16

2.2 The SIR model 17

2.2.1 Transmission rate 19

2.2.2 Recovery rate 19

2.2.3 Basic reproduction ratio 19

2.2.4 Extensions of the SIR model 19

2.3 Spatial models 21

2.3.1 Patch models 21

2.3.2 Distance transmission 22

2.3.3 Network models 22

3 Disease propagation in the light of human mobility **29**

3.1 Article 29

3.2 Observations 47

4 Influence of commuting movements on influenza propagation **48**

4.1 Introduction 48

4.2 Spatial autocorrelation of influenza incidence 49

4.2.1 Investigating the link between influenza propagation and commuting: spatial autocorrelation and model design . . 50

4.2.2 Observations and Perspectives 51
4.3 Influence of the network structure on the propagation 55
4.3.1 Relation between the structure of commuting networks and similarities between epidemic courses 57
4.3.2 Article 58
4.3.3 Observations, Supplementary informations and Perspectives 79
4.3.4 Structure of the network and similarities between epidemic propagation 81
4.3.5 Interpretation and perspectives 85
4.4 Conclusion 87

5 Districts role in the system dynamics 88
5.1 Linearized model 89
5.1.1 Linearization 89
5.2 Kernels 91
5.2.1 Kernel definition 91
5.2.2 Kernel analysis 93
5.3 Perspectives of use for the kernels 99
5.4 Conclusion 100

6 Analysis of the system global dynamics 102
6.1 Next generation operator 104
6.2 Eigenvalues and eigenvectors of the next generation operator 105
6.2.1 Isolated districts 105
6.2.2 Partial analyses for isolated areas 106
6.2.3 Perspectives 111
6.3 Conclusion 112

7 Technical development 113
7.1 Management of the simulations 114

7.1.1 OpenMOLE and simulations on grid 114
7.1.2 From stochastic to deterministic 115
7.2 Management of the results 116

8 Conclusion **117**

8.1 Influence of commuting structure on influenza propagation 118
8.1.1 Role of commuting in the propagation 118
8.1.2 Age related commuting 118
8.1.3 Underlying structure: attractors and basins of attraction 119
8.2 New methodology for network analysis 120
8.2.1 Early propagation 120
8.2.2 Global dynamics 121

Chapter 1

Preface

Following directly the evolutions of the industrial revolution, last century has seen a tremendous development of transportation means. The unprecedented increase of speed and volume of travel has led to deep modifications in human mobility [134]: in developed countries, for example, the number of kilometers daily crossed by an individual has increased by over 1000-fold in the last century [143]. The evolution of air travel in particular has been spectacular: the number of air travellers has grown at a rate of 9% a year since 1960 and is expected to keep growing at 5% for at least 10 years [136]. Similarly, due to the globalization of economy, shipping traffic has greatly progressed, increasing by 27% since 1993 [145]. Due to the development of international tourism and migrations, more and more people move temporarily or permanently on increasing distances: in 1990, 500 million persons crossed annually the border of a state (unpublished data of the World tourism organization). These evolution generate increasing contacts between distant populations [42]: as an example, the rate of contacts between Europe and Americas has constantly increased during past century and is nowadays higher than it has ever been.

Consequences on human health

The increase of mobility has major consequences on human health: a growing number of people are exposed to unknown pathogens [66] and more diseases become potential pandemics [134]. Similarly, the propagation of several diseases has been facilitated and accelerated by the development of transportation network and the evolution of human movements. The worldwide spread of cholera for example, has been directly linked to human travel: first, the pathogen has been introduced to the Baltic, Canada, USA and Mexico in the 1830s by the explorations of Russian troops, Irish immigration and Canadian exploration [35]. Then, during the 19th century and the beginning of the 20th, each pandemic of cholera increased in extent and intensity, following the expand of global transport system [126]. Human migration and air travel were also proven to be one of the main supports of HIV early propagation, which allowed its spread to between 100000 and 300000 people on the five continents between the start of the epidemic, in the mid-1970s and 1980 [99]. On smaller scales, the intensity of movements between them and urban center shapes the propagation of HIV to rural areas [93]. More recently, SARS, which emerged in the province of Guang-Dong, China, in November 2002 [118] propagated through airplane network to several countries worldwide [30] before the public health measures set by WHO stopped its diffusion in July 2003. Similarly, air traffic has been proven to be the main vector of international influenza propagation [10].

To halt the propagation of these infectious diseases, we need to understand how mobility affects the propagation and what type of interventions could be relevant. Among the numerous methods used to preview future evolutions of epidemics, modeling has been one of the most extensively used [52] and has proven useful in cases of real outbreaks, like the 2001 foot-and-mouth epidemic [51, 86]. Within this scope, we dedicated the first part of this thesis to the systematic review of mobility data uses in epidemiological models, which will be presented in the third chapter. This review showed that, due to the complexity of human mobility, its integration in models requires a strict selection of the

movements that will be inserted: the scale of study, the mean of transmission of the disease, its generation time limit the range of movements of interest for the model. In most models of transmission, choice has been made to focus on airplane or commuting movements: if airplane has been extensively studied and proven to influence international propagation of infectious diseases, the relevance of commuting movements has been less explored. Therefore, we dedicated the following of this thesis to the analysis of commuting movement influence on the propagation of diseases in France. Due to its mode of transmission and to the existence of many datasets documenting its propagation, we used influenza as a case study for this analysis.

The case study of influenza

Despite a low fatality rate, influenza is a major threat [115]: due to its genetic variability and the speed at which it spreads [100], it infects a large number of people every year, so that almost 20% of children and 5% of adults are affected by influenza each year [112]. Influenza epidemic occur on a regular basis, with recurrent pandemic outbreaks [34], like the 1918 pandemic, which killed around 40 million people in a year [104].

The speed of transmission of influenza, and its frequency of occurrence, which allows for the constitution of large databases detailing its propagation, make influenza a good example to understand how human mobility affects infectious disease transmission. For these reason, in the second part of this work, we used the case of influenza propagation in France to study how commuting movements influence infectious diseases propagation.

We started this work by analyzing the spatial autocorrelation of influenza-like illness surveillance data provided by the Sentinelles network and confirmed that they exhibited a spatial structure linked to commuting movements. Given this result, we investigated the influence of the commuting structure on disease propagation: to do so, we developed several tools to evaluate the similarity of

epidemic propagations and study the role of the networks nodes in the diffusion. In a first part, we developed a metapopulation model of propagation based on commuting movements: to analyze its simulation results, we defined two criterions, based on the similarities between epidemics, that were used to demonstrate the existence of recurring propagation patterns. Chapter four will present this work.

In a second part, we proposed a linearized version of this model and performed its analytical study, using tools based on the spectral analysis of the model equivalent matrix. This study gave results on the global dynamic of the system and on the specific role of different cities in the propagation. These two works will be exposed in chapters five and six.

The studies performed in this thesis required the analysis of large datasets and the storage of large amounts of data. To complete them, we faced a technological challenge: the solutions that were developed to overcome it will be presented in the seventh chapter of this thesis.

Chapter 2

Background & definitions

In this chapter, we will briefly describe mathematical models that have been developed to simulate the transmission of infectious diseases. This notably includes the SIR model, on which most of our work was based. We will present how spatial information can be added to the model and show some examples of spatial models.

In a second part, we will focus on networks, a mathematical object that is nowadays often used in epidemic models to describe spatial relationships. Some methods to recognize two specific types of networks, namely the scale-free and the small world networks, will also be presented, as these techniques will be used in following chapters.

2.1 Why do we use models ?

Nowadays, several tools are available to fight the propagation of an epidemic: distribution of curative or preventive treatment, vaccination, isolation of infected individuals, closing of public spaces... However, depending on the disease and its characteristics, like the mode of transmission, the probability for a contact to be infectious, the time of incubation or the time before remission, do not have the same impact on disease propagation (in terms of number of infected

for example): the optimal strategy may ask for the use of one specific technique, or a mix of several. Testing all combinations is impossible: to overcome this issue, one can used mathematical models to understand disease propagation and decide which policy should be adopted to fight it. Different problematics can be addressed by the use of models: first, they can be used to make purely theoretical analysis to understand the influence of a parameter or the structure of the model on the size of the epidemic. When data documenting the propagation of epidemics are available, different models can be compared to data to determine which one gives the best approximation of reality and describes the most accurately the mechanisms of propagation. Once the best model has been found, it can be used to test different methods to halt epidemic propagation and select the most efficient.

When a model is constructed, one has to adopt a trade-off between accuracy and transparency: accuracy usually improves by adding more complexity to the model, but in the process, its transparency decreases. Conversely, simplifying the model to ease the understanding of its behavior and the role of each parameter decreases the proximity of the model to reality. Therefore, simplifying assumptions on the system behavior have to be made to develop a model on the behavior of the system.

In next part, we will present one of the simplest epidemiological model existing, the SIR model, and some of its parameters. Then, we will take an interest into the integration of spatial information in this model.

2.2 The SIR model

The SIR model divides the population according to their infectious state, between susceptibles (individuals who have not been exposed to the disease and can be contaminated), infected (currently colonized by the pathogen) and recovered (individuals previously infected and cured). During the simulation, individuals can pass from a state to another. The evolution of the system is

assured by the transition of susceptibles in the infected compartment as they make contact with infected individuals who transmit them the disease and by the transition of infected in the recovered compartment when they are cured (Figure 2.1)

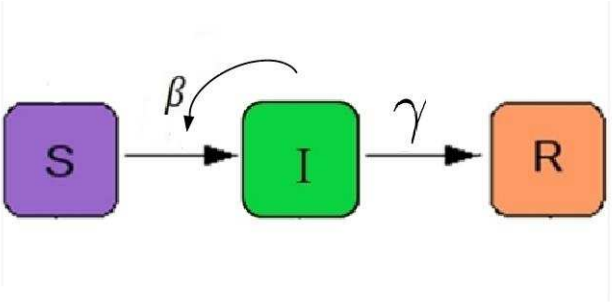

Figure 2.1: **The SIR model**

The SIR model makes the simplifying assumptions that the compartments are homogeneous: in the susceptible compartment, all individuals have the same susceptibility to the disease. In the same way, all infected individuals have the same infectivity and the same probability to transmit the disease when encountering a susceptible. All infected recover at the same rate and get off the infectious compartment at the same speed. It is also supposed that all pairs of individuals have the same probability of establishing a contact.

The system dynamics are regulated by the following equations:

$$\begin{cases} \frac{dS}{dt} = -\beta \frac{SI}{n} \\ \frac{dI}{dt} = \beta \frac{SI}{n} - \gamma I \\ \frac{dR}{dt} = \gamma I \end{cases} \tag{2.1}$$

β and γ, respectively named rate of contact and recovery rate, describe the natural history of the infection. With them can be defined the reproduction ratio, or R_0, which describes the capacity of the disease to invade the population or not.

2.2.1 Transmission rate

The rate of contact, often noted β, is the product of two parameters: the contact rate, c and the transmission probability, p. The contact rate, is the probability for two individuals, from any compartment, to make a contact: it is constant, and equal for all pairs of individuals. Similarly, the transmission probability, the chance for a susceptible to be infected when it makes a contact with an infected individual, is constant and equal for all pairs of susceptible-infected.

2.2.2 Recovery rate

The recovery rate, often noted γ characterizes the speed of recovering of an infectious individual: it is constant through time and equal for all individuals. Its reciprocate, $\frac{1}{\gamma}$ determines the average length of infectious period.

2.2.3 Basic reproduction ratio

The basic reproduction ratio, R_0, is the number of individuals to which a single infected can transmit the disease in a fully susceptible population. When $R_0 > 1$, each infected has more than 1 descendant and the disease tends to propagate to a growing number of individuals, but when $R_0 < 1$, it tends to quickly stop its propagation and disappear from the population. In the SIR model, the basic reproduction ratio can easily be calculated with the transmission rate and the recovery rate, with following equation:

$$R_0 = \frac{\beta}{\gamma} \tag{2.2}$$

2.2.4 Extensions of the SIR model

The natural history of some diseases cannot be reproduced with the SIR model: other compartmental models have been proposed to give a more accurate description of a larger set of diseases [88]. For example, in the SEIR model (E standing for "exposed"), an additional compartment, which members are in-

fected but not not infectious is added: this structure enables the modeling of a latency period. To simulate diseases for which infection does not confer immunity, the model SIS has been developed: after recovery, individuals return in the susceptible class instead of entering the recovered one. In case of temporal immunity, the SIR model can be used by adding to it a lost of immunity rate, describing the speed at which individuals go out of recovery class to re-enter the susceptible one.

The compartments of the model can also be divided to take into account the

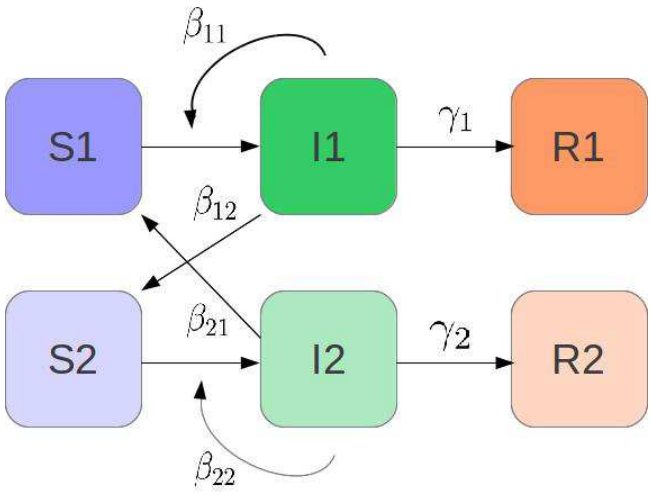

Figure 2.2: **An example of compartmental SIR model**

heterogeneity of the population (age, gender, susceptibility to the disease...) (Figure 2.2). For example, the infectious class can be divided between symptomatic and asymptomatic individuals: to model the different behavior of these classes, different parameters will be used to simulate their interaction with the susceptible class. The existence of heterogeneity of contacts in the population can also motivate a division : for sex transmitted diseases, for example, a division of classes between genders and sexual orientations can be adopted in order to represent accurately the potential interactions.

Geography is an important source of heterogeneity in the interactions between people, as individuals at a small distance have a higher probability of

interacting. Moreover, the division of the territory in cities, linked by a networks of highways and public transportation, shapes the way people interact. Several solutions have been developed to integrate this spatial information in models: in next part, we will present some of them and focus on networks, the mathematical objects used in some models a,d that will be used in this thesis.

2.3 Spatial models

There are diverse ways of taking into account the heterogeneities induced in the interactions between individuals by their spatial repartition. Depending on the disease and its mode of transmission, the relevant information, that should be integrated in the model, will be different. The three models that we present here, are three different ways of using spatial information, but focusing on different aspects of it, namely the division in groups, the distance and the network of interactions.

2.3.1 Patch models

Patch models have been the first occurrence of spatial models used for epidemiology [20, 64, 84]: they have been extensively used to simulate the propagation of measles and helped in the comprehension of some specific patterns observed in the district-level dataset describing its propagation in England since 1948. In patch models, the population is divided in sub-groups: the force of infection exerted on an individual depends on the distance between its patch and the others and on the prevalence of infection in all patches. The same force is exerted on all individuals of the same patch.

In patch models, the focus is set on the notion of community, represented by the patch, which can be understood as a city or a neighborhood. These model rely on the simplifying hypothesis that the probability of interacting with another individual mainly depends on the community of living: the place of residence induces the affiliation to a group and all individuals gathered in the

same group have equal probability of making contact.

2.3.2 Distance transmission

In distance transmission models, instead of simulating the evolution of an entire class, each individual is modeled separately. An individual can infect any other in a range surrounding it, with a probability decreasing with distance. Such models are called individual-based, or IBM, and they have notably been used to model the propagation of foot-and-mouth disease [51, 87]: in these cases, individuals were farms and exerted a force of infection on their neighbors caused by the transportation of bovines and cattle between them.

Distance transmission gives an accurate description of the decreasing of the force of infection caused by distance, but doesn't account for the fact that distance is not the only factor shaping the network of contacts. As an example, cities in which many businesses are implanted attract more commuters than others, increasing the probability for their residents to meet outsiders.

2.3.3 Network models

When distance models and patch models represent the interactions between individuals with general laws depending on the distance or the individuals characteristics, network models focus on the contacts between individuals or groups. To represent them, a mathematical object which study have been initiated by graph theory [20, 70, 142] and social sciences [94, 130, 140] is used: this structure is composed of nodes and edges connecting them and can be used to model a population and the presence or absence of contacts between individuals or subgroups. These models have been extensively used to understand the transmission of a disease due to social contacts [45, 48, 139] or sexual contacts [106].

Despite social scientists, who were more interested by the reason behind the connections, epidemiologists using networks models have focused on the structure of networks and have been using several tools and measures to describe them.

Common measures

Given the complexity of real-life networks, describing them in a meaningful way is a difficult challenge: for several decades, researchers have introduced parameters, based on networks measurable properties, in order to extract important information from this complexity [4, 109, 111]. Here, we will present the definition of the parameters that have been used in this thesis.

- Degree The degree of a node is the number of its neighbors: in directed networks (i.e, networks in which edges have a direction), degree is subdivided between the number of incoming edges and of outgoing edges. To describe the entire network, the degree distribution, often noted $P(k)$ has been introduced: it is defined as the probability for a node to have k neighbors. Degree distribution has been often used to describe social networks, especially for epidemiological purposes, as the number of contacts of a node gives an important information on its role in the network and its capacity to transmit an infectious disease. The heterogeneity of degrees has a great influence on the probability for an infection diffusing on the network to become an epidemic [13, 32]: epidemics are less frequent when all nodes have the same amount of neighbors. Degree correlation, defined as the propensity of a node to have a link with a node of same degree, has also been proven to influence epidemic propagation: the high correlation between the degrees of connected nodes in networks of sexual contacts is for example a key element to understand HIV propagation [67]. Speed and extent of an epidemic are also influenced by degree correlation [19].

- Distance The distance d_{ij} between two nodes i and j is defined as the minimal number of steps needed to go from one node to the other, following the edges of the network. A pathway between i and j of minimal length is called a shortest path: there are often many shortest path for each pair of nodes. From this notion can be defined the average path length, the mean distance between 2 nodes of the network: it measures the typical distance between any

pair of nodes. The notion of distance can be of crucial importance to understand disease propagation: percolation approaches have shown that a short average minimal distance can have many consequences on the diffusion on a network. When only a short number of steps are needed to reach a node from any starting point, diseases can spread more quickly [105].

Distance can also be used to characterize the importance of a node in a network: the more shortest path pass through the node, the more it is central to the network. This notion is quantified with the parameter of betweenness centrality, which measures the number of shortest paths passing through a node. Identifying the nodes with high betweenness is important for disease control, as due to their central place in the network, they are likely to be infected early in an epidemic and to transmit the disease to many others: they are thus key targets for intervention [18].

- Clustering Clustering evaluates the propensity of nodes to form communities inside a network. It can be measured with the clustering coefficient, which is the probability that 2 nodes connected to a third one will be connected one with each other (2.3).

$$\phi = \frac{3 * Number\ of\ triangles\ in\ the\ network}{Number\ of\ connected\ triples} \tag{2.3}$$

where a connected triple designates a group of three nodes where at least one is connected to the others. Clustering coefficient can also be defined for a single node i: in this case, it evaluates how the nodes connected i are connected to each others.

This parameter gives an insight on the mixing among the network: in highly clustered networks, nodes tend to aggregate in separated communities, with few contacts between them. This parameter is an important information to predict the diffusion of an epidemic on the network: in highly clustered networks, the size of clusters and quantity of contacts between them strongly influence the final size of the epidemic [81]. Moreover, clustering as been proven to be the

main determinant of epidemic growth rate: R_0 tends to be lower in highly clustered networks, diminishing the epidemic growth rate [103].

Types of networks

Several types of networks have been used in epidemiological models: we will briefly present here the 4 more commonly used, namely lattices, random, scale-free and small-world networks.

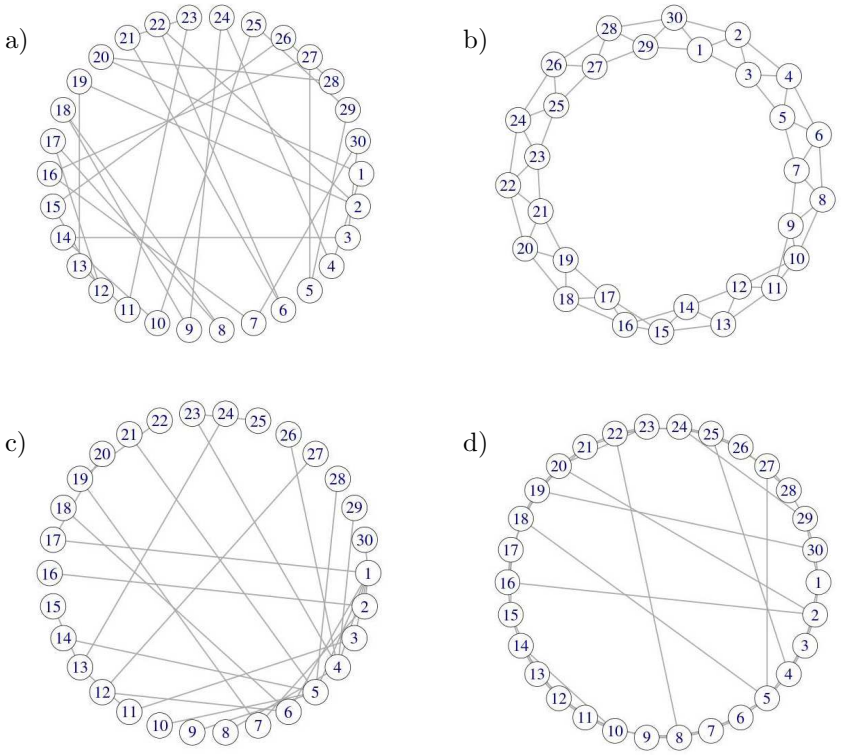

Figure 2.3: **Examples of networks**
a) A random network, b) A lattice, c) A scale-free network, d) A small-world network

- Random networks In these networks, connections between nodes are randomly formed, without consideration of nodes spatial position (Figure 2.3-a): therefore, degree distribution follows a gaussian law and the heterogeneity in the degrees of nodes is low. Due to the randomness of bounding, random networks

also lack clustering and their average path length is usually low.

If some works have been done to understand the propagation of epidemics on random networks, like [81] who showed that the initial transmission rate was inferior in a random network than in a random mixing model, random networks are more commonly used as a comparison point in studies using other types of networks, in order to highlight the influence of their characteristics on the diffusion .

- Lattices Lattices are constructed on a fixed grid of nodes, which all have the same number of contacts, usually 4 or 8, depending on the number of dimensions the lattice is built on (Figure 2.3-b). As connections highly depend on the spatial position of nodes, clustering is often high in lattices. Connections are only local and lattices lack of long distance edges, connecting distant clusters: therefore, the average path length is high. Due to the absence of long-distance connection, the propagation of an epidemic on a lattice is purely wave-like, with an initial growth smaller than the one observed on random networks [64].

Lattice have notable been used to study plant diseases, like in the well known example of forest fire [9]. In these cases, geographical constraints limit the number of possible contacts to the immediate neighbors of an individual, making lattice an appropriate choice of model for the network of contacts.

- Scale free networks In many observed networks, the distribution of degree is far from homogeneous: when most individuals have few contacts, a small number of people have many: this repartition is mainly caused by the natural process of social contacts generation, where individuals tend to bound preferentially with people having already many contacts. Such process of contact generation creates specific networks, called scale-free networks (Figure 2.3-c), in which degree distribution follows a power-law [15].

In scale-free networks, hyper-connected individuals play a disproportionate role in epidemic propagation: such people are both at great risk of being infected and highly probable to transmit the disease to others. As an example,

the persistence of sexually transmitted diseases in populations were most individuals are involved in monogamous relationships is caused by the presence of super spreaders, highly susceptible to these diseases [72]. Similarly, during the SARS epidemic, most infections were caused by super spreaders [123]. The identification of super spreaders and their targeted vaccination can significantly decreases the attack rate of an epidemic [116].

- Small world networks Small world networks have a structure similar to a lattice, with addition of some long distance links. They are characterized by a strong clustering coefficient and a small average shortest distance, caused by the presence of long distance links which connect clusters of neighbors, interconnected nodes. This type of network captures both the local component of interactions and the possibility of rare long-distance contacts that can be encountered in social interactions: the spread of an epidemic on a small world network will associate a wave-like behavior and occasional jumps to uninfected areas [105].
A network can be identified as small-world if it exhibits both of its characteristic, namely an equal or higher average shortest path length and a higher clustering coefficient than a random network with the same number of edges and vertices [76]. This can be measured using a parameter, S, such as:

$$S = \frac{C_N}{l_N} * \frac{l_{rand}}{C_{rand}} \tag{2.4}$$

where C_N and C_{rand} are the clustering coefficient of the network and a random network, and l_N and C_N their average shortest path length. If a network can be considered as small-world, this parameter will be superior to 1.

*****

In this chapter, we have seen that several structures have been proposed to insert some spatial information in an epidemiological model. The choice of a model notably depends on the mode of transmission of the disease. In next

chapter, we will see how these structures have been used to represent a particular type of spatial information: the human mobility.

*****

This chapter has been based on:

- The review "Networks and epidemics models", by M.Keeling and K.Eames, J. R. Soc. Interface, (2005) 2, 295-307

- The review "Networks and the epidemiology of infectious disease", by L.Danon, A.P.Ford, T.House and al, Interdiscip Perspect Infect Dis (2011) 284909

- The review "Large-scale spatial-transmission models of infectious disease", by S.Riley, Science, (2007) 316(5829), 1298-1301.

- The book "Modeling infectious diseases in humans and animals", by M.Keeling and P.Rohani. Princeton University Press, (2011)

Chapter 3

Disease propagation in the light of human mobility

Human mobility is a complex system [59]: movements at different scales, frequent or rare, predictable or random form a dense network, difficult to model. To integrate mobility data in epidemiological models, simplifying assumptions have to be made: frequently, models are focused on one specific type of movements, or one scale. During the first part of this thesis, we have realized a systematic review of existing epidemiological work using models of mobility. This work gave us an insight on the structure of existing datasets concerning human mobility and the techniques developed to insert this information in epidemiological models.

3.1 Article

Review: Human mobility and the spread of infectious diseases

Review : human mobility and the spread of infectious diseases

Segolene Charaudeau

1 Introduction

Disease propagation is intrinsically a spatial phenomenon : geography and urban development shape the repartition of individuals on the territory, and their interactions, impacting pathogens transmission. Within this scope, movements of agents can influence disease propagation, either by facilitating it, or by diffusing the pathogen to uninfected populations. Therefore, epidemic forecast crucially depends on our capacity to model the movement of individuals on differents scales and to understand its influence on infectious diseases propagation. Several solutions have been proposed to address this issue, from simple compartmental models to complex systems : with technological progresses in terms of storage capacity and calculation time, epidemiologists have been able to developp more and more complex spatially explicit models, integrating more precise data. This review aims at studying how mobility data have been used in epidemic models to describe disease propagation. In the first part will be proposed a small historical outline of the joint evolution of transportation means and disease propagation. The second part will briefly present the type of datasets describing human mobility that have been used to parameterize epidemiological models and the third part will present the techniques and solutions that have been proposed to model different types of movements.

2 Historical perpective

Disease propagation has radically changed during history, following the deep changes in transportation means. During pre-industrial times, populations were relatively isolated : by foot and by cart movements were the main vectors of spatial diffusion. It has for example been proven that, in 14^{th} century, the propagation of Black Death from southern to nothern Europe occured at a speed of 200 to 400 miles an hour, which is coherent with a transmission via by foot or by horse travels [1].
The progressive development of new means of transportation, and the rise of long distance travel has modified the propagation of many diseases and, in some cases, has even exposed entirely susceptibles populations to new pathogens. From 15^{th} to 18^{th} century, the rush for new land discovery led by Europeans exposed many people to unknown pathogens, like smallpox or measles, brought to Americas during their exploration, or like malaria and tropical diseases to which European travellers were ewposed during African colonization [2]. On a more daily basis, the rise of long distance trade also facilitated the propagation of many

diseases : trade caravans and religious pilgrimage were for example privileged means for plague and smallpox propagation [3]. The development of transport and communications even modified disease propagation at the smaller scale of a single country : in United States, the propagation of cholera between 1832 and 1866 evolved from a wave-like propagation to a propagation oriented towards main cities as railway transportation became more and more efficient [4].
Even if transportation means have improved during all last 5 centuries, making long distance travel more frequent, the most important changes have occurred during the last fifty years, where global traffic has increased in a spectacular way. Indeed, the volume and speed of traffic have grown so much, that human mobility in high-income countries has increased by 1000 fold in the last 40 years [3]. The extremely quick development of air travel has been a major cause of this evolution : indeed, since 1960, the number of air passengers has grown of 9% each year [5]. Last century migration of people from countryside to cities, and the quick growth of urban areas it has induced, is also an important concern for disease propagation. As 52% of world population now lives in urban context [6], the density of population in the biggest cities increases, multiplying the contacts between individuals, which sets a favorable context for disease transmission. Understanding human mobility inside urban environments thus becomes another important issue in the fight against epidemics, as it sets up the network of interaction between individuals.

3 Data on human mobility

Gathering data on mobility can be done using two different approaches, either by focusing on the locations between which movements occur, or by following the individuals. Following several individuals in their daily movements necessitate either a complex technological infrastructure or a tedious reporting lying entirely on the good will and the memory of participants. The first approach can be implemented more easily, as the a priori selection of the locations of interest permits to avoid the issue of following individual trajectories. Historically, the second approach have thus been favoured, as technological limitations made it difficult to implement the first. However, with the developpement of technologies, and especially of geolocalization, more and more datasets were constituted with the first approach : collected information is more detailled and less rigid, as unusal trajectories including unplanned locations can be registered with this method. However, information is also more complex and its interpretation is harder than for datasets constitued with a selected set of locations. In following part, we will present datasets on human mobility constituted by both methosds : first we will present datasets focused on fixed locations and then datasets constituted of individual movements.

3.1 Movements between fixed locations

Human movements form a complex, diverse system, composed of travels for diverse causes (migration, business travels, students exchanges, vacation...) and using diverse means of transportation (train, plane, boat...). Due to this diversity, studying mobility is complex and obtaining exhaustive data difficult : given

the variety of individual behaviors, most datasets are constructed by focusing on a single mean of transportation, or a single type of movement and approximate the diversity by describing its main, often repetitive, patterns. This is for example the case of many public and commercial datasets of transportation companies, that were used in the first epidemiological models studying mobility. At first quite scarce, due to limitations caused by technology, those datasets became more and more complex. Airplane travel and commuting, both involving massive and repetitive movements, have been the movements most used for epidemiological modelling.

3.1.1 Airplane mobility

At first limited, datasets describing airplane movements provided by commercial companies became gradually more complete, notably by including more airports and by keeping a more exhaustive track of travellers. This allowed for a better comprehension of the complex, multi-scalar structure of the airplane network : due to the low impact of geographical constraints on its development, airplane network exhibits a a small-world structure, with a scale-free distribution of outgoing degree [7]. It also presents a high degree of clustering : most airports are gathered in communities, which communicate through specific airports, through which pass all outgoing traffic.Thus, most connected nodes in the network are not necessarily the most central ones. Indeed, most of these nodes are hubs in their community, but not global hubs. Non-hubs nodes have mainly connections within their community, and passengers leaving them have to pass through the local hub to travel out. Identifying the nodes connecting communities is of a crucial importance to understand travelling patterns and diffusion phenomena on the network.

3.1.2 Commuting movements

More sensitive to geographical constraints, commuting networks don't exhibit the same structure as airplane network, but some common pattern can be observed, like the power-law distribution of nodes degree. This structure can be explained by the process by which commuting networks are formed. Indeed, as some locations attract more commuters than others : main cities tend to have significantly more schools and workplaces and drain most of the travellers. The comparison of several commuting networks, in Germany, Italy and Sardinia, with themselves a few years before showed that they actually evolved by adding new links between connected cities [8, 9]. Creating networks by successively adding links between highly connected nodes tend to create networks which degree distribution follows a power law [10]. However, despite this characteristic, commuting networks don't exhibit a scale free structure : the creation of new links between cities is constrained by the means of transport used for commuting, and by geography (the number of incoming individuals that can pass through a limited space is itself limited). Real-life commuting networks are thus not scale-free [8, 11, 12].

3.2 Individualized movements

During past years, the development of technologies has gradually permitted a more and more precise follow-up of identified individuals, and a better insight into individual behavior. The record of mobile phone data [13–15], dollar bills movements [16] or public transportation use [17, 18] gave a distribution of distances crossed by humans, and evidenced the predominance of local movements on long distance travel. The distribution of distance covered by travel exhibited a power-law behavior and individuals mainly moved in a restricted area [13, 16, 17]. Using the "Where's Georges ?" study of US dollar bill circulation, [16] also showed that human movements was composed of long periods of stagnation in a restricted perimeter, interspersed with long distance jumps. This behavior can satisfyingly be described with a truncated Levy flight model, an improved version of the French mathematician Paul Levy model of random walk with a power law distribution distances, where unlikely larger movements are suppressed. However, this random description neglects an important characteristic of human movement : its reproducibility. The analysis of mobile phone trajectories showed that individuals tend to execute 93% of their movements in a restricted set of locations that are frequently visited [14, 15]. However, despite the individuality of each one movement pattern, they merge in a single spatial distribution, making truncated Levy flight a good approximation of human movement [13].
At each time, the choice of a destination for an individual is influenced by both the distance between its current location and the different spots of interest, and the attractiveness of these spots. Indeed, the workplaces, schools and residential areas of a city don't attract the same number of individuals : some of these spots act as major hubs, draining most of the moving individuals [18]. A study on 3 months data of mobile phone activity in Boston city showed that urban movement was both repetitive and focused on the most attractive spots of the city [19], making it mostly predictable. Most movements are either morning or evening movements of commuters going from or back home : both of them can be described as simultaneously diffusive and directed, as they diffuse away from repulsive spots (either home or workplace) towards attractive ones [20]. However, distance remains a decisive factor in the choice of a destination, and individuals privilege short distance travels [21].

4 Mobility networks in epidemic models

Epidemiologists have not waited for the completion of databases on human movement to study the impact of mobility on disease transmission : metapopulation models, first introduced in ecological litterature to analyze problems such as extinction or recolonization [22–24], have quickly been adapted to epidemiological problematics, notably to study childhood diseases, like measles [25–30]. Despite their simplicty, these models gave interesting results on the role played by mobility on propagation, and explained observed phenomenons such as synchronization of epidemics between cities in pre-vaccine area and its disappearance after vaccination [29, 30], waves of infection [27] or persistence of measles [26].
Progressively, with the constitution of detailled databases describing the movements of individuals at different scales, models have been developed to describe

specific situations. At first, the precision of data was limited by thecnology : movements were aggregated in few flows, occuring on large scales. First models were thus addressing the issue of movements on large to global scales. With the progression of technology, the precision of data increased, describing movements on smaller scales : epidemiological models have gradually included these information, simulating infectious disease propagation on more local scales.

4.1 Modelling disease propagation on a global scale

First implementations of models including mobility information were based on very aggregated datasets, in which few flows between a restricted number of destinations were documented. Gradually, larger datasets, including more locations have been used. In parallel of this evolution, the improvement of computer technology has permitted the development of more complex compartmental models, in which travellers were identified, instead of modelling the influence of travel as a force of infection, as it was previously done.
The flows described in first datasets were large scale ones : progressively, more local flows, like commuting, have been described and included in models, at first in models focusing on national or local scales, then in multi-scalar models.

4.1.1 First models on restricted datasets

At first, a metapopulation approach was employed to build the models, using the same methods as the ones of already existing patch models. Due to the limits of the technology available at the time, first datasets were imprecise and averaged the movement of individuals in some major tendencies : models thus included a small number of patches. The work on global influenza propagation through airline network of [31] was one of the first implemention of a metapopulation model using transportation data. This SIR metapopulation model included 52 cities of different continents, linked by a network of interaction calibrated with data from the international air transport statistics. The evolution of the epidemic was directed by a difference equation, parameterized using national data of 1968 influenza morbidity and, despite the limited number of cities included in the model, it was able to successfully mimic the real course of the epidemic. Several works later re-used this first model, and adapted it to other transportation data [32, 33], to other set of cities [34, 35] or to study the propagation of other diseases, like smallpox [36]. The reproduction of simulations conducted by [31] with updated airline transportation data from 2000 by [33] showed how much the increase of worlwide air traffic deeply modified influenza diffusion : the disease propagated much faster between countries, and reached each hemispheres simultaneously, contrary to what [31] observed with transportation data from 1968.

4.1.2 Larger datasets

Following the increase of computational capacities and the gathering in 2002 of a large dataset on movements of airplane travellers by the International Air Transport Association (IATA), new models were developped to study more exhaustively the influence of air travel on worlwide disease propagation [37–40]. These works introduced the use of stochastic Langevin equations to have an

insight into the diversity of possible epidemic pathways : the high variance exhibited by simulations results highlighted the sensitivity of epidemic diffusion to stochastic fluctuations. At first, the model was parameterized to simulate SARS epidemics [37] : propagation was found to be higly replicable, due to the heterogeneity of airports connectivity. Later, its adaptation to other diseases, like influenza [38] showed that this behavior also depended on the characteristics of the disease. Indeed, H5N1 propagation on the same network exhibited different behaviors, as the replicability of epidemics depended on the city it started in : epidemics initiated in poorly connected airports were more replicable than the ones starting in the hubs of the network. However, some preferential pathways of propagation were also evidenced, highlighting the existence of a backbone of main transmission axes in the network. This result was later improved in [40], which refined this backbone definition, by introducing seasonnality in transmission parameters. The IATA database recorded information on the number of travellers, but also on the mean duration of a stay abroad : both information were used in previous models, but no individual follow of the travellers was introduced in the model and the memory of individuals original city was not kept.

4.1.3 Identification of travellers

[41] proposed a model introducing a follow-up of travellers, who eventually returned to their city of origin : and confirmed the influence of the network connectivity on the propagation. The model was calibrated with the data of the official Airline Guide (OAG) of 2010 and included a precise description of the variations induced by seasonal climate change on influenza transmissibility, which permitted an updated prediction of influenza transmission patterns, notably of the differences encoutered in the 2 hemispheres. Cities in northern hemisphere, more connected, experience influenza simultaneously, and they share a common influenza season, while epidemics are less ssynchronized in southern hemisphere, due to the smaller interconnection of airports. Patterns of seasonality were also found to have an important influence on the timing of influenza apparition in every city, as the timing of southern and nothern hemisphere contamination depends on the location of first cases.

4.1.4 Role of Commuting movements

The crucial role of airplane network in the global propagation of influenza evidenced by previous models has been confirmed in 2009, during the H1N1 pandemic : a strong correlation between data of international flows from Mexico and imported H1N1 cases was found, confirming the role of airplane travel in the diffusion of the disease [42]. However, even if airplane travel is indubitably important to understand disease propagation, these models neglect more local movements, which also have an influence on the dynamics of epidemics [43]. Daily movements from workplace to home constitute a large part of human mobility, and their influence on epidemic propagation, mainly influenza and smallpox, have also been widely studied [44–46]. In [46], Viboud et al. performed a comparison of US 1972 to 2002 mortality from pneumonia or influenza data to several measure of domestic transportation, among which workflows, internal air flows and long-distance travel, and found that commuting was the

best predictor for influenza propagation, confirming the important role played by commuting movements.
As commuting often happen on shorter scales than airplane travel, obtaining an exhaustive description of its structure is often complicated : to reconstruct commuting movements with incomplete datasets, [46] introduced gravity networks models, as developped by transportation theory. Based on the propension of individuals to choose their commuting destination on distance and size of the communities, the modelled flow C_{ij} between 2 communities i and j depends on the numbers of inhabitants in i, N_i and j, N_j, and on the distance between the cities d_{ij} [47].

$$C_{ij} = \theta \frac{N_i^{\tau_i} N_j^{\tau_j}}{d_{ij}^{\rho}} \tag{1}$$

Parameters τ_i, τ_j and ρ calibrate the influence of each variable on the intensity of commuting flow and θ is a proportionality constant : they are fixed using datas on commuting.

4.1.5 Multiscalar model

Commuting movements on a global scale were first intoduced in the computational platform GLEam [48] : using data from the IATA base and commuting data gathered in 29 countries, Balcan and al. developped a model including 3362 subpopulations over the world, with the objective of studying the respective influence of these movements on international influenza propagation. If the IATA database contituted an exhaustive description of airplane movements, commuting data could'nt be gathered in every country : missing data were simulated using a gravitationnal model. The international propagation was found to be little influenced by commuting movements : as the final size of the epidemic on each continent was not modified in the abscence of commuting, air travel alone seems to be sufficient to successfully predict the final attack rate of influenza on each continent. However, some differences appeared at smaller scales : in the abscence of commuting, epidemics in American desynchronized and the propagation inside the country was slowed down. If both types of movements thus seems to influence disease propagation, the scale at which it can be relevant to include them seems to be different. Airplane transportation is sufficient to study international propagation, while commuting movements, as they influence the relative timing of epidemics episodes, seems to be important to study more local scales.

4.2 Diffusion on a local scale

Most of the time, national or more local movements are described in aggregated datasets : however, more and more individualized data are available, specially to describe local movements, inside a single city for example. The techniques used to integrate these data in models differ depending on their nature : if using aggregated data gathered on national scale can be done with techniques similar as the one encountered for global scale, using individualized data calls for the developpement of other methods.

4.2.1 Propagation at national scale

Many models describing propagation on national scale are compartmental, like global scale models. However, as the number of people in these contexts is smaller than in global models, other methods, demanding more computation, and that would have been impossible to simulate for larger populations, have also been proposed.

- Compartmental models To study more specifically propagation on smaller scales (national or regional), models have been developped, using data on commuting movements and demographic data on the population studied. They demonstrated the influence of commuting networks structure on local propagation. The typical structure of commuting networks, characterized by the presence of highly connected hubs and a strong clustering based on spatial proximity, causes an important variance in nodes connectivity : due to this diversity, the probability for a disease to diffuse to the entire network and its speed strongly depends on the initial city infected. In United States, for example, [46] showed that the synchronization of epidemic episodes in different cities depended on the connectivity of the initially infected city. The repartition of movements on the territory, notably of long distance movements, are also determinant for the propagation : despite their scarcity, long distance movements are nevertheless crucial for epidemic survival, as they enables diffusion of the disease to distant patches [44].
In most works using gravity models to represent commuting, the coupling between subpopulations linked by workflows is modelled by the share of a part of their force of infection [45, 46] : propagation happens as if the pathogen was directly diffusing between the cities, proportionnaly to the number of commuters exchanged, while the commuters themselves are not identified. However, the individuality of commuters is an important characteristic of commuting, as each individual daily repeats the same movement back and forth between its home and its workplace. Models representing commuting as a force of infection do not consider this individuality, and often overestimates epidemics spatial synchrony and the number of cases at their peak [49]. A way to avoid this issue would be to use compartmental models separating individuals depending on their commuting movements.
If metapopulation approach have been often used [45, 46, 50, 51]. the recent progress of data gathering, and the completion of precise datasets also enabled the development of individual based models, which had been known and used for more than 30 years [52], but not at their fully potential [53] due to the scarcity of available data.

- IBM models In most individual based models, individuals are gathered in groups of interaction, like a school or a household, where privileged transmission is supposed to occur.Each individual is, depending on its personal characteristics, such as gender or age, allocated to a household, a workplace and sometimes a leisure space. Demographic data and descriptions of the spatial repartition of residences and workplaces are used to allocate individuals in a way mimicking the population of interest. The interpretation of data, and their translation into a model design, can be a crucial issue in the conception of IBM models. The unavailability of some informations leads to the necessity of making assumptions

to generate the missing data : for example, [54] and [55] faced the same issue to generate the workplaces of their models. The only information they had was the average workplace size : [54] chose to make all workplaces the same, when [55] chose to generate a set of workplaces using a gaussian distribution for their sizes. Given the impact that the heterogeneity of the network of interaction has on disease propagation [38,46], such a choice can lead to differences in the results of simulations. Moreover, [56,57] highlighted the importance of using accurate and precise data to describe a population with their work on influenza propagation in Europe : the simulations shown that the speed of an influenza epidemic in a country strongly depended on the demographic structure of its population and its characteristics (vacation planning, matrix of contacts between generations, employment rate...). The use of erroneous data to generate the population can cause false predictions on the characteristics of epidemic propagation.
The simulation of community contacts can also be a puzzling issue : usually, individuals are supposed to make one third of their contact at home, one third at work and one third in the global community. As data on community contacts are often harder to gather, assumptions have to be made on how these kind of contacts occur, and with who. [58] and [59], while both studying influenza propagation in South East Asia, made different choices to give an account of community transmission. In [59], community contact could occur between any 2 individuals, depending on the distance between them, while [58] chose to place individuals in communities where they fully interacted with other members randomly. [54], which studied influenza propagation, proposed to mix both approaches considering that individuals were interacting with people in their neighborhood and little less with every other people. The differences in modeling led to different conclusions in terms of public health strategies : [59] planned that an influenza epidemic could be contained as long as $R_0 < 1.9$, when [58] concluded to a limit of $R_0 < 2.4$.
In addition to the different ways of interacting considered by previous models, some work tried to include contacts during transportation, by making people interact on intermediary points between their departure and their arrival [60,61]. However, as data about transportation were missing, strong assumptions had to be made to calibrate the model, which can be seen as not parsimonious.
The time scale of influenza propagation allows the approximation of a static host population : aging of individuals and birth of new susceptibles can be ignored. However, the study of persistent or long time scale diseases, like hepatitis, tuberculosis or HIV requires to take into account the renewal of the population. Such a model was proposed by [62] to study hepatitis A propagation : birth and mortality rate in the studied Italian areas were used to simulate the creation and suppression of individuals, depending on their age. New-born individuals were placed in appropriate households : demographic data on the composition of households were used to do so realistically.

4.2.2 Propagation at urban scale

At a smaller scale, [63] modelled the behavior and daily travels of Portland inhabitants, using an individual-based second by second microsimulation. A static network of interactions was constructed using dynamical movements of individuals, themselves generated from data about transportation networks, location of work or leisure places and composition of the population. A contact

was created between 2 individuals if they stayed in the same location, even for a brief time : the more they remained in the same place, the more the contact was considered intense. The obtained contact network was a small world one, but the number of contacts might have been overestimated, as no difference was made between different environments, assuming that people interact in the same way at home, at work or during their leisure. Aggregating the contacts occurring at different times in one graph might also have induced an overestimation of epidemic speed.

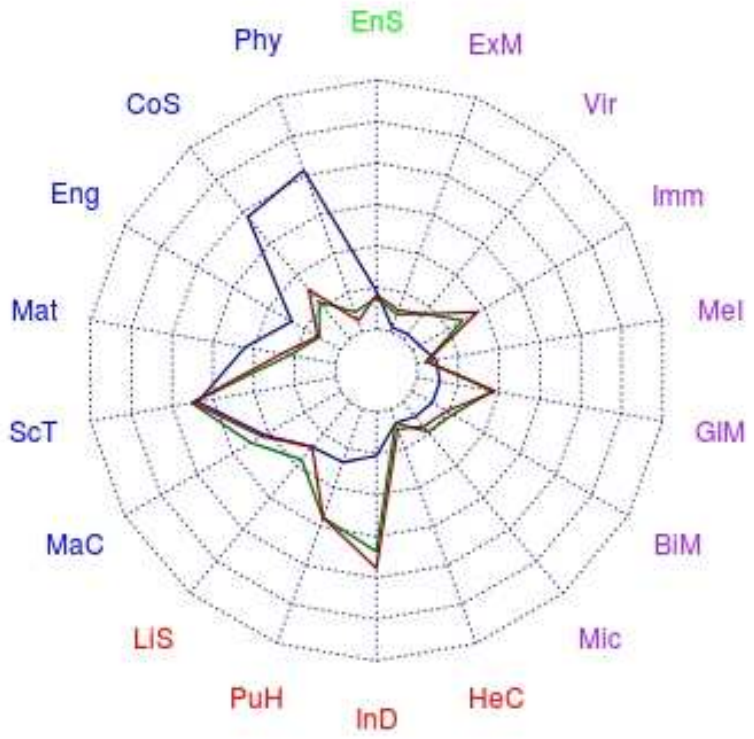

FIGURE 1 – **Citations of the works of Eubank and al., Ferguson and al and Longini and al.**
Research areas to which belongs papers citing [63] (in blue), [59] (in green) and [58] (in red). Only the research areas to which 0.01% of citing papers of at least one article belonged were selected (EnV : Environmental sciences, Phy : Physics, CoS : Computer Science, Eng : Engineering, Mat : Mathematics, ScT : Science and technology, MaC : Mathematics and Computational Biology, LiS : Life Sciences and Biomedicine, PuH : Public health, InD : Infectious Diseases, HeC : Health care, Mic : Microbiology, BiM : Biochemistry and molecular biology, GIM : General Internal medicine, MeI : Medical informatics, Imm : Immunology, Vir : Virology, ExM : Experimental medicine). Research areas were related to physics and computer sciences (blue), epidemiology (red), medicine (purple) and environmental sciences (green).

This work remains a unique example of the use of such detailled dataset

in epidemiological models : the method has been essentially considered as a technological breakthrough. Among the 877 papers which cited the work of Eubank and al, 49.7% were physics or computer sciences papers, but only 22.3% were epidemiological one (Figure 1). The contemporary works of Ferguson et al. [59] and Longini et al. [58], which also used an idividual-based approach but less detailled descriptions of individuals movements, had a comparable numbers of citations but more impact for epidemiology (Figure 1), with respectively 42.1% and 41.8% of citations in either epidemiological or medicine papers.

4.2.3 Comparison of different models

The introduction of IBM models in 2004 has not met full support from epidemiologists : the high complexity of this model, which needs multiple assumptions on human behavior and are highly parameterized, seemed unnecessary to many, who tried to show that simple compartment models could be used instead to study the same situations [64]. Indeed, despite their realistic possibility of modelling each individual separately, the complexity of IBM models, and the high computational time they ask for, make them maladjusted to some problematics, for which metapopulation models can be a better choice.

They are thus mainly used to study small scales propagations : in a city, a country or sometimes a continent, where gathering complete data about human behavior can be easier, and where the smaller number of individual makes extensive simulations easier. On the other hand, metapopulation models can be used to model wider areas and study worldwide propagation, as their smaller time of execution allows to execute more replications of stochastic simulations. However, the results obtained with these models are less detailed than previous ones.

The 2 approaches have been compared by [65], with a confrontation of influenza propagation in Italy simulated with GLEaM and an agent-based model. Starting from the same initial conditions, the 2 models simulated very similar epidemics, exhibiting the same timing of epidemic peaks in different cities. However, due to its lack of precise contact structure, the metapopulation model induced more interaction between individuals and the epidemic it generated affected more people than the one simulated with agent-based model.

4.3 Challenges

Due to the lack of precise data, models including mobility often have to ignore some aspects of the movements, that could have an impact on the epidemic studied. Movements are not constant in time : seasonnal migrations, vacations or week-ends can modify the intensity of travellers flows and their destinations. Holidays for example strongly modify the behavior of individuals : children stop school interactions, and a significant number of people leave their residence to gather with others in vacation places. These changes in the network of interaction between individuals can have repercussions on the propagation of the disease [57].

Moreover, the movements usually included in epidemiological models, like commuting or airplane travel do not fully describe human regular travelling patterns. At local scale, sporadic, spontaneous random movements of individuals from one city to another can also be considered for example [44]. When those movements

are included in a local model, the speed of an influenza epidemic increase from up to 25%.
Travellers are also often supposed to exhibit the same travel behavior, whatever their age, gender or departure location. However, differences in travel behavior can influence the speed of an epidemic propagation : introducing a differentiation between high and low frequency travellers in worldwide epidemic modelling doesn't change the average speed of propagation but strongly modify its variance [66], giving a better exploration of possible behaviors. Nevertheless, few data about travel patterns are currently available, making it hard to include realistic travel behavior in epidemic modelling.

4.4 Confrontation of models to data

A large range of situations have been addressed by epidemiological models : from airplane travels to the internal mobility of the city of Portland, many different networks of movements have been modelled to simulate the propagation of different infectious diseases, including influenza, SARS or smallpox. In some cases, surveillance data describing the propagation of the disease studied on the scale considered are available : the confrontation of simulation results to these data have permitted to confirm the interest of using mobility movements to simulated disease diffusion.
The role of airplane travel on the propagation of several infectious diseases, including SARS and influenza has been showed by several comparasons between simualtion results and surveillance data. To confirm the predictions of GLEaM on influenza international diffusion, [67] compared their simulation results to the number of cases registered in 7 countries and 9 american states during the 2001-2002 influenza season, and to the number of acute respiratory infection cases of tje same period. They found the imulations to be a good predictor of the epidemic timing, which confirmed the pertinence of using airplen movements to predict international propagation of influenza. During the 2009 H1N1 pandemic, [42] used data on the ongoing epidemic to compare the date on which first cases of influenza where declaired in several countries to the intensity of air travel between these countries and Mexico. A strong correlation was found between the two variables. Similar results have been obtained on SARS : [68] compared the number of cases predicted in 20 countries by a model based on airpalne flows to the number of cases reported in these countries in the WHO database : once again, simulations results gave a good prediction of propagation.
Despite their common use in epidemiological models, less studies have been done to prove the pertinence of commuting movements to predict disease propagation. Such a study was realized by [46] , who performed a Mantel test to evaluate the correlation between the synchrony of influenza temporal series in american states to commuting flows and other measures of distance, like airflows or geographical distance. Among all distances used to perform the test, commuting flows were the best predictor of similarity of epidemic timing between states. This analysis confirmed that commuting movements have an influence on influenza propagation. To our knowledge, no other study has been done to compare surveillance data to commuting flows.

5 Conclusion

In this review, we highlighted the importance of taking mobility movements into account when studying epidemic propagation. The intrinsic complexity of human mobility makes it difficult to study, but some of its specific components have been well studied and their influence on propagation at different scales proven. At the worldwide scale, the impact of airplane movements have been widely studied and confirmed with data from 2009 influenza pandemic. The influence of commuting movements on pathogen diffusion on a smaller scale has also been well analyzed : this result has been less confronted to data analysis, but some work like [46] still confirms it.
Current progress in data gathering and modeling tools have permitted the developpment of models studying propagation on an even smaller case. However, even if the models seem to confirm the intuition that daily short-distance movement are vectors of propagation, these results have not yet been confirmed with data analysis.
Despite the progress accomplished during last decades to integrate human mobility in epidemiological, some challenges remain, to determine if some other aspects of movements, currently not taken into account, could influence epidemic propagation.

Références

[1] J D Murray. Mathematical Biology. 1993.

[2] M E Wilson. The traveller and emerging infections : sentinel, courier, transmitter. *Journal of Applied Microbiology*, 94(S) :1S–11S, 2003.

[3] M E Wilson. Travel and the emergence of infectious diseases. *Emerging Infectious Diseases*, (1) :39–46, 1995.

[4] A Cliff and P Haggett. Time, travel and infection. *British Medical Bulletin*, 69 :87–99, 2004.

[5] P Upham, C Thomas, D Gillingwater, and D Raper. Environmental capacity and airport operations : current issues and future prospects. *Journal of Air Transport Management*, 9(3) :145–151, 2003.

[6] World Bank data, 2011.

[7] R Guimera, S Mossa, A Turtschi, and L A N Amaral. The worldwide air transportation network : Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22) :7794–7799, 2005.

[8] A De Montis, S Caschili, and A Chessa. Time evolution of complex networks : commuting systems in insular Italy. *Journal of geographical Systems*, 13(1, SI) :49–65, March 2011.

[9] R Patuelli, A Reggiani, S P Gorman, P Nijkamp, and F-J Bade. Network analysis of commuting flows : A comparative static approach to German data. *Networks & Spatial economics*, 7(4) :315–331, 2007.

[10] A-L Barabasi and R Albert. Emergence of scaling in random networks. *Science*, 286(5439) :509–512, October 1999.

[11] A De Montis, M Barthelemy, A Chessa, and A Vespignani. The structure of interurban traffic : A weighted network analysis. *Environment and Planning B-Planning & Design*, 34(5) :905–924, September 2007.

[12] W-S Jung, F Wang, and H E Stanley. Gravity model in the Korean highway. *EPL*, 81(4), 2008.

[13] M C Gonzalez, C A Hidalgo, and A-L Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196) :779–782, 2008.

[14] C Song, Z Qu, N Blumm, and A-L Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968) :1018–1021, 2010.

[15] C Song, T Koren, P Wang, and A-L Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10) :818–823, October 2010.

[16] D Brockmann, L Hufnagel, and T Geisel. The scaling laws of human travel. *Nature*, 439(7075) :462–465, January 2006.

[17] B Jiang, J Yin, and S Zhao. Characterizing the human mobility pattern in a large street network. *Physical Review E*, 80(2, Part 1), 2009.

[18] G Chowell, J M Hyman, S Eubank, and C Castillo-Chavez. Scaling laws for the movement of people between locations in a large city. *Physical Review E*, 68(6, Part 2), 2003.

[19] S Phithakkitnukoon, T Horanont, G Di Lorenzo, R Shibasaki, and C Ratti. Activity-Aware Map : Identifying Human Daily Activity Pattern Using Mobile Phone Data. In A Salah, AA and Gevers, T and Sebe, N and Vinciarelli, editor, *Human Behavior Understanding*, volume 6219 of *Lecture Notes in Computer Science*, pages 14–25, HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY, 2010. SPRINGER-VERLAG BERLIN.

[20] M Padgham. Human Movement Is Both Diffusive and Directed. *PLoS One*, 7(5), 2012.

[21] C Kang, X Ma, D Tong, and Y Liu. Intra-urban human mobility patterns : An urban morphology perspective. *Physica A-Statistical Mechanics and its applications*, 391(4) :1702–1717, 2012.

[22] Jordi Bascompte and Ricard V Solé. *Modeling spatiotemporal dynamics in ecology*. Springer New York, 1998.

[23] Ilkka Hanski and Michael E Gilpin. *Metapopulation biology : ecology, genetics, and evolution*. Academic press San Diego, 1997.

[24] David Tilman and Peter M Kareiva. *Spatial ecology : the role of space in population dynamics and interspecific interactions*, volume 30. Princeton University Press, 1997.

[25] B M Bolker and B T Grenfell. SPACE, PERSISTENCE AND DYNAMICS OF MEASLES EPIDEMICS. *Philosophical transactions of the Royal Society of London Series B-Biological Sciences*, 348(1325) :309–320, 1995.

[26] B T Grenfell and J Harwood. (Meta)population dynamics of infectious diseases. *Trends in Ecology & Evolution*, 12(10) :395–399, October 1997.

[27] B T Grenfell, O N Bjornstad, and J Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865) :716–723, 2001.

[28] M J Keeling and B T Grenfell. Disease extinction and community size : Modeling the persistence of measles. *Science*, 275(5296) :65–67, January 1997.

[29] A L Lloyd and R M May. Spatial heterogeneity in epidemic models. *Journal of Theoretical Biology*, 179(1) :1–11, March 1996.

[30] P Rohani, D J D Earn, and B T Grenfell. Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, 286(5441) :968–971, October 1999.

[31] L A Rvachev and I M Longini. A mathematical-model for the global spread of influenza. *Mathematical Biosciences*, 75(1) :3–23, 1985.

[32] A Flahault, T Blanchon, Y Dorleans, L Toubiana, J F Vibert, and A J Valleron. Virtual surveillance of communicable diseases : a 20-year experience in France. *Statistical methods in medical research*, 15(5) :413–421, October 2006.

[33] R F Grais, J H Ellis, and G E Glass. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *European Journal of Epidemiology*, 18(11) :1065–1072, November 2003.

[34] B S Cooper, R J Pitman, W J Edmunds, and N J Gay. Delaying the international spread of pandemic influenza. *PLoS Medicine*, 3(6) :e212, 2006.

[35] R F Grais, E J H, A Kress, and G E Glass. Modeling the spread of annual influenza epidemics in the US : the potential role of air travel. *Health care management science*, 7(2) :127–34, 2004.

[36] R F Grais, J H Ellis, and G E Glass. Forecasting the geographical spread of smallpox cases by air travel. *Epidemiology and Infection*, 131(2) :849–857, October 2003.

[37] L Hufnagel, D Brockmann, and T Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42) :15124–15129, October 2004.

[38] V Colizza, A Barrat, M Barthelemy, and A Vespignani. The modeling of global epidemics : Stochastic dynamics and predictability. *Bulletin of Mathematical Biology*, 68(8) :1893–1921, November 2006.

[39] V Colizza, A Barrat, M Barthelemy, and A Vespignani. 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(7) :2015–2020, 2006.

[40] V Colizza, A Barrat, M Barthelemy, A-J Valleron, and A Vespignani. Modeling the worldwide spread of pandemic influenza : baseline case and containment interventions. *PLoS Medicine*, 4(1) :e13, January 2007.

[41] E Kenah, D L Chao, L Matrajt, M E Halloran, and I M Longini Jr. The Global Transmission and Control of Influenza. *PLoS One*, 6(5), 2011.

[42] K Khan, J Arino, W Hu, P Raposo, J Sears, F Calderon, C Heidebrecht, M Macdonald, J Liauw, A Chan, and M Gardam. Spread of a Novel Influenza A (H1N1) Virus via Global Airline Transportation. *New England journal of medicine*, 361(2) :212–214, 2009.

[43] C Viboud, M A Miller, B T Grenfell, O N Bjornstad, and L Simonsen. Air travel and the spread of influenza : important caveats. *PLoS Medicine*, 3(11) :e503 ; author reply e502, November 2006.

[44] L Danon, T House, and M J Keeling. The role of routine versus random movements on the spread of disease in Great Britain. *Epidemics*, 1(4) :250–258, 2009.

[45] I M Hall, J R Egan, I Barrass, R Gani, and S Leach. Comparison of smallpox outbreak control strategies using a spatial metapopulation model. *Epidemiology and Infection*, 135(7) :1133–1144, October 2007.

[46] C Viboud, O N Bjornstad, D L Smith, L Simonsen, M A Miller, and B T Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772) :447–451, 2006.

[47] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis : theory and extensions*, volume 3. Vsp, 1990.

[48] D Balcan, V Colizza, B Goncalves, H Hu, J J Ramasco, and A Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51) :21484–21489, 2009.

[49] M J Keeling, L Danon, M C Vernon, and T House. Individual identity and movement networks for disease metapopulations. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19) :8866–8870, 2010.

[50] A Lunelli, A Pugliese, and C Rizzo. Epidemic patch models applied to pandemic influenza : Contact matrix, stochasticity, robustness of predictions. *Mathematical Biosciences*, 220(1) :24–33, 2009.

[51] C Rizzo, A Lunelli, A Pugliese, A Bella, P Manfredi, G Scalia Tomba, M Iannelli, M L Degli Atti, and EPICO Working Grp. Scenarios of diffusion and control of an influenza pandemic in Italy. *Epidemiology and Infection*, 136(12) :1650–1657, 2008.

[52] L R Elveback, J P Fox, E Ackerman, A Langworthy, M Boyd, and L Gatewood. INFLUENZA SIMULATION-MODEL FOR IMMUNIZATION STUDIES. *American journal of epidemiology*, 103(2) :152–165, 1976.

[53] J Koopman. Controlling smallpox. *Science*, 298(5597) :1342–1344, November 2002.

[54] T C Germann, K Kadau, I M Longini, and C A Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15) :5935–5940, 2006.

[55] S Riley. Large-scale spatial-transmission models of infectious disease. *Science*, 316(5829) :1298–1301, 2007.

[56] S Merler and M Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B - Biological Sciences*, 277(1681) :557–565, 2010.

[57] S Merler, M Ajelli, A Pugliese, and N M Ferguson. Determinants of the spatiotemporal dynamics of the 2009 H1N1 pandemic in Europe : implications for real-time modelling. *PLoS Computational Biology*, 7(9) :e1002205, September 2011.

[58] I M Longini, A Nizam, S F Xu, K Ungchusak, W Hanshaoworakul, D A T Cummings, and M E Halloran. Containing pandemic influenza at the source. *Science*, 309(5737) :1083–1087, 2005.

[59] N M Ferguson, D A T Cummings, S Cauchemez, C Fraser, S Riley, A Meeyai, S Iamsirithaworn, and D S Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056) :209–214, September 2005.

[60] L Perez and S Dragicevic. An agent-based approach for modeling dynamics of contagious disease spread. *International Journal of Health Geographics*, 8, 2009.

[61] F Rakowski, M Gruziel, L Bieniasz-Krzywiec, and J P Radomski. Influenza epidemic spread simulation for Poland - a large scale, individual based model study. *Physica A-Statistical Mechanics and its applications*, 389(16) :3149–3165, 2010.

[62] M Ajelli and S Merler. An individual-based model of hepatitis A transmission. *Journal of Theoretical Biology*, 259(3) :478–488, 2009.

[63] S Eubank, H Guclu, V S A Kumar, M V Marathe, A Srinivasan, Z Toroczkai, and N Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988) :180–184, 2004.

[64] J Arino, F Brauer, P van den Driessche, J Watmough, and J Wu. Simple models for containment of a pandemic. *Journal of the Royal Society Interface*, 3(8) :453–457, 2006.

[65] M Ajelli, B Goncalves, D Balcan, V Colizza, H Hu, J J Ramasco, S Merler, and A Vespignani. Comparing large-scale computational approaches to epidemic modeling : Agent-based versus structured metapopulation models. *BMC Infectious Diseases*, 10, 2010.

[66] T D Hollingsworth, N M Ferguson, and R M Anderson. Frequent travelers and rate of spread of epidemics. *EMERGING INFECTIOUS DISEASES*, 13(9) :1288–1294, September 2007.

[67] Duygu Balcan, Bruno Gonçalves, Hao Hu, José J Ramasco, Vittoria Colizza, and Alessandro Vespignani. Modeling the spatial spread of infectious diseases : The GLobal Epidemic and Mobility computational model. *Journal of computational science*, 1(3) :132–145, 2010.

[68] V Colizza, A Barrat, M Barthelemy, and A Vespignani. Predictability and epidemic pathways in global outbreaks of infectious diseases : the SARS case study. *BMC Medicine*, 5, November 2007.

3.2 Observations

Taking mobility movements into account is of crucial importance to understand infectious disease propagation: epidemiologists have for the moment mainly focused on two aspects of human movements, namely commuting and airplane travel and both of them have been shown to impact spatial diffusion of disease like SARS or influenza [30, 56, 91].

The international airplane network has been widely studied in the literature. Indeed, in our review, more than $\frac{2}{3}$ of our papers were either including or focusing on the airplane network and its influence on disease propagation: no other networks has been so extensively studied. Moreover, several analysis have compared the results of simulations on models including the airplane network to surveillance data and confirmed the simulations were good predictors of the timing of the international epidemic propagation.

Commuting networks on the other hand, despite their frequent use in epidemiological models, have been less studied. Few analysis have been done to confirm the pertinence of using commuting flows to understand disease propagation. In the following chapter, we will address this issue, using the example of influenza propagation in France as a case model.

*****

Chapter 4

Influence of commuting movements on influenza propagation

4.1 Introduction

In last decade models investigating the correlation between human mobility and infectious disease propagation, the worldwide air transportation network has been the most extensively studied system (see the previous chapter). The international propagation of diseases like SARS or influenza has been shown to be largely driven by flows of airplane travelers [30, 91]. However, in the complex network of human mobility, airplane traffic only covers long-range movements: in countries like France, or Germany, where most daily movements occur on smaller scales, airplane has significantly less importance than other means of transportation, like train or car. Modeling mobility on those scales thus necessitates to take into consideration other types of movements. If human mobility on local scale is composed of very diverse movements, the collection of GPS data on car geolocalization in Italy [58] has shown that the majority of

regular highway traffic is composed of commuting movements, making them a good approximation of mobility on national scale. As more and more data on commuting have been gathered in past years, they have been introduced in an increasing number of models studying infectious disease propagation on local to intermediate scales. However, the correlation between disease transmission and commuting has not been extensively studied yet: except for [137] work on United States influenza propagation, which showed a spatial autocorrelation linked to commuting data in influenza propagation data, no more work has been done to investigate this link. In this chapter, we will first focus on the analysis of the specific case of influenza propagation in France and see if it shows a correlation with commuting network. Then, we will present a model that have been developed to simulate influenza propagations based on flows of commuting, and make some observations on the simulation results.

These results will be analyzed in the second part of this chapter. In first chapter, we have seen that the structure of the network can influence disease diffusion, its speed and the pathways it follows: we will therefore explore how the patterns observed in the simulations are related to the underlying network structure.

Both of these works are presented in the article "Commuter mobility and the spread of infectious diseases: application to influenza in France", joint to this chapter.

4.2 Spatial autocorrelation of influenza incidence

In this paragraph will be presented a study that we proposed to demonstrate the influence of commuting on the propagation of influenza in France. To perform this analysis, we used the database on influenza propagation gathered by the Sentinelles network, a network of physicians who have been recording influenza incidence in a weekly basis since 1984. This study was presented in the article joint to this chapter, and we hereby summarize the first part of this article,

where the autocorrelation calculus was presented.

4.2.1 Investigating the link between influenza propagation and commuting: spatial autocorrelation and model design

The Sentinelles network gathers over 1300 physicians distributed between French territory: as the network covers regularly the territory, the influenza incidence it records can be expressed at NUTS3 level (which corresponds to the French administrative division of department). We compared able to compare the incidence temporal series of the 26 epidemics to the commuting flows between departments. In order to account for the different commuting behavior of workers and students, we used two separated networks, each of them describing one type of commuting flows.

To perform the analysis of the spatial autocorrelation of Sentinelles data, we performed Mantel's test, as presented in [137] and calculated Moran's index for both networks of commuting. Moran's index, which evaluates spatial autocorrelation at a specific time, was found significantly positive throughout the epidemic, with both network of commuting. In both cases, the index was positive 2 to 3 weeks before the national peak of the epidemic. Mantel's test, which compared the synchrony between the temporal series of incidence in the departments to the matrices of commuting, was also significant. Both results confirmed the existence of a spatial structure in influenza incidences. The progression of Moran's index (increasing during the first phase of the epidemic, when influenza was transmitted from department to department and decreasing when the incidence started decreasing in some departments) and the correlation between the synchrony of incidence evolution in different departments and commuting flows, established the existence of a correlation between epidemic spread and commuting flows.

To investigate this relation, we developed a deterministic metapopulation model of influenza propagation, where the natural history of the disease was modeled with a SEIR process. To generate a population similar to the French one, we used data gathered at LAU1 level in a census conducted in 1999: we chose to divide individuals in 5 ages classes, including 2 classes of children and 1 class of adult, to take into consideration the different patterns of contacts depending on the type of commuting an individual was involved into. We considered 3 types of contacts between individual: household contacts between different age classes in the same district, commuting contact occurring at school or workplace between individual of the same age and community contact occurring between neighbors.

Extensive simulations were performed using this model by introducing successively a single infected in each district. The results of the simulations were analyzed with Mantel's test and Moran's index and the same results as for Sentinelles data were obtained, revealing the existence of a spatial autocorrelation in simulated incidence.

4.2.2 Observations and Perspectives

Observation of simulated propagations

The model exposed in this article was used to simulate epidemics initiated by the introduction of a single infected (either a child or an adult) in each district. The observation of the spatial propagation obtained in each cases evidenced a recurring pattern of propagation that could be found whatever district the disease was inoculated into. Indeed, the epidemic started with a first phase of wave like propagation, where the disease was mainly transmitted to surrounding districts. Progressively, the front line became less circular, as the epidemic started diffusing in some preferential directions: in most cases, 10 to 12 weeks after the introduction of influenza, the epidemic reached areas not geographically linked to the one already infected. When disconnected areas were reached, the

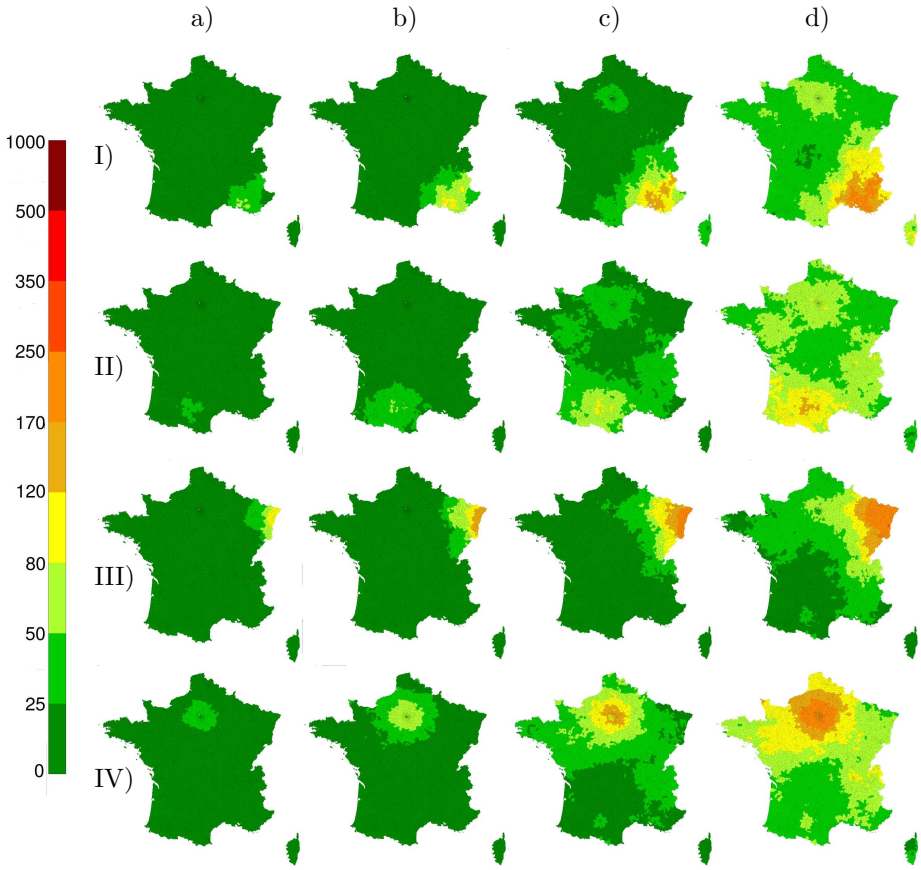

Figure 4.1: **Simulations with a model based on commuting data**
Maps of the incidence for 100000 individuals of influenza in every district 11 weeks (a), 12 weeks (b), 13 weeks (c) and 14 weeks (d) after the seeding. Each epidemic was seeded by introducing a single infected child in a different district.

epidemic started propagating from them, in a wave like progression similar to the encountered at the beginning. The three phases propagation, with a local diffusion followed by a phase of quick non wave like expansion, was encountered in every simulations. An illustration of this observation is given in Figure 4.1 , with the example of four epidemics seeded in randomly chosen districts. As it can be seen on maps of week 10 and 12, the propagation of influenza during the first weeks of the epidemic has been mainly local in 4 cases. Week 13 of epidemics I, II and IV show an example of the disrupted propagation explained before, as the influenza suddenly appears in Paris or Lyon, without a wave like propagation from the seeding district. Same behavior can be observed on epidemic III, with a sudden diffusion to the West of the country, while the

epidemic has been confined in eastern districts before.

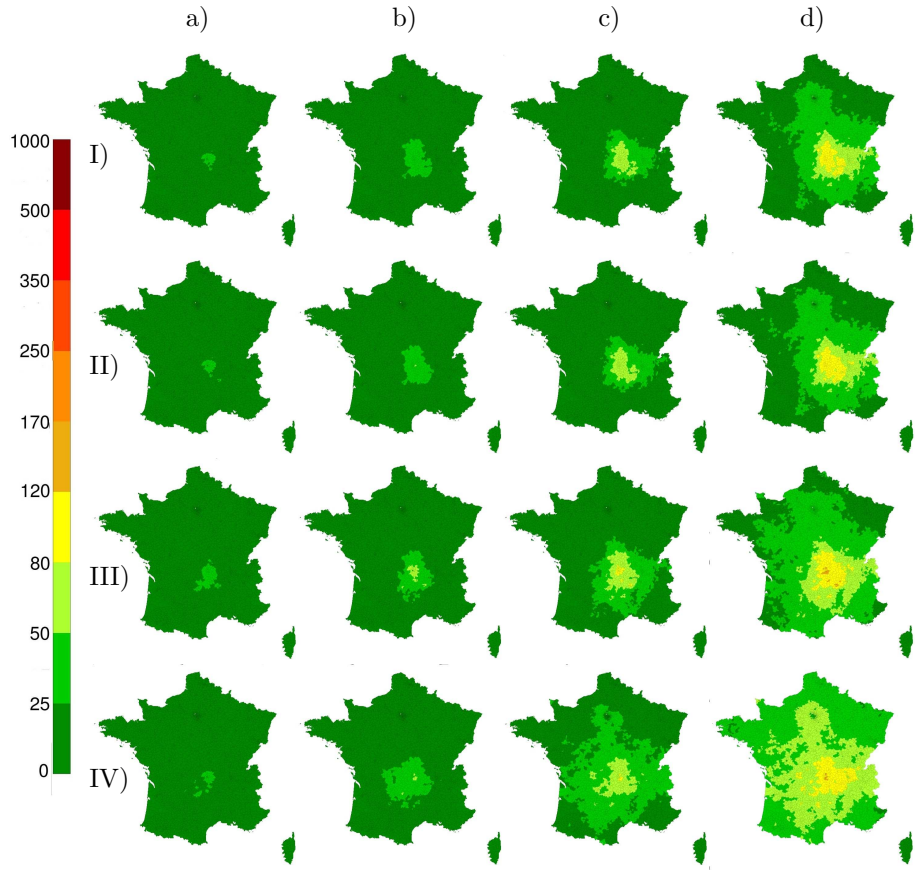

Figure 4.2: **Four simulations initiated in Puy de Dome and following different pathways**
Maps of the incidence for 100000 individuals of influenza in every district 11 weeks (a), 12 weeks (b), 13 weeks (c) and 14 weeks (d) after the seeding. The epidemics were initiated in four neighbor districts of Puy de Dome, but followed different pathways

The location of the first infected plays an important role in the spatial propagation. The observation on the simulations showed that besides the existence of a similar global behavior, influenza diffusion followed recurring patterns: in most cases, the direction of Paris was privileged, as either the epidemic front line was distorted in this direction or Paris was infected even if the front line had not reached it yet. Important French cities, like Lyon, Marseille or Toulouse played a similar role for epidemics starting in their vicinity (epidemics I and II of figure 4.1 are illustrations of this phenomenon). The location of the first infected district has a major influence on the pathways followed by the propa-

gation and thus on the whole course of the epidemic. Indeed, despite the local behavior of the propagation in the first phase of diffusion, epidemics starting in neighboring districts could exhibit a very different behavior, as shown on Figure 4.2. The four epidemics presented in this figure have been initiated in neighboring districts of the same department, Puy de Dome, located near the city of Clermont Ferrand. However, despite the vicinity of these districts, epidemics initiated in them exhibited different behaviors. While epidemics I and II followed similar pathways, first diffusing in the direction of Lyon, then of Paris and Marseille, epidemic III diffused simultaneously towards Lyon and Montpellier, another southern city. Epidemic IV diffused in the direction of Lyon and Paris and reached Paris before the other epidemics did. Despite their neighboring seeding districts, influenza did not diffuse in the same direction in each epidemic and the distortion of the epidemic front line did not occur in the same directions.

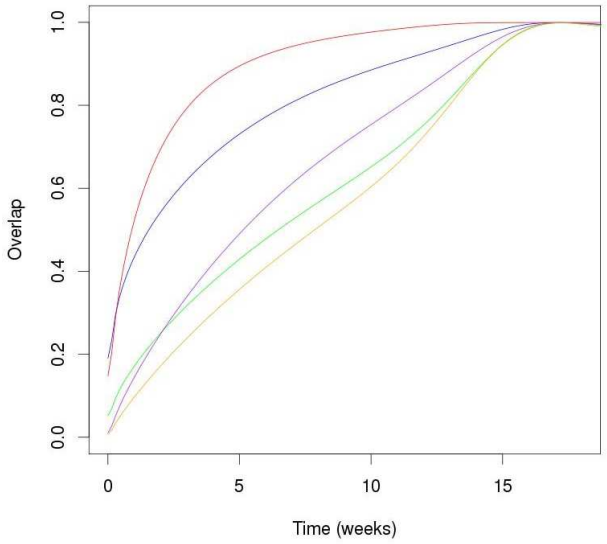

Figure 4.3: **Example of overlap curves**
Overlap curves measuring the similarity between epidemic propagations in case of seeding in different couples of districts

The parameter of overlap [29], which compares the incidence in each district at a fixed time step for two epidemics was used to quantify the similarity between

epidemics propagation. This parameter varies between 0 and 1, and is maximal when the incidence is the same in each district in both situations. We compared the courses of epidemics seeded in different couples of districts: for every couple, overlap increased from 0 to 1 but with a different growth rate (Figure 4.3). The diversity of overlap behaviors was coherent with observations made on epidemic propagations: on average, overlap growth rate was higher for couples of districts at small distance, in line with the observations of a mainly local initial propagation. In the same time, the variance of overlap initial values for couples at small distance was high, which was coherent with the observation that a great diversity of propagation pathways could be followed by epidemics starting in neighboring districts. Different behaviors were also exhibited for couples of distant districts, for which overlap started to increase more or less quickly: this result was coherent with the simulations observed, in which the timing of epidemic arrival in an area depended of the district of seeding.

Observing these specific patterns in the epidemics diffusing on the network created by the flows of commuting ask the question of the existence of a link between the structure of the network and the patterns. Commuting network are both small world , as many social networks are (like acquaintances network [6], scientific collaboration network [108] or Internet network [5]), in which the shortest path between any couple of nodes is small, and scale-free: in the following part, we will relate the patterns of propagation observed on the simulations to this specific structure.

4.3 Influence of the network structure on the propagation

The networks shaped by commuting movements are complex structures, composed of 495891 edges for the work network and 282883 for the school one (Figure 4.4-a,b). They both have a strongly clustered structure, with a significantly high local clustering coefficient for both networks (0.46 for school commuting

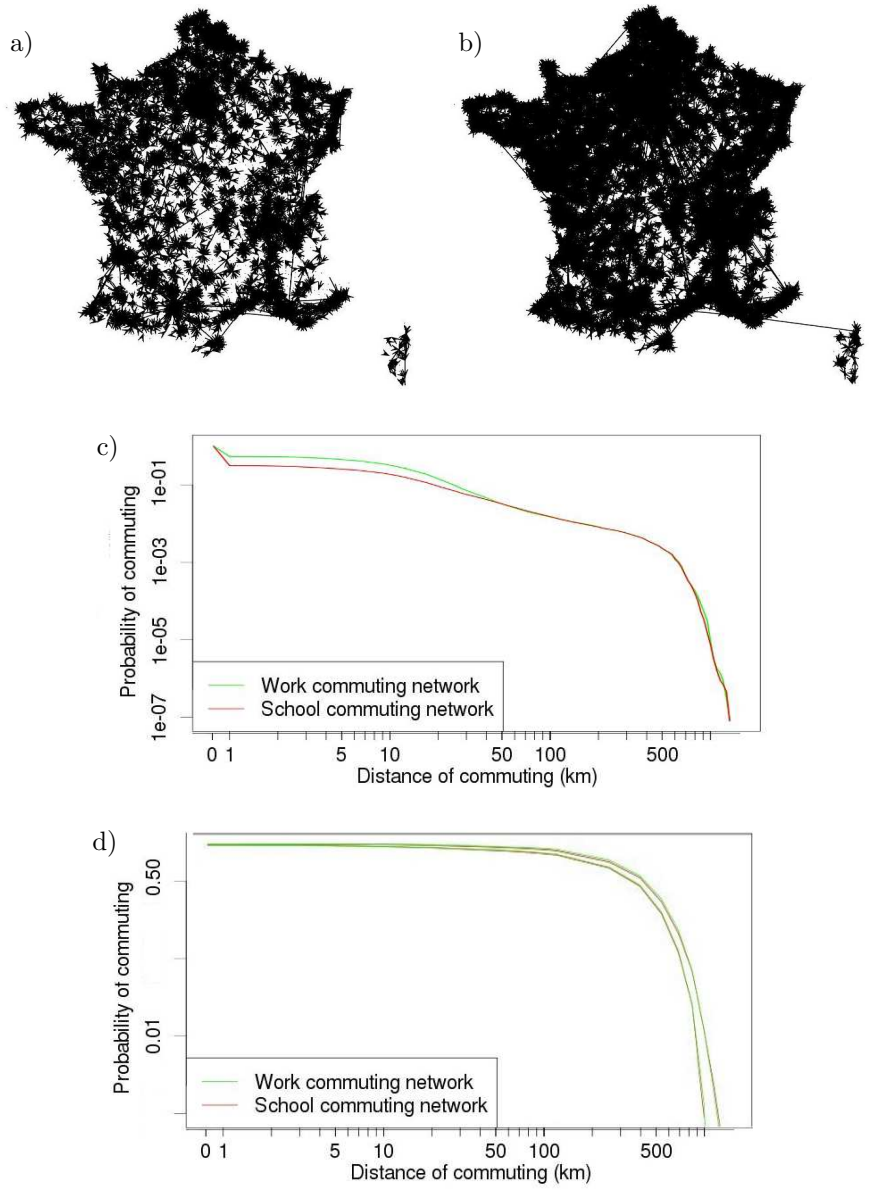

Figure 4.4: **Distance crossed by commuters**
a) School commuting network b) Work commuting network (to ease the reading, only the commuting paths followed by 50 people or more were represented. Probability for a commuter to cross a certain distance in its commuting travel. In both networks, less than 2% of individuals cross more than 100km in regular commuting networks c) and randomized ones d). On Figure d) are represented the intervals in which 95% of distributions are found

and 0.38 for work commuting). As most individuals commute in the vicinity of their home (Figure 4.4), clustering is mainly local and communities gather neighboring districts. Given this structure of commuting networks, we ventured

the hypothesis that the similarity of patterns observed between some epidemics at their beginning and during their course was linked to the structure of the networks' clusters. To test this hypothesis, we developed two criterions, based on the measure of similarity between epidemic propagations, that will be presented in next part of this chapter. In the following part, will be exposed our investigation of the correlation between the distribution of these criterions and the structure of the networks. A part of this analysis was exposed in the article "Commuter mobility and the spread of infectious diseases: application to influenza in France".

4.3.1 Relation between the structure of commuting networks and similarities between epidemic courses

In order to evidence the impact of local communities on influenza propagation, we generated randomized network of commuting, by reshuffling the arrival district of each commuting link. The networks generated by this process were small-world networks (S=4.7 for school and 2.86 for work), with the same incoming and outgoing degree for each district, but the distribution of distances crossed by commuting was different, as more people were involved in long distance commuting (Figure 4.4-b). After randomization, although still small world, the networks were less clustered (with an average clustering coefficient of 0.16 for school network and 0.18 for work): the local communities observed in regular commuting networks were not present after their randomization . As we had done previously, we simulated epidemic propagations on randomized networks by introducing a single infected in every district. The attack rate was not modified and progression of national incidence was the same in those simulations: however, the spatial propagation of influenza was different and the patterns of diffusion previously observed were not present here. We evaluated the similarity of epidemic propagation between epidemics starting from couples of random districts and found the same result for each one of them, independently of the distance between them, which was not the case for epidemics simulated with

non-randomized commuting networks.

To test the previously ventured hypothesis, we developed a criterion comparing the propagation of epidemics starting in 2 different districts on regular networks and randomized networks. This criterion compared for how long the propagations on commuting networks presented more similarity than the ones on randomized networks, in which the local communities had been suppressed.

We observed that this parameter was positively correlated with the number of commuters exchanged by the 2 districts and negatively correlated with the distance between them. Moreover, we found that, in case of districts more than 100 km away, the criterion reached its minimum, meaning that epidemics seeded in these spatial units showed from the first day less similarity than epidemics simulated on randomized networks. For couples of less distant districts, the criterion was highly variable: epidemics initiated in the most populated French cities, which are hubs in commuting networks, had an important similarity with epidemics initiated in their vicinity for several weeks. When we clustered French districts based on this criterion, we observed that each district could be included in a cluster with one or two of these cities. French territory was so divided in 49 communities of districts gathered around a major city: epidemics seeded in a community would propagate following a pathway similar to epidemics seeded in the local hub.

4.3.2 Article

Article: "Commuter mobility and the spread of infectious diseases: application to influenza in France".

Commuter mobility and the spread of infectious diseases: application to influenza in France

Ségolène Charaudeau[1,2,*], Khashayar Pakdaman[2], Pierre-Yves Boëlle[1]

1 INSERM, UMR S 707, Paris, France; Université Pierre et Marie Curie - Paris 6, Paris, France

2 Institut Jacques Monod, Paris, France; Université Denis Diderot, Paris, France

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

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.

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.

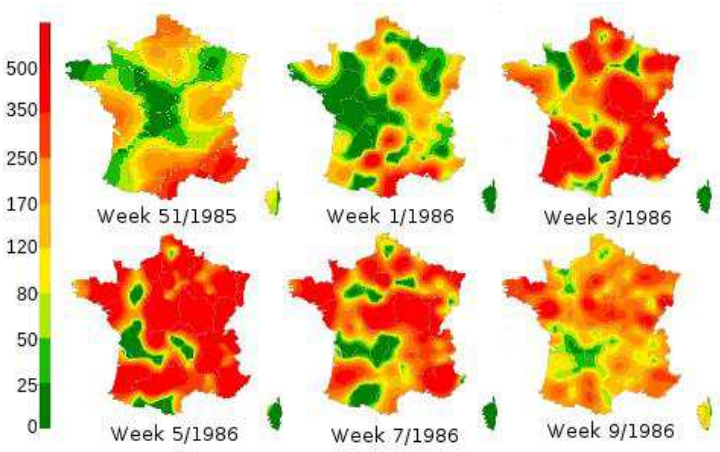

Figure 1. Spatial spread of influenza like illness in France Incidence for 100000 inhabitants as monitored by the Sentinelles network during season 1985-1986. Maps are 2 weeks apart

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 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 com-

munities 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).

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

where $\lambda^X_{a,D}(t)$ 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) :

$$\lambda^H_{a,D}(t) = \beta_H \frac{\sum_{a'} M^D_H(a,a') \times (I^A_{a',D}(t) + I^S_{a',D}(t))}{\sum_{a'} M^D_H(a,a') \times N_{a',D}(t)} \tag{2}$$

where β_H was the pairwise rate of contact leading to transmission in the household. $I^A_{a',D}(t)$ and $I^S_{a',D}(t)$ 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. 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

$$\Theta(t) = \left(\begin{array}{ll} & \sqrt{\frac{\sum_j I_j^I(t)}{N} \frac{\sum_j I_j^{II}(t)}{N}} \\ + & \sqrt{(1 - \frac{\sum_j I_j^I(t)}{N}) \times (1 - \frac{\sum_j I_j^{II}(t)}{N})} \end{array} \right) \times \sum_j \sqrt{\Pi^I(t) \times \Pi^{II}(t)} \quad (6)$$

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.

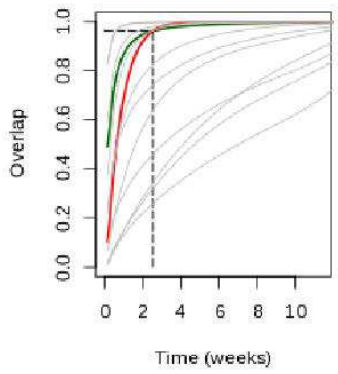

Figure 2. Measuring similarity in spread above randomness C_1 Lines correspond with overlap measures for a given pair of district at different times after introduction of a single infected. For a particular pair (green line), we also present the overlap measure obtained using reshuffled networks for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

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.

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.

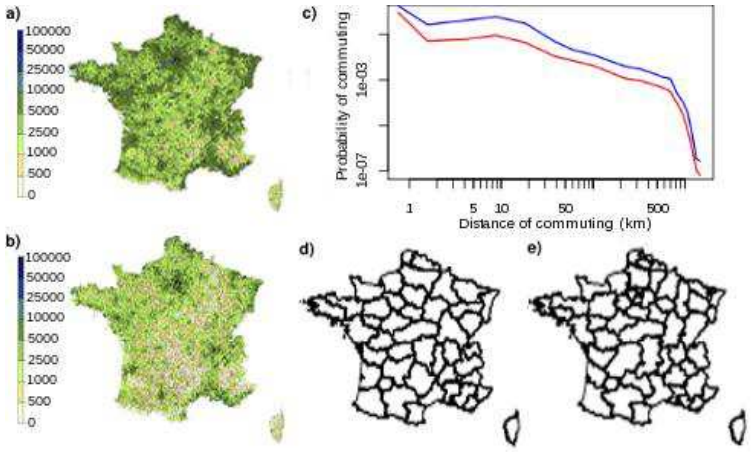

Figure 3. Commuter mobility in France (a,b)Total number of individuals leaving each district via work commuting (a) and school commuting (b). (c) Proportion of commuters and travelled distance in the school network (red) and the work network (green). (d,e) Clusters identified in the work (d) and schoool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

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 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),

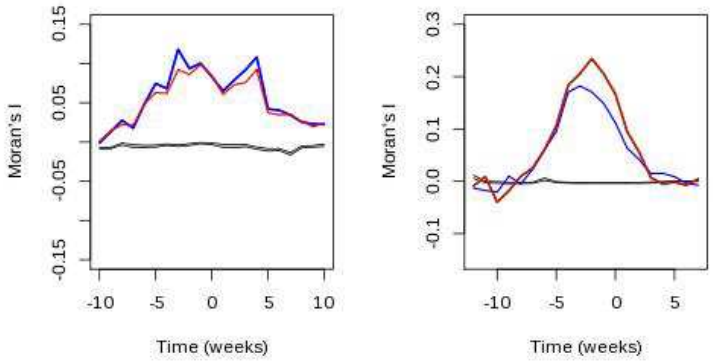

Figure 4. Autocorrelation in incidence for observed and simulated epidemics (a) Mean value of Moran's Index computed on the 26 epidemics from the Sentinelles network, and (b) on 100 simulated epidemics. In each case, the blue line uses work commuting based weights or school (red line). Gray areas corresponds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

confirming the existence of a spatial auto-correlation 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 ($r = -0.916 \pm 0.040$, 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 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 ± 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: $r^2 = 0.852 \pm 0.022$.

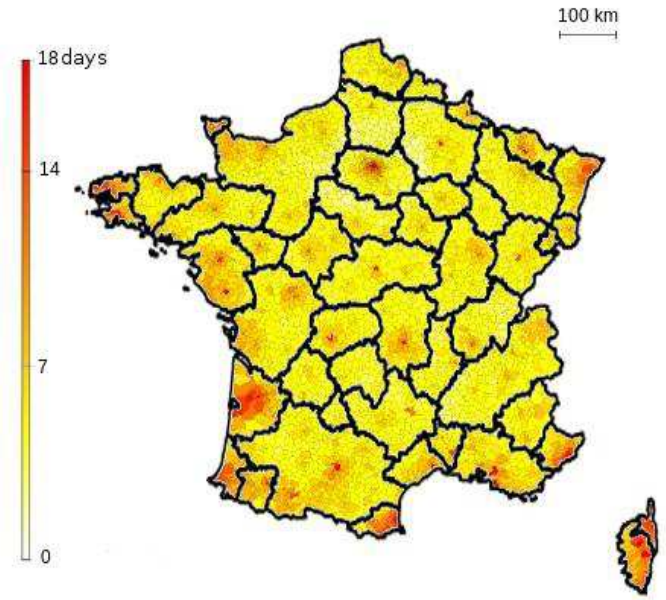

Figure 5. Typical pathways according to initial infective location For each district, C_1 values were averaged over all neighbors less than 100 km away. Basins of attraction were identified by clustering.

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 ± 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 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

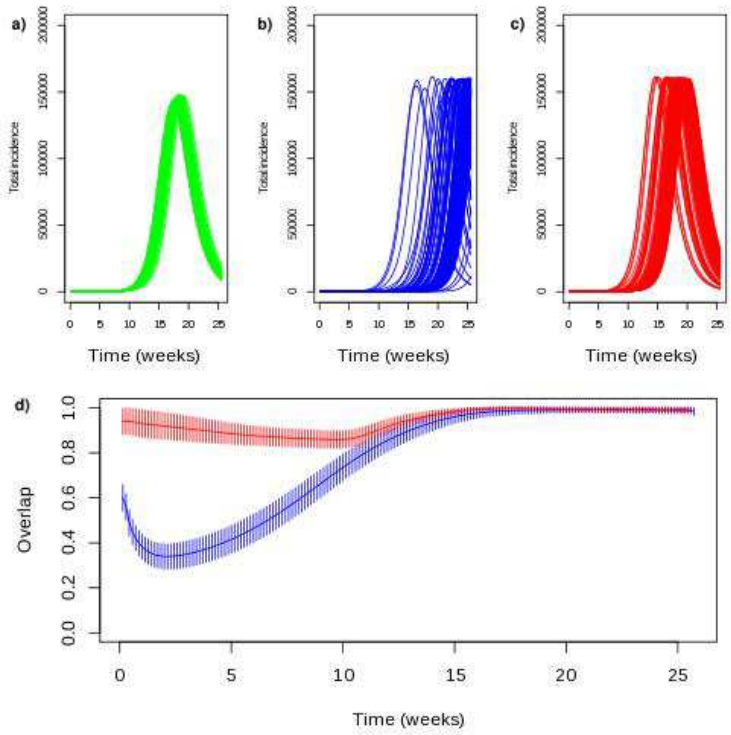

Figure 6. School and work commuting networks and the spatial spread of epidemics (a,b,c) ILI epidemic curves using all commuting networks (a), only work commuting (b) and only school commuting (c). Epidemics were started form 1000 randomly chosen districts. (d) Overlap between epidemics using work (blue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

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

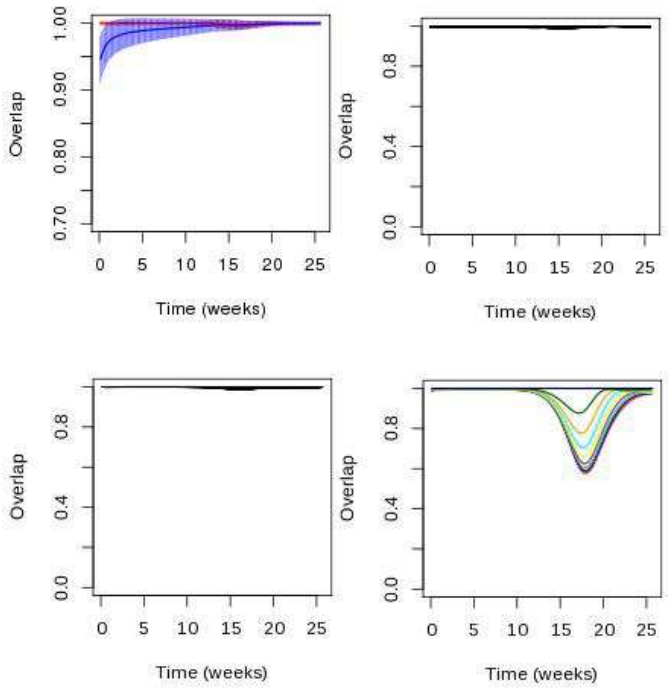

Figure 7. Sensitivity analysis Overlap between epidemics simulated with first model and epidemics propagating only by school (red) or work (blue) commuting (a) , with epidemics for which asymptomatic adults do not have a reuced genration time (b), with epidemics simulated with different parameters of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

(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.

Acknowledgments

S Charaudeau was funded by the Fondation pour la recherche médicale (FRM; www.frm.org/).

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 Influenza 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 influenza? Nature Medicine 12: 497–499.

7. Merler S, Ajelli M (2010) The role of population heterogeneity and human mobility in the spread of pandemic influenza. 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 influenza. 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 influenza 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 influenza 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 influenza. 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

influenza. 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 influenza 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 influenza. 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 influenza 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 influenza: 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 influenza. PLoS computational biology 8: e1002425.

27. Polgreen PM, Chen Z, Segre AM, Harris ML, Pentella MA, et al. (2009) Optimizing Influenza Sentinel Surveillance at the State Level. American journal of epidemiology 170: 1300–1306.

4.3.3 Observations, Supplementary informations and Perspectives

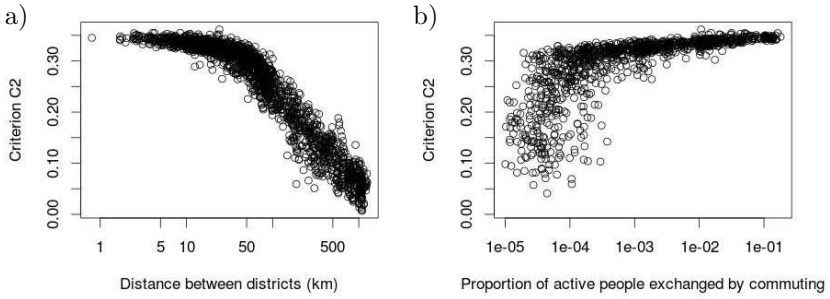

Figure 4.5: **Comparison of criterion C_2 to the distance between districts and the flows of commuters exchanged by them**
Criterion C_2 evaluated for 2000 randomly chosen pairs of districts compared to the distance between districts (a) and the rate of active people from the districts commuting between them, either for work or school purposes (b). C_2 was negatively correlated with distance but positively correlated with the rate of commuters

The study conducted with criterion C_1 showed districts could be gathered in communities from which the pathway followed by an epidemic would be the same whatever the district of seeding: the structure of commuting networks, both small world networks with many local communities, was identified as a cause of this division. This analysis explained the the diversity of overlap patterns at the start of epidemics. However, as criterion C_1 focused on the beginning of epidemics, later similarities between epidemics could not be investigated using it. We thus introduced a second criterion, C_2, that we defined as the integral of overlap curve on the 180 days of simulation.

As criterion C_1, C_2 was negatively correlated with the distance ($r = -0.962 \pm 0.010$, spearman correlation) and positively correlated with the rate of commuters exchanged by the districts ($r = 0.842 \pm 0.041$) (Figure 4.5). Both distance and volume of commuting contributed to C_2: indeed, partial correlation between C_2 and volume of commuters conditional on distance was 0.481.

Despite its decrease with distance, we noticed that for some areas criterion C_2 was significantly high compared to the distance between them (Figure 4.6).

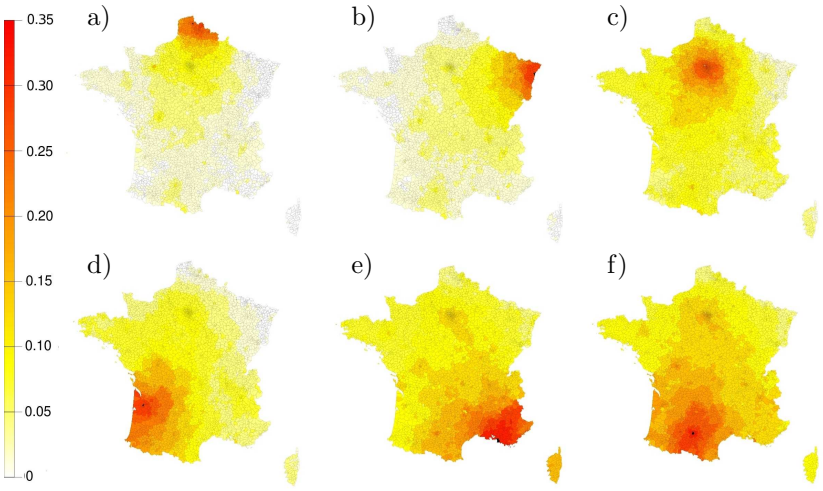

Figure 4.6: C_2 **for six districts and all other districts**
Criterion C_2 evaluated for six hubs, namely Lille (a), Strasbourg (b), Paris (c), Bordeaux (d), Marseille (e) and Toulouse (f)

Areas surrounding hubs had among them a criterion C_2 greater than with other districts located at approximatively the same distance: moreover, we noticed that the criterion was higher for some couples of hubs, like Toulouse and Lyon, than for others couples at the same distance, like Toulouse and Poitiers. This result confirmed that epidemic propagation did not happen solely by wave-like diffusion and was in line with observations made on the simulations, in which we noticed the attraction exerted by some hubs, like Paris on epidemics depending on their district of seed. Criterion C_2 would indeed be higher for a hub and a district if the disease seeded in this district diffused quickly to the hub.

Analyzing more specifically the values of criterion C_2 for each couples of hubs would be an interesting lead for future work: such study would permit to identify the existence of specific links between some hubs, which would indicate an important attractivity of each of them on the other. It could be interesting for prevision of future epidemics diffusion, as it could be a tool to predict preferential directions of propagation, depending on the current localization of the disease.

4.3.4 Structure of the network and similarities between epidemic propagation

In previous part, the analysis of simulation results with criterions C_1 and C_2 evidenced the existence of a spatial structure of influenza propagation: the diffusion being partly led by commuting movements, we sought how this structure could be linked to the commuting network. Given the negative correlation of both criterions with the geographical distance between district, we started this analysis by evaluating their correlations with the distance between nodes on the network. As the notion of distance didn't take into account the weights of edges, which play an important role in commuting networks, we then looked for other parameters to measure the proximity of two nodes and calculated their correlation with criterions C_1 and C_2. This study will be presented in the following chapters.

Shortest path

The pathway followed by an influenza epidemic depends both on the location of its first infected and the connectivity of seeding district. The positive correlation between both C_1 and C_2 with the rate of commuters exchanged by two districts indicated that epidemics are highly probable of following the same pathway when they are initiated in districts between which an important flow of commuters exists. However, as we also observed strong similarities between epidemics initiated in not directly linked districts, the existence of a direct link was not sufficient to explain the distribution of the criterions. To further investigate the link of the distribution of criterions with the structure of the network, we evaluated the correlation between both C_1 and C_2 with the parameter usually employed to measure proximity on networks, distance (Figure 4.7- a,b). We found strong correlations between distance and both C_1 ($r = -0.667 \pm 0.039$) and C_2 ($r = -0.681 \pm 0.042$).

Despite the high correlations observed, the coefficient of determination of

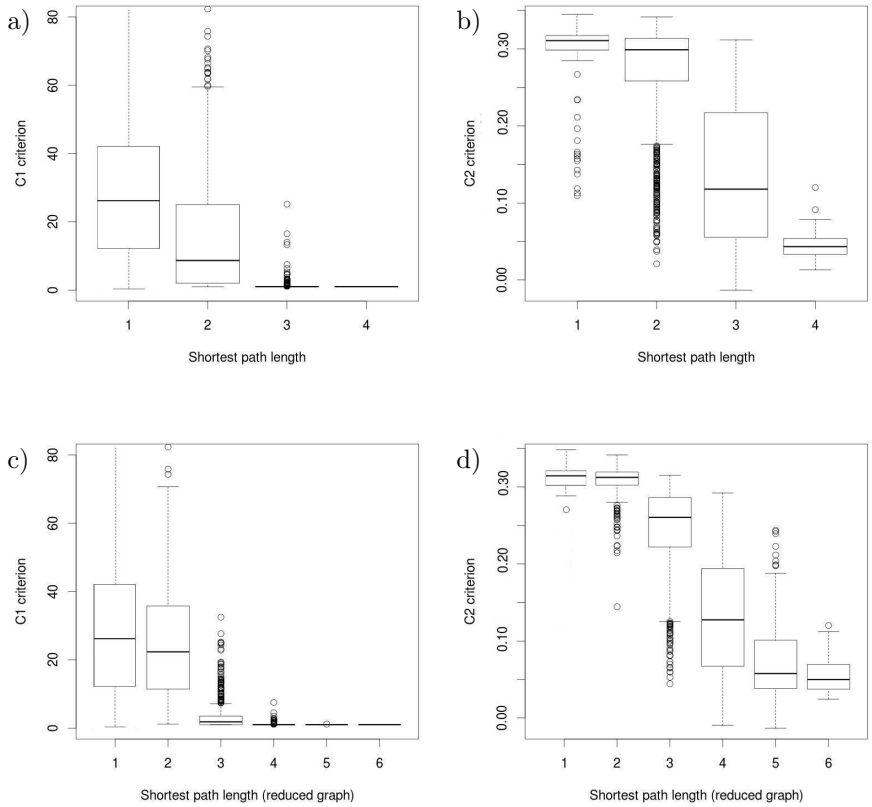

Figure 4.7: **Influence of shortest path length on criterion C_1 and C_2** Boxplots of criterion C_1 (a) and C_2 for couples of districts at distance 1, 2 3 and 4. Average value of both criterions decreases when the distance increases. Similar boxplots have been traced for C_1 and C_2 with couples classified by their distance on a reduced commuting network, where vertices representing the movement of less than 10 people have been ignored.

distance on C_1 and C_2 was low ($r^2 = 0.194$ and 0.470 respectively), indicating that the variations observed were not fully explained by distances.

Commuting networks expose a small-world structure: parameter S [76] was indeed significantly positive for both networks ($S = 9.864 \pm 0.005$ for school commuting and 5.328 ± 0.007 for work). In consequence, the average shortest path length is small in both networks: for 96% couples, shortest path length is inferior to 2. Therefore, distance does not provide a discriminant information on couples of nodes. Moreover, as 79.4% of edges in work commuting network and 87.0% in the school commuting one represent less than 10 people traveling,

most couples for which distance is low are joined by a shortest path including one of these weak links. The probability of influenza transmission through such path is small, due to the few commuters travelling through it: the small distance between districts thus didn't imply a strong proximity in terms of disease transmission.

Therefore, we re-evaluated correlations between the criterions and distances, after suppression in the networks of the edges representing the movement of less than 10 commuters. As previously, strong correlations were observed between distances and both C_1 ($r = -0.856 \pm 0.029$) and C_2 ($r = -0.868 \pm 0.015$) (Figure 4.7 - c,d). The coefficient of determination of new distance on C_1 and C_2 was higher than previously ($r^2 = 0.361$ and 0.667 respectively).

This result confirmed that common distance was not sufficient to understand the distribution of criterions C_1 and C_2: therefore, we looked for other measures of distance on networks, in which commuting traffics would be taken into consideration.

Flows of commuters

- Measuring distances in a transportation network The weights of the edges in commuting networks measure the intensity of travellers flow between districts: following a common approximation in transportation theory, they can be considered as capacities or maximum flows. This approximation allows for the generalization of measures commonly used in unweighted networks, notably distance [110].

To extend the concept of distance, we need to introduce a new notion: pathway transport rate. For a given pathway P, constituted of nodes $(n_1, ..., n_p)$ and weighted links $(f_{n_1 \to n_2}, ..., f_{n_{p-1} \to n_p})$, transport rate is defined as the ratio of its capacity ($c(P) = max_{1 \leq i \leq p-1} f_{i \to i+1}$) to its length. Assuming that transportation through each edge takes one unit of time, transport rate can be understood as the quantity of flow transmitted through this path by unit of

time. An equivalent of distance for weighted networks, l', can thus be deduced from this notion [22], with equation 4.1:

$$l'(m,o) = \frac{\sum f_{i\to j}}{N} * \frac{1}{max_{P(i\to j)} r(P)} \tag{4.1}$$

where N is the number of nodes in the network, $r(P)$ the transportation rate of path P and $max_{P(m\to o)} c(P)$ the maximal transport rate that can be obtained with a pathway going from node m to node o.

This expression requires the evaluation of all pathways relying nodes and can therefore be hard to compute for large networks: this difficulty can be circumvented using the concept of maximal flow.

- Maximal flow and Ford Fulkerson algorithm The maximal flow problem is a well known problematic of optimization theory, involving the search of the maximal volume that can be passed through a network from a source node to a sink. The capacities of edges constitute the constraints of this problem, as they limit the quantity of flow that can be transmitted between nodes. From this notion we can define the maximum average simultaneous transport rate between two nodes i.e the average transport rate when the maximal flow is transmitted between the nodes, given by equation 4.2:

$$r_{average}(m,o) = \frac{1}{F_{max}(m,o)} * \sum_{P\in S(m,o)} r(P) * c(P) \tag{4.2}$$

where $F_{max}(m,o)$ is the maximal flow from m to o and S(m,o) the set of pathways followed to transmit it.

We can then redefine the distance previously defined by equation 4.1 using this new notion:

$$l'(m,o) = \frac{\sum f_{i\to j}}{N} * \frac{1}{r_{average}(m,o)} \tag{4.3}$$

The problem of maximal flow being well explored, 4.1 is easier to evaluate than the previous one: in this thesis, the algorithm of Ford Fulkerson [55] was used

to determine the maximum flows and the pathways permitting its transmission.

- Flows and similarities between districts High correlation was found between redefined distance and C_1 ($r = 0.921 \pm 0.026$) and C_2 ($r = 0.952 \pm 0.042$). 4.8 The coefficient of determination of refined distance on both criterions was

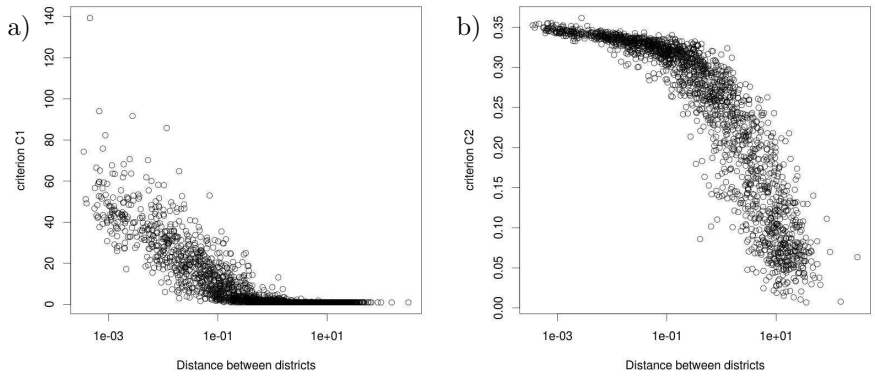

Figure 4.8: **Influence of distance on criterion C_1 and C_2**

higher than previously ($r^2 = 0.65$ and 0.75 respectively). This large values indicated that the similarities of epidemic propagations were strongly linked to the structure of commuting networks. Intensity of commuting traffic between nodes gives particularly a good explanation of the differences observed in the similarities between the whole propagation, measured by C_2.

4.3.5 Interpretation and perspectives

Influenza propagation is greatly influenced by the structure of commuting networks: indeed, the hubs of the network play a central role in the diffusion. Hubs are surrounded by basins of attraction, from which any epidemic will be attracted to them. Therefore, the similarity between epidemics initiated in different districts of the same basin of attraction will be higher than with epidemics starting from outside.

The division of districts between basins of attraction is directly linked to the structure of the network: indeed, the high correlation between criterions C_1 and

C_2 and the weighted distance between districts indicates that high similarities of early and total propagation were found for districts between which an important flow of commuters was exchanged. Districts gathered in the same basin of attraction are thus highly connected district. Due to the predominance of short-distance links, highly connected districts are mostly short-distanced ones: basins of attraction thus gather districts located in the same geographical area.

The three stages behavior observed on simulations is fully explained by the structure of commuting networks: the conjunction of predominant local commuting movements and attractiveness of hubs causes the first phase of wave-like local propagation when the existence of privileged links between hubs evidenced by C_2 analysis explains the sudden jumps of the epidemic.
The variability of initial pathways observed on simulation can also be explained by the structure of the network: the division of districts in spatially connex basins of attraction, linked to the strength of commuters flows between districts, explains the differences of patterns of propagation observed for epidemics starting in short- distanced districts.

This behavior is typical of propagations on scale-free networks such as commuting networks [16]: due to their central role in propagation, identifying the network's hubs is of crucial importance to design adequate public health policies [116]. However, for complex networks including a large number of nodes and edges, finding those hubs can be a tricky issue: the criterions defined in this part can be used to pinpoint hubs from simulation results and to find their basins of attraction, an important information to predict future propagation of epidemics.

4.4 Conclusion

In this chapter, we showed that epidemic in France exhibited a spatial structure, determined by the mouvements of commuting, which proves the interest of using commuting data to understand epidemic propagation on a national/ regional level.

To understand the consequences of mobility structure on the propagation of infectious diseases, we developed a metapopulation model based on commuting movement: we observed that simulations exhibited a three-phases behavior of propagation typically encountered in propagations on scale-free networks as commuting networks. Using two criterions based on the analysis of similarity between epidemic propagations, we showed that the network hubs could be retrieved from propagation data. Those hubs have a central role in propagation, as they attract and redistribute the epidemic: French territory could be divided in basins of attraction surrounding hubs, in which the epidemic propagation was attracted by the central hub. Thus, we showed that epidemic propagation followed recurring pathways, depending on their seeding district and proposed tools to predict the pathway an epidemic will follow depending on its current location. The structure in basins of attraction was shown to be directly linked to the intensity of commuters flows between districts.

*****

In this chapter, we proposed an analysis of the structure of influenza propagation based on criterions measuring the similarity between results of epidemic simulations. Those results gave an insight of the role of different districts in the propagation: to quantify this result, we developed a linearized version of our model and performed an analysis based on a study of its next generation operator. This analytical study will be presented in chapters five and six.

Chapter 5

Districts role in the system dynamics

In the previous chapter, an extensive exploration of potential epidemic propagation on French territory was performed and we evidenced the existence of preferential pathways of diffusion. The observation of the simulations also showed the presence of an apparently recurring scheme in propagation: whatever district influenza was seeded into, the epidemic first propagated mainly to neighboring districts, until it reached one of the network hubs. Once a first hub was reached, the disease then spread to other hubs, in some cases not geographically linked to the initial area of infection and was then transmitted to their neighbors. The study performed on criterions C_1 and C_2 was coherent with this observation: the small values of criterion for couples of distant districts confirmed that at the beginning, epidemic propagation only occurred on a local scale. The central role played by hubs was also confirmed by the high values of the criterion obtained for them and their surroundings.

Previous work by Barthelemy and al [16] had proven that in scale free networks, as the commuting networks, hubs provides transmission on a large scale and diffusion between other nodes occurs locally. Propagation on these struc-

tures has been shown to be highly predictable as soon as the epidemic reaches the main hubs [32], only the timing of this event and the paths followed are variable, depending on the node of seed [16]. To understand the diversity of possible patterns of propagation, an analysis of diffusion in the first stages of an epidemic is thus required.

Given the complexity of the model, making an analytical study of early dynamics is complex. However, during the initial phase of the epidemic, only few individuals are infected: the depletion of susceptibles can be neglected and it can be assumed that infectious individuals make contact with susceptibles only. Given this approximation, we can consider the number of susceptibles to be constant and linearize the equations of the model: new equations give a good approximation of the model dynamics in its first stages and its analytical study can be performed more easily [43].

5.1 Linearized model

5.1.1 Linearization

The linearized model can be summarized with equation 5.1.

$$X(t+dt) = \begin{pmatrix} I_{3704*5} & -S_{3704*5} * M \\ \\ 0 & T_{I'} + T_I * S_{3704*5} * M \end{pmatrix} * X(t) \qquad (5.1)$$

In previous equation, vector $X(t)$ describes the state of the system at time t, namely the number of susceptibles and infected in each districts and age classes, and keeps memory of the previous states of the system during the past 50 time steps (equation 5.2). The limit of 50 for the memory sized was fixed as it corresponds to a generation time of 10 days, which has less than $10^{-6}\%$

chances to occur.

$$X(t) = \begin{pmatrix} S_{0,0}(t) \\ \vdots \\ S_{0,4}(t) \\ S_{1,0}(t) \\ \vdots \\ S_{3703,4}(t) \\ \left.\begin{matrix} I_{0,0}(t-(T_H-1)dt) \\ \vdots \\ I_{0,0}(t) \end{matrix}\right\} T_H \\ \left.\begin{matrix} I_{0,1}(t-(T_H-1)dt) \\ \vdots \\ I_{0,1}(t) \end{matrix}\right\} T_H \\ \left.\begin{matrix} I_{3703,4}(t-(T_H-1)dt) \\ \vdots \\ I_{3703,4}(t) \end{matrix}\right\} T_H \end{pmatrix} \tag{5.2}$$

In equation 5.1, matrix S_{3704} is a diagonal matrix which diagonal is constituted of the number of susceptibles in each class. Given the linear approximation made for the first stages of the epidemic, this matrix is constant.
The evolution of the system state is accomplished by the product of vector $X(t)$ by the matrix M (equation 5.3).

$$M = \begin{pmatrix} & \beta_W N_W T_W^t W W^t T_W^* \\ + & \beta_S N_{S_1} T_{S_1}^t S S^t T_{S_1}^* \\ + & \beta_S N_{S_2} T_{S_2}^t S S^t T_{S_2}^* \\ + & \beta_C N_C T_C^t C C^t T_C^* \end{pmatrix} * (0.1 p_H + 0.9 p_O) + (\beta_H H_{3704} N_H T_H^*) * p_H \tag{5.3}$$

where W, S and C are respectively the matrices of commuting for work, school and community. The different generation times of individuals are represented with matrices p_H and p_O, calibrated to follow the gamma distribution of generation times in household and in the community. N_X are diagonal matrices which diagonal components are the number of people encountered via commuting or household contacts. Finally, T_X and T_X^* are simple transition matrices.

The similarity between the model and its linearized version was verified using overlap on 1000 simulations starting from randomly chosen initial states (Figure 5.1). Overlap was found to be superior to 0.99 for the first 6 weeks of simulation, before quickly decreasing, as the diminution of the susceptibles pool became too important for the approximation of constant susceptibles to stay accurate. The paths followed by epidemic propagation with both models being the same during first weeks, the linear model could thus be used for an analytical study of the epidemic behavior in its beginning.

5.2 Kernels

5.2.1 Kernel definition

The variance in epidemic propagation in the first weeks of an epidemic being largely correlated with the district of seed [16], we used the matrix of next

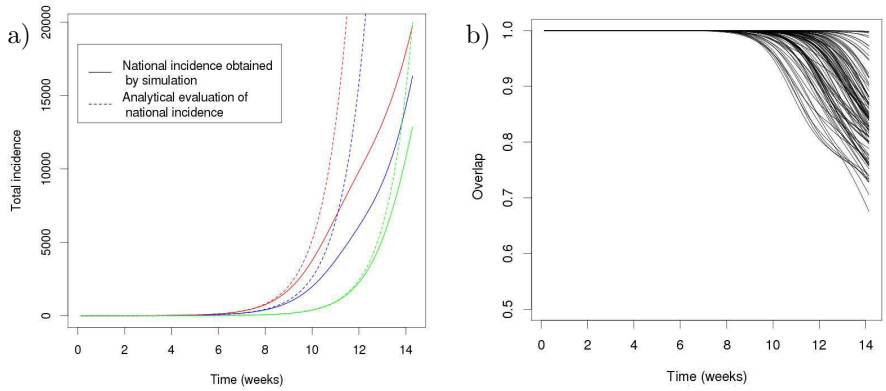

Figure 5.1: **Simulations with the commuting model compared to simulations with its linearized version**
a) Comparison of national incidence through time in 3 epidemics simulated with both the commuting model (continued line) and its linearized version (dotted line). Curves in the same colors are from simulations generated by starting in the same district and the same age class. b) Overlap between incidence of epidemics generated with the commuting model and the linear model. During the first 8 weeks of simulation, the overlap was greater than 0.99, indicating an important similarity between the propagation in the 2 cases

generation to determine the number of descendants that a single infected in each district and each age class could generate through time. This way, we defined a function k, giving the expected number of descendants of an individual of age a and from district d infected n days ago, as

$$k(a, d, n) = M \times X_{a,d,n} \tag{5.4}$$

where $X_{a,d,n}$ is a null vector, except for its $(50 \times 5 \times d + 50 \times a + 50 - n)$ component, equal to 1.

Function k is called the kernel of the model [43]: this tool of modeling was used to perform several analysis on the influence of each district and age class on the intensity and the extension of first stages propagation.

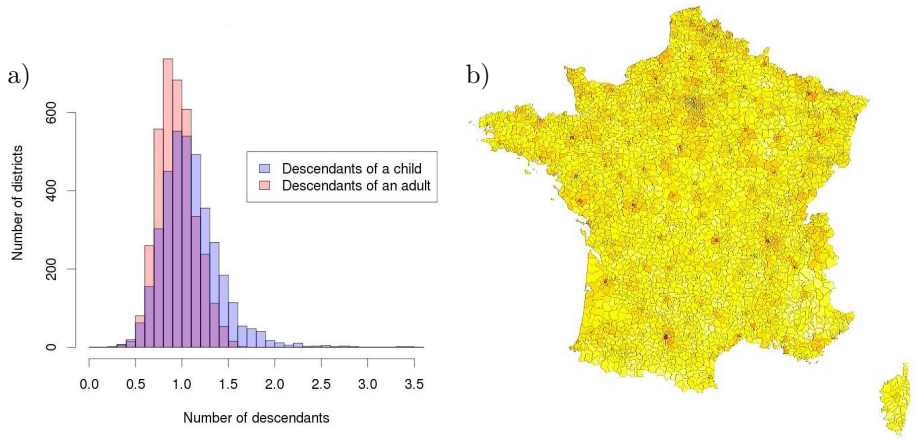

Figure 5.2: **Descendants of a single infected**
a) Distribution of the number of descendants for an infectious of a child age class (blue) and an adult age class (red). The dispersion was higher in cases of children, with a long tail of distribution for high values. On average, adults had less than 1 descendants, but children had more than 1. b) Number of descendants of a child for each district. The higher values were found in districts colored in red, which happened to be the hubs identified in previous study.

5.2.2 Kernel analysis

Number of descendants

The age class of an individual had a strong influence on the number of susceptibles to which it could transmit influenza: while elders and babies could infect a very limited number of people (respectively 0.1 and 0.3 on average) and weren't able to have at least one descendant in any district, adults and children of both age classes had a higher capacity of transmission. Children were also found to infect more individuals than adults: moreover, as a child had on average 1.10 ± 0.01 descendants, these age classes had the capacity to amplify the number of infected, while adults, whom number of descendants was 0.94 ± 0.02 , didn't (Figure 5.2-a). Number of descendants distribution was quite different in both cases: with a variance of 0.03, in case of an infected adult the distribution was massed around its average value and had a maximum of 1.6. However, for children, the distribution has a higher variance (0.1) and a maximum of 3.4, and, despite the one of adults, didn't exhibit a symmetrical structure, as it showed a long tail. The number of descendants of a child infected was found

to be superior to 2 in 42 districts, all of them being hubs of the commuting networks.

The highest numbers of people infected by a single individual, for both children an adult age classes, were found in the districts identified as hubs (Figure 5.2-b). In major cities, as Lyon, Toulouse or Marseille, a child could have more than 3 descendants, up to 3.4. This characteristic could notably be explained by the positive correlation between this parameter and the number of children inside the district ($p = 0.59 \pm 0.02$): as hubs tend to have more inhabitants than other cities, the quantity of children living there is also higher.

In certain districts, we observed that an individual from any age class infected less than 1 person: those districts would not be able to grow an epidemic from an initial single infected. However, in the previous chapter, we noted that the introduction of a single individual in any district led to a national infection: the network of connections linking the infection in different districts thus had the ability to support the epidemics seeded in districts were individuals had few descendants.

First-generation descendants localization

Due to the scale-free structure of both commuting network, most districts have many connections with other: a high proportion of districts were thus infected whatever the location of the initial individual. On average, 1159.5 ± 10.05 districts were infected by an adult first infectious, and 2133.9 ± 20.93 by a child (Figure 5.3-a). However, many links in the networks, especially the one of school commuting, are weak links, through which the probability of infection is small: therefore, for some of the districts considered as infected in previous calculus, the probability of transmission is actually small, sometimes inferior to 0.001%. If we only consider the districts where the probability of transmission was superior to 1%, the quantity of districts reached was considerably reduced, falling to 7.4 ± 0.10 on average for an infectious of the adult age class and 11.4 ± 0.13

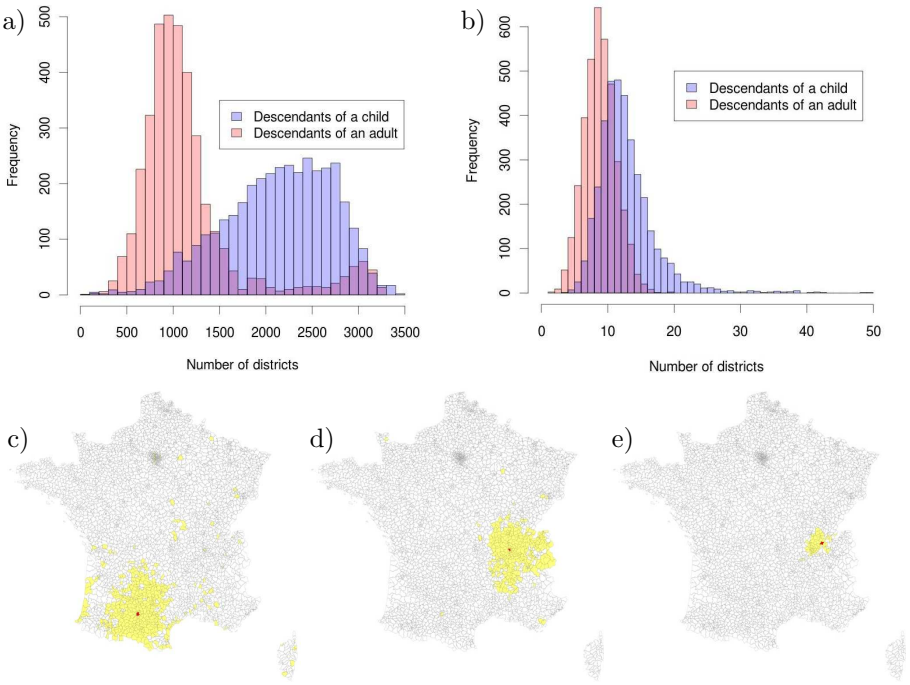

Figure 5.3: **Localization of first-generation descendants**
a) Distribution of the number of districts infected in case of a first infected who is a child (in blue) or an adult (in red). b) Same distribution, but considering only the district for which the probability of transmission is greater than 1%. c,d,e) Maps of the districts infected by respectively Toulouse (c), Lyon (d) and a randomly chosen districts. Only districts with a probability of infection greater than 1% were represented. While Lyon and Toulouse are hubs of the connectivity network, the randomly chosen district is not.

for one in a child age class (Figure 5.3-b). As for the number of infected, the distribution of infected districts exhibited the symmetrical structure of a gaussian distribution for an adult first infected and showed a long tail in case of a child.

Once again, the maximal number of districts reached from a single individual with a probability greater than 1% were found for the hubs of the network (on the other hand, we didn't find any specific pattern when considering the total number of districts reached, without distinction on the probability of influenza transmission). An illustration of this is given with Figure 5.3-c,d,e: we represented the districts having more than 1% chance of being infected by 2 hubs, namely Toulouse (Figure 5.3-c) and Lyon (Figure 5.3-d), and a randomly chosen district (Figure 5.3-d). This example illustrates the width of the area

reached by an infected placed in a hub compared to the one reached by an average infected.

This result was coherent with the previous one: on average, an infected individual placed in a hub would infect people on more districts, thus having in total more descendants than the average infected. The number of person infected in the district of origin was not different in average nodes and in hubs, despite the higher quantity of inhabitants of the latter: the number of people infected outside of the seeding district was therefore higher in case of a seed placed in a hub than in any other node. Hubs play both a role of amplification of the epidemic, increasing the number of infected and a role of distribution, as they enable the disease to reach a large number of places.

Moreover, we observed that the average distance between a district and the one it infects was higher in the case of a hub, meaning that they can export the disease to not neighboring areas.

Antecedents of each type of infected

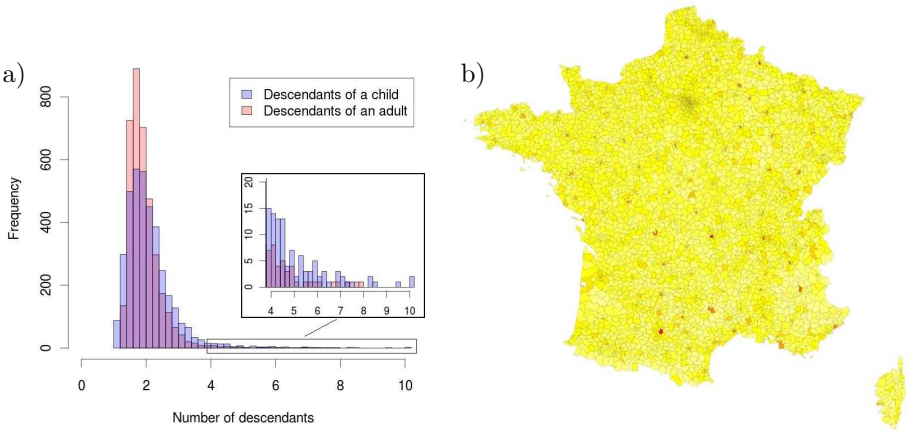

Figure 5.4: **Antecedents of the infected of each district**
a) Distribution of the number of people infected in a district by a distribution of one infected per district in a child age class (in blue) or in th adult age class (in red). The average number of infected was similar in both distributions, but with children infected, the variance of the distribution was higher. b) Number of infected in each district in cases of children as first infected. The higher values were found in hubs, here colored in red.

The location of an individual didn't influence only the number of people it

could infect, it also had an impact on its probability of being infected. Using the kernels, we looked at how many individuals were infected in each district if we put an infected child or an infected adult in each age class (due to the low capacity of disease transmission of babies and elders, we did not introduced an infected in those age classes): the total number of infected on average slightly differed depending on the age class of the first infected, as the average number of infected was 1.94 ± 0.01 in case of an adult and 2.09 ± 0.02 in case for a child. Both distributions had a long tail on their upper values, but more districts had an important number of infectious when the initial infected was a child than when it was an adult (Figure 5.5-a). The network of infection from adults therefore diffused influenza equally between districts, while the network from children concentrated the disease in some privileged districts. Similarly as what had been observed in the study of descendants, we found that hubs were the districts where most infections occurred (Figure 5.5-b). Moreover, we also looked from which places each district was infected and found that hubs were the places for which the probability of being infected by many different districts was the highest, i.e an infected in any randomly chosen district has a more important probability to have descendants in a hub than in any other district.

Hubs combine an important capacity of amplification, as any infected in a hub will have many descendants, in and outside it, and a central position in the epidemic, as there is a high probability for their individuals to be infected by sick people of other districts. This result could explain how epidemics can be bred from districts which individuals transmit the disease to less than 1 person on average: if the disease succeed in being exported to another district, it is highly probable that a hub will be infected and will play a role of amplificator for the epidemic.

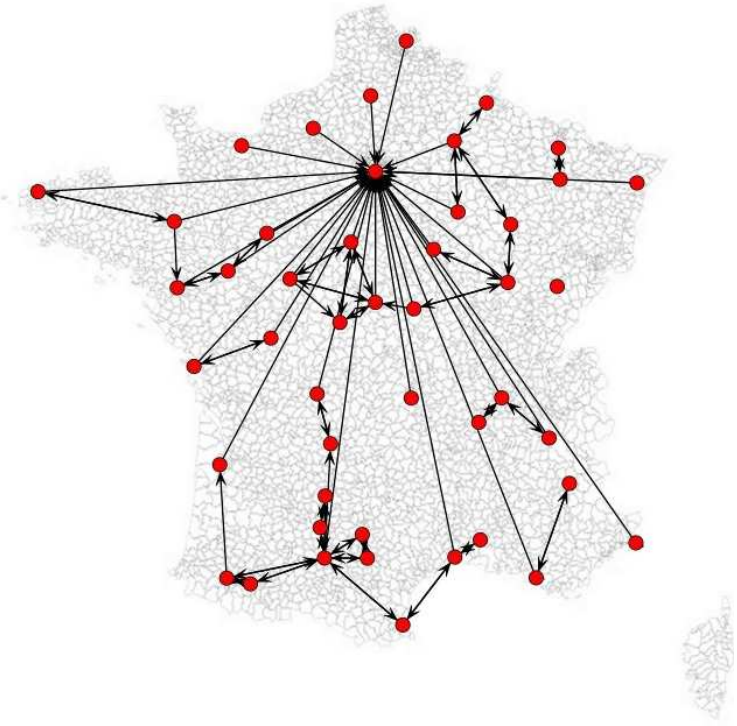

Figure 5.5: **Privileged links between hubs**

Linked hubs

The kernels also gave an insight of the intensity of the connection between two districts: the most descendants an infected from a district had in another one, the more connected they could be considered. We selected the 50 cities which were receiving the most infectious from the rest of the country, and observed how they were connected. On Figure 5.5 are represented the 50 cities (all of them were hubs): an arrow is represented between two nodes if one infected in one of them has more than 1% chance of having descendants in the other.

This analysis underlined the centrality of Paris, linked to most of the other hubs: this result was coherent with the simulations of epidemics, on which we had observed that Paris was a privileged direction of propagation for almost any district of seed. The hyper-connectivity of Paris, combined with its important number of inhabitants, would explain why it was among the first cities infected in each simulations. The organization of French urban network, historically

centered on Paris, could explain why Paris is the most connected of all hubs. Apart from this case, most links were between non distant hubs, which could be explained by the local nature of commuting. Local clusters could also be observed, in which local hubs had privileged contacts with a neighboring bigger hub. Some of these clusters could be observed in the vicinity of the important cities of Lyon, Toulouse and Marseille, three among the most populated French cities.

The same analysis could be done on all districts, to find the hub they are the most linked to. It could help to identify the direction in which an epidemic will be attracted, depending on its current localization and be an interesting tool to predict epidemic propagation.

5.3 Perspectives of use for the kernels

Given the linearity of the equations, the kernels can be used to evaluate the state of the system at a time step t depending on its previous states time steps.

$$X(t) = \sum_{n \in [1,\dots,n_{max}]} \sum_{(a,d) \in X(t-n)} k(a,d,n) \tag{5.5}$$

This property could be used on datasets describing the evolution of incidence in an epidemic to find the localization of its seeds. French influenza epidemics have multiple seeds: due to the communication of the country with its European partners and international communication with the rest of the world by airplane transportation, new influenza cases are re-introduced on the territory throughout the entire influenza season. In the incidences recorded by the Sentinelles network, the aggregation of the data on week and department drown the information of these re-introductions. Kernels could thus be used on incidence data to reconstruct previous time steps with the vectors of incidence of an epidemic from the week of its official start. With this work, the main places

of re-introduction of influenza in France could be identified: this information could be used to design public health policies targeting these places as a priority in order to limit the arrival of new influenza cases during the influenza season.

5.4 Conclusion

In this chapter, we have linearized the model defined in the previous chapter: the linear equations were used to define kernels of propagation, which give the number of descendants of an individuals depending on its age and district of residence.

These kernels were used to study the influence of individuals on the propagation of an epidemic: we found that the hubs of the network play a double role. On one hand, they attract the epidemic, as their residents can be contaminated by infected from many other districts. On the other hand, they amplify the epidemic and redistribute it, as an infected from a hub has many descendants, distributed on a large number of districts. The attractiveness of these districts and their propensity to amplify the epidemic compensate the low capacity of some other districts in the network, which inhabitants have less than 1 descendant when infected. An analysis of the contacts between hubs also highlighted that Paris was the main attractor of the epidemic, as infected in most hubs have descendants in Paris. We also observed that hubs located in the same geographical area formed little clusters, surrounding the most populated hub of the area. This analysis was coherent with the observations performed on the simulations of the model in the previous chapter.

We also found that children were the main vectors of influenza transmission: on average, an infected child has more descendants than an infected adult from the same district. This result confirmed the results obtained in chapter 2.

*****

In this chapter, we used the linearized model to study the influence on epi-

demic propagation of each district individually: we have seen that only some of them have the capacity of amplifying an epidemic. However, this study does not give information on the behavior of the whole system: the following chapter will address this issue.

Chapter 6

Analysis of the system global dynamics

In the previous chapter, we developed a linear model to approximate the behavior of the model at the beginning of a simulation. From this model, we defined for each category of individual, a kernel giving the number of descendants of an infected in this category through time. An analysis of kernels highlighted the importance of the networks hubs, which play the role of attractors and amplificators for the epidemic. This analysis gave us an insight on the specific influence of each district separately: in this chapter, we will study the general behavior of the model and study its capacity in sustaining an epidemic whatever district influenza is seeded into.

To do so, we will take an interest in R_0, the basic reproduction number of the model, which is defined as the average number of new cases generated by a single typical infected individual. R_0 is one of the most important concept in epidemiology [71]: originally developed for demographic studies [44, 131], it was introduced in epidemiological research to study vector-borne disease as malaria [128] then directly transmitted disease [90] and is now widely used to study infectious disease models. It is among the first quantities estimated in case of

an emerging disease outbreak and is used to help in the design of interventions for recurring infections. R_0 is mainly used as a threshold: indeed, from its definition, it is clear that if $R_0 < 1$, each infected individual infects less than one new individual and the infection will disappear from the population. On the contrary, if $R_0 > 1$, the pathogen will be able to invade the susceptible population. The magnitude of R_0 also gives an insight of the speed of the number of infected growth.

Initially defined for simple homogeneous models, in which R_0 is directly linked to the probability for a contact between a susceptible and an infected to be infectious, the concept has been extended for complex multi-compartment models and several methods have been proposed to evaluate the R_0 of such models. Notably, [43] defined the next generation operator, from which R_0 is the spectral radius. The next generation operator depends on the equations of the linear model: to explain its calculation, let us consider the following linear system:

$$\Delta x(t) = A \times x(t) \tag{6.1}$$

where x is a n-dimensional vector representing the number of infectious individuals in each category of the system at time step t. If the spectral bound (i.e the maximal real part of all eigenvalues), here noted $s(A)$, is inferior to 0, the system is stable and all variable decay to 0. On the contrary, if $s(A) > 0$, the system is unstable, meaning that an epidemic can be generated with this system. The next generation theorem [41] states that, for any decomposition $A = F - V$ where $s(-V) < 0$, $V^{-1} \geq 0$ and $F \geq 0$,

$$\begin{cases} s(A) < 0 & \Leftrightarrow & \rho(FV^{-1}) < 1 \\ s(A) > 0 & \Leftrightarrow & \rho(FV^{-1}) > 1 \\ s(A) = 0 & \Leftrightarrow & \rho(FV^{-1}) = 1 \end{cases} \tag{6.2}$$

Here, $\rho(X)$ defines the spectral radius of the matrix X and $F \geq 0$ requires that all elements of F are greater than or equal to 0. F and V can be chosen to

respect the conditions of the previous theorem and to have an epidemiological interpretation [77]. A classical decomposition is to set F as the matrix giving the rate at which new individual are infected and V as the one describing the internal movement of individuals between the classes (due to recovery). When F and V are chosen as such, FV^{-1} forms the next generation operator and $\rho(FV^{-1}) = R_0$ [43].

In the previous chapter, we developed a linearized version of our model: we will see in next part how we can define the next generation operator given these equations.

6.1 Next generation operator

From the previous chapter equation, we can extract a subset of linear equations describing the evolution of the vector of infectious in each class:

$$X_I(t + \Delta t) = T_I^{'} + T_I \times S_{3704\times5} \times M \times X_I(t) \tag{6.3}$$

where $X_I(t)$ is the part concerning the number of infectious individuals of vector $X(t)$ defined in the previous chapter. This expression can be related to previous system:

$$\begin{aligned} X_I(t + \Delta t) &= X_I(t) + \Delta X_I(t) \\ &= X_I(t) + A \times X_I(t) \end{aligned} \tag{6.4}$$

We can therefore relate the matrix M to matrix A:

$$A = T_I \times S_{3704\times5} \times M - (Id - T_I^{'}) \tag{6.5}$$

We find here the decomposition proposed previously, with $F = T_I \times S_{3704\times5} \times M$ and $V = Id - T_I^{'}$: R_0 of our model can be evaluated with the spectral radius of FV^{-1}.

In next part, we will study the eigenvalues of FV^{-1}. To simplify the calculation, we only kept one time step for each infected class, suppressing the 'memory'

of the system: in this case, we have $T_I^{'} = 0$, thus $V = Id$. The next generation operator is then equal to the next generation matrix $T_I \times S_{3704\times5} \times M$: the calculation of R_0 can be completed with an analysis of the eigenvectors of the matrix, which throw here light on the dynamics of the system.

6.2 Eigenvalues and eigenvectors of the next generation operator

The complete next generation operator is a $((3704 \times 5) \times (3704 \times 5))$ matrix: given the size of this matrix, evaluating its eigenvectors and eigenvalues was a complex task, requesting important calculation resources. Therefore, instead of performing the calculation for the entire matrix, we designed smaller models, including a restricted number of either specifically or randomly chosen districts. Small models were constructed in the same way as the global one, but including only some of the 3704 districts: all edges bounded by at least a node outside the subset were suppressed from the commuting matrices and we considered all commuting happened in the subset considered, in the same proportions than the one found in the complete commuting networks. The calculation was performed on 100 different subsets of between 100 and 400 districts and similar results were obtained in each cases.

6.2.1 Isolated districts

As a first analysis, we isolated each district from the others and considered all children and workers were commuting inside their district of origin. We evaluated the next generation operator for each of them and calculated their spectral radius: we found that its variance among districts was low. Indeed, the average R_0 was 2.29, with a variance of 3×10^{-3}: this result can easily be understood, as the only difference between the models restricted to a single district was the number of inhabitants and the rate of contacts between age classes. The number of inhabitants had little impact on R_0: given the formulation of the force

of infection in school, work and community, it was not influenced by the size of the population. Only the inside-household part of the force of infection was modified, but only in function of the repartition of the population between age classes. Similarly, the rate of contacts between age classes only modified the force of infection in households and not the 3 others.

This result shows that an epidemic can be generated in each district individually, while in the previous chapter, the analysis of kernels showed that in some districts, the number of descendants of an infected was inferior to 1, whatever the age of the first infected, making them unable to breed an epidemic without communication with other districts. The heterogeneity of connections and degrees increases the heterogeneity in the number of descendants of a first infected between districts, which is very limited in the absence of commuting. The presence of external commuting thus increases the difference between districts, reducing for some of them their capacity to generate an epidemic.

6.2.2 Partial analyses for isolated areas

Complete commuting networks

Following previous analysis, we constructed several models based on different subsets of districts: as a start, we built these models by selecting the districts of a department and its surrounding departments. We then calculated the next generation operators of these models and studied their eigenvalues and eigenvectors. The figure 6.1 shows an example of result obtained for a subset centered on the department of Bouches du Rhone, where the French city of Marseille, second most populated city, is located: similar results were obtained for the various subsets we analyzed.

In each cases, we found that, among the multiple eigenvalues of the next generation operator, several were superior to 1 (Figure 6.1-a): this result influenced our analysis of eigenvectors. In classical analysis of the next generation matrix eigenvectors, only the dominant eigenvector is analyzed as the state of

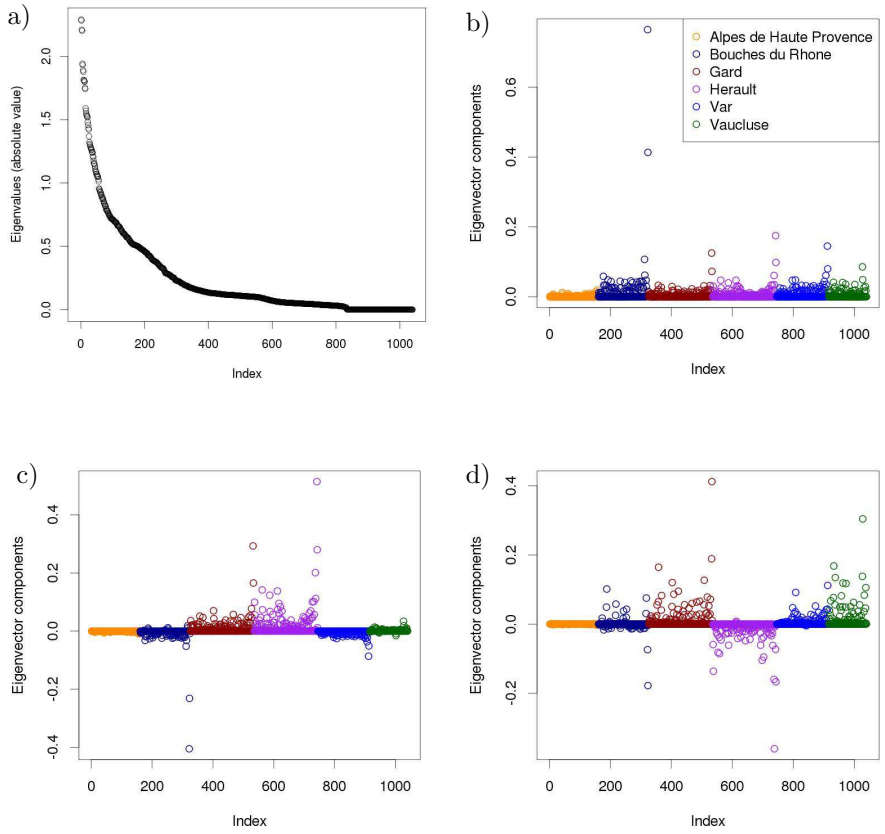

Figure 6.1: **Eigenvalues and some eigenvectors for the subset of districts surrounding Marseille**
a) Eigenvalues of the next generation operator for a model restricted to the districts of the 5 departments surrounding the city of Marseille b) Dominant eigenvector of the same matrix: the highest component are found for the 2 children age class of the most important cities of this subset of districts c) and d) Two others eigenvectors, associated with eigenvalues superior to 1. High component are found in a mix districts from different departments

the system tends to follow the distribution of infectious individuals encountered in it [88]. To explain this, let us reconsider the system of linear equations previously introduced: $\Delta x(t) = A \times x(t)$. If A is a $(n \times n)$ matrix, it has n eigenvalues λ_k, to which n eigenvectors V_k composing a base of the space $\mathbb{C}^n$ are associated. Therefore, for each initial vector X_I, n coefficient $\{i_1, \ldots, i_n\} \in \mathbb{C}^n$ such as

$X_I = \sum_{k=1}^{n} i_k V_k$ can be found. Therefore, we can write:

$$\begin{aligned} A \times X_I &= \sum_{k=1}^{n} i_k \lambda_k V_k \\ A^m \times X_I &= \sum_{k=1}^{n} i_k \lambda_k^m V_k \end{aligned} \tag{6.6}$$

If we suppose that λ_1 is the dominant eigenvalue, we can see that, when n grows, λ_k^m can be neglected before λ_1^m for all $k \neq 1$: the evolution of the system thus tend towards a repartition of infectious individuals proportional to V_1. In our case, as multiple eigenvalues are superior to 1 and cannot be neglected before the dominant eigenvalue, a better understanding of the system dynamics can be obtained by studying the subspace generated by the base of eigenvectors associated with eigenvalues superior to 1. Therefore, instead of studying the dominant eigenvector only as it is usually done, we performed an analysis of the subspace generated by eigenvectors associated to eigenvalues superior to 1.

In each case, we found that all components of the dominant eigenvector were positive and real: in other eigenvectors, both positive and negative real components were found (it is demonstrated in [43] that due to the construction of the next generation matrix, the dominant eigenvector is the only positive eigenvector). For each subset of districts, we found that the highest components of the dominant eigenvector were the ones associated to the children age class of the main hubs of the subset of districts (Figure 6.1-b): in the example of the subset centered on Bouches du Rhone, the highest component was found for Marseille. The components associated to children in other important cities, which were smaller than Marseille and were not among the cities identifies in the previous chapter as main amplificators of the system, also had values higher than the rest of districts but smaller than the one of Marseille. Dominant eigenvectors in other subsets showed similar results. If the largest values observed in hubs could be explained by their more important population, the fact that the highest components are associated to children highlights that classes of children are more affected by the epidemic.

In the other eigenvectors, we observed that highest components could be

found for districts in different departments (Figure 6.1-c,d), which indicates a correlation between the evolution of incidence in different departments.

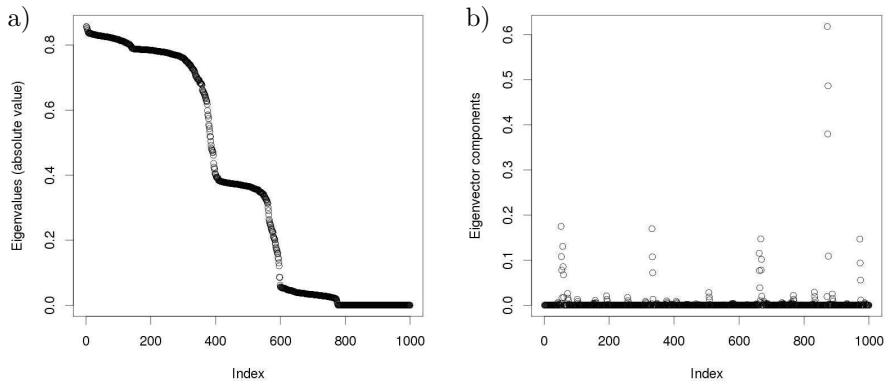

Figure 6.2: **Eigenvalues and dominant eigenvectors for a random subset of districts**
a) Eigenvalues of the next generation operator for a model restricted to a random subset of 200 districts b) Dominant eigenvector of the same matrix: the highest components were found for local hubs and populated cities

To complete this analysis, we performed the same calculation on random subsets of districts, of approximately the same size. Similar results were obtained, with a continuum of eigenvalues and a dominant eigenvector in which maximal components were found for the most important cities of the subset, whether a local hub or a global one: an example of these is given in figure 6.2. On the subsets we tested, we noted that eigenvalues were lower when no global hub was present in the subset: in some cases including neither global nor local hubs, the dominant eigenvalue was inferior to 1 (Figure 6.2). We didn't perform an exhaustive analysis to prove this result: further work would be necessary to understand how the presence or absence of major hubs affect the eigenvalues of the next generation operator.

However, to investigate further the role of hubs, we performed a first test, by suppressing all outgoing links from the children in hubs and analyzed the next generation operators in these cases.

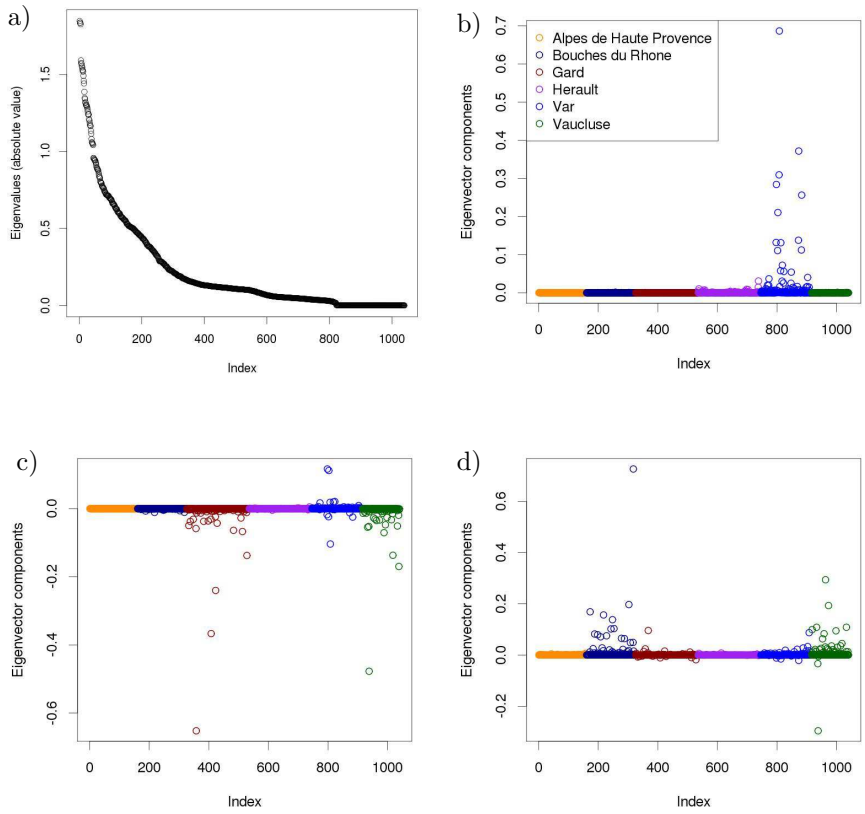

Figure 6.3: **Eigenvalues and some eigenvectors for the subset of districts surrounding Marseille without hubs**
a) Eigenvalues of the next generation operator for a model restricted to the districts of the 5 departments surrounding the city of Marseille and where the children of hubs can't infect any susceptible b) Dominant eigenvector of the same matrix: the highest components are regrouped in the same department. c,d) 2 eigenvectors associated with eigenvalues superior to 1: the highest components are found on districts belonging to the same department

Isolation of the children in hubs

We performed this experiment on the sub models centered on a department and its surrounding departments. In these models, we selected the districts for which the components where significantly higher than the others in the dominant eigenvector and suppressed the outgoing links for the children in them. Then, we evaluated their next generation operator and analyzed their eigenvalues and eigenvectors. On figure 6.3 are shown the results obtained for the sub model centered on the department of Bouches du Rhone.

In each case, we noticed a diminution of the next generation operator eigenvalues (Figure 6.3-a): the dominant eigenvalue was smaller and less eigenvalues were superior to 1. Preventing the children from the main cities from transmitting influenza causes a decrease of the system basic reproduction ratio, which slows down the transmission of the epidemic.

The eigenvectors were also modified: the highest components of the dominant eigenvector were regrouped on the districts of a single department (Figure 6.3-b). Similarly, in all eigenvectors associated with an eigenvalue superior to 1, the high components were gathered on the districts of 1 or 2 departments. In the absence of infections caused by the children of hubs, the evolution of the epidemic in each district is mainly correlated with its evolution in districts of its department. This result would need further investigation to be proven, but it would confirm that the transmission of influenza between clusters of districts is mainly conducted by the hubs and especially by their children.

6.2.3 Perspectives

Given the similarity of observations made on each model based on a random subset of districts we tested, which structure reproduces exactly the one of the entire model, from the characteristics of commuting network to the way individuals interact, we believe that the results obtained on small models can be extended to the global model. Actually performing this analysis could be interesting to understand the model dynamics: indeed, dominant eigenvectors could be used to identify groups of districts in which the evolution of incidence is correlated: the clusters identified could be compared to the ones found in chapter 2. In the calculation of eigenvectors performed on sub-models, we observed that the highest components of dominant eigenvectors were the districts belonging to one or more departments, thus defining connex clusters, as were the clusters identified in chapter 2. We can suppose that a similar result would be obtained for the complete matrix, which would confirm the existence of higher similarities between non distant districts caused by the predominance of local commuting.

The study of the dominant eigenvector could also be used to confirm the role of amplificator of children age classes and hubs.

In previous part, we have seen that suppressing the capacity of hubs of transmitting the disease caused a decrease of R_0 and all eigenvalues and diminished the correlation of incidence between districts belonging to different departments: the same test could be performed on the complete matrix. This would indicate that the transmission of the disease between clusters is mainly driven by hubs.

6.3 Conclusion

In this chapter, we proposed an analytical study of the model presented in the previous chapters, based on the calculation of the eigenvalues and eigenvectors of its next generation operator. An analysis of smaller models, including only a subset of districts showed that children in hubs were the most affected by the disease. With the dominant eigenvectors, we observed that districts could be clustered in connex groups, in which the evolution of incidence is correlated. The values of dominant eigenvectors components also indicated the existence of correlation between some clusters. When we suppressed the capacity of children in hubs to transmit the disease to others, the correlations previously observed between clusters significantly decreased.

*****

The results presented in chapter four to six were obtained by a work of extensive simulations, using big datasets, and analysis of their results, composed of thousands of text files. To perform this study, the scientific work of defining methods and criterions has to come with a technical work, to implement algorithms able to compute the simulations and analysis in a reasonable time. This work will be exposed in next chapter.

Chapter 7

Technical development

To complete the work presented in chapter four to six, we had to work with big datasets: each commuting networks contained several hundred of thousand entries and the composition of households was registered for 3704 districts, each containing individuals from five different age classes. We performed our simulations on a model composed of 3704 districts and five age classes, dividing the susceptibles in 18520 compartments. The simulations were realized on a time frame of 180 days, with a time step of $\frac{1}{5}$ of a day. Simulating epidemics thus required the manipulation of big files of data and to dispose of important memory space, as the simulations required to remind the state of the systems and its previous states on 50 time steps. Due to these constraints, we had to develop technical solutions to manage the issues encountered with the time and memory needed for simulations: they will be presented in the first part of this chapter. Once the simulations were performed, we were also confronted to problems caused by the size of the data generated. Indeed, to analyze the results, we had to memorize an important quantity of data for each simulation: the treatment and the storage of these data have quickly become an issue. Once again, we had to think of solutions to overcome this problem: they will be exposed in the second part of this chapter.

7.1 Management of the simulations

At first, we developed a stochastic model and wanted to execute an extensive exploration of eventual epidemic pathways, by realizing several replications of epidemics starting in each of the 3704 district and each of the five age classes. However, due to the size of the data and the number of information that the system had to keep in memory for each epidemic, the time needed for a simulation was important (eight to ten minutes). Even if we had at our disposal a cluster of computers, on which 20 nodes were available for this work, the time needed to complete all simulations would have enormous (from 40 to 102 months). To reduce calculation time, we developed a second version of our model, that could be launched on a grid of computers, on which the calculation power was higher. This didn't solve our problems: each simulation generated a file of size 2MO: all files together thus represented 3TO. The time needed to get these files back from the grid was once again enormous: to overcome this issue, we had to reduce the number of simulations we wanted to realize.

7.1.1 OpenMOLE and simulations on grid

During this thesis, the Institute of Complex Systems (ISC) gave us an access to the ISC-PIF grid of computers, which gathers several thousands of servers, dispatched in different countries: we therefore gained access to an important capacity of calculation. To interface so many different servers, ISC has developed OpenMOLE, a work flow engine designed to enabled the launch of applications programmed in different languages (among which C, C++, Java, R...) in parallel on different computers from the grid, whatever their location. To use the grid, the program to be launched in parallel has to be inserted in an OpenMOLE capsule: the start of parallel simulations and their repartition on available computers is done by an OpenMOLE program.

We developed an OpenMOLE program to encapsulate our program and launch the simulations in parallel. The time needed to realize the exploration of

potential epidemic pathways decreased, as many more computers were available to perform the simulations. However, we encountered another issue: as simulations were performed on distant computers, the result file had to be sent back. Due to the size of the files, this task asked for a longer time than the simulations themselves. Therefore, we chose to make the simulations on the cluster at our disposal, but to reduce the number of simulations realized.

7.1.2 From stochastic to deterministic

To reduce the number of simulations, we first decided to use a deterministic version of our model instead of the stochastic one. Indeed, in this thesis, we aimed at analyzing the average behavior of epidemics and identifying the principal pathways used by influenza to diffuse among districts. We were not interested in the variations around the average behavior: the analysis could thus be performed with a deterministic model. This decisions permitted to divide the number of simulations required by 100.

When we performed a first set of simulations, we observed that it wasn't necessary to simulate propagations created by individuals from each 5 age classes: as babies and elders mainly have contacts with children and adults of their own district, a first infected of these age class will mostly transmit the disease inside its district. The infection will be transmitted to other district by infected children and adults: thus the pathways followed by the epidemic will be similar to the one used when the first infected is an adult or a child. Similarly, it is not necessary to launch simulations from both children age classes: the network of commuting used by both classes being the same, the patterns of propagation started by individuals from them are also the same. Therefore, we decided to only simulate epidemics starting from the class of adults or from one class of children.

7.2 Management of the results

Even if we performed less simulations than initially planned, the size of the files generated by simulations remained an issue. To perform the analysis of our results, we needed to perform all simulations using the matrices of commuting and several randomized versions of these matrices. Simulations performed with one couples of matrices, with all possible initial states, generates an ensemble of files of size 17GO. To storage all files needed for the analysis, we needed bigger storage space than the one we had: this issue was easily overcome by the purchase of external devices of storage. However, this solution also brought another problem: as data were stored on different devices, the treatment of results needed to interface these different spaces. We chose to perform all analysis on a local server and to create automatic routines to copy the files from their storage space when needed and delete them after. A part of the thesis was thus dedicated to the development of such programs, using the languages bash and expect, to interface my computer, the cluster, where simulations were performed and the FTP servers were stored.

*****

Chapter 8

Conclusion

Last decades have known a spectacular improvement of computational capacities: nowadays, data storing and analysis are possible in an unprecedented extent. In chapter three, we showed that epidemiologists have taken advantage of this evolution by constructing more and more complex systems to register precise data on human mobility, that were integrated in models giving a more and more accurate description of human movement. If the realism of models is indubitably improved by this process, they also gain complexity through it: understanding how the structure of networks influence propagation and if their usage is pertinent become more and more difficult.

For many infectious disease, the set up of extensive surveillance plans have permitted the construction of large datasets describing their propagation with an increasing spatial and temporal precision: extracting the spatial structure of these data and comparing it to the complex networks of contacts is of crucial importance to understand the mechanisms underlying the propagation. Moreover, the design of suitable public health policies requires for an extensive analysis of this structure [36, 83]. In this context, we are in great need of new tools, allowing the analysis of complex, large-size datasets, to extract their structure and rely it to contact networks.

Commuting data are a good example of this issue: despite their frequent

use in epidemologic models [48, 49, 96], few evidence has been given that they give a description of movements useful to understand epidemic propagation and few analysis have been made to describe how they structure diffusion. In this thesis, we performed such analysis, based on the study of influenza propagation in France: to perform this work, we defined several tools, providing new methodologies for the analysis of observed epidemics and mobility networks.

8.1 Influence of commuting structure on influenza propagation

8.1.1 Role of commuting in the propagation

Using the time series of influenza-like illness recorded by the Sentinelles network since 1984, we found that the magnitude and timing of epidemics in departments connected either by work or school commuting were correlated.

Thus, influenza propagation shows a spatial structure, related to commuting movements, in coherence with the results exposed in [137]. This result evidence the pertinence of using commuting data to understand the propagation of infectious diseases in countries where the patterns of contact are similar to France.

Since we showed that commuting movements influence influenza propagation, we took an interest in the consequences this influence had on propagation: we looked for the existence of specific behaviors of propagation induced by the structure of commuting movements. To perform this analysis, we generated data of influenza propagation with a metapopulation model based on commuting movements.

8.1.2 Age related commuting

In simulations, we found that age-related differences in commuting had an impact on propagation: ignoring either adult commuting or children commuting,

which constituted different networks, modified the timing of simulations. Indeed, the interaction of both networks tended to homogenize epidemic curves, when the timing of epidemics simulated after the removal of either of them was greatly variable. As many models including commuting only take into account one of these networks [10, 102], this results highlights the importance of using data on both networks in epidemic models.

8.1.3 Underlying structure: attractors and basins of attraction

On all simulated epidemics was observed a three-times recurring pattern, starting with a phase of wave-like local diffusion followed by quick transmission of the disease to French most populated cities as soon as the disease had reached one of them. Wave-like propagation from these cities to their neighborhoods followed.

Using both criterions measuring the similarity between epidemic propagations and analysis of the kernels of the linearized model, we showed that a subset of populated cities, hyper-connected nodes of the commuting network, played a role of attractor and amplificator of the epidemic. Due to this dual role, they were both at high risk of being infected and transmitting the disease to a large set of other cities. All French cities could be associated to at least one of this hub, to which it preferentially transmitted the disease. We could thus divide the territory in clusters, centered on a global or local hub, to which the epidemic was attracted when starting anywhere in the cluster. Suppressing the capacity of hubs to transmit the disease to other districts greatly decreased the communication between the clusters, showing that the diffusion of influenza between them is largely caused by transmission from the hubs. Using similar tools, we evidenced the existence of preferential transmission between hubs, inducing the existence of preferential pathways of propagation between the clusters. Among all hubs, Paris played a focal role, as it was connected to all other hubs.

The propagation is greatly influenced by the structure of commuting net-

work: indeed, the behavior we observed in the simulations is typical of propagations on scale-free networks as commuting networks [16]. The propagation is mainly driven by the hyper-connected hubs, which both attract and redistribute the epidemic. Paris, the most connected hub of the network, plays a central role in the propagation and is among the first infected in all simulations. On another hand, the important local clustering of the networks induces a wave-like local propagation between non-hubs neighboring districts. Identifying those patterns of propagation and particularly the hubs, is of great importance to design adequate public health policies [116], as the targeted vaccination of hubs could significantly reduce the attack rate of the epidemic, when vaccination of random cities would have only limited effect. Identifying recurring patterns of propagation can also be used to predict the hubs in which influenza will diffuse in real time in order to vaccinate them preferentially.

Finding hubs and their basins of attraction is thus an important challenge to understand propagation on a network. This research can be turned in a tricky issue due the complexity of social network, but the methods we developed in this thesis to treat the example of influenza propagation on commuting network can be used to overcome this issue and seek for hubs and their basins from propagation data.

8.2 New methodology for network analysis

In this thesis, we developed both methods to be applied on simulation results and analytical methods to directly study a model of propagation. These tools give results on the early behavior and the general dynamics of the epidemic.

8.2.1 Early propagation

Due to the predominance of short distance commuting, most propagation, including early propagation, occur on very short distance. Long-distance transmission mainly occurs when the epidemic reaches a hub, from which it is trans-

mitted to other districts.

Criterion C_1 and kernels were defined to analyze this early propagation, based on similarities between epidemics propagation and the localization of an infected descendants depending on its district of residence. As early propagation tend to be directed towards hubs, this analysis permitted to pinpoint them and to find their basins of attraction.

8.2.2 Global dynamics

Despite the importance of local propagation, influenza diffusion also occurs on longer distances, mainly between hubs: due to the existence of strong commuting links between some of them, influenza diffusion follows preferential pathways.

To analyze these pathways, we defined criterion C_2, based on the measurement of epidemic similarity and the next-generation operator, which also gave us an insight on the model global dynamics.

Bibliography

[1] World Bank data, 2011.

[2] M Ajelli, B Goncalves, D Balcan, V Colizza, H Hu, J J Ramasco, S Merler, and A Vespignani. Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. *BMC Infectious Diseases*, 10, 2010.

[3] M Ajelli and S Merler. An individual-based model of hepatitis A transmission. *Journal of Theoretical Biology*, 259(3):478–488, 2009.

[4] R Albert and A L Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, January 2002.

[5] R Albert, H Jeong, and A L Barabasi. Internet - Diameter of the World-Wide Web. *Nature*, 401(6749):130–131, September 1999.

[6] L A N Amaral, A Scala, M Barthelemy, and H E Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences of the United States of America*, 97(21):11149–11152, October 2000.

[7] Julien Arino, Fred Brauer, P van den Driessche, James Watmough, and Jianhong Wu. A final size relation for epidemic models. *Mathematical Biosciences and Engineering*, 4(2):159–175, 2007.

[8] P Bajardi, C Poletto, J J Ramasco, M Tizzoni, V Colizza, and A Vespignani. Human mobility networks, travel restrictions, and the global spread of 2009 H1N1 pandemic. *PLoS One*, 6(1):e16591, 2011.

[9] Per Bak, Kan Chen, and Chao Tang. A forest-fire model and some thoughts on turbulence. *Physics letters A*, 147(5):297–300, 1990.

[10] D Balcan, V Colizza, B Goncalves, H Hu, J J Ramasco, and A Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489, 2009.

[11] D Balcan, H Hu, B Goncalves, P Bajardi, C Poletto, J J Ramasco, D Paolotti, N Perra, M Tizzoni, W den Broeck, V Colizza, and A Vespignani. Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. *BMC Medicine*, 7, September 2009.

[12] D Balcan and A Vespignani. Phase transitions in contagion processes mediated by recurrent mobility patterns. *Nature Physics*, 7:581–586, 2011.

[13] D Balcan and A Vespignani. Invasion threshold in structured populations with recurrent mobility patterns. *Journal of Theoretical Biology*, 293:87–100, January 2012.

[14] Duygu Balcan, Bruno Gonçalves, Hao Hu, José J Ramasco, Vittoria Colizza, and Alessandro Vespignani. Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *Journal of computational science*, 1(3):132–145, 2010.

[15] A-L Barabasi and R Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, October 1999.

[16] M Barthelemy, A Barrat, R Pastor-Satorras, and A Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of Theoretical Biology*, 235(2):275–288, 2005.

[17] Jordi Bascompte and Ricard V Solé. *Modeling spatiotemporal dynamics in ecology*. Springer New York, 1998.

[18] David C Bell, John S Atkinson, and Jerry W Carlson. Centrality measures for disease transmission networks. *Social networks*, 21(1):1–21, 1999.

[19] M Boguna, R Pastor-Satorras, and A Vespignani. Epidemic spreading in complex networks with degree correlations. In A PastorSatorras, R and Rubi, M and DiazGuilera, editor, *Statistical mechanics of complex networks*, volume 625 of *Lecture Notes in Physics*, pages 127–147. Spanish Govt, DGICYT; Generalitat Catalonia, CIRIT; Univ Barcelona; Ctr Especial Recerca, Phys Complex Syst, SPRINGER-VERLAG BERLIN, 2003.

[20] B M Bolker and B T Grenfell. SPACE, PERSISTENCE AND DYNAMICS OF MEASLES EPIDEMICS. *Philosophical transactions of the Royal Society of London Series B-Biological Sciences*, 348(1325):309–320, 1995.

[21] Béla Bollobás. *Random graphs*, volume 73. Cambridge university press, 2001.

[22] Markus Brede and Fabio Boschetti. Analysing Weighted Networks: An Approach via Maximum Flows. In J Zhou, editor, *Complex Sciences, Part 1*, volume 4 of *Lecture Notes of the Institute for Computer Sciences Social Informatics and Telecommunications Engineering*, pages 1093–1104. SPRINGER-VERLAG BERLIN, 2009.

[23] D Brockmann, L Hufnagel, and T Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, January 2006.

[24] G Chowell, J M Hyman, S Eubank, and C Castillo-Chavez. Scaling laws for the movement of people between locations in a large city. *Physical Review E*, 68(6, Part 2), 2003.

[25] M L Ciofi degli Atti, S Merler, C Rizzo, M Ajelli, M Massari, P Manfredi, C Furlanello, G Scalia Tomba, and M Iannelli. Mitigation measures for pandemic influenza in Italy: an individual based model considering different scenarios. *PLoS One*, 3(3):e1790, 2008.

[26] A Cliff and P Haggett. Time, travel and infection. *British Medical Bulletin*, 69:87–99, 2004.

[27] V Colizza, A Barrat, M Barthelemy, A-J Valleron, and A Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Medicine*, 4(1):e13, January 2007.

[28] V Colizza, A Barrat, M Barthelemy, and A Vespignani. The modeling of global epidemics: Stochastic dynamics and predictability. *Bulletin of Mathematical Biology*, 68(8):1893–1921, November 2006.

[29] V Colizza, A Barrat, M Barthelemy, and A Vespignani. 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(7):2015–2020, 2006.

[30] V Colizza, A Barrat, M Barthelemy, and A Vespignani. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC Medicine*, 5, November 2007.

[31] V Colizza, M Barthelemy, A Barrat, and A Vespignani. Epidemic modeling in complex realities. *Comptes rendus Biologie*, 330(4):364–374, 2007.

[32] V Colizza and A Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations. *Journal of Theoretical Biology*, 251(3):450–467, 2008.

[33] B S Cooper, R J Pitman, W J Edmunds, and N J Gay. Delaying the international spread of pandemic influenza. *PLoS Medicine*, 3(6):e212, 2006.

[34] N J Cox and K Subbarao. Global epidemiology of influenza: past and present. *Annual review of medicine*, 51(1):407–421, 2000.

[35] Philip D Curtin. *Death by migration: Europe's encounter with the tropical world in the nineteenth century.* Cambridge University Press, 1989.

[36] L Danon, A P Ford, T House, C P Jewell, M J Keeling, G O Roberts, J V Ross, and M C Vernon. Networks and the epidemiology of infectious disease. *Interdiscip Perspect Infect Dis*, 2011:284909, 2011.

[37] L Danon, T House, and M J Keeling. The role of routine versus random movements on the spread of disease in Great Britain. *Epidemics*, 1(4):250–258, 2009.

[38] A De Montis, M Barthelemy, A Chessa, and A Vespignani. The structure of interurban traffic: A weighted network analysis. *Environment and Planning B-Planning & Design*, 34(5):905–924, September 2007.

[39] A De Montis, S Caschili, and A Chessa. Time evolution of complex networks: commuting systems in insular Italy. *Journal of geographical Systems*, 13(1, SI):49–65, March 2011.

[40] Frank R DeLeo, Michael Otto, Barry N Kreiswirth, and Henry F Chambers. Community-associated meticillin-resistant Staphylococcus aureus. *The Lancet*, 375(9725):1557–1568, 2010.

[41] Pauline den Driessche and James Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical biosciences*, 180(1):29–48, 2002.

[42] Jared M Diamond and Doug Ordunio. *Guns, germs, and steel*. Norton New York, 1997.

[43] Odo Diekmann and Johan Andre Peter Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. Wiley, 2000.

[44] Loots I Dublin and Alfred J Lotka. On the True Rate of Natural Increase: As Exemplified by the Population of the United States, 1920. *Journal of the American Statistical Association*, 20(151):305–339, 1925.

[45] W John Edmunds, C J O'callaghan, and D J Nokes. Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 264(1384):949–957, 1997.

[46] L R Elveback, J P Fox, E Ackerman, A Langworthy, M Boyd, and L Gatewood. INFLUENZA SIMULATION-MODEL FOR IMMUNIZATION STUDIES. *American journal of epidemiology*, 103(2):152–165, 1976.

[47] Sven Erlander and Neil F Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[48] S Eubank, H Guclu, V S A Kumar, M V Marathe, A Srinivasan, Z Toroczkai, and N Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[49] N M Ferguson, D A T Cummings, S Cauchemez, C Fraser, S Riley, A Meeyai, S Iamsirithaworn, and D S Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209–214, September 2005.

[50] N M Ferguson, D A T Cummings, C Fraser, J C Cajka, P C Cooley, and D S Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448–452, 2006.

[51] N M Ferguson, C A Donnelly, and R M Anderson. The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. *Science*, 292(5519):1155–1160, 2001.

[52] N M Ferguson, M J Keeling, W J Edmunds, R Gant, B T Grenfell, R M Anderson, and S Leach. Planning for smallpox outbreaks. *Nature*, 425(6959):681–685, October 2003.

[53] A Flahault, El Vergu, L Coudeville, and R F Grais. Strategies for containing a global influenza pandemic. *Vaccine*, 24(44-46):6751–6755, November 2006.

[54] Antoine Flahault and Alain-Jacques Valleron. A method for assessing the global spread of HIV-1 infection based on air travel. *Mathematical Population Studies*, 3(3):161–171, 1992.

[55] Lester Randolph Ford and Delbert R Fulkerson. A simple algorithm for finding maximal network flows and an application to the Hitchcock problem. 1955.

[56] Christophe Fraser, Christl A Donnelly, Simon Cauchemez, William P Hanage, Maria D Van Kerkhove, T Déirdre Hollingsworth, Jamie Griffin, Rebecca F Baggaley, Helen E Jenkins, Emily J Lyons, and Others. Pandemic potential of a strain of influenza A (H1N1): early findings. *science*, 324(5934):1557–1561, 2009.

[57] T C Germann, K Kadau, I M Longini, and C A Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15):5935–5940, 2006.

[58] Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Fabio Pinelli, Chiara Renso, Salvatore Rinzivillo, and Roberto Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB JournalThe International Journal on Very Large Data Bases*, 20(5):695–719, 2011.

[59] M C Gonzalez, C A Hidalgo, and A-L Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[60] R F Grais, J H Ellis, and G E Glass. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *European Journal of Epidemiology*, 18(11):1065–1072, November 2003.

[61] R F Grais, J H Ellis, and G E Glass. Forecasting the geographical spread of smallpox cases by air travel. *Epidemiology and Infection*, 131(2):849–857, October 2003.

[62] R F Grais, E J H, A Kress, and G E Glass. Modeling the spread of annual influenza epidemics in the US: the potential role of air travel. *Health care management science*, 7(2):127–34, 2004.

[63] D M Green, I Z Kiss, and R R Kao. Modelling the initial spread of foot-and-mouth disease through animal movements. *Proceedings of the Royal Society B - Biological Sciences*, 273(1602):2729–2735, November 2006.

[64] B T Grenfell, O N Bjornstad, and J Kappey. Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723, 2001.

[65] B T Grenfell and J Harwood. (Meta)population dynamics of infectious diseases. *Trends in Ecology & Evolution*, 12(10):395–399, October 1997.

[66] R Guimera, S Mossa, A Turtschi, and L A N Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7794–7799, 2005.

[67] S Gupta, R M Anderson, and R M May. NETWORKS OF SEXUAL CONTACTS - IMPLICATIONS FOR THE PATTERN OF SPREAD OF HIV. *AIDS*, 3(12):807–817, 1989.

[68] I M Hall, J R Egan, I Barrass, R Gani, and S Leach. Comparison of smallpox outbreak control strategies using a spatial metapopulation model. *Epidemiology and Infection*, 135(7):1133–1144, October 2007.

[69] Ilkka Hanski and Michael E Gilpin. *Metapopulation biology: ecology, genetics, and evolution*. Academic press San Diego, 1997.

[70] Frank Harary. *Graph theory and theoretical physics*. Academic P., 1967.

[71] J A P Heesterbeek and K Dietz. The concept of Ro in epidemic theory. *Statistica Neerlandica*, 50(1):89–110, 1996.

[72] H W Hethcote. Immunization model for a heterogeneous population. *Theoretical Population Biology*, 14(3):338–349, 1978.

[73] T D Hollingsworth, N M Ferguson, and R M Anderson. Will travel restrictions control the international spread of pandemic influenza? *Nature Medicine*, 12(5):497–499, 2006.

[74] T D Hollingsworth, N M Ferguson, and R M Anderson. Frequent travelers and rate of spread of epidemics. *EMERGING INFECTIOUS DISEASES*, 13(9):1288–1294, September 2007.

[75] L Hufnagel, D Brockmann, and T Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42):15124–15129, October 2004.

[76] Mark D Humphries and Kevin Gurney. Network small-world-ness: a quantitative method for determining canonical network equivalence. *PLoS One*, 3(4):e0002051, 2008.

[77] Amy Hurford, Daniel Cownden, and Troy Day. Next-generation tools for evolutionary invasion analyses. *Journal of the Royal Society Interface*, 7(45):561–571, 2010.

[78] B Jiang, J Yin, and S Zhao. Characterizing the human mobility pattern in a large street network. *Physical Review E*, 80(2, Part 1), 2009.

[79] W-S Jung, F Wang, and H E Stanley. Gravity model in the Korean highway. *EPL*, 81(4), 2008.

[80] C Kang, X Ma, D Tong, and Y Liu. Intra-urban human mobility patterns: An urban morphology perspective. *Physica A-Statistical Mechanics and its applications*, 391(4):1702–1717, 2012.

[81] M J Keeling. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society B - Biological Sciences*, 266(1421):859–867, 1999.

[82] M J Keeling, L Danon, M C Vernon, and T House. Individual identity and movement networks for disease metapopulations. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19):8866–8870, 2010.

[83] M J Keeling and K T D Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, September 2005.

[84] M J Keeling and B T Grenfell. Disease extinction and community size: Modeling the persistence of measles. *Science*, 275(5296):65–67, January 1997.

[85] M J Keeling and P Rohani. Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters*, 5(1):20–29, January 2002.

[86] M J Keeling, M E J Woolhouse, R M May, G Davies, and B T Grenfell. Modelling vaccination strategies against foot-and-mouth disease. *Nature*, 421(6919):136–142, January 2003.

[87] M J Keeling, M E J Woolhouse, D J Shaw, L Matthews, M Chase-Topping, D T Haydon, S J Cornell, J Kappey, J Wilesmith, and B T Grenfell. Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science*, 294(5543):813–817, October 2001.

[88] Matt J Keeling and Pejman Rohani. *Modeling infectious diseases in humans and animals*. Princeton University Press, 2011.

[89] E Kenah, D L Chao, L Matrajt, M E Halloran, and I M Longini Jr. The Global Transmission and Control of Influenza. *PLoS One*, 6(5), 2011.

[90] M D Kermark and A G Mckendrick. Contributions to the mathematical theory of epidemics. Part I. In *Proc. R. Soc. A*, volume 115, pages 700–721, 1927.

[91] K Khan, J Arino, W Hu, P Raposo, J Sears, F Calderon, C Heidebrecht, M Macdonald, J Liauw, A Chan, and M Gardam. Spread of a Novel Influenza A (H1N1) Virus via Global Airline Transportation. *New England journal of medicine*, 361(2):212–214, 2009.

[92] J Koopman. Controlling smallpox. *Science*, 298(5597):1342–1344, November 2002.

[93] Emmanuel Lagarde, M Schim Van Der Loeff, Catherine Enel, B Holmgren, R Dray-Spira, Gilles Pison, Jean-Pierre Piau, Valérie Delaunay, Souleymane MBoup, Ibrahima Ndoye, and Others. Mobility and the spread of human immunodeficiency virus into rural areas of West Africa. *International Journal of Epidemiology*, 32(5):744–752, 2003.

[94] Samuel Leinhardt. *Social networks: A developing paradigm*. Academic Press New York, 1977.

[95] A L Lloyd and R M May. Spatial heterogeneity in epidemic models. *Journal of Theoretical Biology*, 179(1):1–11, March 1996.

[96] I M Longini, A Nizam, S F Xu, K Ungchusak, W Hanshaoworakul, D A T Cummings, and M E Halloran. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087, 2005.

[97] I M Longini Jr., M E Halloran, A Nizam, Y Yang, S Xu, D S Burke, D A T Cummings, and J M Epstein. Containing a large bioterrorist smallpox attack: a computer simulation approach. *International Journal of Infectious Diseases*, 11(2):98–108, March 2007.

[98] A Lunelli, A Pugliese, and C Rizzo. Epidemic patch models applied to pandemic influenza: Contact matrix, stochasticity, robustness of predictions. *Mathematical Biosciences*, 220(1):24–33, 2009.

[99] Jonathan M Mann. AIDS: A worldwide pandemic'. *Current topics in AIDS*, 2:1–10, 1989.

[100] William Hardy McNeill. *Plagues and peoples*. Anchor, 1976.

[101] S Merler and M Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B - Biological Sciences*, 277(1681):557–565, 2010.

[102] S Merler, M Ajelli, A Pugliese, and N M Ferguson. Determinants of the spatiotemporal dynamics of the 2009 H1N1 pandemic in Europe: implications for real-time modelling. *PLoS Computational Biology*, 7(9):e1002205, September 2011.

[103] Joel C Miller. Spread of infectious disease through clustered populations. *Journal of the Royal Society Interface*, 6(41):1121–1134, 2009.

[104] C E Mills, J M Robins, and M Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904–906, 2004.

[105] C Moore and M E J Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5, Part b):5678–5682, 2000.

[106] Martina Morris. Sexual networks and HIV. *AIDS (London, England)*, 11:S209, 1997.

[107] J D Murray. Mathematical Biology. 1993.

[108] M E J Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):404–409, January 2001.

[109] M E J Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[110] M E J Newman. Analysis of weighted networks. *Physical Review E*, 70(5, Part 2), November 2004.

[111] Mark Newman, Albert-László Barabási, and Duncan J Watts. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[112] Zambon M Nicholson KG, Wood JM. Influenza. In *Lancet*, pages 362:1733–45, 2003.

[113] for National Statistics Office. Travel trends 2004: a report on the international passenger survey., 2005.

[114] M Padgham. Human Movement Is Both Diffusive and Directed. *PLoS One*, 7(5), 2012.

[115] Peter Palese. Influenza: old and new threats. *Nature medicine*, 10:S82—-S87, 2004.

[116] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200–3203, 2001.

[117] R Patuelli, A Reggiani, S P Gorman, P Nijkamp, and F-J Bade. Network analysis of commuting flows: A comparative static approach to German data. *Networks & Spatial economics*, 7(4):315–331, 2007.

[118] J S M Peiris, Y Guan, and K Y Yuen. Severe acute respiratory syndrome. *Nature medicine*, 10:S88—-S97, 2004.

[119] L Perez and S Dragicevic. An agent-based approach for modeling dynamics of contagious disease spread. *International Journal of Health Geographics*, 8, 2009.

[120] S Phithakkitnukoon, T Horanont, G Di Lorenzo, R Shibasaki, and C Ratti. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In A Salah, AA and Gevers, T and Sebe,

N and Vinciarelli, editor, *Human Behavior Understanding*, volume 6219 of *Lecture Notes in Computer Science*, pages 14–25, HEIDELBERGER PLATZ 3, D-14197 BERLIN, GERMANY, 2010. SPRINGER-VERLAG BERLIN.

[121] F Rakowski, M Gruziel, L Bieniasz-Krzywiec, and J P Radomski. Influenza epidemic spread simulation for Poland - a large scale, individual based model study. *Physica A-Statistical Mechanics and its applications*, 389(16):3149–3165, 2010.

[122] S Riley and N M Ferguson. Smallpox transmission and control: Spatial dynamics in Great Britain. *Proceedings of the National Academy of Sciences of the United States of America*, 103(33):12637–12642, 2006.

[123] Steven Riley, Christophe Fraser, Christl A Donnelly, Azra C Ghani, Laith J Abu-Raddad, Anthony J Hedley, Gabriel M Leung, Lai-Ming Ho, Tai-Hing Lam, Thuan Q Thach, and Others. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science*, 300(5627):1961–1966, 2003.

[124] C Rizzo, A Lunelli, A Pugliese, A Bella, P Manfredi, G Scalia Tomba, M Iannelli, M L Degli Atti, and EPICO Working Grp. Scenarios of diffusion and control of an influenza pandemic in Italy. *Epidemiology and Infection*, 136(12):1650–1657, 2008.

[125] M G Roberts and J A P Heesterbeek. A new method for estimating the effort required to control an infectious disease. *Proceedings of the Royal Society B - Biological Sciences*, 270(1522):1359–1364, 2003.

[126] Leonard Rogers. *Fevers in the Tropics*. H. Frowde, Hodder & Stoughton, 1908.

[127] P Rohani, D J D Earn, and B T Grenfell. Opposite patterns of synchrony in sympatric disease metapopulations. *Science*, 286(5441):968–971, October 1999.

[128] Ronald Ross. *The prevention of malaria*. Dutton, 1910.

[129] L A Rvachev and I M Longini. A mathematical-model for the global spread of influenza. *Mathematical Biosciences*, 75(1):3–23, 1985.

[130] John Scott. *Social network analysis*. SAGE Publications Limited, 2012.

[131] Francis R Sharpe and Alfred J Lotka. L. A problem in age-distribution. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 21(124):435–438, 1911.

[132] C Song, T Koren, P Wang, and A-L Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, October 2010.

[133] C Song, Z Qu, N Blumm, and A-L Barabasi. Limits of Predictability in Human Mobility. *Science*, 327(5968):1018–1021, 2010.

[134] A J Tatem, D J Rogers, and S I Hay. Global transport networks and infectious disease spread. In DJ Hay, SI and Graham, A and Rogers, editor, *Advances in parasitology, Vol 62: Global mapping of infectious diseases: Methods, Examples and emerging applications*, volume 62 of *Advances in Parasitology*, pages 293–343. ELSEVIER ACADEMIC PRESS INC, 525 B STREET, SUITE 1900, SAN DIEGO, CA 92101-4495 USA, 2006.

[135] David Tilman and Peter M Kareiva. *Spatial ecology: the role of space in population dynamics and interspecific interactions*, volume 30. Princeton University Press, 1997.

[136] P Upham, C Thomas, D Gillingwater, and D Raper. Environmental capacity and airport operations: current issues and future prospects. *Journal of Air Transport Management*, 9(3):145–151, 2003.

[137] C Viboud, O N Bjornstad, D L Smith, L Simonsen, M A Miller, and B T Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science*, 312(5772):447–451, 2006.

[138] C Viboud, M A Miller, B T Grenfell, O N Bjornstad, and L Simonsen. Air travel and the spread of influenza: important caveats. *PLoS Medicine*, 3(11):e503; author reply e502, November 2006.

[139] Jacco Wallinga, W John Edmunds, and Mirjam Kretzschmar. Perspective: human contact patterns and the spread of airborne infectious diseases. *TRENDS in Microbiology*, 7(9):372–377, 1999.

[140] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[141] D J Watts, R Muhamad, D C Medina, and P S Dodds. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32):11157–11162, 2005.

[142] Douglas Brent West and Others. *Introduction to graph theory*, volume 2. Prentice hall Englewood Cliffs, 2001.

[143] M E Wilson. Travel and the emergence of infectious diseases. *Emerging Infectious Diseases*, (1):39–46, 1995.

[144] M E Wilson. The traveller and emerging infections: sentinel, courier, transmitter. *Journal of Applied Microbiology*, 94(S):1S–11S, 2003.

[145] M Zachcial and C Heideloff. ISL Shipping Statistics Yearbook 2003 (Inst. of Shipping Economics and Logistics, Bremen, Germany). 2003.