1.

Modélisation, Simulation multi-niveau pour l'optimisation de politiques de vaccination

TRAN Thi Cam Giang, Jean Daniel ZUCKER, Marc CHOISY, Yann CHEVALEYRE

03/06/2015

1 State of the art

1.1 Epidemiology (and monitoring)

1.1.1 Epidemiology

The public health problems are some of the emerging issues in the entire world. They directly influence human health, the health of one person, the health of a community. In particular, any news about infectious diseases for children has always been a subject of concern to parents as well as everyone. Hence, in the world, a discipline "epidemiology" has risen to study the factors, causes, and effects of infectious diseases.

This thesis is proposed in a context in which many public health serious events have occurred in the world : SRAS in 2003, avian influenza in 2004 or swine flu in 2009, etc. In more details, at the start of 2014, the World Health Organization (WHO) officially stated global measles epidemic outbreak. In the first three months of the year 2014, there were about 56,000 cases of measles infections in 75 countries, particularly in southeast Asia and in Vietnam. More specially, in the United States, measles cases significantly rose in 2014, 14 years after national leaders stated that the disease had been disappeared within the country. There were 288 cases of measles reported in the U.S between Jan. 1 and May 23, 2014. To explain this sudden outbreak of the measles, epidemiologist found out that the reasons are strongly relative to international travel by unvaccinated people (particularly U.S. residents), and incomplete vaccination as in figure [**?**] : In the same year as the measles, Ebola epidemic is the largest in human history, affecting almost all countries in West Africa. To date, the WHO have reported a total of 28,575 suspected cases and 11,313 deaths. However, it is numbers in documents understated for the magnitude of the outbreak. Now, in Liberia, the government was officially stated Ebola-free in May 2015, however new cases were found in late June and early July.

The world suffered two large epidemics in the same year, this has stressed the importance of the epidemiological phenomena anticipation when diseases occur. Many studies proposed by the WHO, the Pasteur Institute and the

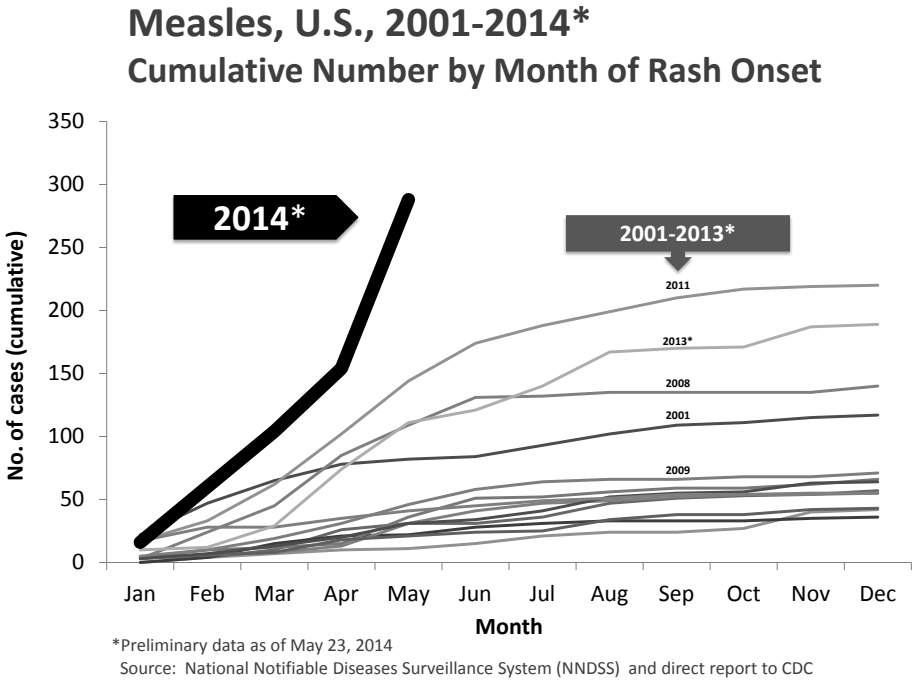

Figure 1.1: Measles, U.S., 2001-2014: Cumulative Number by Month of Rash Onset

Inserm in the field of environmental security have tried to understand disease phenomena and spread of disease over a territory, to better manage when diseases occur. These researches consist of mathematical or statistical studies via surveillance networks [14]. This is one of the axes of the UMMISCO laboratory's research themes (IRD UMI 209).

1.1.2 Control

Pathogenic microorganisms such as bacteria, viruses, parasites or fungi are key factors causing infectious diseases. The diseases can be spread directly or indirectly from one person to another, through a mediate environment or contaminated tools. As far as directly infectious diseases are concerned, meaning diseases directly transmitted from one person to another, we have some normal policies to prevent the spread of diseases such as vaccines, anti-viral medications, and quarantine. In this thesis, we focus on vaccinations in the human community. Firstly, a vaccine is understood as a biological preparation that provides active acquired immunity to a particular disease for our body. After having been vaccinated, we transport microorganisms in weakened or killed form of the microbe into our body. The body's immune system produces the right antibodies to recognize the germs as a threat, destroy them and keep a record of them. Because of that, when the disease occurs, our immune system can recognize and destroy with a better chance of success any of these germs that it later encounters. The administration of vaccines is called vaccination. Vaccination has greatly helped human beings. The vaccination of influenza, Human Papillomavirus (HPV) and chicken pox have been particularly appreciated. Smallpox is a particular example. It

was a big black point in human history during the closing years of the 18th century. Smallpox killed an estimated 400,000 Europeans annually and among the people that luckily survived, a third had been blinded by the disease. However, the WHO officially stated the eradication of smallpox in 2011 [47, 16]. In addition, many infectious diseases are clearly restricted such as influenza, polio, measles and tetanus from much of the world. Thus, one big question proposed is why many infectious diseases still exist in the world though we have produced vaccines for most infectious diseases. In order to answer this question, first of all, we have to answer to some following small questions:

Question	**Answer**	**Why?**
Are vaccines safe?	YES	Vaccines are generally quite safe
Are there vaccines for all infectious?	NO	For example: dengue
Are all vaccines free?	NO	Funding problem
Are all people vaccinated before a requested age for each disease?	NO	Funding/geographic/cultural problems

Table 1: Vaccine state

With the four answers above, we can say that the human still faces up to infectious diseases. A thorough knowledge of the disease is essential in order to implement large-scale proper infection control measures and prevention campaigns. Granted that the disease transmission methods depend on the characteristics of each disease and the nature of the microorganism that causes it. In this thesis, we will investigate popular infectious diseases with transmission by direct contact. This transmission requires a close contact between an infected person and a susceptible person, such as touching an infected individual, kissing, sexual contact with oral saliva, or contact with body lesions. Therefore, these diseases usually occur between members of the same household or close friends and family. In particular, this thesis will mostly focus on measles. Because measles is a highly contagious, serious disease caused by a virus. It is a typical infectious disease with direct transmission. In 1980, approximate 2.6 million people was killed each year before we had the widespread vaccination policies. It spreads very fast by coughing and sneezing in human communities via close interpersonal contact or direct contact with secretions. Its main symptoms consist of high fever, cough, runny nose and red eyes. These first symptoms usually take from 10 to 12 days after exposure to an infectious person, and lasts 4 to 7 days [39]. In fact, now there is no proper treatment for measles to totally prevent the spread of measles except routine measles vaccination policy for children. According to the report by the WHO, since 2002 measles was eradicated from U.S. However, today measles vaccination has not been extensively popularized in the entire world. Beside the obtained results, for example, in 2013, there was about 84% of the world's children having received one dose of measles vaccine, and during 2000-2013, measles vaccination prevented an estimated 15.6 million deaths; we have had to face up about 145700 measles deaths globally- estimated 400 deaths every day or 16 deaths every hour in 2013. Measles becomes one of the leading causes of death among young children in the world, although now we are having a big stock of safe and readily available measles vaccines.

Mass policy (or the routine measles vaccination policy for measles) that vaccinates the maximum number of children before a certain age, is the oldest (started from the 1950s in the rich countries) and is now the most used.

The policy has obtained clear results : a clear decrease of the incidence in most countries. However, the problem of this vaccination policy is too expensive, really ineffective and quite impossible to implement in poor countries, especially in Africa because of both financial and logistical problems. (e.g. the WHO project "Extended Program on Immunization" in Vietnam for the measles extinction before 2012 failed). In addition, when a vaccination policy is performed in a country, there is only one policy deployed, but in modeling, we can realize many policies and assess their results.

In short, measles is still a common and often fatal disease in the world. We still very much need to model the transmission dynamics of measles and investigate the effect of vaccination on the spread of measles in the entire world. More largely, we need to give new optimal vaccination policies in artificial intelligence in order that these policies may become more effective, less expensive, and take into account the spatial dimension for all popular infectious diseases.

1.2 Dynamiques/structures spatiales (théorie métapopulations, réseaux, etc. . .)

- For directly transmitted infectious diseases by virus and bacteria, susceptible individuals are not only infected by infected individuals in the same location, but also by other infected individuals due to the movement of individuals between populated regions. This is one very important part in the domain studying the geographical spread of infectious diseases. We care for host population characteristics, then characteristics of spatial spread of an infectious disease among populations. Through these characteristics, we find optimal policies to minimize the number of infected individuals in a community. In fact, there are many studies about the interactions among populations. However, we can divide the spatial structure of populations into two main levels: "inter-city level" and "intra-city level". At the inter-city level (or called "micro-level"), we use differential equations to control its models. At the "intra-city level" (also called "macro-level") in which we provide connections between the populations, simulate the intra-city traffic. We consider the effect of travel through the connections between population regions as a means of spreading a virus [43].

- We have two basic models considered in the "macro-level", the model has no explicit movement of individuals and the models describes enough travels and movements of individuals among populations and even takes into account the resident population as well as the current population of individuals [49]. A population may be simplified as a city, community, or some other geographical region. Population travel (e.g. among animals and among people by foot, birds, mosquitoes and in particular, people travel by air from one city to another), is the main reason why diseases can spread quickly among very distant cities such as SARS disease in 2003. Therefore, the term "metapopulation" arrived in the ecological literature in 1969 by Levins [33, 26]. A metapopulation is a population of a set of spatially discrete local populations (or subpopulations in short) with mutual interaction [33]. In the metapopulation in which a subpopulation can only go extinct locally and be recolonized by another after it is emptied by extinction [6, 24, 33] and migration between subpopulaitons

is significantly restricted. In a metapopulation, if recolonization rates are smaller than extinction rates, then total extinction of all local population will easily be reached. The persistence time of the metapopulation is measured as the time until all subpopulations go extinct. According to Harrison (1991) [26] there are four types of spatially dynamic populations : classic Levins metapopulation, mainland-island metapopualtion, patchy population and non-equilibrium populations.

- The first metapopulation model was proposed in 1969 by Levins. It is called the classic Levins Metapopulation [33]. Wilson in 1980 [50] stated that in this classic model "A nexus of patches, each patch winking into life as a population colonizes it, and winking out again as extinction occurs."

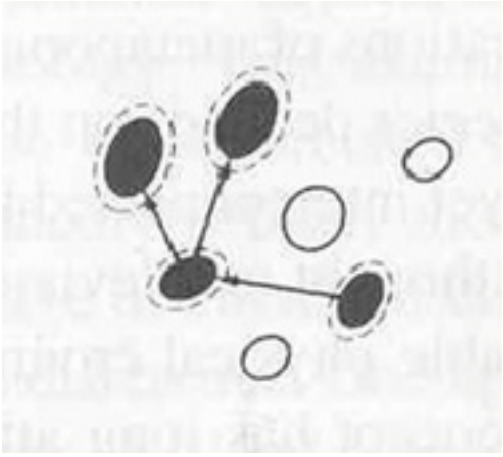

Figure 1.2: Classic Levins Metapopulation Model [27]

All subpopulations in this classic model are relatively small. The levels of interaction among individuals within a subpopulation is much higher than between subpopulations.

- The second model is the mainland-island metapopulation in which there are some small "island" subpopulations within dispersal distance of a much larger "mainland" subpopulation.

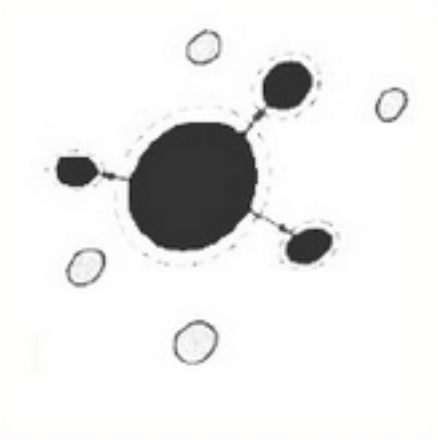

Figure 1.3: Mainland-Island Metapopulation [27]

It is evident that smaller sub-populations have a high probability of local extinction, but the mainland population will hardly become extinct. The migration from the mainland to the islands is independent of the islands white or filled, but is propagated for the connected islands. Therefore, if the mainland population has a low individual density and there is no immigration, then population growth rate is positive. Inversely, if island populations are in the same conditions as the mainland, then its population growth rate is negative. Thus, the islands would go down to extinction if there are no immigrants.

– The third model is patchy population. The local populations exist in a big habitat population and the dispersal rate between sub-populations is high.

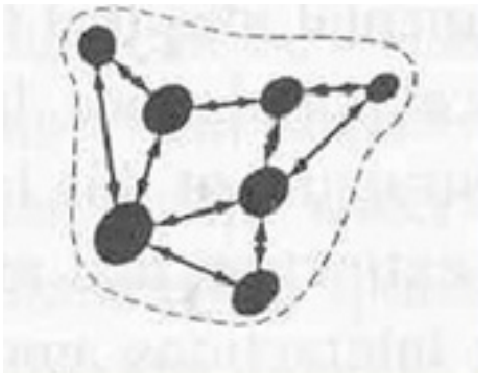

Figure 1.4: Patchy population [26]

Here we can find that the population structure is grouped and the interaction among them is frequent. However, this model is not referred as a concept for metapopulation and most researchers do not consider this a meta-population either.

– The final model is the non-equilibrium population. The local populations are patches, its local extinctions are much greater than its recolonisation.

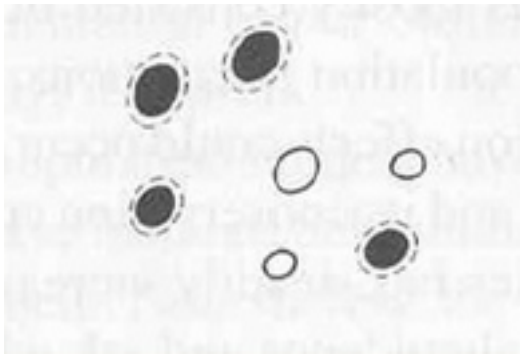

Figure 1.5: Non-equilibrium population [26]

It is obvious that white patches are rarely or never recolonized. Therefore, this model is not considered as a functional metapopulation. We can find this model in forested agricultural fields.

We already have four metapopulation models. In order to model the metapopulations mentioned above, we have three main model to implement : spatially-implicit model, spatially-explicit model and spatially-realistic model. For the first model, this is the type of model used in Levins (1969) [33] in which supposing that all local populations are connected with each other and they have independent local fluctuations. At any one time, we save track of the proportion of local populations and we do not take care the distance between them and the population size of each subpopulation. This model are mathematically and conceptually easy to implement. But this model can only answer some metapopulation problems because it ignores so many variables of a metapopulation. This model should be used for metapopulation close to a steady state.

For the second model, the spatially-explicit model is more complex than the first model. Subpopulations may be filled or vacant. Local populations only have interactions with the nearest neighbors. Subpopulations are organized as cells on a grid and migration among them depends on population density. We also only consider presence or absence of a species in each subpopulation. The advantage of this model is easy to model because of same local behaviors from subpopulation to subpopulation. However, we cannot simply describe the state of the metapopulation through filled subpopulations. Finally, the spatially-realistic model uses GIS to realize attributes, geometric coordinates, etc ... to a metapopulation. The first author using this model is Hanski in 1994 [25]. His model was defined as the incidence function (IF) model. This model is more realistic, and we can estimate quantitative predictions about metapopulation fluctuation. However, in fact, this model is very complicated, and many geographic data have to be estimated. Hence, the metapopulation concept start to no longer exist.

In the scope of this thesis, we focus on a metapopulation model that is result of combination between the spatially-explicit model and the patchy population. In general, this a simple spatial model, but is one of the most applicable model to descrire spread of diseases in human communities. This metapopulation consists of distinct subpopulations, each of which fluctuates independently, together with interaction limited by a coupling parameter ρ. These subpopulations may be filled or empty and contact with any neighbours.

1.3 Epidemiological models

It is known that, there are many current models that are used to model complex systems in nature, in ecology system and in epidemiology. Mathematical models in epidemiology are a typical example. These models permit us to present behavior of diseases and disease process in mathematics. However, explaining the transmission of infectious diseases is a difficult problem for an epidemiologist. Because there are many different interacting factors causing the outbreak of diseases such as the environment, the climate, the geography, the culture,...Hence, the role of the epidemiologist is how to model the characteristics and the transmission process of an infectious disease. Researchers have proposed compartmental models in epidemiology by dividing the population into “compartments” that illustrate health states of human through individuals. These compartmental models are called the epidemic models too. The first benefit of these models is to model the transmission process of a communicable disease

through compartments. Then, we can predict the properties of the disease dynamics such as the estimated number of infected individual, the time of persistence of disease, further that where and when we can implement vaccination policies to have both a minimum number of vaccined individuals and the minimum number of infected individuals in a given population. Let image that now in your country, there is an infectious disease as measles, a baby can be infected. According to the process of infection of disease, firstly this baby was born, he is fine and he is not infected yet by the measles but he may be infected in the future. We say that he belongs to the susceptible group (in short, S). Then, his mother takes him to a supermarket, there he see so many people, he is really infected through any way. He starts having a high fever, he may have to pass this state from 3 days to 5 days. In this period, he is really infected but he cannot infected others. We say that he belong to the exposed group (in short, E). After that, he start decreasing the temperature, but at the same time, he begins having red rashes on the back of the ears, after a few hours, on the head, on the neck and finally most of the body. This period appears from five to eight days after the exposed step. This duration is very sensible. The baby is completely infected and he can infected others if they see him. He belongs to the infected group (in short, I). Finally, he passes to the final period, he comes back good state. We say that he belongs to the recovered group with immunity (in short, R).

Around these four main health groups presenting the process of infection propagation in community, there are many epidemic models proposed. We give here the development of epidemic models by focusing on acute infections, assuming the pathogen causes illness for a periods of time followed by (typically lifelong) immunity. The first simplest model is the S-I-R model created by W. O. Kermack and A. G. McKendrick in 1927. The authors categorized hosts within groups as described above **S**usceptible (if not yet exposed to the pathogen), **I**nfected (if currently infected by the pathogen) and **R**ecovered (if they have successfully cleared the infection). From the simplest SIR model, in order to accord each infectious disease and real property of disease, scientists have modified it, made it different multiform. However, in shape of this thesis, we concentrate on the SEIR model (as the figure 1.6) that fit many currently infectious diseases in the world. Each patient must pass four health steps : susceptible stage, incubation stage, infectious stage and recovered stage.

In this model, the host population (N) is divided into four classes : susceptible S(t), exposed E(t), infected I(t) and recovered R(t). We have :

$$N(t) = S(t) + E(t) + I(t) + R(t)$$

- Class S(t) : contains the number of individuals not yet with the disease at time t, or those susceptible to the disease.
- Class E(t) : contains the number of individuals who are in the exposed or latent period of the disease.
- Class I(t) : contains the number of individuals who have been infected with the disease and are capable of spreading the disease to those in the susceptible category.
- Class R(t) : contains the number of individuals who have been infected and then removed from the disease,

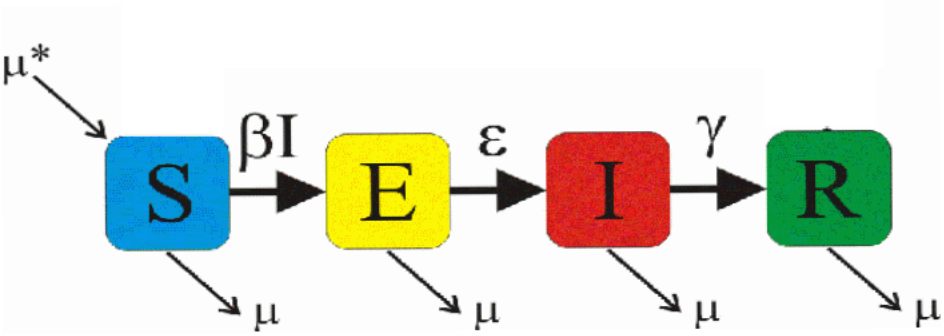

Figure 1.6: SEIR model

either due to immunization or due to death. Individuals of this class are not able to be infected again or to transmit the disease infection to others.

The conceptual descriptions of the model can be represented by a flow diagram above. The flow diagram for the SEIR model uses arrows to present the movement between the S and I classes, the E and I classes and the I and R classes. Here, individuals are born susceptible, die at a rate μ, become infected with the force of infection λ that is a function among the contact rate β, the number of infected individual I and the population size N, infectious after a latency period of an average duration of $1/\sigma$ and recover at the rate γ.

The SEIR model is investigated by ordinary differential equations (ODE) that are deterministic [30]. The value of variable states is only determined by parameters in the model and by sets of previous states of these variables. Moreover, the epidemic models are often proposed for one single population [30]. In the scope of this thesis, we propose a deterministic model for many subpopulations in a metapopulation. The standard SEIR model (susceptible-exposed-infective-recovered) has been strongly developed for the dynamics of directly infectious disease [5]. For disease-based metapopulation models, we give here a suit able new version of the SEIR equation that would be as follows:

Consider a metapopulation of n sub-populations. In a subpopulation i of size N_i, disease dynamics can be deterministically described by the following set of differential equations [3]:

$$\frac{dS_i}{dt} = \mu N_i - \lambda_i S_i - \mu S_i \tag{1.1}$$
$$\frac{dE_i}{dt} = \lambda_i S_i - \mu E_i - \sigma E_i \tag{1.2}$$
$$\frac{dI_i}{dt} = \sigma E_i - \mu I_i - \gamma I_i \tag{1.3}$$
$$\frac{dR_i}{dt} = \gamma I_i - \mu R_i \tag{1.4}$$

where S_i, E_i, I_i and R_i are the numbers of susceptible, exposed, infectious and recovered in this sub-population i respectively. Individuals are born susceptible, die at a rate μ, become infected with the force of infection λ_i, infectious after a latency period of an average duration of $1/\sigma$ and recover at the rate γ. In a case the infectious contact rate is constant, the equilibrium values of the variables S, E, I and R can be expressed analytically (see late part). The force of infection depends not only on the total population size N_i and the number of infected I_i in subpopulation i, but also in other sub-populations [30]:

$$\lambda_i = \sum_j \rho_{ij}\kappa_j \log\left[1 - \sum_{k=1}^{M}\left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right)\right] \tag{1.5}$$

where $c_{i,k}$ $(0 \leqslant c_{ij} \leqslant 1)$ is the probability that a susceptible individual native from i being in contact with another infected individual native from k gets infected. ξ_{jk} $(0 \leqslant \xi_{ij} \leqslant 1)$ refers to the probability that an individual y meeting x in C_j comes from C_k. κ_j is the average number of contacts per unit of time a susceptible will have when visiting subpopulation j. $\rho_{i,j}$ $(0 \leqslant \rho_{ij} \leqslant 1)$ is denoted as the probability that an individual from subpopulation i visits subpopulation j, of course, $\sum_{j=1}^{M} \rho_{ij} = 1$. See appendix for detail on the construction of this equation. We can verify that in the limit case on one single subpopulation in the metapopulation ($i = j$ and $n = 1$), we have:

$$\lambda_i = -\kappa_i \log(1 - \frac{I_i}{N_i} \times c_{ii}) \tag{1.6}$$

Consider that the average number of contacts per unit of time κ_i is seasonally forced [2] and seasonality is an annually periodic function of time [23]. As a result, for the subpopulation i :

$$\kappa_i(t) = \kappa_{i0}\left[1 + \kappa_{i1}\cos\left(\frac{2\pi t}{T} + \varphi_i\right)\right] \tag{1.7}$$

where t is the time, κ_{i0} and κ_{i1} are the mean value and amplitude of the average contact rate κ_i at which a susceptible will have when visiting subpopulation i per unit of time, T and φ_i are the period and the phase of the forcing. With the annual sinusoidal form of the average contact rate, we really have the sinusoidally forced SEIR metapopulation model.

In detail, the deterministic model performs the same way for a given set of initial conditions. It doesn't

have randomness, dynamics, and don't present dynamic of diseases in nature. Thus, stochastic models have been proposed. A stochastic model is always more realistic than a deterministic one. These models have stochastic and variable states are not described by unique values, but by probability distributions. It is why we will use the stochastic models to predict extinction probability of disease in spatial context[30].

1.4 Algorithms of stochastic simulation

Stochastic simulation works on variables that both are random and can be changed with certain probability. Today, these stochastic models have been used widely in many domain because of some reasons as following : before, in order to model chemically reacting systems, in simple way, we solved a set of coupled ordinary differential equations (ODEs) [36] of deterministic approaches. Basically, these approaches use the law of mass action that shows a simple relation between reaction rate and molecular component concentrations. We start with a given set of initial molecular concentrations, the law of mass action permits us to see the component concentrations over time. The states of a reaction are a homogeneous, free medium. The reaction rate will be directly scaled with the concentrations of the elements. Most systems can use the traditional deterministic approaches to simulate. It is evident that many systems such as some biochemical systems consist of random, discrete interactions between individual elements. However, in the case, these systems becomes smaller and smaller, the traditional deterministic models may not be accurate. It is the reason for that the fluctuations of these systems can be simulated exactly by applying stochastic models, particularly as well as the Stochastic Simulation Algorithms (SSA) [18, 19].

The SSA uses Monte Carlo (MC) methods to study the time evolution of the jump process. Because the basis feature of the Monte Carlo simulation is insensitive to the dimensionality of the problem, and the work grows linearly with the number of reaction channels in the model. The SSA describes time-evolution statistically correct trajectories of finite populations in continuous time by solving the corresponding stochastic differential equations. Using the stochastic models can solve three questions. (1) These models take account the discrete character of the number of elements and the evidently random character of collision among elements. (2) They coincide with the theories of the dynamic and stochastic processes. (3) They are a good idea to describe "small systems" and "instable systems". The main idea of the stochastic models is that element reactions are essentially random processes. We don't know certainly how a reaction occur at a moment. We also call it the process stochasticity. In particular, in the process stochasticity, we talk about demographic and environmental stochasticities in epidemic models. The demographic stochasticity is strongly controlled by population size such as the birth and death rates, contamination,etc. But, the environmental stochasticity is just affected by environmental factors what we can not govern. Therefore, there are many epidemic models that forcus on exploration of demographic stochasticity. Demographic stochasticity is considered as fluctuation in population processes that are based the random nature of events at the level of the individual. Each event is related to one baseline probability fixed, individuals are presented in differing fates due to chance. In addition to the demographic stochasticity, the number of infectious, susceptible, exposed and recovered

individuals is now required to be an integer. Modeling approaches that incorporate demographic stochasticity are called event-driven methods. These methods require explicit consideration of events. The first approach published by Daniel T.Gillespie in 1976 [18] is an exact stochastic simulation approach for chemical kinetics. The Gillespie stochastic simulation algorithm (SSA) has become the standard procedure of the discrete-event modelling by taking proper value of the available randomness in such a system. The methods modelling the event-driven model demands explicit presentation of events. For the standard SEIR model, we have to consider the nine events that can occur, each causing the numbers in the relative groups to go up or down by one. Table 2 lists all the events of the model, occurring in subpopulation i of a metapopulation:

Events	Rates	Transitions
birth	μN_i	$S_i \leftarrow S_i + 1$ and $N_i \leftarrow N_i + 1$
death of a susceptible	μS_i	$S_i \leftarrow S_i - 1$
death of an exposed	μE_i	$E_i \leftarrow E_i - 1$
death of an infected	μI_i	$I_i \leftarrow I_i - 1$
death of an immune	μR_i	$I_i \leftarrow I_i - 1$
infection	$\lambda_i S_i$	$S_i \leftarrow S_i - 1$ and $E_i \leftarrow E_i + 1$
becoming infectious	σE_i	$E_i \leftarrow E_i - 1$ and $I_i \leftarrow I_i + 1$
recovery	γI_i	$I_i \leftarrow I_i - 1$ and $R_i \leftarrow R_i + 1$

Table 2: Events of the stochastic version of the model of equations, occurring in subpopulation i.

To implement this SEIR stochastic model, there are many different methods, thought most researchers often use the method of Gillespie in 1977. Starting from the initial states, the stochastic simulation algorithms simulate the trajectory in population processes by repeatedly answering the following two questions and updating the states.

- When (time τ) will the next reaction fire?
- Which (reaction channel index μ) reaction will fire next?

So the key parameters in the SSA model are τ and μ. To calculate these distributions, we set $a_0(x) = \sum_{j=1}^{M} a_j(x)$. The time τ, given $X(t) = x$, that the reaction will fire at $t + \tau$, is the exponentially distributed random variable with mean $\frac{1}{a_0(x)}$,

$$P(\tau = s) = a_0(x) exp(-a_0(x)s),$$

and the index μ is the integer random variable of that firing reaction with probability :

$$P(\mu = j) = \frac{a_j(x)}{a_0(x)}$$

In each step, the SSA generates random numbers and calculates τ and μ according to the probability distribution 1.4 and 1.4. Below we will show some methods that simulate exactly stochastic models. In this part, we will review the variant formulations of Gillespie's stochastic simulation algorithm (SSA) about the overview and the computational cost of each algorithm, then approximate simulation methods, and finally hybrid and multi-scale methods. We will also point out which algorithm that is most efficient, which algorithm that we have used in this thesis.

1.4.1 Exact stochastic simulation

The key property for a discrete event simulation of a Markovian system basically samples a time for the next event from this distribution and selecting the reaction that occurs at that time. The Markovian simulation methods have the basic steps of the widely used kinetic Monte Carlo (KMC) method. First persons introduced the method (also known as the KBL algorithm) are Young and Elcock in 1966 [51]and independently Lebowitz, Bortz and Kalos in 1975 [7]. However, Gillespie really is the person who made kinetic Monte Carlo popular in the chemical and biochemical domains, calling the algorithm the Stochastic Simulation Algorithm (SSA) in his seminal articles [18, 19]. Below we will review his two papers.

1. First reaction method (FRM) [GILLESPIE 1976] [19]: The first reaction method was proposed by Gillespie in 1976. It models demographic stochasticity in the more intuitive and slower way to the deterministic model. The obtained result is "fluctuations in population processes that arise from the random nature of events at the level of the individual" [30]. The following pseudo-code provides a clean implementation of Gillespie's first reaction method :

   If supposing that generate one random number takes C_{rand} time. Then the FRM takes nC_{rand} time per step. It is a big problem for that Gillespie proposed a new improved algorithm in 1976. The FRM has three main disadvantages : (1) generating random numbers is relatively slow, (2) the FRM generates a cycle too many random numbers in the case where the simulation time is big and the random number generator will have tend on saturation when it generates too many numbers, (3) the FRM is difficult for indexing the events to effectively implement the update step.

2. Direct method (DM) [GILLESPIE 1977] [18]: The Direct Method was proposed by Gillespie in 1977. The main

Algorithm 1 Gillespie's first reaction method in 1976 - MONOPOPULATION

0. Label all species $X_1, ..., X_k$.
1. Label all possible events E_1,...,E_n.
2. For each event determine the rate at which it occurs, R_1,...,R_n.
3. **While** ($t < t_{end}$ and $R_N = \sum_{v=1}^{n} R_v \neq 0$) **then**
4. **For** $m = 1, n$ **do**
5. Generate one random number $U(0,1)$: $RAND_m$
6. At the event m calculate the time until the next event is $\delta t_m = \frac{-1}{R_m} log(RAND_m)$
7. **end for**
8. Find the event, p , that happens first (has the smallest δt)
9. The time is now updated, $t \rightarrow \delta t_p$
10. Update $\{X_i\}$ following the event p.
11. Return to Step 3.

objective is to present the stochastic simulation for chemically reacting systems. The following pseudo-code of this method is :

Algorithm 2 Direct method of Gillespie in 1977 - MONOPOPULATION [18]

0. Label all species $X_1, ..., X_k$.
1. Label all possible events E_1,...,E_n.
2. For each event determine the rate at which it occurs, R_1,...,R_n.
3. **While** ($t < t_{end}$ and $R_N = \sum_{v=1}^{n} R_v \neq 0$) **then**
4. **For** $m = 1, n$ **do**
5. Calculate R_m and $R_m = \sum_{v=1}^{m} R_v$
6. **end for**
7. Generate uniformly distributed random numbers (r_1, r_2)
8. Determine when ($\tau = ln(1/r_1)/R_N$ and which ($min\{p|R_p \geq r_2 R_N\}$) reaction will occur
9. Set $t = t + \tau$
10. Update $\{X_i\}$
11. Return to step 3

We can find that on each step, the Direct Method has to generate two random numbers. Supposing that generate one random number takes C_{rand} time. Hence, on each step, the DM takes $(2C_{rand} + O(n))$ where $O(n)$ is time to search the index p of the next reaction channel. For this reason, the DM is more efficient than the FRM.

The Gillespie algorithm plays an very important role and has become a fundamental method in computational systems biology. Hence, many efforts have been proposed to improve its efficiency. The key step in the DM is to choose the next reaction channel to implement. This step applies a linear search with a complexity $O(n)$ where n is the number of occurred event in the system. Many methods have focused on this step to improve and give more efficient formulations. For example, Maksym in 1988 [37] separated the set of reactions into subsets with a complexity of $O(n^{1/2})$. In 1995, Blue et al. [4] extended the division approach of Maksym, where a $K - level$ method results in a search time proportional to $n^{1/K}$. Then, in taking K to the limit, they applied a binary tree structure and obtained the complexity $O(logR)$. Don't stop improving, the Next Reaction Method (NRM) by Gibson and Bruck is more known to the systems biology domain [17].

3. Next Reaction Method (NRM) [GIBSON2000] [17] In 2000, Gibson and Bruck successful transformed the

algorithm FRM into an equivalent but more efficient new structure. The Next Reaction Method applies just a single random number per iteration. Moreover, the initiation times of the reactions can be set as the firing times of independent, unit rate Poisson processes with internal times given by integrated propensity functions. It is evaluated steady faster than the FRM and more efficient than DM in special cases as the system includes many species and loosely coupled reaction channels. The NRM is presented as follows :

Algorithm 3 Next Reaction Method (NRM) [GIBSON2000]

1. Initialize
   (a) set initial numbers of species, set $t = 0$, generate a dependency graph G
   (b) calculate the propensity function $R_j(x)$, for all j
   (c) for each j, generate a putative time τ_j according to an exponential distribution with parameter $R_j(x)$
   (d) store the τ_j values in an indexed priority queue P

2. Let μ be the reaction whose putative time τ_μ stored in P is least. Set $\tau = \tau_\mu$ 3. Update the states of the species to reflect execution of reaction μ. Set $t = \tau$ 4. For each edge (μ, α) in the dependency graph G
   (a) update R_α
   (b) if $\alpha \neq \mu$, set $\tau_\alpha = R_{\alpha,old}/R_{\alpha,new}(\tau_\alpha - t) + t$
   (c) if $\alpha = \mu$, generate a random number r and compute τ_α according to an equation similar to Step 8 of the DM $\tau_\alpha = \frac{1}{R_\alpha(x)} log(\frac{1}{r}) + t$
   (d) replace the old $R_\alpha$value in P with the new value.

Go to step 2.

The data structure presented in the NRM is a dependency graph. Because a propensity function R_j should be modified when a given reaction is implemented. A node in the graph is correspondent to a reaction channel. A directed edge of the reactions A_i and A_j points out that the execution of A_i really affects the molecules in A_j. Due to the the dependency graph, in step 4, the number of propensity functions recalculated is minimal.

In addition, the indexed priority queue is similar to a heap tree in computer science. It is a tree that includes ordered pairs of the form (i, τ_i), where i is both the reaction channel index and the position in the tree, τ_i is the corresponding time when the new A_i reaction is expected to occur. In the tree, the value τ of each parent is a smaller than that of its children. The top node of the tree is always the minimum value of τ and the order is only vertical. In each step, the nodes of the tree will change its positions according to its value to get the new priority queue.

In short, the NRM solved two additional optimizations : (1) by switching to absolute time, Gibson and Bruck reduced the number of random numbers needed in each step from two to one; (2) because of the use of a dependency graph, the number of propensities needing to be recomputed for every time step is minimum. In estimating the computation time of the NRM, we find that, for every reaction channel, the time until that reaction occurs is computed, maintained in an indexed priority queue, and efficiently implemented as a binary heap. However, the cost for maintaining the priority queue is relatively high. The time to select the next event is done in constant time, but the time to update the propensities is done in logarithmic time. Hence, the

NRM is commonly used for systems with many reaction channels and where relatively few propensities change with each reaction. The disadvantage of the NRM is that diffusion is added to the models and the reaction-diffusion master equation is simulated, so systems arise. For such models, in 2004, Elf et al. [15] proposed a variant of the NRM that is called the Next Subvolume Method (NSM). This method can be referred as a clever association of the ideas of NRM and Maksym's method for intracellular 3D chemical reaction systems.

4. Compare the direct method (DM) to the next reaction method (NRM), which algorithm is most efficient? We find that in the NRM, after the executed first initial step, all sequent time steps only ask one random number to be generated while the DM requests two. Even, the search step for the index μ of the next reaction channel takes $O(M)$ time for DM, but the corresponding task of updating the indexed priority queue takes $log(M)$. It is the reason for that the NRM is always evaluated more efficient for large scale problems. To evaluate the real cost of two methods (NRM and DM) based on the total simulation time, Yang Cao et al. [12] made experiments for both formulations of SSA on a 1.4 GHz Pentium IV Linux workstation. The problem used in their experiments is a stochastic model of the heat shock response of E. Coli [31], which includes 28 variables and 61 reactions. The experimental resultats pointed out that the average simulation time for DM is larger than that for NRM due to its data structure maintenance. In particular, for the loose coupling system where the components (or elements) in a system are interconnected and dependant on each other to the least extent practicable, NRM works better than DM. Yang et al. found that three main factors that strongly affect the CPU cost, are the costs, C_p to calculate M propensities, $C_{a_0}$to calculate the sum of all propensity probabilities, and C_s to search for a event μ. Hence, to reduce these costs, Yang et al. [12] suggested that a new optimization called the Optimized Directed Method (ODM). They found that in a reaction set of a system, some reactions fire much more frequently than others. To reduce the search time C_s, they arranged the index of the reaction ordering, placing the most frequently occurring events first based on how often they fire, combined with a dependency graph, achieves better results than the NRM for moderately large systems. Their optimized direct method (ODM) gives a new search depth smaller than the original method DM. The obtained result is that C_s can be significantly declined. In the next step to reduce the costs C_{a_0} and C_p, the authors used an idea from the method NRM. The ODM only recomputes the propensities for those reaction channels affected by the last reaction. Because of an extra cost used for accessing the dependency graph, so this approach applies only to loose-coupling systems. In conclusion, the obtained results of Yang et al. have shown that the ODM is faster than NRM, in particular, unless the system is very nearly uncoupled. This result broken the held belief for a long time that NRM was the fastest.

Through this result, we can say that the efficiency of the ODM, currently is evaluated to be the fastest known algorithm for stochastic simulation for most biological problems. This method can be negatively impacted by transient shifts in the frequency at which reactions occur, and commonly used in biochemical reaction networks because of the inductive and repressive nature of genetic regulation. Due to these shifts, the ODM defeats

the pre-simulation strategy employed within the ODM to decrease the time complexity of the SSA's reaction selection step and thus degrade performance. In order to decrease this degradation, in 2005, McCollum et al. [38] introduced the sorting direct method (SDM) that improves on ODM. This method eliminates the pre-simulations required by the ODM and permits the simulator to adapt to sharp changes in reaction execution frequencies. The common point of these two methods is to focus on the optimization of the system instead of the method itself by reducing the average number of operations required to obtain the index of the next reaction to fire. This average number of operations is called the search depth that is highly dependent on the biochemical system. For these two methods, the search depth us $O(M)$. Besides, within the paper of McCollum et al. [38], the authors also gave a detailed overview of the difference in the implementation of DM, NRM, ODM and SDM. In 2006, Li and Petzold introduced an alternative formulation of the SSA, named the Logarithmic Direct Method (LDM). In this method, the computational cost is independent of the ordering of the reactions and no need for a pre-simulation. The LDM declined the search depth to $O(logM)$, and pointed out the efficiency of the logarithmic method.

Two years later, in 2008, a different approach is proposed by Hellander [29] where the authors used the uniformity and quasi-Monte Carlo to reduce the number of trajectories needed to compute an approximation of the probability density function (PDF) at the price of a higher cost per trajectory. Another is also found to reduce the number of simulation in [32].

A new search direction is very beneficial to generate ensembles of trajectories in parallel. Because the parallelization of a single trajectory is very hard. Apply clusters to implement SSA is proposed by Li [34]. Then, Li et al. continued executing SSA on the graphics processing unit [36, 35].

1.4.2 Approximate methods

As mentioned above, the methods Direct, First Reaction and Next Reaction are all exact stochastic approaches of the underlying ordinary differential equations. Their advantage give us a really mathematically exact approach to simulate time-to-event model (in condition that the definition of the propensity functions accurately reflects the dynamics of the system). But, their disadvantages are 1) noise in exact simulations only affects the probabilities associated with fates of individuals and the updating of each consecutive event is independent – there is no assumption concerning environmental stochasticity; 2) these exact solutions become too slow and impractical when any one transition rate is large, when there is a big number of subpopulations or one a big number of event in a metapopulation; because the exact algorithms SSA must proceed one reaction at a time and take the task of explicitly simulating each and every reaction event, hence they are much too slow for most practical problems. It is the reason for that, approximate models have been proposed instead of the exact stochastic methods. The approximate approaches ask the question : "How many times does each action channel fire in each subinterval?"

These approaches could be used in larger systems, and made much faster than the exact methods. However,

Algorithm 4 $\tau - leap$ method proposed by Gillespie (2001)[30]

1. Let δt be the time increment between steps, δt is fixed as a constant. 2. Let $M_T(t)$ and $M_R(t)$ be the number of transmission and recovery events by time t. 3. Setting $\delta M_i = M_i(t+\delta t) - M_i(t)\ i = T, R)$, then

$$\begin{aligned} P(\delta M_T = 1|X,Y) &= \frac{\beta XY}{N}\delta t + o(\delta t) \\ P(\delta M_R = 1|Y) &= \gamma Y \delta t + o(\delta t) \end{aligned}$$

These two equations represent the transition probabilities for transmission and recovery events occurring in the time interval δt. 4. For small δt, the increments δM_i are approximately Poisson, such that:

$$\begin{aligned} \delta M_T &\approx Poisson(\frac{\beta XY}{N}\delta t) \\ \delta M_R &\approx Poisson(\gamma Y \delta t) \end{aligned}$$

5. Updating the values of the variables :

$$\begin{aligned} X(t+\delta t) &= X(t) - \delta M_T + \delta M_R \\ Y(t+\delta t) &= Y(t) + \delta M_T - \delta M_R \end{aligned}$$

6. Updating the time, $t = t + \delta t$. 7. Return to Step 4

the exact structure of the Markov chain is no longer simulated in the approximate approaches, hence the validity of the approximations become a main issue. Now we will show some mostly used approximate methods.

1. $\tau - leaping$ method Gillespie (2001) [21] has proposed a new method that decreases the simulation accuracy, but speeds up the stochastic simulation. This is the explicit Poisson $\tau - leap$ method known as an approximate method reduces the number of iterations by treating transition rates as constant over time periods for which this approximation leads to little error [30]. The $\tau - leap$ method applies a Poisson approximation to can "leap over" many fast reactions and approximate the stochastic behavior of the system very well. The $\tau - leap$ method is described as follows :

   The main problem in the $\tau - leap$ method is relative to the value of the time increment between steps, δt. How do we choose the value fixed of δt? δt must satisfy two conditions, large enough so that many reaction events occur in that time and small enough of the leap condition. The leap condition is pointed out that [11]: For the current state x, the value of δt is asked to be small enought that the modification in the state during $[t, t+\delta t]$ will be so small that no propensity function will suffer an appreciable change in its value. Thus, the key to the success of this technique is to choose a leap size large enough to allow many reactions to occur during the leap (reducing computation) and small enough that none of the propensity functions will change significantly in value (causing an error). Cao et al. [9] pointed also out a method to estimating the largest value of δt, the expected change in each propensity function during a leap be limited by $\epsilon a_0(x)$, where ϵ $(0 < \epsilon \ll 1)$ is the error control parameter. In short, the best advantage of the $\tau - leap$ method is to speed up the stochastic simulation for many "not-too-stiff" systems −i.e., systems in which the difference between the characteristic time scales of the fastest and slowest dynamical modes is not too large. However, the number

of firings of each reaction channels during a fixed time step δt is approximated as a Poisson random variable. This Poisson variable can have arbitrarily large sample values. Hence, there exists the possibility that this $\tau - leap$ method will cause one or more reaction channels to fire so many times during δt that number of reactants in each population will be became negative, in particular in systems with multiple timescales (for short the stiff systems). Due its obtained advantage, so this technique has continued to mature, specially in the area of leap-size selection, through the work of a variety of researchers : procedure for determining the maximum leap size for a specified degree of accuracy [22]; a binomial leaping method developed independently Tian et al.[46] and Chatterjee et al. [13]; based on the multinominal distribution by Pettigrew et al. in [40]; the post-leap checks of Anderson [3]; and a further work in this area can be expected. In detail for the binomial leaping method, this method replaces the Poisson random variables with binomial random variables, whose values are naturally bounded. However, the disadvantage of this method appeared when the system is in the state : there are multiple reactions with common consumed reactants, so the issues of the binomial tau-leaping strategy have not yet been fully resolved, and to write a general binomial tau-leaping program that reliably handles all situations that could possibly arise, this task would seem to be a very challenging task. It is the reason for that, Yang et al. (2005) [9] is introduced a modified Poisson tau-leaping procedure that also avoids negative populations and these particular issues, but is easier to implement than the binomial procedure.

1.4.3 Hybrid and multiscale methods

- Key word : Langevin equation, a Langevin equation is a stochastic differential equation that could present the time advance of a subset pf the degrees of freedom. In the epidemiology, the Langevin equations are the equation that describe of the dynamics of the individuals between the different compartments depends on the specific disease considered.

Here, we show an other kind of approximation methods that are based on the validity of different approximations, and are the combination of the deterministic equations and the Langevin equations for subsets of the reactions, the chemical species or both. The results of these methods have pointed out that the speedup obtained applying this hybrid idea can be substantially. For example, Adalsteinsson et al. Combine the deterministic and stochastic approaches in order to develop the software package Biochemical Network Stochastic Simulator (BioNetS) for efficiently and accurately simulating stochastic models of biochemical networks in [1]; the method that applies chemical Langevin equations in [20] and in [44, 45]; Poisson-Runge-Kutta methods [8]; multiscale algorithms like the slow-scale stochastic simulation algorithm [10] and use of the quasi-steady state assumption [41]. Besides, there are many varieties of others, for exemple : Haseltine and Rawlings [28] associate deterministic or Langevin equations for fast reactions with SSA for slow channels. Saliz and Kazessis [42] introduced a hybrid stochastic method that divides the reaction set of the system into fast and slow reaction subsets, applies a chemical Langevin equation for

approximating the fast reactions as a continuous Markov process, and has been used successful to the simulation of the dynamics of a system based on stochastic differential, ordinary differential, and master equations. These methods are one part in the stock of current methods. To find more information, readers can find detailed reviews in Turner et al. [48] and Burrage et al. [8]. In conclusion, so many researches have considered the approximate methods as a strong solution to speed up simulations, however, to exactly answer to question : where general biochemical modelers can know when and when not to apply these methods, these techniques have not matured yet. Inversely, exact stochastic simulation techniques such as “exact method”, “direct method”,.. are inferior computational methods, simulation time is large, but they remain effective, trusted, and exact for simulating stochastic models in bioinformatics. It is why, in my thesis, we use the exact method "Direct Method" of Gillespie to modeling stochastic effects.

1.5 reinforcement learning

2 DIZZYS : Description du modèle (ce qui corretspond au package R)

- comparaison avec ce qui existe déjà en termes de (1) possibilité (ce que l'on peut faire) et de (2) rapidité. En gros il y a un compromis entre flexibilité et rapidité. Il faut que tu montres où ce situe ton package. Par exemple, sous R, à comparer avec “adaptivetau” et “GillespieSSA”. Voir aussi les autres outils qu'il existe (par exemple ceux développés par Petzold http://www.cs.ucsb.edu/~cse/index2.php?publications.php)

- **Kullback-Leibler Divergence or Kolmogorov–Smirnov test to compare the simulation results.**

The SEIR model have successful described hosts within a population as Susceptible (the number of individuals not yet infected with the disease), Exposed (the number of individuals who are infected with the disease but not infectious),Infected (the number of individuals who have been infected with the disease and are capable of spreading the disease), and Recovered (the number of individuals who have successfully cleared the infection). We ignore population demography-births, deaths, and migration. So we have only three transitions:

$$S \to E, E \to I, andI \to R. \tag{2.1}$$

It is obvious that the second and third of these are easier, so we focus on the first transition in which the level of the infectious disease influences strongly the disease transmission rate from a susceptible individual into the infected class. If this first step doesn't occur, all infected individuals can go to the recovered class. This first transition plays an important role in studying fluctuation of disease. Many researches have been tried to find the "infectious period" (the amount of time spent in the infectious class). They have found that for acute infections, the "infectious period" is distributed around some mean values that we can estimate from clinical data. To present enough properties of the "infectious period" by formulations, we have to depend upon three main factors of the disease transmission from S

to E : the prevalence of infecteds, the underlying population contact structure, and the probability of transmission given contact. For a directly transmitted disease, the key factor is the contact between susceptible and infected individuals. Hence, in this section, we will interpret the "infectious period" in more detailed, and realistic contexts.

First of all, we define the force of infection, λ, which is referred as the per capita rate at which susceptible individuals contract the infection. The fact that λ directly scales the number of infectious individuals I. Moreover, for directly infectious diseases, disease transmission demands contact between infecteds and susceptibles. Hence, there are two proposed general possibilities based on how we present the contact structure to change with population size: $\lambda = \beta I/N$ applied when we want to mention frequency dependent (or mass action) transmission, and $\lambda = \beta I$ applied when we want to mention density dependent (or pseudo mass action) transmission, where I is the number of infectious individuals, N is the total population size (N=S+E+I+R), and β is the product of the contact rates and transmission probability. In the shape of this thesis, we assume that the infection process is frequency-dependent, meaning that the force of infection λ is proportional to a proportion of infected: I/N. Infection of susceptibles from one population i can be due to contacts with infected from the same population i or to contacts with infected from another population j. In the next section, we will introduce a new mechanistic derivation of the transmission term in more detailed, and realistic way.

2.1 Infection force

Goal :

Probabilistic derivation of multi-population epidemic model

with $\beta_{ijk} = -\kappa_j \log(1 - c_{ik})$

Definition .1. *During the small time interval δt, each native individual of the city i visits **a single** city j (with probability ρ_{ij}) and will meet **in average** κ_j individuals that come from all cities.*

2.1.1 Notation

Notation :

Here, we present list of sets and events describing the state of the system at time t :

- C_i is the set of all individuals born in sub-population i.
- $V_{i,t}$ is the set of all individuals physically located in sub-population i from time t to time $t+\delta t$. This includes foreigners traveling in subpopulation i at time t, and all natives from subpopulation i which are not traveling abroad at time t.
- S_t, E_t, I_t, R_t are the sets of all individuals respectively susceptible, exposed, infected and recovered at time t. Note that these set include individuals from all subpopulations.

- $S_{i,t}, E_{i,t}, I_{i,t}, R_{i,t}$ are the same sets, restricted to natives of subpopulation i. So formally, $S_{i,t} = S_t \cap C_i$, $E_{i,t} = E_t \cap C_i$, $I_{i,t} = I_t \cap C_i$, and $R_{i,t} = R_t \cap C_i$.
- $Transmit(y, x)$ is an event indicating that individual x gets infected by individual y which was already infected
- $c_{i,k}$ is the probability that a susceptible individual native from i being in contact with another infected individual native from k gets infected.
- κ_j is the average number of contacts per unit of time a susceptible will have when visiting city j.
- ξ_{jk} refers to the probability that an individual y meeting x in C_j comes from C_k.
- $\rho_{i,j}$, the probability that an individual from subpopulation i visits subpopulation j. Of course, $\sum_{j=1}^{M} \rho_{ij} = 1$.

Note that : The coefficient κ should also depend on i, because an individual native from city i meets more people in his own city than abroad ($\kappa_{i,i} > \kappa_{i,j}$).

2.1.2 The background

Let us write a probabilistic formulation of $\frac{dE_i}{dt}$: One general question is always posed "how does the population of exposed individuals of subpopulation i evolve ?". For the sake of simplicity, in the process of transmission of the SEIR model, we focus on the incidence and we assume for now that the latent period and the recovery rate, respectively $\mu = \sigma = 0$. Thus, we write a probabilistic formulation of $\frac{dE_i}{dt}$. Assuming the time is discrete, we have $\frac{dE_i}{dt} \approx \mathbb{E}\left[E_{i,t+1} \setminus E_{i,t}\right]$. Then,

$$\begin{aligned}
\mathbb{E}\left[E_{i,t+1} \setminus E_{i,t}\right] &= \mathbb{E}\left[E_{i,t+1} \cap S_{i,t}\right] \\
&= \sum_{x \in C_i} Pr\left[x \in E_{t+1} \wedge x \in S_t\right] \\
&= \sum_{x \in C_i} Pr\left[x \in S_t\right] * Pr\left[x \in E_{t+1} \mid x \in S_t\right] \\
&= Pr_{x \sim \mathcal{X}_i}\left[x \in E_{t+1} \mid x \in S_t\right] * \sum_{x \in C_i} Pr\left[x \in S_t\right] \\
&= |S_{i,t}| \times Pr_{x \sim \mathcal{X}_i}\left[x \in E_{t+1} \mid x \in S_t\right]
\end{aligned}$$

Assume there are M cities. An individual x of the subpopulation i may be visiting another subpopulation, or staying in its own subpopulation. Applying the law of total probabilities, we get:

$$
\begin{aligned}
Pr_{x\sim\mathcal{X}_i}\left[x \in E_{t+dt} \mid x \in S_t\right] &= \sum_{j=1}^{M} Pr_{x\sim\mathcal{X}_i}\left[x \in E_{t+dt} \wedge x \in V_{j,t} \mid x \in S_t\right] \\
&= \sum_{j=1}^{M} Pr_{x\sim\mathcal{X}_i}\left[x \in E_{t+dt} \mid x \in S_t \wedge x \in V_{j,t}\right].Pr_{x\sim\mathcal{X}_i}\left[x \in V_{j,t}\right] \\
&\quad \sum_{j=1}^{M} Pr_{x\sim\mathcal{X}_i}\left[x \in E_{t+dt} \mid x \in S_t \wedge x \in V_{j,t}\right] \times \rho_{ij}
\end{aligned}
$$

Where $\rho_{i,j} = Pr_{x\sim\mathcal{X}_i}\left[x \in V_{j,t}\right]$, the probability that an individual from subpopulation i visits subpopulation j. Of course, $\sum_{j=1}^{M} \rho_{ij} = 1$.

Study of case where agent x native from subpopulation i visits subpopulation j

Here, we look at the probability that a susceptible $x \sim \mathcal{X}_i$ visiting j gets infected or not after δt time steps. Let $\mathcal{Y}$ be the uniform distribution over $V_{j,t}$. The correct mathematical approach for this would be to assume that for each subpopulation k, the number of people native from k that we meet during δt follows a Poisson process. So both the number of people we meet and the number of infected people we meet during δt should be random variables.

In the approach described in [30], the authors did not do this. They assumed that both the number of people we meet and the number of infected people we meet *are fixed* (otherwise the maths they write would have been different). We will call this the old interpretation of the infection force proposed by "Keeling & Rohani" (for short, OIIF) that we will present it in the following parts.

We introduce an alternative approximation, where we assume that the number κ of people we meet during δt is *fixed*, but each of these people has *some probability* to be infected. This is an *in-between interpretation*, easier than the Poisson process maths, but better than the OIIF. We will call this the new interpretation of the infection force (for short, NIIF).

1. The new interpretation: NIIF

Proposition .1. *Agent x meets* exactly κ_j *other individuals, and each of these individuals has a probability $\frac{|I_{k,t}|}{N_k}$ of being infected, where k is its native subpopulation. Let $y_1 \dots y_{\kappa_j}$ be the individuals that x meets. We get:*

$$
\begin{aligned}
& Pr_{x\sim\mathcal{X}_i}\left[x \in S_{t+\delta t} \mid x \in S_t \wedge x \in V_{j,t}\right] \\
= \; & Pr_{x\sim\mathcal{X}_i, y_1 \dots, y_{\kappa_j}\sim\mathcal{Y}}\left[\bigwedge_{p=1}^{\kappa_j} \neg\left(y_p \in I_t \wedge Transmit(y_p, x)\right) \mid x \in S_t \wedge x \in V_{j,t}\right]
\end{aligned}
$$

So we have:

$$\begin{aligned} & Pr_{x\sim\mathcal{X}_i}\left[x \in S_{t+\delta t} \mid x \in S_t \wedge x \in V_{j,t}\right] \\ = \quad & Pr_{x\sim\mathcal{X}_i,y\sim\mathcal{Y}}\left[\neg\left(y \in I_t \wedge Transmit(y,x)\right) \mid x \in S_t \wedge x \in V_{j,t}\right]^{\kappa_j \delta t} \end{aligned}$$

Moreover, we have:

- the probability so that a susceptible individual x is infected by an infected individual y :

$$\begin{aligned} & Pr_{x\sim\mathcal{X}_i,y\sim\mathcal{Y}}\left[y \in I_t \wedge Transmit(y,x) \mid x \in S_t \wedge x \in V_{j,t}\right] \\ = \quad & \sum_{k=1}^{M} Pr_{x\sim\mathcal{X}_i,y\sim\mathcal{Y}}\left[y \in I_t \wedge Transmit(y,x) \mid x \in S_t \wedge x \in V_{j,t} \wedge y \in C_k\right].Pr_{y\sim\mathcal{Y}}\left(y \in C_k\right) \\ = \quad & \sum_{k=1}^{M} \{Pr_{x\sim\mathcal{X}_i,y\sim\mathcal{X}_k}\left[y \in I_t \mid x \in S_t \wedge x \in V_{j,t}\right] \\ & \quad \times Pr_{x\sim\mathcal{X}_i,y\sim\mathcal{X}_k}\left[Transmit(y,x) \mid y \in I_t \wedge x \in S_t \wedge x \in V_{j,t} \wedge y \in C_k\right] \times Pr_{y\sim\mathcal{Y}}\left(y \in C_k\right)\} \\ = \quad & \sum_{k=1}^{M} \left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right) \end{aligned}$$

$\xi_{jk} = \frac{N_k \rho_{kj}}{\sum_{v=1}^{M} N_v \rho_{vj}}$ refers to the probability that an individual y meeting x in C_j comes from C_k.

- hence, the probability so that a susceptible individual x is not infected by an infected individual y :

$$1 - \sum_{k=1}^{M} \left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right)$$

- thereby, the probability so that a susceptible individual x is not infected after κ_j contacts per unit time δt.

$$\left[1 - \sum_{k=1}^{M} \left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right)\right]^{\kappa_j \delta t}$$

- thus, the probability so that a susceptible individual x becomes infected after κ_j contacts per unit time δt.

$$Pr_{x\sim\mathcal{X}_i}\left[x \in E_{t+\delta t} \mid x \in S_t \wedge x \in V_{j,t}\right] \quad = \quad \left[1 - \sum_{k=1}^{M} \left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right)\right]^{\kappa_j \delta t}$$

We now apply the *log* approximation which consists in approximating $1-(1-u)^v$ by $v\log(1-u)$:

$$Pr_{x\sim\mathcal{X}_i}\left[x \in E_{t+\delta t} \mid x \in S_t \wedge x \in V_{j,t}\right] \quad = \quad -\kappa_j \delta t \log\left[1 - \sum_{k=1}^{M} \left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right)\right]$$

So, the transmission rate per susceptible individual is as follows:

$$\frac{dPr_{x\sim\mathcal{X}_i}\left[x \in E_{t+dt} \mid x \in S_t \wedge x \in V_{j,t}\right]}{dt} \simeq -\kappa_j \log\left[1 - \sum_{k=1}^{M}\left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right)\right]$$

In fact, we use the parameter λ to present this quantity, and it is denoted as the "force of infection" : If there is only one subpopulation i, then

$$\lambda_i = \kappa_j log(1 - \frac{|I_i|}{N_i} \times c_{ii})$$

2. The old Interpretation : OIIF [30]

Proposition .2. *Agent x meets* exactly $\kappa_j \delta t \xi_{jk} \frac{|I_{k,t}|}{N_k}$ *other infected individuals native from subpopulation k. Let $l_k = \kappa_j \delta t \xi_{jk} \frac{|I_{k,t}|}{N_k}$. Let $y_1^k \dots y_{l_k}^k$ be the infected individuals native from k that our individual x meets between t and $t + \delta t$.*

We have the probability so that a susceptible individual x is not infected after having seen l_k individuals between t and $t + \delta t$:

$$\begin{aligned}
& Pr_{x\sim\mathcal{X}_i}\left[x \in S_{t+\delta t} \mid x \in S_t \wedge x \in V_{j,t}\right] \\
= \; & Pr_{x\sim\mathcal{X}_i}\left[\begin{array}{l} \bigwedge \quad \neg\left(Transmit(y_p^k, x)\right) \mid x \in S_t \wedge x \in V_{j,t} \\ k = 1 \dots M \\ p = 1 \dots l_k \end{array}\right] \\
= \; & \prod_{k=1}^{M} Pr_{x\sim\mathcal{X}_i}\left[\bigwedge_{p=1\dots l_k} \neg\left(Transmit(y_p^k, x)\right) \mid x \in S_t \wedge x \in V_{j,t}\right] \\
= \; & \prod_{k=1}^{M} (1 - c_{ik})^{\kappa_j \delta t \xi_{jk} \frac{|I_{k,t}|}{N_k}}
\end{aligned}$$

Then, we plug this back into the previous formula, and we get:

$$Pr_{x\sim\mathcal{X}_i}\left[x \in E_{t+\delta t} \mid x \in S_t \wedge x \in V_{j,t}\right] \quad = \quad 1 - \prod_{k=1}^{M} (1 - c_{ik})^{\kappa_j \xi_{jk} \frac{|I_{k,t}|}{N_k} \delta t}$$

The first order approximation of $1 - \prod_{k=1}^{M}(1 - c_{ik})^{v_k}$ is $\sum_{k=1}^{M} -v_k \log(1 - c_{ik})$. Applying this approximation here, we get:

$$Pr_{x\sim\mathcal{X}_i}\left[x \in E_{t+\delta t} \mid x \in S_t \wedge x \in V_{j,t}\right] \quad \simeq \quad \delta t \sum_{k=1}^{M}\left(-\kappa_j \xi_{jk} \frac{|I_{k,t}|}{N_k} \log\left(1 - c_{ik}\right)\right)$$

Define $\beta_{ijk} = -\kappa_j \log(1 - c_{ik})$, let δt converge to zero, and we get:

$$\frac{dPr_{x\sim\mathcal{X}_i}\left[x \in E_{t+dt} \mid x \in S_t \wedge x \in V_{j,t}\right]}{dt} \simeq \sum_{k=1}^{M} \left(\xi_{jk} \frac{|I_{k,t}|}{N_k} \beta_{ijk}\right)$$

If there is only one subpopulation i, then we fall back to the formula of [30]. We have :

$$\beta_i = -\kappa_i \log(1 - c_i)$$

$$\frac{d}{dt}\mathbb{E}\left[|E_{i,t+dt} - E_{i,t}|\right] \simeq -|S_{i,t}| \left(\frac{|I_i|}{N_i}\beta_i\right)$$

and the force of infection as follows :

$$\lambda_i = \beta_i \frac{|I_i|}{N_i}$$

3. Final Formula We simply have to plug in the probability ρ_{ij} that i visits j. We get, for the new interpretation NIIF:

$$\frac{d}{dt}\mathbb{E}\left[|E_{i,t+dt} - E_{i,t}|\right] \simeq -|S_{i,t}| \sum_j \rho_{ij}\kappa_j \log\left[1 - \sum_{k=1}^{M}\left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right)\right]$$

And for the old interpretation OIIF [30]:

$$\frac{d}{dt}\mathbb{E}\left[|E_{i,t+dt} - E_{i,t}|\right] \simeq -|S_{i,t}| \sum_j \rho_{ij} \sum_{k=1}^{M}\left(\xi_{jk}\frac{|I_{k,t}|}{N_k}\beta_{ijk}\right)$$

In conclusion, in this thesis, we use the interpretation NIIF to present the infection force in meta-population. This interpretation shows enough interactions between individuals in the same city, individuals in different cities, and between cities in meta-population.

2.2 The equilibrium state

As mentioned above, we are interested in investigating the SEIR model with demography. This model is an extension of the epidemic simple SEIR model that allows for birth and death. Because, demographic processes play an important role for exploring the longer-term persistence and endemic dynamics of an infectious disease. In SEIR model with demography, assuming that the time scale of disease propagation is fast enough not to be affected by population births and deaths, the model is considered in the case with a constant birth rate and a constant per-capita death rate (that are independent of the population size). Now, we consider a meta-population of many subpopulations. Each subpopulation is modelled by a SEIR model with demography. Again, the ordinary differential equations for a $subpopulation_i$ in a metapopulation is presented as follows:

$$\frac{dS_i}{dt} = \mu N_i - \lambda_i S_i - \mu S_i \tag{2.2}$$
$$\frac{dE_i}{dt} = \lambda_i S_i - \mu E_i - \sigma E_i \tag{2.3}$$
$$\frac{dI_i}{dt} = \sigma E_i - \mu I_i - \gamma I_i \tag{2.4}$$
$$\frac{dR_i}{dt} = \gamma I_i - \mu R_i \tag{2.5}$$

In simulation, we know that the equilibrium state allow a disease to persist in a population for a long time, and the variables do not change with time. So, an infectious disease in the *subpopulation*$_i$ is available in long term this system is at equilibrium. It means that at which $\frac{dS_i}{dt} = \frac{dE_i}{dt} = \frac{dI_i}{dt} = \frac{dR_i}{dt} = 0$ (*). Thus, we let all ordinary differential equations in the system be equal to zero, then calculate the values of the variables (now denoted by S_i^*, E_i^*, I_i^*, and R_i^*) that satisfy this condition (*). We have these values as follows:

$$S_i^* = N_i \frac{(\gamma+\mu)(\sigma+\mu)}{\beta\sigma} \tag{2.6}$$
$$E_i^* = N_i \mu \left(\frac{1}{\sigma+\mu} - \frac{\gamma+\mu}{\beta\sigma} \right) \tag{2.7}$$
$$I_i^* = N_i \mu \frac{\beta\sigma - (\sigma+\mu)(\gamma+\mu)}{\beta(\sigma+\mu)(\gamma+\mu)} \tag{2.8}$$
$$R_i^* = N_i - S_i^* - E_i^* - I_i^* \tag{2.9}$$

Here, if we set $R_0 = \frac{\beta\sigma}{(\gamma+\mu)(\sigma+\mu)}$, so we have

$$S_i^* = N_i \frac{1}{R_0} \tag{2.10}$$
$$E_i^* = N_i \frac{\mu\sigma}{R_0} (R_0 - 1) \tag{2.11}$$
$$I_i^* = N_i \frac{\mu}{\beta} (R_0 - 1) \tag{2.12}$$
$$R_i^* = N_i - S_i^* - E_i^* - I_i^* \tag{2.13}$$

One normal conditions for all population variables is that the equilibrium values cannot be negative. Therefore, an infectious disease is available in the *subpopulation*$_i$ if $R_0 > 1$. Now, the endemic equilibrium in the system is given by $(S_i^*, E_i^*, I_i^*, R_i^*) = (N_i \frac{1}{R_0}, N_i \frac{\mu\sigma}{R_0}(R_0 - 1), N_i \frac{\mu}{\beta}(R_0 - 1), N_i(1 - \frac{1}{R_0} - \frac{\mu\sigma}{R_0}(R_0 - 1) - \frac{\mu}{\beta}(R_0 - 1))$.

2.3 DIZZYS : Description of package dizzys

Stochastic and analytical methods are widely applied for the analysis of epidemic models. Many simulation software as well as packages are proposed to help scientists observe fluctuations of infectious diseases over time. These tools simulate epidemic models either by dealing with a set of ordinary differential equations (ODEs) or by applying the stochastic simulation algorithm (SSA) of Gillespie. Simple epidemic models work well on these software tools. However, the accuracy, the simulation speed, and the complexity of models that the tools can simulate are three main drawbacks that always prompt us not to stop improving tools to increase efficient implementations available in software tools. Moreover, rather than dynamics of infectious diseases, predicting the potential spread of an infectious disease in a meta-population is the most difficult problem for scientists. To give an exact prediction about propagation of infectious diseases in a meta-population, we need to make simulations in a complex meta-population with many interconnected sub-populations where the meta-population takes into account many factors about the pathogen, the climatic conditions and simultaneously the interactions between sub-populations. Therefore, we introduce the "dizzysNewInfec" package that allows us to exactly simulate and accurately analyze dynamics of an infectious disease in a meta-population of interconnected sub-populations by using two basic and common disease models SIR and SEIR, and by implementing the direct algorithm of Gillespie in 1977 and the adaptive tau leaping to approximate the trajectory of a continuous-time stochastic process. In addition, on the technical aspect, this package integrates C++ in R, we use C++ to build algorithms, and use R to show two-dimensional and three-dimensional interfaces and use the available statistic functions in R to analyze obtained results. Hence, dizzysNewInfec, it speeds up simulations, it is very easy to install, to use and to show trajectories of disease evolution over time in a meta-population of sub-populations.

2.3.1 Introduction

Fundamentally, Kermack-McKendrick gave the first epidemic model to provide a mathematical description of the kinetic transmission of an infectious disease in an unstructured sub-population. Due to this model, today we have known well the SIR and SEIR deterministic epidemic models. These two basic epidemic models are very popularly used by scientists. The reactions in the system are modelled by a set of Ordinary Differential Equations (ODEs) [36]. The deterministic method is the simplest to solve an epidemic model. The main idea of this method is to solve a single differential equation per species of the model. Basically, the deterministic method uses the law of mass action that has applicability in many areas of science. In chemistry, it is also called Fundamental Law of Chemical Kinetics (the study of rates of chemical reactions), introduced by the Norwegian scientists Cato M. Guldberg in 1864-1879 and Peter Waage. The law of mass action shows a simple relation between reaction rate and molecular component concentrations. For a set of initial molecular concentrations given, the law of mass action permits us to see the component concentrations over time. The states of a reaction are a homogeneous, free medium. The reaction rate will be directly scaled with the concentrations of the elements. Most systems can use the traditional

deterministic approaches to simulate. It is evident that many systems such as some biochemical systems consist of random, discrete interactions between individual elements. In fact, we have applied the deterministic model in the epidemiology to solve epidemic models such as the SEIR and SIR models. However, in the case, these systems become smaller and smaller, the traditional deterministic models may not be accurate. In addition, in the deterministic approach, the time evolution of a reacting system is assumed that it is both continuous and deterministic. But in fact, molecular population levels can change only by discrete integer amounts. It is the reason for that this time evolution is not both a continuous process and a discrete process. Indeed, the deterministic approach is impossible to predict the exact molecular population levels at some future times unless we can compute exactly the precise positions and velocities of all molecules in the system. It is the reason for that the fluctuations of these systems can be simulated exactly by applying stochastic models via Stochastic Simulation Algorithms (SSA) [18, 19]. The SSA uses Monte Carlo (MC) methods to study the time evolution of the jump process. Because the basis feature of the Monte Carlo simulation is insensitive to the dimensionality of the problem, and the work grows linearly with the number of reaction channels in the model. The SSA describes time-evolution statistically correct trajectories of finite populations in continuous time by solving the corresponding stochastic differential equations. Using the stochastic models can solve three questions. (1) These models take account the discrete character of the number of elements and the evidently random character of collision among elements. (2) They coincide with the theories of the dynamic and stochastic processes. (3) They are a good idea to describe "small systems" and "unstable systems". The main idea of the stochastic models is that element reactions are essentially random processes. We don't know certainly how a reaction occurs at a moment. We also call it the process stochasticity. Demographic stochasticity is considered as fluctuation in population processes that is based the random nature of events at the level of the individual. Each event is related to one baseline probability fixed, individuals are presented in differing fates due to chance. In addition to the demographic stochasticity, the number of infectious, susceptible, exposed and recovered individuals in a meta-population infected by an infectious disease is now required to be an integer. Modeling approaches that incorporate demographic stochasticity are called event-driven methods. These methods require explicit consideration of events. The first approach published by Daniel T.Gillespie in 1976 [18] is an exact stochastic simulation approach for chemical kinetics. The Gillespie stochastic simulation algorithm (SSA) has become the standard procedure of the discrete-event modelling by taking proper value of the available randomness in such a system. The methods modelling the event-driven model demands explicit presentation of events. Therefore, the "dizzysNewInfec" package permits us to obtain the dynamics of the deterministic and the stochastic dynamics of two basic epidemic models SIR and SEIR. We use the exact algorithm of Gillespie in 1977 and the "adaptive tau-leaping" algorithm in the package. With these two algorithms, each has its private advantages and its private disadvantages. For the exact algorithm, it gives us a really exact approach of simulating population-based time-to-event through two steps with many iterations of 1) searching the time of next event by an exponentially distributed function and 2) searching the nature of next event. Each single event in

the Gillespie's solution is explicitly simulated, so this exact simulation becomes exceedingly slow and impractical in systems where the transition rate grows large over time. Hence, approximate models are born instead of the Gillespie's solution, they are concerned with larger transition rates and with increasing simulation speed while still maintaining reasonable accuracy. The "adaptive tau-leaping" algorithm known as an approximate method reduces the number of iterations by treating transition rates as constant over time periods for which this approximation leads to little error [**?**].

The SEIR epidemic model used in our package have successful described hosts within a population as Susceptible (the number of individuals not yet infected with the disease), Exposed (the number of individuals who are infected with the disease but not infectious), Infected (the number of individuals who have been infected with the disease and are capable of spreading the disease), and Recovered (the number of individuals who have successfully cleared the infection). We ignore population demography-births, deaths, and migration. So we have only three transitions:

$$S \to E, E \to I, I \to R. \tag{2.14}$$

It is obvious that the second and third of these transitions are easier, so we focus on the first transition in which the level of the infectious disease influences strongly the disease transmission rate from a susceptible individual into the infected class. If this first step doesn't occur, all infected individuals can go to the recovered class. This first transition plays an important role in studying fluctuation of disease. Many researches have been tried to find the "infectious period" (the amount of time spent in the infectious class). They have found that for acute infections, the "infectious period" is distributed around some mean values that we can estimate from clinical data. To present enough properties of the "infectious period" by formulations, we have to depend upon three main factors of the disease transmission from S to E : the prevalence of infecteds, the underlying population contact structure, and the probability of transmission given contact. For a directly transmitted disease, the key factor is the contact between susceptible and infected individuals. Hence, in this package "dizzysNewInfe", we interpret the "infectious period" in more detailed, and realistic contexts. We introduce a new formula of the probabilistic derivation of multipopulation epidemic model called NIIF. We study the case where agent x native from subpopulation i visits sub-population j. In a meta-population of sub-populations, during a small interval of time δt, each native individual of the sub-population i visits one single sub-population j (with probability ρ_{ij}) and will see on average K_j. These individuals come from all sub-populations. This is absolutely a new interpretation about the infection between individuals and the propagation of disease between sub-population. In previous research described in [keeling2011], the authors always assumed that both the number of people we meet and the number of infected people we meet are fixed. This assumption simplifies the relations between individuals and between sub-populations. It's the reason for that the formula of the infection force did not present clearly the complex connections between individuals and between sub-populations in a meta-population. In our interpretation, we assume that for each sub-population k, the number

of people native from k that we meet during δt follows a Poisson process. So both the number of people we meet and the number of infected people we meet during δt should be random variables.

In short, the **dizzysNewInfec** package implements both the exact solution and the approximate solution for the SIR and SEIR models. The package installs well the new interpretation of the infection force NIIF for SIR and SEIR meta-population models by integrating the R package and the C++ implementation. We can choose one of these two solutions to simulate when the number of sub-populations in a meta-population increases. Using C++ to perform the algorithms, and R to create interfaces makes the **dizzysNewInfec** package much faster and much easy to use than any pure R package.

2.3.2 Methods

In this section, first we will talk about the deterministic and stochastic SEIR models. Then, we will present transformation the SEIR model into the SIR model through the usage of the two algorithms. We hope that the models and the algorithms should be well understood before obtaining simulation results.

a. Deterministic SEIR model: To describe infectious diseases in a in a spatial context, we consider a meta-population of n sub-populations. In sub-population i of size N_i, disease dynamics can be deterministically described by the following set of differential equations:

$$\frac{dS_i}{dt} = \mu N_i - \lambda_i S_i - \mu S_i \tag{2.15}$$

$$\frac{dE_i}{dt} = \lambda_i S_i - \mu E_i - \sigma E_i \tag{2.16}$$

$$\frac{dI_i}{dt} = \sigma E_i - \mu I_i - \gamma I_i \tag{2.17}$$

$$\frac{dR_i}{dt} = \gamma I_i - \mu R_i \tag{2.18}$$

where S_i, E_i, I_i et R_i are respectively the numbers of susceptible, exposed, infectious and recovered in this sub-population i. Individuals are born susceptible and die at a rate μ, become infected with the force of infection λ_i, infectious after a latency period of an average duration of $1/\sigma$ and recover at the rate γ. In case the infectious contact rate is constant, the equilibrium values of the variables S, E, I and R can be expressed analytically. The force of infection depends not only on the total population size N_i and the number of infected I_i in subpopulation i, but also in other sub-populations :

$$\lambda_i = \sum_j \rho_{ij}\kappa_j \log\left[1 - \sum_{k=1}^{M}\left(\frac{|I_{k,t}|}{N_k} \times c_{ik} \times \xi_{jk}\right)\right]$$

where $\rho_{i,j}$ the probability that an individual from sub-population i visits sub-population j. κ_j is the average number of contacts per unit of time a susceptible will have when visiting city. $c_{i,k}$ is the probability that a susceptible

individual native from i being in contact with another infected individual native from k. ξ_{jk} refers to the probability that an individual y meeting x in C_j comes from C_k. See appendix for detail on the construction of this equation. We can verify that in the limit case on one single subpopulation in the metapopulation ($i = j$ and $n = 1$), we have :

$$\lambda_i = -\kappa_i \log(1 - \frac{I_i}{N_i} \times c_{ii}) \tag{2.19}$$

In the case, we consider that the contact number K_i is seasonally forced [Altizer2006]:

$$K_i(t) = K_{i0} \left[1 + K_{i1} \cos\left(\frac{2\pi t}{T} + \varphi_i \right) \right] \tag{2.20}$$

where K_{i0} and K_{i1} are the mean value and amplitude of the contact number and T and φ_i are the period and the phase of the forcing.

b. Stochastic models: The stochastic model is built by depending upon the deterministic model. More detailed, this stochastic model relies on chance variation in risks of exposure, disease, and other factors. It is referred as an individual-level modeling, because every individual plays an important role in the model, so this stochastic model can consider well most small population size that the deterministic model can not do. To implement this SEIR stochastic model, there are many different methods. The first method "First Reaction Method" is born in 1976 by Gillespie. Then, according to this first method and these two key factors of demographic stochasticity models (event, randomness), many scientists have improved the first method, and created many better algorithms for stochastic simulations. There are two main types of methods, exact methods and approximative methods. The typical approach in exact methods most practitioners use, is the algorithm "Direct Method" of Gillespie(1977) improved from the first approach "First Reaction Method", and in approximative methods, is the "tau-leaping" method. For the Direct Method (Gillespie 1977), the first step estimates the time until the next event, by accumulating the rates of all possible events. Then, by transforming event rates into probabilities, the method randomly selects one of these events. The time and numbers in each class are then updated according to which event is chosen. We repeat this process to iterate model through time. For the "tau-leaping" method, the main crux is the use of Poisson random variables to approximate the number of occurrences of each type of reaction event during a carefully selected time period, τ . According to these two types of algorithms, the common point is both methods use continuous-time Markov process for which the transition rates are constants, isn't a function of time. The future state of the process, is only conditional on the present state, but independent of the past. However, for the exact algorithm, its advantage give us a really exact approach to simulate time-to-event model. This process is repeated to iterate the mode. GibsonBruck(2000) [KeelingRohani2008] modified the first reaction method and created the Next Reaction method that substantially more challenging to program but is significantly faster than even the method when there are a large number of different event types. The Direct, First Reaction and Next Reaction methods are all exact stochastic approaches of the underlying ordinary differential equations. But, its disadvantages are 1)

noise in exact simulations only affects the probabilities associated with fates of individuals and the updating of each consecutive event is independent – there is no assumption concerning environmental stochasticity; 2) these exact solutions become too slow and impractical when any one transition rate is large, when there is a big number of sub-populations or one a big number of event in a meta-population. It is the reason for that, approximate models have been proposed instead of the exact stochastic methods. Gillespie (2001) has proposed a new method that decreases the simulation accuracy, but increases simulation speed. This is the "tau-leap method" known as an approximate method reduces the number of iterations by treating transition rates as constant over time periods for which this approximation leads to little error [KeelingRohani2008] as mentioned above. However, when we use the "tau-leap method", there is a possibility, the number of individual in each class can become negative. According to Keeling2008 [KeelingRohani2008], the authors compared these methods together, they pointed out that when the population size increases, the simulation time of the three methods (First Reaction, Direct Method, tau -leap) is almost augmented. The simulation time of the method "First Method" is maximum, it means that this is the slowest method, but the simulation time of the method "tau-leap" is minimum or the fastest. In the shape of my package, we use the direct method to exactly estimate spread of diseases in meta-population. Because the direct method is one exact approach and its simulation time isn't too slow and too fast. We use the approximate method to speed up simulations in the case where there are many sub-populations in a meta-population.

c. Direct method :

Based on the differential equations above, we give a stochastic version of this model. We use for that a population-based time-to-next-event model based on Gillespie's algorithm [Daniel.T.Gillespie1977]. Table 3 lists all the events of the model, occurring in subpopulation i.

Table 3: Events of the stochastic version of the model of equations, occuring in subpopulation i.

Events	**Rates**	**Transitions**
birth	μN_i	$S_i \leftarrow S_i + 1$ and $N_i \leftarrow N_i + 1$
death of a susceptible	μS_i	$S_i \leftarrow S_i - 1$
death of an exposed	μE_i	$E_i \leftarrow E_i - 1$
death of an infected	μI_i	$I_i \leftarrow I_i - 1$
death of an immune	μR_i	$I_i \leftarrow I_i - 1$
infection	$\lambda_i S_i$	$S_i \leftarrow S_i - 1$ and $E_i \leftarrow E_i + 1$
becoming infectious	σE_i	$E_i \leftarrow E_i - 1$ and $I_i \leftarrow I_i + 1$
recovery	γI_i	$I_i \leftarrow I_i - 1$ and $R_i \leftarrow R_i + 1$

To apply the Direct Method in a meta-population of n sub-populations, we know that at a moment t, there are only one single event fired in one single sub-population during one time unit *tstep*. Hence, we improve the Direct Method for one single population by adding a random number to calculate chance which sub-population to fire. Starting from the initial states, the stochastic simulation algorithms simulate the trajectory in population processes by repeatedly answering the following three big questions and updating the states.

- When (time) will the next event occur?

- Which subpopulation where the event will occur next?
- Which event in the subpopulation will occur next?

Although, it is a small improvement in the original Direct Method, however it turns back much more benefits for a meta-population, it well presents the interactions between sub-populations in the meta-population.

d. Adaptive tau-leaping algorithm In this step, we provide basic concepts for the adaptive tau-leaping algorithm by using the detailed description of Cao [Cao2007].

For the Markov process at time t, to describe a meta-population of n sub-populations, we have: **state set:** X(t) $X(t) := [S_1(t), ..., S_n(t), E_1(t), ..., E_n(t), I_1(t), ..., I_n(t), R_1(t), ..., R_n(t)]$ each variables of $X(t)$ is defined on the non-negative integers. **set of allowable transitions:** $\triangle_j$, for each allowable transition, j, we define a rate λ_j, by using a function independent on t but dependent on the current state $X(t)$, to calculate transition rates given the state ($\lambda(X)$) through the deterministic model, and a vector of n integers, $\triangle_j := [\triangle_{j,1}, ..., \triangle_{j,n}]$, that reflects the change in state if this transition were followed: $X(t) + \triangle_j$.

e. Transformation SEIR model into SIR model : The SIR model used in this package is the SIR model with births and death. By observing this SEIR model, if we give a numerical value for the parameter σ then a SEIR model would have. On the other side, if we give Inf (to infinity) the parameter σ then we have a SIR model with birth and death (because, basically, a SEIR model tends to a SIR model when σ tends to infinity).

2.3.3 Examples

a. Example 1 The deterministic SEIR model with one sub-population by exploiting the 'globSEIRNewInfec' function in the package.

«fig=T,echo=T»= library(dizzysNewInfec) We have the values of parameters and of variables. Here, we have S=E=I=R=NULL and N=1e7. It means that we use N=1e7 to calculate the equilibrium values of variables. obj<- globSEIRNewInfec(typeSIMU="deterministic",duration=10*365,mu=1/(70*365),sigma=1/8,gamma=1/5,phiPHASE=0,nbCO

Use the plot function of the seir class plot(obj,col="red",ylab="number of infectives", xlab="time (day)") @

The obtained result:

Now, we want to continue or to redo this simulation with other values of parameter, we can do it by exploiting the 'globSEIRSimulNewInfec' function in the package. «fig=T,echo=T»= newobj<- globSEIRSimulNewInfec(obj,duration= 20*365,continue=T, append=T,nbCONTACT0=100, nbCONTACT1=0.0, phiPHASE=pi/2) plot(newobj,col= of infectives", xlab="time (day)") @

3 Relation structure/dynamique spatiale et persistence

C'est ce que tu es en train d'explorer pour le moment. Plusieurs questions à explorer. Chaque question constitue un sous-chapitre. A toi de développer et structurer cette partie plus en détails.

4 Contrôle par reinforcement learning:

- comment utiliser ton simulateur pour faire du reinforcement learning. Partie qui reste à développer.

5 Conclusion et discussion générales.

Commence donc à écrire certaines parties dès que tu peux (un peu chaque semaine et de plus en plus au fur et à fur que le temps avance). Pense aussi à bien faire la bibliographie (il faut que tu sois incollable sur le sujet). Les nouveau login et mot de passe de Bibliovie (http://bibliovie.inist.fr) sont 15SCBUMR5290 et 4NX9E5. Ou, on peux utiliser l'account de Giang à UPMC selon les conseil de la site : http://www.jubil.upmc.fr/fr/ressources_en_ligne2/mode_acces_ressour

References

[1] David Adalsteinsson, David McMillen, and Timothy C Elston. Biochemical network stochastic simulator (bionets): software for stochastic modeling of biochemical networks. *BMC bioinformatics*, 5(1):24, 2004.

[2] S. Altizer, A. Dobson, P. Hosseini, P. Hudson, M. Pascual, and P. Rohani. Seasonality and the dynamics of infectious diseases. *Ecol Lett*, 9(4):467–484, Apr 2006.

[3] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[4] James L Blue, Isabel Beichl, and Francis Sullivan. Faster monte carlo simulations. *Physical Review E*, 51(2):R867, 1995.

[5] B. Bolker and B. Grenfell. Space, persistence and dynamics of measles epidemics. *The Royal Society*, 348:309–320, 1995.

[6] B. M. Bolker and B. T. Grenfell. Impact of vaccination on the spatial correlation and persistence of measles dynamics. *Proc Natl Acad Sci U S A*, 93(22):12648–12653, Oct 1996.

[7] Alfred B Bortz, Malvin H Kalos, and Joel L Lebowitz. A new algorithm for monte carlo simulation of ising spin systems. *Journal of Computational Physics*, 17(1):10–18, 1975.

[8] Kevin Burrage and Tianhai Tian. Poisson runge-kutta methods for chemical reaction systems. 1(1):82–96, 2004.

[9] Yang Cao, Daniel T Gillespie, and Linda R Petzold. Avoiding negative populations in explicit poisson tau-leaping. *The Journal of chemical physics*, 123(5):054104, 2005.

[10] Yang Cao, Daniel T Gillespie, and Linda R Petzold. The slow-scale stochastic simulation algorithm. *The Journal of chemical physics*, 122(1):014116, 2005.

[11] Yang Cao, Daniel T Gillespie, and Linda R Petzold. Adaptive explicit-implicit tau-leaping method with automatic tau selection. *The Journal of chemical physics*, 126(22):224101, 2007.

[12] Yang Cao, Hong Li, and Linda Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The journal of chemical physics*, 121(9):4059–4067, 2004.

[13] Abhijit Chatterjee, Dionisios G Vlachos, and Markos A Katsoulakis. Binomial distribution based τ-leap accelerated stochastic simulation. *The Journal of chemical physics*, 122(2):024112, 2005.

[14] Pierre Chauvin. Constitution and monitoring of an epidemiological surveillance network with sentinel general practitioners. *European journal of epidemiology*, 10(4):477–479, 1994.

[15] Johan Elf and Måns Ehrenberg. Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases. *Systems biology*, 1(2):230–236, 2004.

[16] Frank Fenner. [book review: The eradication of smallpox: Edward jenner and the first and only eradication of a human infectious disease herve bazin, andrew morgan, glenise morgan]. *Quarterly Review of Biology*, 76(4):476, 2001.

[17] Michael A Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A*, 104(9):1876–1889, 2000.

[18] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[19] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[20] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[21] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[22] Daniel T Gillespie and Linda R Petzold. Improved leap-size selection for accelerated stochastic simulation. *The Journal of Chemical Physics*, 119(16):8229–8234, 2003.

[23] B.T. Grenfell, B. M. Bolker, and A. Klegzkowski. Seasonality and extinction in chaotic metapopulation. *The royal society*, 259:97–103, 1995.

[24] I. Hanski. Metapopulation dynamics. *Nature*, 396, 1998.

[25] Ilkka Hanski. A practical model of metapopulation dynamics. *Journal of animal ecology*, pages 151–162, 1994.

[26] Ilkka Hanski and Michael Gilpin. Metapopulation dynamics: brief history and conceptual domain. *Biological journal of the Linnean Society*, 42(1-2):3–16, 1991.

[27] Susan Harrison and Andrew D Taylor. Empirical evidence for metapopulation dynamics. *Metapopulation biology: ecology, genetics, and evolution. Academic Press, San Diego, California, USA*, pages 27–42, 1997.

[28] Eric L Haseltine and James B Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *The Journal of chemical physics*, 117(15):6959–6969, 2002.

[29] Andreas Hellander. Efficient computation of transient solutions of the chemical master equation based on uniformization and quasi-monte carlo. *The Journal of chemical physics*, 128(15):154109, 2008.

[30] M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in humans and animals*. Princeton University Press, 2008.

[31] Hiroyuki Kurata, Hana El-Samad, T-M Yi, Mustafa Khammash, and J Doyle. Feedback regulation of the heat shock response in e. coli. In *Decision and Control, 2001. Proceedings of the 40th IEEE Conference on*, volume 1, pages 837–842. IEEE, 2001.

[32] Christian Lécot and Bruno Tuffin. Quasi-monte carlo methods for estimating transient measures of discrete time markov chains. In *Monte Carlo and Quasi-Monte Carlo Methods 2002*, pages 329–343. Springer, 2004.

[33] R. Levins. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the Entomological Society of America*, 15:237–240, 1969.

[34] Li. User's guide for stochkit.

[35] Hong Li and Linda Petzold. Efficient parallelization of the stochastic simulation algorithm for chemically reacting systems on the graphics processing unit. *International Journal of High Performance Computing Applications*, 24(2):107–116, 2010.

[36] Hong Li and Linda R Petzold. Stochastic simulation of biochemical systems on the graphics processing unit. *Bioinformatics. Cité pages 32 et*, 35, 2007.

[37] PA Maksym. Fast monte carlo simulation of mbe growth. *Semiconductor Science and Technology*, 3(6):594, 1988.

[38] James M McCollum, Gregory D Peterson, Chris D Cox, Michael L Simpson, and Nagiza F Samatova. The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. *Computational biology and chemistry*, 30(1):39–49, 2006.

[39] Peter Ludwig Panum. Observations made during the epidemic of measles on the faroe islands in the year 1846. *The Challenge of Epidemiology: Issues and Selected Readings, Pan American Health Organization, New York*, pages 37–41, 1988.

[40] Michel F Pettigrew and Haluk Resat. Multinomial tau-leaping method for stochastic kinetic simulations. *The Journal of chemical physics*, 126(8):084101, 2007.

[41] Christopher V Rao and Adam P Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the gillespie algorithm. *The Journal of chemical physics*, 118(11):4999–5010, 2003.

[42] Howard Salis and Yiannis Kaznessis. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *The Journal of chemical physics*, 122(5):054103, 2005.

[43] L Shaw, W Spears, L Billings, and P Maxim. Effective vaccination policies. *Information sciences*, 180(19):3728–3744, 2010.

[44] Michael L Simpson, Chris D Cox, and Gary S Sayler. Frequency domain analysis of noise in autoregulated gene circuits. *Proceedings of the National Academy of Sciences*, 100(8):4551–4556, 2003.

[45] Michael L Simpson, Chris D Cox, and Gary S Sayler. Frequency domain chemical langevin analysis of stochasticity in gene transcriptional regulation. *Journal of theoretical biology*, 229(3):383–394, 2004.

[46] Tianhai Tian and Kevin Burrage. Binomial leap methods for simulating stochastic chemical kinetics. *The Journal of chemical physics*, 121(21):10356–10364, 2004.

[47] Eugenia Tognotti. The eradication of smallpox, a success story for modern medicine and public health: What lessons for the future? *The Journal of Infection in Developing Countries*, 4(5):264–266, 2010.

[48] Thomas E Turner, Santiago Schnell, and Kevin Burrage. Stochastic approaches for modelling in vivo reactions. *Computational biology and chemistry*, 28(3):165–178, 2004.

[49] P Van den Driessche. Spatial structure: Patch models. In *Mathematical Epidemiology*, pages 179–189. Springer, 2008.

[50] David Sloan Wilson. *The natural selection of populations and communities*. Benjamin/Cummings Pub. Co., 1980.

[51] WM Young and EW Elcock. Monte carlo studies of vacancy migration in binary ordered alloys: I. *Proceedings of the Physical Society*, 89(3):735, 1966.