Développement d'une approche basée sur les modèles dynamiques compartimentaux pour évaluer le bénéfice et l'impact des nouveaux médicaments en population générale : application au cas de l'hépatite C

Arnaud Nucit

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Mathématique*

Présentée et soutenue le *16/12/2016* **par :**

ARNAUD NUCIT

Développement d'une approche basée sur les modèles dynamiques compartimentaux pour évaluer le bénéfice et l'impact des nouveaux médicaments en population générale : application au cas de l'hépatite C

JURY

PASCALE TUBERT-BITTER	Directrice de recherche	Présidente du Jury
JEAN-YVES DAUXOIS	Professeur des Universités	Directeur de thèse
AURÉLIEN LATOUCHE	Professeur des Universités	Rapporteur
FABRICE ROSSI	Professeur des Universités	Rapporteur
NATHALIE SCHMIDELY	Ingénieure	Invitée

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées
Unité de Recherche :
Institut de Mathématiques de Toulouse (UMR 5219)
Directeur de Thèse :
Jean-Yves DAUXOIS

Remerciements

Si vous pouvez lire ces lignes aujourd'hui c'est grâce à un certain nombre de personnes qui m'ont suivi et guidé durant l'accomplissement de ces travaux et sans qui tout ceci n'aurait pas été possible. Prenons donc un peu de temps (si, si) pour remercier ces personnes.

J'aimerais tout d'abord remercier mon directeur de thèse Jean-Yves Dauxois, Professeur des Universités à l'Université de Toulouse-INSA, qui m'a soutenu, guidé, conseillé et encouragé durant ce parcours du combattant. J'aimerais notamment saluer sa motivation et sa patience qui lui ont permis de relire sans relâche les notes (parfois des âneries) que je lui ai envoyées tout au long de ces années. Merci encore pour les allers et retours répétés entre Paris et la ville rose.

Je remercie également Nathalie Schmidely qui m'a permis d'intégrer son équipe au sein de Bristol-Myers Squibb et qui m'a soutenu du début à la fin sur ce projet de thèse. Je ne peux évidemment pas parler de Nathalie sans évoquer mes nombreux collègues de bureau que je ne nommerai pas tant la liste est longue (PEMB/RWR, MDO, RCO, la PV, l'Infomed, les gens de l'accueil, de la cafèt', non sérieux la liste est longue). Je les remercie pour tous ces moments de fun qui m'ont permis à la fois de décompresser et d'avoir des relations sociales autres qu'avec mon ordinateur. Je remercie tout de même très chaleureusement mon équipe d'origine (Anne, Annelore, Astrid, Coralie, Driss, Elisette, Isabelle, Jean-Luc, Karine, Leyla, Marc, Marie-Hélène, Marie-Jo, Mélanie, Pierre, Rosine).

Un très grand merci également aux membres du jury que je n'ai pas cités encore, Pascale Tubert-Bitter, Directrice de Recherche à l'INSERM, Aurélien Latouche, Professeur des Universités au CNAM et Fabrice Rossi, Professeur des Universités à l'Université de Paris 1 Panthéon-Sorbonne pour avoir répondu présent lors de la soutenance. Merci notamment à Pascale Tubert-Bitter d'avoir bien voulu présider ce jury et merci à Aurélien Latouche et Fabrice Rossi d'avoir pris le temps de relire en profondeur mon manuscrit de thèse pour en rédiger les rapports de thèse.

J'aimerais également adresser un grand merci à Julien Randon-Furling, Maître de conférences à l'Université de Paris 1 Panthéon-Sorbonne avec qui j'ai eu le plaisir de travailler durant les derniers mois de la thèse et qui a partagé mon stress durant le rush final.

Je tenais à remercier mes amis pour m'avoir encouragé durant ces quelques années et surtout ces derniers mois qui ont été les plus durs. Merci d'avoir su faire avec mes indisponibilités remarquées (et remarquables?) pour nos sorties, répétitions théâtrales, etc.

J'adresse un énorme merci à mes parents et à ma soeur qui ont toujours cru en moi et qui ne m'ont malheureusement plus revu beaucoup à la maison durant ces années de travail. Tout ceci n'aurait évidemment pas été possible sans votre soutien et votre confiance.

Et enfin je voulais remercier Nelly, qui a un été soutien indéfectible durant toutes ces années, qui a su être patiente et su supporter mes périodes de stress et d'incertitude. Merci d'avoir été là au jour le jour et d'avoir résisté à ma lente dégénérescence mentale en terme de vie à deux. Les mots ne suffisent pas à exprimer ma gratitude.

Abstract

Français

Ce travail de thèse s'articule autour de trois parties distinctes abordant chacune un thème précis lié à l'épidémiologie. La première partie de ces travaux s'inscrit dans le cadre de la propagation de virus via l'utilisation de modèles épidémiques. Dans cette partie, sont analysés différentes méthodes d'estimations paramétriques et y sont étudiés la qualité de ces estimateurs. Une application à des virus informatiques est proposée. La deuxième partie de cette thèse propose une méthode d'estimation de la prévalence actuelle du virus de l'hépatite C en France par l'intermédiaire d'un modèle de rétro-calcul associé à un modèle de Markov modélisant l'histoire naturelle de la maladie. Cette méthode et les résultats qui en découlent sont comparés avec les résultats obtenus via l'approche de référence en France. Enfin, la dernière partie s'intéresse à l'étude de l'impact des nouvelles thérapeutiques anti-hépatite C susceptible d'éradiquer le virus à moyen terme. En assimilant la population d'intérêt à un groupement de graphes aléatoires, la propagation du virus est modélisée à partir d'un modèle de métapopulation construit sur la base de données migratoires où les dynamiques de chaque sous-population sont régies par un ensemble d'équations différentielles déterministes. Ce travail doctoral a été réalisé dans le cadre dune convention CIFRE avec les laboratoires Bristol-Myers Squibb.

Mots clé : modélisation, modèles compartimentaux, statistique, hépatite C, modèles dynamiques, épidémiologie.

English

The works undertaken in this doctoral thesis are conducted in three parts, each one dealing with a specific epidemiology-related domain. The first part of this work deals with the propagation of viruses by using well-known epidemic models. It is mainly focused on the analyze of different estimation methods and on their performance. An application on computer virus is proposed. The second part of this thesis gives an estimation method of the hepatitis C virus prevalence in France based on a back-calculation model in association with a Markov model of the disease's natural history. This method and its results are compared with those generated by the reference approach in France. The last part is focused on the study of the recent anti-hepatitis C therapeutics impact on the population since is has been stated that those could eradicate the virus at middle term. In that optic, based on published migration data and assuming that the population of interest is organized into a set of specific contact networks, a metapopulation is computed in which the dynamics of each sub-population is governed by a set of deterministic differential equations. This doctoral research has been conducted through a CIFRE industrial research agreement with the Bristol-Myers Squibb pharmaceutical company.

Keywords: modeling, compartmental models, statistics, hepatitis C, dynamic models, epidemiology.

Contents

Remerciements III

Abstract IV

List of Tables VII

List of Figures VIII

Introduction 1

1 Preliminaries 4
- 1.1 Theoretical reminders . . . 5
  - 1.1.1 Least squares estimation . . . 5
  - 1.1.2 Maximum likelihood estimation . . . 6
  - 1.1.3 Kaplan-Meier estimation . . . 6
  - 1.1.4 Expectation-maximization algorithm . . . 7
  - 1.1.5 Random graph theory . . . 8
- 1.2 Review of some epidemic models . . . 12
  - 1.2.1 Deterministic epidemic model . . . 12
  - 1.2.2 Stochastic epidemic models . . . 13
  - 1.2.3 Network-based epidemic models . . . 17
  - 1.2.4 Conclusion . . . 19

2 Some estimation problems in epidemic modeling 21
- Résumé . . . 22
- 2.1 Introduction . . . 23
- 2.2 General theory . . . 23
  - 2.2.1 Epidemic model background . . . 23
  - 2.2.2 Inference based on an initial estimation of $I(\cdot)$ . . . 24
  - 2.2.3 Inference based on an initial estimation of the cumulative distribution function of the time until infection . . . 25
- 2.3 Simulation study of the Euclidian parameters . . . 27
  - 2.3.1 Observation of a single population . . . 28
  - 2.3.2 Observation of several independent populations . . . 33
- 2.4 Simulation study of the estimators of the function $I(\cdot)$ . . . 39
- 2.5 Applications on real data sets . . . 43
  - 2.5.1 Application to the Code Red v2 worm propagation . . . 43
  - 2.5.2 Application to the human immunodeficiency virus (HIV) transmission . . . 44
- 2.6 Conclusions . . . 45

3 Some models of back-calculation for the hepatitis C virus infection incidence in France 46
- Résumé . . . 47
- 3.1 Introduction . . . 47
- 3.2 Models and estimation methods . . . 48
  - 3.2.1 Model . . . 48
  - 3.2.2 Maximum likelihood estimation . . . 49
  - 3.2.3 Expectation maximization smoothing algorithm . . . 50
  - 3.2.4 Precision . . . 51

3.3 The CépiDc dataset of death by HCC in France between 1979 and 2012 . . . 52
3.3.1 Data . . . 52
3.3.2 Natural history model . . . 54
3.3.3 Lifetime distribution of an individual who develop and die of HCC . . . 56
3.3.4 Distribution of the time to death from HCC in the overall population . . . 57
3.3.5 Simplified Deuffic et al. model . . . 58
3.3.6 Quality of the models . . . 59
3.4 Results . . . 60
3.5 Discussion . . . 65

4 Network model for the propagation of hepatitis C virus in France 66
Résumé . . . 67
4.1 Introduction . . . 68
4.2 Epidemic models . . . 68
4.2.1 Initial model of the spread of hepatitis C virus . . . 68
4.2.2 Incorporating contact with HIV-infected individuals . . . 71
4.3 Epidemic propagation on a metapopulation model . . . 72
4.4 Simulations . . . 74
4.4.1 Parameter values . . . 74
4.4.2 Results on an abstract three-city network . . . 76
4.4.3 Results on the network of France's 100 largest urban areas . . . 79
4.5 Enhanced HCV model . . . 83
4.5.1 Model structure . . . 83
4.5.2 Parameter values . . . 88
4.5.3 Migration crisis . . . 93
4.5.4 Prevalence of HCV and initial conditions of the model . . . 93
4.6 Results . . . 94
4.7 Discussion . . . 98

Conclusions 99

Bibliography 102

Appendix A Differential equations 110

Appendix B The main French urban areas 111

Appendix C Differential equations - Enhanced model - Genotype 1 112

Appendix D Alternative model 117

List of Tables

(1.1) Main characteristics of the different epidemic models 20

(2.1) Values of the exponential distribution parameter 27
(2.2) Simulation results for the RCS model based on an initial estimation of $I(t)$ 28
(2.3) Simulation results for the NHRS model based on an initial estimation of $I(t)$ 28
(2.4) Simulation results for the RCS model based on an initial estimation of $F(t)$ 30
(2.5) Simulation results for the NHRS model based on an initial estimation of $F(t)$ 30
(2.6) Simulation results for the RCS model with n independent populations (MSE) 33
(2.7) Simulation results for the NHRS model with n independent populations (MSE) 34
(2.8) Simulation results for the RCS model with n independent populations (MLE) 36
(2.9) Simulation results for the NHRS model with n independent populations (MLE) 36
(2.10) Comparaison of the simulation results for the RCS model 37
(2.11) Comparaison of the simulation results for the NHRS model 38
(2.12) Mean integrated squared error results for the RCS model 42
(2.13) Mean integrated squared error results for the NHRS model 42
(2.14) Estimation results for the RCS and the NHRS model applied to the CRv2 dataset 43
(2.15) Estimation results for the RCS and the NHRS model applied to the HIV dataset 45

(3.1) Causes of death extracted from CépiDc 52
(3.2) Annual transition probabilities of the natural history model 56
(3.3) Scale coefficients estimation 60
(3.4) Parameter estimation and confidence interval for the different models 62
(3.5) Mean squared error result for the different models 63
(3.6) Estimation and 95% confidence interval of total HCV chronic carriers in 2004 63

(4.1) Annual origin-destination matrix 75
(4.2) Parameters used in the model to simulate the spread of HCV 76
(4.3) Metavir fibrosis scoring 83
(4.4) Annual rates used in the model to simulate the spread of HCV 90
(4.5) Sources of the annual rates 90
(4.6) Annual rates used in the model to simulate the spread of HCV (by genotype) 91
(4.7) HCV and HIV prevalence in IDU population of some urban areas in 2011 93
(4.8) HCV prevalence by health state in France for the year 2013 94

(B.1) Population of the French urban areas based on census data for the year 2012 111

List of Figures

(1.1) Example of graph . . . 8
(1.2) Graph with a complete subgraph . . . 10
(1.3) Graphs generated with a Watts and Strogatz model . . . 11
(1.4) Diagram of the SIR model dynamics . . . 12
(1.5) SIR Markov chain evaluation at times t, $t + \Delta t$ and $t + 2\Delta t$ . . . 14

(2.1) Expected number of hosts infected against time under RCS and NHRS models . . . 27
(2.2) Simulation results under the RCS model based on an initial estimation of $I(t)$ . . . 29
(2.3) Simulation results under the NHRS model based on an initial estimation of $I(t)$ . . . 29
(2.4) Simulation results under the RCS model based on an initial estimation of $F(t)$ . . . 31
(2.5) Simulation results of β under the NHRS model based on an initial estimation of $F(t)$ . . . 32
(2.6) Simulation results of K under the NHRS model based on an initial estimation of $F(t)$ . . . 32
(2.7) Simulation results for different amount of censoring and different values of n (MSE) . . . 35
(2.8) Simulation results for different amount of censoring and different values of n (MLE) . . . 37
(2.9) Simulation of the RCS and the NHRS models for $\hat{I}(t) = X(t)$ . . . 39
(2.10) Simulation of the RCS and the NHRS models for $\hat{I}(t) = \bar{X}(t)$ . . . 39
(2.11) Simulation results of the Kaplan Meier estimator for the RCS and the NHRS models . . . 40
(2.12) Cumulative number of unique IP addresses infected by the Code Red v2 worm . . . 43
(2.13) Code Red v2 worm propagation dataset . . . 44
(2.14) Cumulative number of AIDS cases in France . . . 44
(2.15) Reported AIDS cases dataset . . . 45

(3.1) Number of reported deaths due to a liver tumor in France extracted from CépiDC . . . 53
(3.2) Adjusted number of HCC deaths related to HCV in France . . . 54
(3.3) Markov model of the natural history of HCV . . . 54
(3.4) Simplified Markov model of the natural history of HCV . . . 55
(3.5) Lifetime distribution . . . 57
(3.6) Simplified Markov model of the natural history of HCV with extra state OCD . . . 58
(3.7) Estimation of the expected number of HCV infections . . . 61
(3.8) Estimation of the expected number of deaths from HCC related to HCV . . . 62
(3.9) Estimation and 95% confidence interval of total HCV chronic carriers in 2004 . . . 63
(3.10) Estimated prevalence of total chronic HCV in France . . . 64

(4.1) Compartmental model of HCV's natural history . . . 69
(4.2) Network of PWID infected by HCV and their sharing partners . . . 70
(4.3) Compartmental model of HCV's natural history with HIV coinfection . . . 71
(4.4) Networks of A and B with additional nodes representing people from A in B and vice versa 73
(4.5) Network of A with the additional nodes corresponding to cities B and C . . . 74
(4.6) Degree distribution of a scale-free network constituted of 10,000 nodes . . . 75
(4.7) Mean evolution of the susceptible population in the whole system . . . 77
(4.8) Mean evolution of the HCV mono-infected population in the whole system . . . 78
(4.9) Mean evolution of the HIV/HCV co-infected population in the whole system . . . 78
(4.10) Mean evolution of the population dynamics in France . . . 80
(4.11) Evolution of the spread of HCV in France with an epidemic start in Paris . . . 81
(4.12) Evolution of the spread of HCV in France with an epidemic start in Forbach . . . 82
(4.13) Compartmental model of fibrosis progression and end-stage liver diseases . . . 83
(4.14) Compartmental model of the infection by the different HCV genotypes . . . 84

(4.15) Compartmental model of fibrosis progression and end-stage liver diseases for patients infected by genotype 3 HCV 85
(4.16) Compartmental models of HCV treatment and liver transplant 86
(4.17) Dynamic model of patients infected by HCV (Genotype 1) 92
(4.18) Mean evolution of the HCV prevalence in France by genotype 95
(4.19) Mean evolution of the HCV prevalence in France for different proportion of treated patients 96
(4.20) Mean evolution of the HCV prevalence in France for different sharing frequencies 96
(4.21) Mean evolution of the HCV prevalence in France for different immigration rates 97

(D.1) Alternative model of the patients infected by HCV (Genotype 1) 117

Introduction

From time immemorial, mankind has been struck by a countless number of deadly infectious diseases devastating the population at the time. Indeed, history is littered with example of epidemics responsible for the annihilation of a non-negligible proportion of the civilization. For example, the Plague of Justinian resulted in the death of an estimated 25 to 50 million people between 541 and 542, the Black Death wiped out an estimated 75 to 200 million people between 1346 and 1353, the Spanish flu killed an estimated 50 to 100 million people between 1918 and 1920 and, more recently, it has been estimated that the human immunodeficiency virus (HIV) epidemic has contributed to the death of about 34 million people since 1981.

To better understand the mechanisms of epidemics, such as vector identification or transmission rate estimation, and to anticipate their evolution, epidemiologists have constantly called upon mathematical modeling, improving the methodology at the same time. Hence, epidemic mathematical models have become the main tool for describing the spread of communicable disease through populations and for evaluating strategies to control the epidemic.

Though Daniel Bernoulli was considered the first to provide an epidemic model for the spread of smallpox in 1766, most of the existing work in this area was conduct during the past century. Indeed, one of the most significant breakthrough into epidemic modeling is the introduction of compartmental models in the early 1900s by Kermack and McKendrick (1927). Compartmental models assume that the population of interest can be stratified into several compartments, typically infected and non-infected, interacting with each other over time. The most well-known compartmental model is the SI model where people can enter two successive health sates: susceptible (usually denoted by the capital letter S) which corresponds to the people susceptible to the infection and infected (denoted by the capital letter I) which represents the infected population. Several adaptations of the SI model such as SIR (where R stands for recovery) or SEIR (where E stands for exposed) have been developed ever since including more complex dynamics.

Widely used for the spread of biological pathogens, epidemic models have also been used to describe the propagation of malicious software on Internet. In 2002, Zou et al. (2002) proposed an adaptation of the classic compartmental model developed by Kermack and McKendrick (1927) to the outbreak of version 2 of the Code Red worm (CRv2) on July 19, 2001.

Going back to biological viruses, the main subject of this doctoral thesis focuses on the spread of hepatitis C virus (HCV). Hepatitis C virus is a blood-borne virus which slowly destruct the liver in a process called fibrosis. With an estimated number of chronic carriers reaching 80 million worldwide, the management of HCV has become a major challenge since the end of the 1900s. In France and in all other developed countries, the population at risk is injecting drug users (IDU). For European countries, the proportion of new HCV infections in the population of IDU has been estimated to 76.5% in 2012 (see European Centre for Disease Prevention and Control (2012)).

With the arrival of new therapeutics (second-generation direct-acting antivirals or 2^{nd}-gen DAAs) on the market, it is expected that the prevalence of HCV should significantly decrease since those new therapeutics admit very high sustained virology response (SVR) rates whatever the genotype of the disease or the patient health state. With an estimated infected population of about 200,000 individuals in France, it is highly necessary to measure the impact of the 2^{nd}-gen DAAs on the infected population and to answer the question: "Are we able to definitively eradicate hepatitis C virus and, if so, when?".

This doctoral thesis, conducted through a CIFRE industrial research agreement with the Bristol-Myers Squibb pharmaceutical company, proposes to explore three different problematics, each one dealing with a specific epidemiology-related domain. The problematics considered in this work are not necessarily directly connected with each other but are closely related to the mathematical modeling of epidemics. Those three parts are preceded by some reminders of the different mathematical tools used in the thesis

(models, estimation methods, etc) and a review of different epidemic models based on the compartmental SIR framework originally introduced by Kermack and McKendrick (1927). Starting with the deterministic compartmental model, we review its stochastic counterpart as well as Markov models. We finally review some extensions of the deterministic compartmental model on contact networks.

The second part deals with the propagation of viruses by using well-known epidemic models. Based on a logistic model, Staniford et al. (2002) introduced the random constant scanning (RCS) model to describe the propagation of Internet worms assuming a scanning rate (worm's contact rate) constant over time. Few years later, Kirmani and Hood (2010) proposed to remove this limitation and develop a more realistic model by introducing the nonhomogeneous random scanning (NHRS) model where the expected number of hosts infected during [0,t] is given by:

$$I(t) = \frac{N}{1 + \psi \exp\left(-\int_0^t \beta(u)\,\mathrm{d}u\right)}$$

with N being the size of the population considered and β the scanning rate. Similarly, Kirmani and Hood (2010) offered an analytic expression for the cumulative distribution function (cdf) of the time until infection of an initially uninfected host (T) for their NHRS model. The objective of this work is to tackle some problems of estimation of $\beta(t)$ and the cdf of T by studying various estimation methods, such as the least squares, maximum likelihood, and Kaplan-Meier methods and analyzing their performance. An application on the datasets of the outbreak of version 2 of the Code Red worm (CRv2) and of the outbreak of HIV in France since 1982 is proposed. The work undertaken in this part is currently under minor revision for Communications in Statistics - Simulation and Computation.

The third part of the thesis is focused on the estimation of the prevalence of HCV in France. As new therapeutics emerge, establishing successful strategies to eliminate the virus requires an accurate estimation of the size of the infected population. Hence, we propose to estimate this prevalence by first estimating the past incidence of the disease in France. To do this, Deuffic-Burban et al. (1999) used a back-calculation method based on the weighted least squares method and associated with a Markov model of the diseases natural history. Using a dataset of the annual number of death due to hepatocellular carcinoma (HCC), a direct consequence of HCV, Deuffic-Burban et al. (1999) proposed a parametric model for the past incidence of HCV such as:

$$f(t;\alpha,\beta,\gamma) = \begin{cases} \dfrac{\exp(a(t-t_0))}{b + \beta^{-1}\exp(a(t-t_0))}, & \text{for } t < 1990 \\ 0.6 \times \dfrac{\exp(a(t-t_0))}{b + \beta^{-1}\exp(a(t-t_0))}, & \text{otherwise} \end{cases}$$

with $a = 4 \times \alpha \times \beta^{-1}$ and $b = \beta^{-1} \times \exp(a(\gamma - t_0))$ where α is the slope of the curve, β the asymptotic plateau value and γ the year of mid-epidemic. In 2014 and for the particular case of Taiwan, McEwan et al. (2014) proposed to apply the back-calculation approach used for the HIV past incidence reconstruction introduced by Becker and Marschner (1993). This approach, based on the expectation-maximization-smoothing (EMS) algorithm developed by Silverman and Nychka (1990), assumes that all infections are independent random variables following Poisson distributions. The main goal of our work is to propose an adaptation of the approach developed by McEwan et al. (2014) to the French national data of the observed number of deaths due to HCC. Along with this, we introduce some parametric models based on well-known epidemic models from the literature to give an alternative approach of the work of McEwan et al. (2014). After a large work of formatting the data, to ensure that we only focus on HCC deaths related to HCV, comparisons are made with the results generated by the reference approach in France. The work undertaken in this part has been submitted to Statistics in Medicine.

The fourth part of this thesis is focused on the construction of a deterministic compartmental model of the spread of HCV in France. The main objective of this work is to estimate the impact of the new therapeutics anti-HCV on the whole infected population of France. To do this, we distinguish two parts in this work. In the first part, we consider the transmission of HCV and its natural history in a simple compartmental model. A particular attention is however given on HCV/HIV co-infected individual. Then, we subdivide the whole country into several sub-populations, corresponding to the first hundred largest French urban areas, and assume that the sub-populations interact with each other via an original metapopulation model and French migration data. As IDU represent the population where

the large majority of new HCV infections occur, we focus our model on the IDU population. To take into account the heterogeneity in contact pattern of the IDU population, each sub-population is assumed to be organized according to a contact network. First conclusions are drawn from the results obtained with this simple model, first on a toy-country with just three cities and then on the network of France's 100 largest urban areas.

The second part of this work allows for an enhancement of the previous compartmental model by adding new features such as the natural history of HCV (fibrosis stages, decompensated cirrhosis, etc), the main genotypes of the disease (stratified between genotype 3 or non genotype 3) and the possibility for an individual to cess the use of injecting drugs. Last, to track the evolution of people leaving the IDU population after cessation, an alternative compartmental model is computed. Assuming those people does not participate in the force of infection anymore, no transmission rate is considered in the alternative model.

Last, we want to emphasize the fact that only the work related to the doctoral thesis is presented in the following chapters. Indeed, those last three years also provided several opportunities to apply statistical modeling in various contexts, such as clinical trials (survival analysis) or health economics (Markov modeling), which are not reported here.

1 Preliminaries

Français

Ce chapitre a pour but, d'une part, de donner quelques rappels théoriques sur les concepts mathématiques et les méthodologies employées dans cette thèse de doctorat et, d'autre part, d'introduire quelques modèles épidémiques. Concernant ces dits modèles, nous nous focalisons en particulier sur la structure épidémique Susceptible-Infected-Recovered (SIR) afin de présenter quelques modèles bien connus et d'en introduire des plus complexes. Dans le but de mettre en lumière les similitudes structurelles de ces modèles, nous en présentons, de manière informelle, les équations de dynamique des populations.

English

This chapter is meant to give some theoretical reminders on the mathematical concepts and methodologies addressed in this doctoral thesis on the one hand, and to introduce several epidemic models on the other. Regarding those epidemics models, we focused on the particular Susceptible-Infected-Recovered (SIR) disease spreading framework to give an overview of some well-known models and to introduce more complex ones. To highlight the structural similarities of those models, we introduce, in an informal manner, the equations governing the population dynamics.

Contents

1.1 Theoretical reminders **5**
1.1.1 Least squares estimation 5
1.1.2 Maximum likelihood estimation 6
1.1.3 Kaplan-Meier estimation 6
1.1.4 Expectation-maximization algorithm 7
1.1.5 Random graph theory 8
1.2 Review of some epidemic models **12**
1.2.1 Deterministic epidemic model 12
1.2.2 Stochastic epidemic models 13
1.2.3 Network-based epidemic models 17
1.2.4 Conclusion 19

Résumé

Ce premier chapitre offre une présentation des modèles et concepts abordés tout au long de ce travail de thèse. L'essentiel des travaux de recherche qui ont été entrepris dans cette thèse se sont orientés dans trois directions différentes et ont ainsi fait appel à des méthodologies variées. Aussi nous nous proposons dans cette partie de donner une description sommaire des méthodes utilisées.

Ce chapitre est divisé en deux parties. La première partie se focalise sur la description de quelques rappels théoriques employés tout au long de ces travaux. Nous présentons ainsi brièvement les différentes méthodes paramétriques ou non-paramétriques statistiques utilisées dans le Chapitre 2. à savoir la méthode des moindres carrés, la méthode du maximum de vraisemblance et celle de Kaplan-Meier.

Nous décrivons ensuite l'algorithme d'espérance maximisation (EM) et son extension l'algorithme d'espérance maximisation lissé (EMS pour expectation maximization smoothing) principalement employés au Chapitre 3 pour l'estimation de l'incidence passée du virus de l'hépatite C.

Enfin, nous donnons ensuite une rapide description de la théorie des graphes aléatoires et de son application en épidémiologie. Après en avoir présenté quelques-unes des principales propriétés, plusieurs modèles de graphes célèbres sont décrits. On retrouve en particulier les modèles d'Erdős-Rényi, de Watts et Strogatz et de Barabási-Albert.

La seconde partie de ce chapitre offre une description sommaire des modèles compartimentaux utilisés en épidémiologie. En se basant sur le schéma infectieux du modèle à trois compartiments SIR (susceptible, infected, recovered), nous donnons quelques éléments de théorie de ces modèles et présentons les équations qui régissent les dynamiques de population. Nous introduisons ainsi le modèle classique déterministe et sa contrepartie stochastique. Nous introduisons également les modèles épidémiques Markoviens en temps discret et en temps continu.

Enfin, nous traitons de l'extension du modèle déterministe classique sur un graphe aléatoire. Concrètement, il s'agit d'un graphe aléatoire généré via une distribution de probabilité sur lequel sont appliquées les équations différentielles du modèle déterministe classique. La principale différence avec le modèle classique est que les individus ne peuvent plus entrer en contact, et donc infecter, n'importe quel autre individu. Chaque individu est limité à son nombre de voisins (ou degré). Dans cette partie nous faisons la distinction entre l'approche homogène où il est supposé que tous les individus possèdent le même nombre de voisins et l'approche hétérogène où cette dernière hypothèse est relaxée. Cette dernière approche est utilisée au Chapitre 4.

Cette seconde partie se conclue sur un tableau comparatif reprenant les caractéristiques principales des différents modèles épidémiques.

1.1 Theoretical reminders

In this section we introduce some fundamental information that will be used as reference for the different works undertaken in this doctoral thesis. We successively give an overview of several estimation methods, of the random graph theory and present some well-known contact networks.

1.1.1 Least squares estimation

The method of least squares, first published by Adrien-Marie Legendre in 1805 (see Legendre (1805)) but credited to Carl Friedrich Gauss in 1795, is a well-known estimation method. This method consists in estimating the parameters of a mathematical model by minimizing the sum of the squares of the errors between the observed data and the data predicted by the model. In other terms, we look for the value of the parameter which describes at best the data.

For illustrative purposes, we propose to introduce the least squares method in the context of simple linear regression where a response variable Y is explained by a covariable X independent of Y via a linear relationship. Let us denote by $(x_1, y_1), \ldots, (x_n, y_n)$ an observed sampling of n covariables and their responses and by $f(x, \beta)$ a parametric model, also called regression function, such as:

$$y_i = f(x_i, \beta) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ for } i = 1, \ldots, n \tag{1.1}$$

where β_0 and β_1 are the parameters of the model and ε_i a disturbance term. The least squares estimator of β is the values of β_0 and β_1 which minimize the quantity:

$$S = \sum_{i=1}^{n} \big(y_i - f(x_i, \beta)\big)^2 = \sum_{i=1}^{n} \varepsilon_i^2. \tag{1.2}$$

To do this, one must respectively look for the values of β_0 and β_1 which nullify the partial derivatives of (1.2) with respect to β_0 and to β_1.

The least squares method is probably the most popular estimation method in statistics and has been used in various fields for different kinds of problems in terms of complexity. For example, when dealing with linear regression, one might consider the fact that the response Y is explained by more than one covariable. In such a case, the least squares method can be extended to the general linear model and, for p covariables, one can rewrite (1.1) such as:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \text{ for } i = 1, \ldots, n.$$

That estimation method can also be extended when dealing with a non-linear relationship between the response and the covariables. In that case, the estimation process is far more complex and is usually solved via iterative refinement.

1.1.2 Maximum likelihood estimation

Another widely used estimation method addressed in this section is the maximum likelihood estimation. First introduced by Ronald Fisher between 1912 and 1922 (see Fisher (1912, 1922)), even if it has been used earlier, amongst others, by Carl Friedrich Gauss and Pierre-Simon Laplace, the maximum likelihood estimation is a parametric estimation method which basically consists in maximizing the likelihood function of a model, i.e the agreement between the model with the data.

First, let us introduce the concept of likelihood. The likelihood function $L(\theta; x)$ is a function of the parameter θ of a model given the data x that measures the occurrence plausibility of x for each possible value of θ. This function can be written such as:

$$L(\theta; x) = \mathbb{P}(x|\theta).$$

For a set of n independent and identically distributed continuous (resp. discrete) random variables $X_1, \ldots, X_n$ which joint probability density function (resp. probability mass function) is $f(X_1, \ldots, X_n|\theta)$ with an unknown value of θ, the likelihood function is:

$$L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta) \tag{1.3}$$

where $x_i, i = 1, \ldots, n$, is an observation of the random variable X_i. Then, based on (1.3), we define the maximum likelihood estimator of θ such as:

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i|\theta).$$

In practice, it is even more convenient to maximize the log-likelihood function since log is a monotonic increasing function. Thus, the maximum likelihood estimator of θ can be defined such as:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \ln\left(f(x_i|\theta)\right).$$

1.1.3 Kaplan-Meier estimation

The Kaplan-Meier estimation, introduced by Edward L. Kaplan and Paul Meier in 1958 (see Kaplan and Meier (1958)), is a non-parametric estimation method used in survival analysis. Widely used in medical research, the Kaplan-Meier estimator, also referred to as product-limit estimator, $\hat{S}(t)$ offers an estimation of the quantity $S(t)$ which is the probability that a given individual has lifetime exceeding time t. However, the Kaplan-Meier estimator is not limited to medical research and has also great applications in reliability engineering where this estimator is used to measure the time-to-failure of machine parts.

For illustrative purposes, let us go back to the medical field and assume that we follow the times until death of a population of size N. We denote by t_i the occurrence time of the i^{th} death such as $t_1 \leqslant \cdots \leqslant t_N$. Similarly, we denote by n_i the number of individuals still alive prior to time t_i and d_i the number of individuals dead at t_i. Then, the non-parametric estimator of $S(t)$ is defined as:

$$\hat{S}(t) = \prod_{t_i \leqslant t} \frac{n_i - d_i}{n_i}.$$

A great advantage of the Kaplan-Meier estimator is the fact that it can take into account censored data and deal with incomplete observation, as is often the case in a real-life setting. In such a case, the number n_i of individuals at risk at t_i equals to n_{i-1} minus the number of individuals who died and those lost to follow-up between t_{i-1} and t_i.

Being a non-parametric estimation method, the Kaplan-Meier estimation presents some advantages as well as disadvantages. Indeed, since there is no need to assume that the life data of interest follows a specific probability distribution, we avoid the potential bias brought by this assumption. However, the confidence interval associated to such an estimation method tends to be wider than those obtained with the parametric estimation. For the estimation of confidence intervals in the non-parametric analysis, one can refer to the formula introduced by Greenwood in 1926 (see Greenwood (1926)) which gives an estimation of the variance of Kaplan-Meier estimator such as:

$$\widehat{\text{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i \leqslant t} \frac{d_i}{n_i(n_i - d_i)}.$$

1.1.4 Expectation-maximization algorithm

Let us now give an overview of the expectation-maximization (EM) algorithm. Introduced by Dempster et al. (1977), the EM algorithm is an iterative parametric estimation method, based on the maximum likelihood estimation technic, which estimates the parameters of a statistical model depending on both observed and unobserved variables. Each iteration of this algorithm is based on two steps. The first step (E step) consists in estimating the unobserved data given the observed data and the value of the parameters estimated during the previous iteration of the algorithm. Such a step is achieved by computing the conditional expectation of the log-likelihood function. The second step (M step) consists in estimating and updating the value of the parameters by maximizing the expectation obtained during the E step. The EM algorithm is frequently used for data imputation, data clustering or image reconstruction.

As an example, let us consider a set X of observed variables and a set Z of unobserved variables which realizations are denoted with lowercase letters. For more convenience, we assume that Z is a discrete variable (similar argument holds for continuous variables). By denoting θ the vector of the parameters of the model, we wish to estimate θ such that the marginal likelihood $L(\theta; x)$ is maximum. We also denote by $p(x, z|\theta)$ the joint probability distribution of X and Z knowing θ such as $p(x|\theta)$ is the marginal probability distribution of X knowing θ. By considering the log-likelihood, one can write:

$$l(\theta; x) = \ln\left(L(\theta; x)\right) = \ln\left(p(x|\theta)\right) = \ln\left(\sum_z p(x, z|\theta)\right) \tag{1.4}$$

for which the use of the maximum likelihood estimation method (see Section 1.1.2) to estimate θ can be quite impractical. Hence, by taking an arbitrary distribution $q(z)$ for the unobserved variables Z, one can rewrite (1.4) such as:

$$\begin{aligned}
l(\theta; x) &= \ln\left(\sum_z q(z) \cdot \frac{p(x, z|\theta)}{q(z)}\right) && \text{(1.5a)}\\
&\geqslant \sum_z q(z) \cdot \ln\left(\frac{p(x, z|\theta)}{q(z)}\right) && \text{(1.5b)}\\
&= \sum_z q(z) \cdot \ln\left(p(x, z|\theta)\right) - \sum_z q(z) \cdot \ln\left(q(z)\right) && \text{(1.5c)}\\
&= \mathbb{E}_{q(z)}\big(\ln\left(p(x, z|\theta)\right)\big) - \sum_z q(z) \cdot \ln\left(q(z)\right) && \text{(1.5d)}
\end{aligned}$$

where the passage from (1.5a) to (1.5b) is assured by the Jensen's inequality (see Jensen (1906)). Equation (1.5d) is said to be a lower bound of the marginal log-likelihood $l(\theta; x)$. Thus, by maximizing (1.5d), one should maximize $l(\theta; x)$ at the same time.

However, to ensure that the $l(\theta; x)$ increases when (1.5d) is maximized, we need to find the optimal lower bound of $l(\theta; x)$, i.e find a value for $q(\cdot)$ such that (1.5d) equals $l(\theta; x)$ for a given $\theta^{(0)}$. One can show that this optimal lower bound is reached when $q(z) = p(z|x; \theta^{(0)})$. Thus, one can write:

$$l(\theta; x) \quad = \quad \mathbb{E}_{p(z|x;\theta^{(0)})}\big(\ln\left(p(x, z|\theta)\right)\big) - \sum_z p(z|x; \theta^{(0)}) \cdot \ln\Big(p(z|x; \theta^{(0)})\Big). \tag{1.6}$$

Since the second term of the right-hand member of (1.6) does not depend on θ, the M step of the EM algorithm simply consists in maximizing $\mathbb{E}_{p(z|x;\theta^{(0)})}\big(\ln\left(p(x, z|\theta)\right)\big)$, i.e we have:

$$\hat{\theta}^{(1)} = \arg\max_{\theta} \mathbb{E}_{p(z|x;\theta^{(0)})}\big(\ln\left(p(x, z|\theta)\right)\big).$$

Each estimation $\hat{\theta}^{(t)}$ is then reused in a new iteration to compute $\hat{\theta}^{t+1}$ until convergence of the algorithm.

In the context of image reconstruction, Silverman and Nychka (1990) proposed a variant of the EM algorithm called expectation-maximization-smoothing (EMS) algorithm. This algorithm is based on the standard EM algorithm but add a smoothing step after each M step of the EM algorithm. Indeed, this extra step, meant to limit noisy estimations, consists in recalculating the estimator by considering its neighbors affected by different weighs. In their work, Silverman and Nychka (1990) used a weighted moving average such as:

$$\hat{\theta}_s^{(t)} = \sum_{l=0}^{L} \omega_l \times \tilde{\theta}_{s+l-L/2}^{(t)}, \quad s = 1, \ldots, S$$

where $\tilde{\theta}^{(t)}$ is the estimation of θ obtained from the M step at the t^{th} iteration, L is the order of the moving average and the ω_l are symmetric weights such as $\sum_l \omega_l = 1$. The following convergence criteria is defined:

$$\frac{\left\|\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}\right\|_2}{\left\|\hat{\theta}^{(t)}\right\|_2} < \varepsilon$$

where ε can be chosen sufficiently small and $\|\cdot\|_2$ is the L^2 norm in $\mathbb{R}^S$.

1.1.5 Random graph theory

This section is meant to give some basic knowledge on the random graph theory. First let us introduce the principle of graph. A graph is a mathematical structure describing relations between objects. Denoted by $\mathbb{G}(V, E)$, a graph is constituted of a set V of points, also called nodes or vertices, connected with each other by E edges (or arcs). One can refer to Figure 1.1 for an example of graph where V and E are:

$$V = \{V_1, V_2, V_3, V_4, V_5, V_6\},$$

$$E = \{\{V_1, V_3\}, \{V_3, V_5\}, \{V_2, V_3\}, \{V_2, V_4\}, \{V_2, V_6\}, \{V_4, V_6\}\}.$$

In particular, the graph of Figure 1.1 is said to be undirected since no distinction is made between two nodes connected with a given edge, i.e no orientation is given for the different edges.

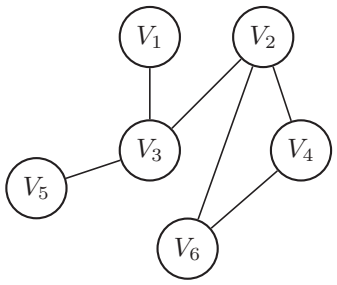


Figure 1.1: Example of graph.

Let us now consider random graphs. A random graph, denoted $\mathbb{G}(n, p)$, is defined as a set of n nodes connected which other by edges generated via a probability distribution p. Random graphs have been increasingly used to model complex networks structures describing natural and/or societal relationships between individuals (human beings, animals, computers, etc) and to understand the mechanisms of various phenomena in the populations considered (the spread of a disease through a population, the spread of rumors through a social network, etc).

As the mechanics of those phenomena can highly differ, different network structures arose in the literature. Indeed, depending on the probability distribution considered, several authors proposed different ways to generate contact networks. Among the most famous contact networks, one can refer to the Erdős-Rényi model, the Watts and Strogatz model and the Barabási-Albert model.

1.1.5.1 Some basic concepts

Let us introduce some basic concepts related to the graph theory. The following concepts do not represent an exhaustive list of the main characteristics of the graph theory but are needed for a better understanding of the models introduced thereafter. For more in-depth details on the graphe theory, one can refer to Bondy and Murty (1976).

One of the main fundamental characteristics of graphs is the degree assigned to the nodes. This notion defines the number of edges incident to a given node. Basically, it corresponds to the number of neighbors (or partners) of a given node. By referring to Figure 1.1, one can see that the node V_1 has a degree of 1 while V_2 has a degree of 3. In the particular case of random graphs, the network is generated via a probability distribution since each node is assigned a degree according to a probability.

The next concept introduced in this section is the concept of path. A path, sometimes referred to as walk, is a finite or infinite alternative sequence of nodes and edges between two given nodes. Infinite path is an extension of finite path where there is neither a starting nor an ending node. In all the following, we only consider finite paths. Denoted by W, the walk between two given nodes V_0 and V_k is expressed such as:

$$W = V_1 E_{1,2} V_2 \cdots V_{k-1} E_{k-1,k} V_k$$

where $E_{1,2}$ is the edge between nodes V_1 and V_2. Particularly, W is said to be of length k. One can refer to Figure 1.1 for an illustration of a path of length 2 between V_2 and V_5. One can also observe two different paths (length 1 and 2) between V_2 and V_6.

This simple example leads to a recurrent problematic in graph theory which is the search for the shortest path. Assuming that the graph of Figure 1.1 is unweighted, i.e no weight is associated to the different edges, the shortest path between V_2 and V_6 is $W = V_2 E_{2,6} V_6$ which is the path with the fewest edges. For weighted graphs, the search for the shortest path between two nodes uses far more complex algorithms to minimize the sum of weights for all the paths between these two nodes.

The last notion addressed in this section is the notion of clustering coefficient. Being a measure of how much the nodes are gathered, it is an important parameter to study neighborhood in social networks and is widely used when dealing with small world networks (see Section 1.1.5.3). However, one can discern two notions of clustering coefficient, the global cluster coefficient associated to the graph and the local cluster coefficient associated to a node. The global cluster coefficient only considers the number of triplets in the graph which is the number of three connected nodes (a node with two partners). This definition differs from the definition of a triangle which is a closed triplets. As an example, and by referring to Figure 1.1, one can see that the nodes V_2, V_4 and V_6 form a triangle. This coefficient is calculated with the following formula:

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of vertices}}.$$

With 1 triangle and 8 triplets, the global cluster coefficient of the graph displayed on Figure 1.1 is 0.375.

As for the local cluster coefficient, this quantity gives an indication on the closeness of a node's neighbors. Basically, this coefficient measures how close the neighbors of a given node are to being a clique, i.e how close the neighbors of a given node are to forming a complete subgraph. By referring to Figure 1.2, one can notice that the subgraph formed by the neighbors of V_2 is complete (each node is connected with each other). For a given node u, the local cluster coefficient is:

$$C_u = \frac{2E_u}{k_u(k_u - 1)}$$

where k_u is the degree of u and E_u is the number of edges between the k_u partners connected to u. From Figure 1.2, one can see that V_2 as 3 partners and that the number of edges between those is also 3 (the blue highlighted edges). Thus, the local cluster coefficient of V_2 is 1. By now referring to Figure 1.1, we find a local cluster coefficient for V_2 of 1/3.

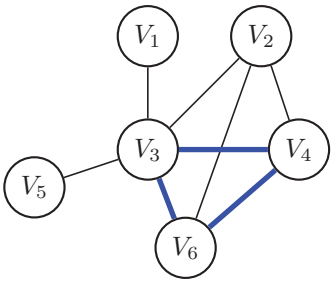


Figure 1.2: Graph with a complete subgraph. The clique formed by the neighbors of V_2 is highlighted by a set of blue thick edges.

1.1.5.2 Erdős-Rényi model

The first model addressed in this section is the Erdős-Rényi model introduced by Paul Erdős and Alfréd Rényi in 1959 (see Erdős and Rényi (1959)). The main assumption of this model is the fact that each node can independently be connected to each other with a probability p, or in other words each edge has an independent probability p of existing.

From this assumption, the probability $p(k)$ that a given node, in a graph generated by the Erdős-Rényi model $\mathbb{G}(n, p)$, has exactly k partners is given by:

$$p(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

which is the binomial distribution function. Indeed, each node in the graph has exactly C_{n-1}^k ways to choose k edges from among the $n-1$ other nodes, the k edges have a probability p^k of existing and no connection is allowed with $n-1-k$ other nodes.

By now referring to the properties of a binomial distribution, when n goes to infinity, and for a very small p, the binomial distribution converges towards the Poisson distribution:

$$\frac{\langle K \rangle^k e^{-\langle K \rangle}}{k!}$$

where $\langle K \rangle$ is the mean degree of the graph such as $\langle K \rangle = (n-1)p$.

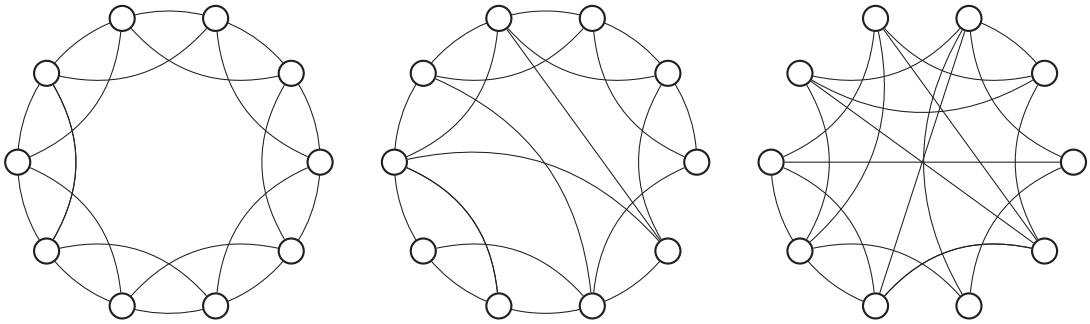
The Erdős-Rényi model is usually used for its relatively understanding simplicity but lacks of accuracy when trying to model real phenomena, especially when connections between nodes aren't independent which leads to a low clustering coefficient.

1.1.5.3 Watts and Strogatz model

Let us now focus on the model introduced by Duncan J. Watts and Steven Strogatz in 1998 (see Watts and Strogatz (1998)). Mainly developed to face the limitation of the Erdős-Rényi model regarding local clustering, the Watts and Strogatz model was built to generate graphs with the small-world properties such as a high clustering coefficient and short average path lengths.

However, the construction of such graphs is slightly more complex than graph generated by the Erdős-Rényi model. Indeed, the algorithm for small-world networks generation can be described in two steps. First, assuming that each node has an even and determinist mean degree $\langle K \rangle$, let us consider a regular ring lattice constituted of N nodes connected with their $\langle K \rangle/2$ nearest neighbors on both sides. For illustrative purposes, one can refer to Figure 1.3(a) for a regular ring lattice with 10 nodes connected with their 2 nearest neighbors on both sides. Next, each existing edge $E_{i,j}$ (between node n_i and n_j) is rewired with a probability β to be replaced by the edge $E_{i,k}$ (the node n_k is chosen with uniform probability among all other nodes) such as $k \neq i$ and $k \neq k'$ where $E_{i,k'}$ is an already existing edge. Each edge is rewired

just once in the process. Figures 1.3(a) to 1.3(c) give an illustration of this process for different values of β.



(a) $\beta = 0$ **(b)** $0 < \beta < 1$ **(c)** $\beta = 1$

Figure 1.3: Graphs generated with the Watts and Strogatz model for different values of β. Regular ring lattice generated via a Watts and Strogatz model with $\beta = 0$ (Figure (a)), small-world graph generated via a Watts and Strogatz model with $0 < \beta < 1$ (Figure (b)) and random graph generated via a Watts and Strogatz model with $\beta = 1$ (Figure (c)).

Considering the degree distribution of a graph generated by a Watts and Strogatz model, it corresponds to a Dirac delta function centered at $\langle K \rangle$ when $\beta = 0$ (regular ring lattice) and it follows a Poisson distribution when $\beta = 1$ (random graph). For the particular case where $0 < \beta < 1$, the probability distribution can be expressed as:

$$p(k) = \sum_{n=0}^{f(k,\langle K \rangle)} \binom{\langle K \rangle/2}{n} \cdot (1-\beta)^n \beta^{\langle K \rangle/2-n} \cdot \frac{(\beta\langle K \rangle/2)^{k-\langle K \rangle/2-n}}{(k-\langle K \rangle/2-n)!} \cdot e^{-\beta\langle K \rangle/2}$$

for $k \geqslant \langle K \rangle/2$ where $f(k, \langle K \rangle) = \min(k - \langle K \rangle/2, \langle K \rangle/2)$. One can refer to the work of Albert and Barabási (2002) for the full explanations of this result.

Contrary to the Erdős-Rényi model, the Watts and Strogatz model allows for a better description of real phenomena since the work of Milgram (1967) has highlighted the fact that the human society was organized according to a small-world type framework. However, this model still lacks of accuracy since the degree distribution is relatively homogeneous, i.e all nodes have approximatively the same degree. Moreover, this model does not allow the addition of new nodes over time, i.e the number of nodes in the graph is constant along the modeling process.

1.1.5.4 Barabási-Albert model

The last model addressed in this section is the model introduced by Albert-László Barabási and Réka Albert in 2002 (see Albert and Barabási (2002)). Developed to generate scale-free networks, the Barabási-Albert model allows for the inclusion of two valuable features which are the possibility for the network to grow over time and the consideration of preferential attachment, an important mechanism of scale-free networks. Preferential attachment, in the field of contact networks, is a process consisting in distributing new partnerships to nodes according to how many partnerships they already have, i.e an overly connected node (high degree) has better chance to gain new partners than nodes with a limited number of partners (low degree). Those graphs are the most widely used to model various kinds of interactions between individuals (disease spread, social networks, Internet, etc).

The construction of scale-free networks starts with a small set m_0 of connected nodes. At each time step, a node with m edges ($m \leqslant m_0$) is added to the graph and connects to m different nodes among the m_0 original nodes. Then, for each edge of the new node among the m possibilities, a partner is selected according to the probability:

$$p_i = \frac{k_i}{\sum_j k_j}$$

where k_i is the degree of the node i and $i \in \{1, \dots, m_0\}$. In other words, new nodes tend to connect with the nodes with the higher degree. As for the degree distribution, scale-free networks are characterized by a power law distribution such as:

$$p(k) \sim k^{-\gamma}$$

where $2 < \gamma < 3$ in most applications.

Compared to the previous models, the Barabási-Albert model has the advantage to deal with heterogeneous populations, constituted of a few nodes with a lot of connections (referred to as hubs) and of a majority of nodes which have far less connections. Moreover, this model can describe the evolution of a dynamic graph over time since the number of nodes in the network is not constant anymore. However, due to the influence of high-degree nodes in the creation of new connections, scale-free networks tend to admit lower a clustering coefficient than small-world networks.

1.2 Review of some epidemic models

In this section, we introduce some well-known epidemic models. By focusing on the Susceptible-Infected-Recovered (SIR) disease spreading framework, we give an analytical description of some frequently used transmission models from the classic deterministic approach to recent works on networks. For the sake of simplicity, we assume that the size of the population N stays constant over time and that there is no birth or death rate.

1.2.1 Deterministic epidemic model

Deterministic compartmental models, also referred to as mass action models, are the simplest epidemic models for the study of infectious diseases dynamics. Based on a set of ordinary differential equations (ODE), they capture the interactions between the susceptible and the infected populations with respect to the homogeneous mixing assumption. This assumption considers that individual heterogeneities related to social structures or geographic locations are negligible. Moreover, as individuals are randomly mixed, the probability of contact between two individuals directly depends on the proportion in the whole population of their own group (susceptible, infected or recovered). Mainly used when a large population is involved, deterministic compartmental models have been extensively used in the past century to describe the dynamics of various infectious diseases such as rubella, measles, the foot-and-mouth disease or the respiratory syncytial virus (see Anderson (1983), Babad et al. (1995), Ferguson et al. (2001) and Moore et al. (2014) respectively).

Hence, based on the work of Kermack and McKendrick (1927) (see also Anderson and May (1992)), one can write the dynamics of the deterministic compartmental model for an SIR-framework epidemic as:

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

where $S(t)$, $I(t)$ and $R(t)$ represent the number of individuals in each compartment at time t. Figure 1.4 shows the dynamics of the SIR model for a closed population.

Figure 1.4: Diagram of the SIR model dynamics

By assuming that all the susceptible individuals are in contact with all the infected individuals and that a susceptible individual becomes infected after a "successful" contact with an infected individual, the number of new infections per unit time is:

$$\beta I(t) \times S(t)$$

where β is the transmission rate of the disease per unit time. We deduce that:

$$S(t + \Delta t) = S(t) - \beta S(t) I(t) \times \Delta t.$$

By now assuming that an infected individual recovers from the infection at a rate γ per unit time, one can write:

$$R(t + \Delta t) = R(t) + \gamma I(t) \times \Delta t.$$

Hence, when $\Delta t \to 0$, we obtain:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{S(t + \Delta t) - S(t)}{\Delta t} = -\beta S(t) I(t)$$

and

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{R(t + \Delta t) - R(t)}{\Delta t} = \gamma I(t).$$

Since N is known and assumed constant, we deduce that:

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = -\frac{\mathrm{d}S(t)}{\mathrm{d}t} - \frac{\mathrm{d}R(t)}{\mathrm{d}t}.$$

It follows that:

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = \beta S(t) I(t) - \gamma I(t).$$

Thus, we obtain:

$$\begin{aligned} \frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\beta S(t) I(t), \\ \frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \beta S(t) I(t) - \gamma I(t), \\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \gamma I(t). \end{aligned} \tag{1.7}$$

The population dynamics of the deterministic compartmental model are entirely defined by (1.7). In the next section, we propose to study more complex methods by focusing on stochastic epidemic models to allow some variability in the population dynamics.

1.2.2 Stochastic epidemic models

First introduced by McKendrick (1926) (see also Bartlett (1949)), stochastic epidemic models are based on random processes. Rather used when dealing with small communities, stochastic models are mostly employed when variations have a strong impact on the population dynamics. In this section, we first consider epidemic Markov models based on either a discrete-time Markov chain (DTMC) or a continuous-time Markov chain (CTMC). Then, we study the stochastic counterpart of the model described in Section 1.2.1. Most of this section is based on the work of Allen (2008).

1.2.2.1 Discrete-time Markov chain approach

Models based on a discrete-time Markov chain can be described as a state-transition process with a time variable t assumed to be discrete. Those models have been used to model the transmission of various infectious diseases such as hepatitis A virus, measles or AIDS (see Geng et al. (1998), Bishai et al. (2012) and Ogunmola (2014) respectively).

In the particular case of an epidemic based on the SIR framework, the Markov chain is subject to the following constraint:

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

where $S(t)$ (resp. $I(t)$, $R(t)$) is a discrete random variable which represents the number of susceptible (resp. infected, recovered) individuals at time t. Since each random variable depends on the two others, the model's state is fully defined by $S(t)$ and $I(t)$. Therefore, we only need to define the joint probability of $(S(t), I(t))$:

$$p_{(s,i)}(t) = \mathbb{P}\big(S(t) = s, I(t) = i\big).$$

By dealing with a Markov chain and assuming that time is uniform with a fixed time step of Δt, our model is defined according to the Markov property:

$$\begin{aligned}\mathbb{P}\Big(\big(S(t+\Delta t), I(t+\Delta t)\big) = (s_{t+\Delta t}, i_{t+\Delta t}) | \big(S(0), I(0)\big) = (s_0, i_0), \dots, \big(S(t), I(t)\big) = (s_t, i_t)\Big) \\ = \mathbb{P}\Big(\big(S(t+\Delta t), I(t+\Delta t)\big) = (s_{t+\Delta t}, i_{t+\Delta t}) | \big(S(t), I(t)\big) = (s_t, i_t)\Big).\end{aligned}$$

Hence the number of susceptible (resp. infected) individuals at $t + \Delta t$ only depends on the number of susceptible (resp. infected) individuals at t.

We now assume that Δt can be chosen sufficiently small to only allow one transition from a state to another. In particular, we write:

$$p_{(s,i),(s+k,i+j)}(t+\Delta t, t) = \mathbb{P}\Big(\big(S(t+\Delta t), I(t+\Delta t)\big) = (s+k, i+j) | \big(S(t), I(t)\big) = (s,i)\Big),$$

the transition probability from $(S(t), I(t)) = (s, i)$ to $\big(S(t + \Delta t), I(t + \Delta t)\big) = (s + k, i + j)$ where

$$(k,j) = \begin{cases} (-1,1) & \text{if a susceptible individual gets infected,} \\ (0,-1) & \text{if an infected individual recovers,} \\ (0,0) & \text{otherwise.} \end{cases}$$

By assuming that the Markov chain is time-homogeneous, i.e the transition probabilities are time independent, one can write:

$$p_{(s,i),(s+k,i+j)}(t+\Delta t, t) = p_{(s,i),(s+k,i+j)}(\Delta t).$$

From now on, we want to determine the equation governing the Markov chain. Referring to the representation of the Markov chain time evolution given in Figure 1.5, we have to determine the expression of $p_{(s,i),(s-1,i+1)}(\Delta t)$ and $p_{(s,i),(s,i-1)}(\Delta t)$.

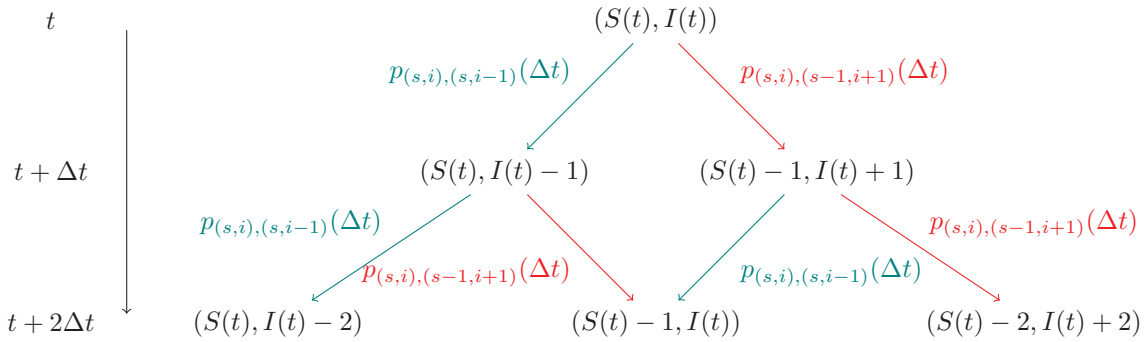


Figure 1.5: SIR Markov chain evaluation at times t, $t + \Delta t$ and $t + 2\Delta t$.
Red arrows represent the infection process while green arrows represent the recovery process.

Regarding the probability that an infection occurs in $[t, t + \Delta t]$, one can write:

$$p_{(s,i),(s-1,i+1)}(\Delta t) = i \times \mathbb{P}(\text{an infected individual infects a susceptible individual in } [t, t+\Delta t]). \quad (1.8)$$

Since an infected individual has to contact a susceptible individual for the infection to transmit, one can rewrite (1.8) as:

$$\begin{aligned}p_{(s,i),(s-1,i+1)}(\Delta t) \quad = \quad & i \times \mathbb{P}\Big(\{\text{an infected individual contacts a susceptible individual in } [t, t+\Delta t]\} \\ & \cap \{\text{the infection occurs}\}\Big).\end{aligned}$$

Hence, we have $p_{(s,i),(s-1,i+1)}(\Delta t) = i \times \dfrac{s}{N} \times \beta \Delta t$ where β is the transmission rate of the infection. Similarly, by assuming that $p_{(s,i),(s,i-1)}(\Delta t)$ is the probability that a recovery occurs in $[t, t + \Delta t]$, one

can write:

$$\begin{aligned} p_{(s,i),(s,i-1)}(\Delta t) &= \mathbb{P}(\text{a recovery occurs in } [t, t+\Delta t]) \\ &= i \times \mathbb{P}(\text{an infected individual recovers in } [t, t+\Delta t]) \\ &= i \times \gamma \Delta t \end{aligned}$$

where γ is the recovery rate of the infection. Hence, we have:

$$p_{(s,i),(s+k,i+j)}(\Delta t) = \begin{cases} \dfrac{\beta i s \Delta t}{N}, & (k,j) = (-1,1) \\ \gamma i \Delta t, & (k,j) = (0,-1) \\ 1 - \left(\dfrac{\beta i s}{N} + \gamma i \right) \Delta t, & (k,j) = (0,0) \\ 0, & \text{otherwise.} \end{cases}$$

Thus, we obtain the following equation:

$$\begin{aligned} p_{(s,i)}(t+\Delta t) &= \left(1 - \left(\frac{\beta i s}{N} + \gamma i \right) \Delta t \right) \times p_{(s,i)}(t) \\ &\quad + \frac{\beta (i-1)(s+1) \Delta t}{N} \times p_{(s+1,i-1)}(t) \\ &\quad + \gamma (i+1) \Delta t \times p_{(s,i+1)}(t). \end{aligned} \tag{1.9}$$

Equation (1.9) describes the evolution of the Markov chain between t and $t + \Delta t$. This equation is also referred as Chapman-Kolmogorov equation. In the next section, we allow the Markov chain to evolve in continuous time.

1.2.2.2 Continuous-time Markov chain approach

Continuous-time Markov chains share the same characteristics as discrete-time Markov chains except for the definition of time. Since epidemic models are mostly defined in continuous time, CTMC models are rather used than their discrete counterpart. Many authors chose to deal with continuous-time Markov chain to model the transmission of infectious diseases like the Eyam plague, AIDS, influenza or measles (see Ragget (1982), Ball and O'Neill (1993), Dushoff et al. (2004) and Cauchemez and Ferguson (2008) respectively).

The model is still subject to the following constraint:

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

where $S(t)$ (resp. $I(t)$, $R(t)$) is a discrete random variable which represents the number of susceptible (resp. infected, recovered) individuals at time t.

The state of the system can be described by $(S(t), I(t))$ with joint probability:

$$p_{(s,i)}(t) = \mathbb{P}\big(S(t) = s, I(t) = i\big).$$

For $0 \leqslant t_0 < t_1 < \cdots < t_n < t_{n+1}$, the Markov property assumes that:

$$\begin{aligned} &\mathbb{P}\Big(\big(S(t_{n+1}), I(t_{n+1}) = (s_{t_{n+1}}, i_{t_{n+1}})\big) | \big(S(t_0), I(t_0)\big) = (s_{t_0}, i_{t_0}), \ldots, \big(S(t_n), I(t_n)\big) = (s_{t_n}, i_{t_n})\Big) \\ &\qquad = \mathbb{P}\Big(\big(S(t_{n+1}), I(t_{n+1}) = (s_{t_{n+1}}, i_{t_{n+1}})\big) | \big(S(t_n), I(t_n)\big) = (s_{t_n}, i_{t_n})\Big). \end{aligned}$$

Let us now describe in an informal manner the different transitions of the model. Assuming that $t_{n+1} - t_n$ is sufficiently small to allow only one transition from a state to another, we write:

$$p_{(s,i),(s+k,i+j)}(t_{n+1}, t_n) = \mathbb{P}\big((S(t_{n+1}), I(t_{n+1})) = (s+k, i+j) | (S(t_n), I(t_n)) = (s,i)\big)$$

the transition probability from $(S(t_n), I(t_n)) = (s,i)$ to $\big(S(t_{n+1}), I(t_{n+1})\big) = (s+k, i+j)$ where

$$(k,j) = \begin{cases} (-1,1) & \text{if a susceptible individual gets infected,} \\ (0,-1) & \text{if an infected individual recovers,} \\ (0,0) & \text{otherwise.} \end{cases}$$

As stated in Section 1.2.2.1, we assume that the Markov chain is time-homogeneous. Thus, one can write:

$$p_{(s,i),(s+k,i+j)}(t_{n+1},t_n) = p_{(s,i),(s+k,i+j)}(t_{n+1}-t_n).$$

More rigorously, by assuming that only one transition can occur instantaneously, infinitesimal transition probabilities are assumed to be:

$$p_{(s,i),(s-1,i+1)}(\Delta t) = \frac{\beta is}{N}\times\Delta t + o(\Delta t) \quad \text{and} \quad p_{(s,i),(s,i-1)}(\Delta t) = i\gamma\times\Delta t + o(\Delta t),$$

where $\lim\limits_{\Delta t\to 0}\dfrac{o(\Delta t)}{\Delta t} = 0$.

Thus the process evolution can be defined by the following infinitesimal transition probability:

$$p_{(s,i),(s+k,i+j)}(\Delta t) = \begin{cases} \dfrac{\beta is\Delta t}{N} + o(\Delta t), & (k,j) = (-1,1) \\ \gamma i\Delta t + o(\Delta t), & (k,j) = (0,-1) \\ 1 - \left(\dfrac{\beta is}{N} + \gamma i\right)\Delta t + o(\Delta t), & (k,j) = (0,0) \\ o(\Delta t), & \text{otherwise.} \end{cases}$$

Hence by writing $\mathbb{P}(S(t+\Delta t), I(t+\Delta t))$ according to $\mathbb{P}(S(t), I(t))$ we get the following equation:

$$\begin{aligned} p_{(s,i)}(t+\Delta t) &= \frac{\beta(i-1)(s+1)\Delta t}{N}\times p_{(s+1,i-1)}(t) + \gamma(i+1)\Delta t\times p_{(s,i+1)}(t) \\ &\quad + \left(1 - \left(\frac{\beta is}{N} + \gamma i\right)\Delta t\right)\times p_{(s,i)}(t) + o(\Delta t). \end{aligned}$$

Then by subtracting $p_{(s,i)}(t)$ and dividing by Δt, we get:

$$\begin{aligned} \frac{p_{(s,i)}(t+\Delta t) - p_{(s,i)}(t)}{\Delta t} &= \frac{\beta(i-1)(s+1)}{N}\times p_{(s+1,i-1)}(t) + \gamma(i+1)\times p_{(s,i+1)}(t) \\ &\quad - \left(\frac{\beta is}{N} + \gamma i\right)\times p_{(s,i)}(t) + \frac{o(\Delta t)}{\Delta t}. \end{aligned}$$

At the limit, when Δt tends to 0, we have:

$$\frac{\mathrm{d}p_{(s,i)}(t)}{\mathrm{d}t} = \frac{\beta(i-1)(s+1)}{N}\times p_{(s+1,i-1)}(t) + \gamma(i+1)\times p_{(s,i+1)}(t) - \left(\frac{\beta is}{N} + \gamma i\right)\times p_{(s,i)}(t). \tag{1.10}$$

The differential equation (1.10), referred to as Forward Kolmogorov Equation, describes the evolution of the Markov chain between t_n and t_{n+1}. In the following section, we propose to study an extension of the continuous time Markov chains to continuous values.

1.2.2.3 Stochastic differential equations approach

We propose in this section to describe the stochastic counterpart of the deterministic compartmental model introduced in Section 1.2.1. We will not deal in-depth with the details but will just give a quick overview of the compartmental model based on stochastic differential equations (SDE). This approach, based on a diffusion process, is a direct extension of the continuous-time Markov chain studied in Section 1.2.2.2 with continuous values instead of discrete states. As stochastic models are generally used to describe the dynamics of small size populations, the approach with SDE offers a good approximation for large size populations as long as the population is finite (see Traulsen et al. (2012)). Stochastic differential equations models have been used to model infectious diseases such as AIDS (see Dalal et al. (2007) or Xu et al. (2007)), measles or malaria (see Kassem and Ndam (2008) or Bhadra et al. (2011) respectively).

From Equation 1.10 obtained in the previous section, Allen (2008) (see also Dargatz (2007)) determined an expression of the probability density function $p_{(s,i)}(t)$ in the form of a partial differential equation being:

$$\begin{aligned}\frac{\partial p_{(s,i)}(t)}{\partial t} &= \frac{\partial}{\partial s}\left(\frac{\beta is}{N}\times p_{(s,i)}(t)\right) - \frac{\partial}{\partial i}\left(\left(\frac{\beta is}{N}-\gamma i\right)\times p_{(s,i)}(t)\right) + \frac{1}{2}\frac{\partial^2}{\partial s^2}\left(\frac{\beta is}{N}\times p_{(s,i)}(t)\right) \\ &\quad -\frac{\partial^2}{\partial s\partial i}\left(\frac{\beta is}{N}\times p_{(s,i)}(t)\right) + \frac{1}{2}\frac{\partial^2}{\partial i^2}\left(\left(\frac{\beta is}{N}+\gamma i\right)\times p_{(s,i)}(t)\right).\end{aligned} \tag{1.11}$$

By denoting $x = \begin{pmatrix} s \\ i \end{pmatrix}$, $\mu(x) = \begin{pmatrix} -\frac{\beta is}{N} \\ \frac{\beta is}{N} - \gamma i \end{pmatrix}$ and $\Sigma(x) = \begin{pmatrix} \frac{\beta is}{N} & -\frac{\beta is}{N} \\ -\frac{\beta is}{N} & \frac{\beta is}{N} + \gamma i \end{pmatrix}$, one can rewrite Equation (1.11) such as:

$$\frac{\partial p_x(t)}{\partial t} = -\frac{\partial}{\partial x}\Big(\mu(x)p_x(t)\Big) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\Big(\Sigma(x)p_x(t)\Big). \tag{1.12}$$

Equation (1.12) is referred to as Fokker-Planck equation for the probability density $p_x(t)$ of the random variable $X(t) = \begin{pmatrix} S(t) \\ I(t) \end{pmatrix}$. According to Dargatz (2007) and Allen (2007), Equation (1.12) is the Markov process solution of the following stochastic differential equation:

$$\mathrm{d}X(t) = \mu(X(t))\mathrm{d}t + B(X(t))\mathrm{d}W(t)$$

where $W(t)$ is a Wiener process[1] and $\Sigma = B^t B$. Thus, by denoting $W_1(t)$ and $W_2(t)$ two independent Wiener processes, one can write:

$$\begin{aligned}\mathrm{d}S(t) &= -\frac{\beta S(t)I(t)}{N}\,\mathrm{d}t + B_{11}(X(t))\mathrm{d}W_1(t) + B_{12}(X(t))\mathrm{d}W_2(t) \\ \mathrm{d}I(t) &= \left(\frac{\beta S(t)I(t)}{N} - \gamma I(t)\right)\mathrm{d}t + B_{21}(X(t))\mathrm{d}W_1(t) + B_{22}(X(t))\mathrm{d}W_2(t)\end{aligned}$$

which describes the time evolution of the stochastic compartmental model.

Next, we propose to study an extension of deterministic compartmental models which can handle more sophisticated features such as population heterogeneities and relieve the assumption of homogeneous mixing. Indeed, by assuming that individuals are organized into a specific framework, one can investigate the spread of a given epidemic on a different population-scale.

1.2.3 Network-based epidemic models

Over the last few decades, important modifications were provided to the basic epidemiological models built on the SIR framework. Indeed, many papers exhibited the need to consider population heterogeneities, interactions between individuals and partnership duration which can play a major role in the propagation of an epidemic Anderson (1988); Ghani et al. (1997); Rolls et al. (2012a)

To deal with more complex epidemiological mechanisms and to extend the proprieties of the classic deterministic SIR model, one can refer to network-based epidemic models which assume that the population of interest is organized according to a specific graph. In this section, we first consider the homogeneous mean field approach which assumes that every individual in the population admits the same mean number of partners. By mean number we directly refer to the parameters of the probability distribution which generates the graph. As the graph is generated just once, all the individuals are assumed to possess the same fixed number of partners (the mean of the probability distribution). This contrasts with the classic deterministic compartmental model for which it is assumed that each individual in the population can enters in contact with anybody else having thus the possibility to infect anyone (full mixing assumption). Then we focus on the heterogeneous mean field approach for which each individual has its own number of partners. Each degree is assumed to be assigned independently.

In the following, we assume that both approaches considered are applied on a random graph of size N generated according to a degree distribution denoted $p(k)$ where k is the degree (number of partners) of a given node. We denote by $\langle K\rangle$ the mean degree of the graph.

[1] The Wiener process is a continuous-time stochastic process such that $\Delta W = W(t+\Delta t) - W(t) \sim \mathcal{N}(0, \Delta t)$.

1.2.3.1 Homogeneous mean field approach

The homogeneous mean field approach is based on the assumption that everyone in the network is connected to a number $\langle K \rangle$ of partners (see Youssef and Scoglio (2011)). This approach simply restrains the homogeneous mixing assumption of the classical deterministic model to the partners of a given individual. First, by considering the SIR framework introduced earlier, we denote by $S(t)$ (resp. $I(t)$ and $R(t)$) the number of susceptible (resp. infected and recovered) individuals in the network at time t. In particular, we still have $S(t) + I(t) + R(t) = N$.

For more convenience, let us work in terms of proportions rather than absolute numbers. Hence, by denoting proportions of the different quantities with lowercase letters, we have:

$$s(t) = \frac{S(t)}{N}, \quad i(t) = \frac{I(t)}{N}, \quad \text{and} \quad r(t) = \frac{R(t)}{N}.$$

As everyone in the population is connected to a number $\langle K \rangle$ of partners, each infected individual can infect $\beta\langle K \rangle$ individuals per unit time. Thus, the proportion of new infections in unit time per infective is $\beta\langle K \rangle \times s(t)$. Hence, we have:

$$\frac{\mathrm{d}s(t)}{\mathrm{d}t} = -\beta\langle K \rangle s(t) i(t).$$

Similarly, by referring to the deterministic compartmental approach, one can write:

$$\frac{\mathrm{d}r(t)}{\mathrm{d}t} = \gamma i(t).$$

From what precedes, we deduce that the governing equations are:

$$\begin{aligned} \frac{\mathrm{d}s(t)}{\mathrm{d}t} &= -\beta\langle K \rangle s(t) i(t), \\ \frac{\mathrm{d}i(t)}{\mathrm{d}t} &= \beta\langle K \rangle s(t) i(t) - \gamma i(t), \\ \frac{\mathrm{d}r(t)}{\mathrm{d}t} &= \gamma i(t). \end{aligned}$$

This first approach based on contact networks offered an interesting way to deal with social structures. However, such a model assumes that everyone in the population has the same number of partners ($\langle K \rangle$) which might be highly improbable in real-life setting. To account for different number of partners, let us introduce the heterogeneous mean field approach which assigns a degree to each individual in the population according to a probability distribution.

1.2.3.2 Heterogeneous mean field approach

To overcome the limitation of the previous approach and give a better approximation of the spread of an epidemic on a network, one can focus on the heterogeneous mean field approach (also called heterogeneous mixing) (see Pastor-Satorras and Vespignani (2002); Sun et al. (2014)). This approach assumes that individuals with the same degree have the same behavior. Hence, for each degree k, individuals are divided into three compartments $S_k(t)$, $I_k(t)$ and $R_k(t)$ representing the susceptible, infected and recovered individuals with degree k respectively. In particular, one can write:

$$S(t) + I(t) + R(t) = \sum_k \Big(S_k(t) + I_k(t) + R_k(t) \Big) = N.$$

Let us respectively denote by $s_k(t)$, $i_k(t)$ and $r_k(t)$ the proportions of susceptible, infected and recovered individuals with degree k. As the probability for a given susceptible individual to become infected depends on the number of its infected partners, the proportion of newly infected individuals with degree k in unit time is proportional to $\beta k \times s_k(t)$.

We denote by $\theta_k(t)$ the probability for a node with degree k to have infected partners at time t. Pastor-Satorras and Vespignani (2002) established the following result:

$$\theta_k(t) = \sum_{k'} p(k'|k) i_{k'}(t)$$

where $p(k'|k) = \dfrac{k'p(k')}{\langle K \rangle}$ is the probability that a node with degree k has a partner with degree k'. From what precedes, one can write:

$$\frac{\mathrm{d}s_k(t)}{\mathrm{d}t} = -\beta k s_k(t)\theta_k(t).$$

Similarly, based on the previous approach, one can write the differential equation governing the proportion of recovered individuals with respect to their degree:

$$\frac{\mathrm{d}r_k(t)}{\mathrm{d}t} = \gamma i_k(t).$$

We easily deduce the following set of differential equations:

$$\begin{aligned}
\frac{\mathrm{d}s_k(t)}{\mathrm{d}t} &= -k\beta s_k(t)\theta_k(t), \\
\frac{\mathrm{d}i_k(t)}{\mathrm{d}t} &= k\beta s_k(t)\theta_k(t) - \gamma i_k(t), \\
\frac{\mathrm{d}r_k(t)}{\mathrm{d}t} &= \gamma i_k(t).
\end{aligned}$$

This approach relieved the hypothesis of homogeneous mixing assumed in the previous section by grouping individuals with the same degree in consideration with the heterogeneous connectivity of the network. However, no specific information is revealed on the individuals.

1.2.4 Conclusion

We conclude this section by a summary table of the different models introduced above. As exhibited in Table 1.1, we focused on five characteristics, population scale, time and variable definitions, the possibility to handle randomness, the number of contacts and complexity.

Regarding the population scale, deterministic models tend to suit better to large populations since the homogeneous mixing assumption greatly reduces the impact of individual variability on the population dynamics. On the contrary, individual variability is better captured by stochastic models such as epidemic models based on Markov chains. Stochastic models can also be valuable to model the very beginning of the epidemic where there is a limited number of infected individuals. In the particular case of the heterogeneous mean field approach, the deterministic nature of the model should be preferred for large populations. However, generating a graph describing the structure of a large population can be quite unmanageable. In Chapter 4, we adopt this approach but consider several smaller graphs connected with each other to overcome this limitation.

Other characteristics are quite straightforward. The time definition is continuous for all the models but the discrete-time Markov chain. Same goes for the variable definition which is considered discrete for both Markov chain based models. Considering randomness, the network based models described in Section 1.2.3 are deterministic but the graphs are generated randomly. One can note that, for the heterogeneous mean field approach, each generation of the graph gives different degree distributions while the mean degree used for the homogeneous mean field approach stays the same. As for the number of contacts, non-network based models assume that each individual can enter in contact with everyone in the population (hence $N-1$ individuals) while homogeneous and heterogeneous mean field approaches respectively assume a constant number $\langle K \rangle$ of partners and a number of contacts based on the degree of each individual. The complexity of the models directly refers to the theory addressed in the previous sections.

Table 1.1: Main characteristics of the different epidemic models.

Methodology	Population scale	Time definition	Variable definition	Handle randomness	Number of contacts	Complexity
Deterministic differential equations	Large	Continuous	Continuous	No	$N-1$ (fixed)	Limited
Discrete-time Markov chain	Small	Discrete	Discrete	Yes	$N-1$ (fixed)	Intermediary
Continuous time Markov chain	Small	Continuous	Discrete	Yes	$N-1$ (fixed)	Intermediary
Stochastic differential equations	Small	Continuous	Continuous	Yes	$N-1$ (fixed)	Challenging
Homogeneous mean field approach	Large	Continuous	Continuous	No	$\langle K \rangle$ (fixed)	Limited
Heterogeneous mean field approach	Large/Small[†]	Continuous	Continuous	Yes/No[‡]	Degree-based (random)	Intermediary

[†] Depends on the reliability of the graph structure, large graphs tend to be more complicated to handle. [‡] The dynamics are deterministic but the structure of the graph is generated randomly.

2 Some estimation problems in epidemic modeling

Français

Dans le domaine des virus informatiques, un nombre grandissant de modèles mathématiques ont été publiés durant la dernière décennie. Cette partie propose d'étudier plusieurs méthodes d'estimation appliquées à deux modèles paramétriques, les modèles "random constant scanning" et "nonhomogeneous random scanning", dans le but de comparer les performances des différents estimateurs à partir de simulations de populations aléatoires. Nous concluons ce travail en appliquant ces différentes méthodes d'estimation à deux jeux de données réelles liés à la propagation de vers sur Internet d'une part et la propagation du VIH en France d'autre part.

English

In the field of computer virus, a growing number of mathematical models have been published during the last decade. In this part, we consider several estimation methods under two parametric models, the random constant scanning model and the nonhomogeneous random scanning model, to propose a comparison of the estimators' performances based on the simulation of random populations. We conclude our work with an application of the different estimation methods to two real datasets dealing with worm propagation on the Internet on the one hand, and dealing with the HIV propagation in France on the other.

Contents

Résumé **22**
2.1 Introduction **23**
2.2 General theory **23**
2.2.1 Epidemic model background 23
2.2.2 Inference based on an initial estimation of $I(\cdot)$ 24
2.2.3 Inference based on an initial estimation of the cumulative distribution function of the time until infection 25
2.3 Simulation study of the Euclidian parameters **27**
2.3.1 Observation of a single population 28
2.3.2 Observation of several independent populations 33
2.4 Simulation study of the estimators of the function $I(\cdot)$ **39**
2.5 Applications on real data sets **43**
2.5.1 Application to the Code Red v2 worm propagation 43
2.5.2 Application to the human immunodeficiency virus (HIV) transmission 44
2.6 Conclusions **45**

Résumé

Dans ce second chapitre, nous nous proposons de comparer les performances de plusieurs méthodes d'estimation statistique appliquées à deux modèles épidémiques paramétriques. Les deux modèles considérés, le modèle "random constant scanning" (RCS) développé par Staniford et al. (2002) et le modèle "nonhomogeneous random scanning" (NHRS) développé par Kirmani and Hood (2010), ont été développés pour modéliser la propagation d'un pathogène dans une population susceptible. Construits sur le modèle logistique développé par Verhulst (1838), ces deux modèles diffèrent essentiellement au niveau de la définition du paramètre β, le taux de contact, qui est supposé dépendant du temps pour le modèle NRHS et constant pour le modèle RCS.

En notant $X(t)$ le nombre d'individus infectés dans l'intervalle de temps $[0,t]$, Kirmani and Hood (2010) ont démontré dans leurs travaux que le nombre moyen d'individus infectés sur $[0,t]$ pouvait s'exprimer de la façon suivante :

$$I(t) = \frac{N}{1 + \psi \exp\left(-\int_0^t \beta(u)\,\mathrm{d}u\right)},$$

et que la fonction de répartition de T, le temps avant infection d'un individus non-infecté donné, s'exprimait de telle sorte que l'on ait :

$$\mathbb{P}(T \leqslant t) = \frac{1 - \exp(-B(t))}{1 + \psi \exp(-B(t))}, \quad t \geqslant 0.$$

De ces deux résultats, nous procédons, dans un premier temps, à l'estimation directe des paramètres euclidiens des modèles RCS et NHRS. En se basant sur une estimation initiale de la fonction $I(t)$, estimée par $\hat{I}(t) = X(t)$ ou $\hat{I}(t) = \bar{X}(t)$ selon le nombre de populations simulées considéré, puis de la fonction de répartition $F(t)$, les paramètres des modèles sont estimés via les méthodes des moindres carrés et du maximum de vraisemblance. Les performances de chacune de ces méthodes d'estimation sont comparées entre elles via le calcul des biais relatifs et des écarts-types relatifs de chacun des paramètres. On notera en particulier que l'approche basée sur l'estimation initiale de $F(t)$ permet la prise en compte de temps de censure. Différents pourcentages de censure sont considérés dans ces travaux.

De ces premières estimations, il résulte que la méthode du maximum de vraisemblance permet d'obtenir de meilleures performances en terme de biais relatif et d'écart-type relatif vis-à-vis de la méthode des moindres carrés et ce pour les deux modèles. Sans surprise, on remarque une nette amélioration de ces performances lorsque augmente la taille de la population et/ou le nombre de populations simulées. Les différences de performance entre les deux méthodes sont nettement moins marquées lorsque l'on considère la censure.

Dans un second temps, nous étudions l'estimation de la fonction $I(t)$ pour chacun des modèles via les méthodes des moindres carrés, du maximum de vraisemblance et de Kaplan-Meier. Les performances de ces estimations sont cette fois-ci comparées via le mean integrated squared error (MISE).

De façon similaire, les résultats obtenus dans cette seconde phase d'estimation pour les modèles RCS et NHRS ont globalement permis de mettre en évidence la supériorité de la méthode par maximum de vraisemblance vis-à-vis des autres méthodes. Une fois encore, dans le cas de temps censurés, ces différences de performance sont sensiblement moins marquées et la méthode des moindres carrés démontre même de meilleurs résultats lorsque le pourcentage de temps censurés considéré est grand.

Nous concluons ces travaux par une application des modèles RCS et NHRS sur deux jeux de données réelles. La première base de données considérée est la base caractérisant la propagation du virus Code Red v2 sur Internet en Juillet 2001 utilisée par Staniford et al. (2002) pour la caractérisation du modèle RCS. Cette application a permis de mettre en évidence une relative similitude entre les différentes méthodes d'estimations. Des différences sont en revanche apparues dans le cas du modèle NHRS. En effet, de moins bonnes performances ont été obtenues pour la méthode des moindres carrés.

La seconde base de données considérée caractérise la propagation du VIH en France. Contrairement à la première base de données, aucune des méthodes n'a réellement pu démontrer de bonnes performances. Nous en concluons une faible adéquation des modèles RCS et NHRS à ce second jeu de données.

2.1 Introduction

The simplest but still useful model of the growth of an epidemic is the logistic model first introduced by Verhulst (1838) and later studied by Pearl and Reed (1920). It is based on the assumption that in a closed homogeneous population consisting of N individuals who are either susceptible to infection or already infected, $X'(t)/X(t)$, the per capita growth rate of the number infected by time t, is directly proportional to $(1-X(t))/N$, where $X(t)$ is the number of individuals infected during the time period [0,t]. Such a way to allocate the population into different compartments directly refers to the Mathematical theory introduced by Kermack and McKendrick (1927). To allow for the fact that evolution of an epidemic is a stochastic, rather than a deterministic process, $X(t)$ must be replaced by its expected value $I(t) = \mathbb{E}(X(t))$. Despite its simplicity, the logistic model has continued to be useful in some practical situations including internet worm propagation (see Liljenstam et al. (2003); Vojnović and Ganesh (2008); Zou et al. (2002)). The random constant scanning (RCS) model of worm propagation, applicable to internet worms which propagate themselves through random scanning of IP addresses to be infected, is, in essence, a logistic model. Staniford et al. (2002) fitted this model to the total number of inbound scans seen during $[0, t]$, for $0 < t \leqslant 16$, on port 80 at the Chemical Abstracts Service during the initial outbreak of version 2 of the Code Red' worm (CRv2) on July 19, 2001. As the name suggests, the RCS model of worm propagation, assumes that the worm's scanning rate is constant in time. To remove this limitation and develop a more realistic model, Kirmani and Hood (2010) developed a nonhomogeneous random scanning (NHRS) model which allows the worm's contact rate to vary during worm propagation. Kirmani and Hood (2010) proved that, for the NHRS generalization, the expected number of hosts infected during [0,t] is given by:

$$I(t) = \frac{N}{1 + \psi \exp\left(-\int_0^t \beta(u)\,\mathrm{d}u\right)}$$

of Kirmani and Hood (2010). Interestingly, Kirmani and Hood (2010) also showed that, for their NHRS model, the cumulative distribution function (cdf) of T, the time until infection of an initially uninfected host, is as follows:

$$\mathbb{P}(T \leqslant t) = \frac{1 - \exp(-B(t))}{1 + \psi \exp(-B(t))}, \quad t \geqslant 0$$

of Kirmani et al. (2010). This important result paves the way for statistical inferences regarding the contact function $\beta(t)$. The objective of the present chapter is to tackle some problems of estimation of $\beta(t)$ and the cdf of T. We propose and study various estimation methods, including the least squares, maximum likelihood, and Kaplan-Meier methods, for the contact rate $\beta(t)$ (including the RCS special case of $\beta(t) = \beta$) and the cdf $F(t) = \mathbb{P}(T \leqslant t)$. This chapter is organized as follows. First, Section 2 introduces the theoretical background of the models and reviews the different estimation methods. Section 3 presents the estimation of the contact rate $\beta(t)$ based on different population sizes and parameter values. Similarly, Section 4 presents the estimation of the cumulative distribution function F. Section 5 provides an application of estimation methods to the real dataset initially used by Staniford et al. (2002). Finally, Section 6 presents our conclusions.

2.2 General theory

2.2.1 Epidemic model background

Our work falls in line with the work of Kirmani and Hood (2010). In the following we will be using notations and results established by Kirmani and Hood (2010) in their Section 2.1. Consider the spread of a computer virus in a closed population of N hosts. We denote by $X(t)$ the number of individuals infected during the time period $[0, t]$. We write $I(t) = \mathbb{E}(X(t))$.

Let us denote by $p(t, t + \Delta t)$ the probability that a given individual is contacted by an infected individual during the infinitesimal time period $]t, t + \Delta t]$. We assume that the probability is the same for each individual and

$$p(t, t + \Delta t) = \frac{\beta(t)}{N}\Delta t + o(\Delta t)$$

where $\beta(t)$ denotes the contact rate at time t. By assuming this, we do not allow the local preference scanning (see Zou et al. (2006)). Note that such a process is an increasing pure jump Markov process.

As previously done by Kirmani and Hood (2010), we consider two parametric models for the contact rate function $\beta(t)$:

- the RCS model with $\beta(t) = \beta$ for all $t \geqslant 0$
- the NHRS model with $\beta(t) = K\beta^K t^{K-1}$ for all $t \geqslant 0$,

where β and K are positive constants. In the following, we will assume that all contacts between individuals are independent.

Hence, referring to the paper of Kirmani and Hood (2010), the dynamic of the infection is governed by the following differential equation:

$$I'(t) = \beta(t)I(t)\left(1 - \frac{I(t)}{N}\right)$$

for all $t > 0$. The solution of this equation is:

$$I(t) = \frac{N}{1 + \psi \exp\left(-\int_0^t \beta(u)\,\mathrm{d}u\right)} \tag{2.1}$$

for all $t > 0$ and where $\psi = \dfrac{N - I_0}{I_0}$ with $I_0 = I(0)$.

From now on we will deal with two different kinds of observation of such a process $t \to X(t)$: either only one path is observed or n independent paths are observable (for example different regions with no or very limited connections). In the first case, we will estimate $I(t)$ by $\hat{I}(t) = X(t)$ whereas in the second case we will use $\hat{I}(t) = \bar{X}(t)$ where $\bar{X}(t)$ is the mean number of infected individuals at time t among the n populations.

In the following subsections, we introduce the least squares, the maximum likelihood and the Kaplan-Meier estimators of some quantities of interest for the RCS and NHRS models.

2.2.2 Inference based on an initial estimation of $I(\cdot)$

First let us write

$$B(t) = \int_0^t \beta(u)\,\mathrm{d}u, \text{ for all } t > 0.$$

From equation (2.1), we can rewrite $I(t)$ as

$$I(t) = \frac{N}{1 + \psi \exp\left(-B(t)\right)}. \tag{2.2}$$

It follows that

$$B(t) = \ln\left[\frac{N - I_0}{I_0}\frac{I(t)}{N - I(t)}\right]. \tag{2.3}$$

Suppose that we are under a NHRS model where $B(t) = \beta^K t^K$. From (2.3), one can give an estimation of $B(t)$ such as:

$$\hat{B}(t) = \ln\left[\frac{N - I_0}{I_0}\frac{\hat{I}(t)}{N - \hat{I}(t)}\right] \tag{2.4}$$

which, if the NHRS assumption is fulfilled, should be close to the theoretical expression $B(t) = \beta^K t^K$. By taking the logarithm of both expressions $B(t) = \beta^K t^K$ and (2.4), and by denoting $a_0 = K\ln(\beta)$ and $a_1 = K$, one can give an estimator of (a_0, a_1) by using the least squares criterion:

$$(\hat{a}_0, \hat{a}_1) = \underset{(a_0,a_1)}{\arg\min} \sum_{i=1}^{m} \left\{\ln\left(\ln\left(\frac{N - I_0}{I_0}\frac{\hat{I}(t_i)}{N - \hat{I}(t_i)}\right)\right) - (a_0 + a_1 \ln(t_i))\right\}^2 \tag{2.5}$$

where $t_1, \ldots, t_m$ are the different observed times of infection. If no times are tied then $m = N - I_0$. Thus we can write an explicit expression of $\hat{a}_0$ and $\hat{a}_1$ and obtain the least squares estimators of β and K denoted by $\hat{\beta}_{\mathrm{ls}_1} = \exp\left(\dfrac{\hat{a}_0}{\hat{a}_1}\right)$ and $\hat{K}_{\mathrm{ls}_1} = \hat{a}_1$ respectively. Finally

$$\hat{I}_{\mathrm{ls}_1}(t) = \frac{N}{1 + \psi \exp\left(-\hat{B}_{\mathrm{ls}_1}(t)\right)}$$

where

$$\hat{B}_{\mathrm{ls}_1}(t) = \hat{\beta}_{\mathrm{ls}_1}^{\hat{K}_{\mathrm{ls}_1}} t^{\hat{K}_{\mathrm{ls}_1}} \tag{2.6}$$

is an estimation of $I(t)$ based on a least squares criterion under the NHRS model.

2.2.3 Inference based on an initial estimation of the cumulative distribution function of the time until infection

Let us denote by T the time until infection of a susceptible individual. Kirmani and Hood (2010) have shown that under the NHRS model we have

$$F(t) = \mathbb{P}(T \leqslant t) = \frac{1 - \exp(-B(t))}{1 + \psi \exp(-B(t))}, \text{ for all } t \geqslant 0. \tag{2.7}$$

Since I_0 hosts are infected at $t = 0$, a total number of $N - I_0$ infection times can be observed within an interval $[0, t]$. Thus, based on the observation of the infection times denoted $(T_1, \ldots, T_{N-I_0})$, one can give an estimation of $F(t)$. Unlike the method of Section 2.2.2, such an approach can handle censorship since we can estimate a cumulative distribution function with censored data using the estimator introduced by Kaplan and Meier (1958).

Indeed, let us assume that some infection time T_i may not be observed. Thus, the observation for an individual i is given by

$$\begin{cases} u_i = \min(T_i, C_i), \\ \delta_i = \bullet_{\{T_i \leqslant C_i\}} \end{cases}$$

where C_i is a random variable which modelizes the censoring time. One can find in Klein and Moeschberger (2003) a pedagogical presentation of statistical inference with censored lifetime data.

Since $X(t)$ is the number of infected hosts at time t, one can write

$$X(t) = I_0 + \sum_{i=1}^{N-I_0} \delta_i \cdot \bullet_{\{u_i \leqslant t\}}.$$

The Kaplan-Meier estimation of $F(\cdot)$ based on the observation of these censored data (see Kaplan and Meier (1958) or Klein and Moeschberger (2003)) will allow us to give parametric and nonparametric estimations of some quantities of interest.

2.2.3.1 Nonparametric inference

The Kaplan-Meier estimator of $F(\cdot)$ can be written as

$$\hat{F}(u) = 1 - \prod_{T_i \leqslant u} \left(1 - \frac{\Delta X(T_i)}{N - I_0 - O(T_i)}\right), \text{ for } u \geqslant 0 \tag{2.8}$$

where $O(t) = \displaystyle\sum_{i=0}^{N-I_0} \bullet_{\{u_i < t\}}$. Note that $N - I_0 - O(T_i)$ is the number at risk at time T_i. From equation (2.7), we can write $B(\cdot)$ as a function of $F(\cdot)$. Then one can compute a nonparametric estimator of the cumulative contact rate:

$$\hat{B}_{\mathrm{km}}(u) = \ln\left(\frac{1 + \psi \hat{F}(u)}{1 - \hat{F}(u)}\right) \text{ for } u \geqslant 0 \tag{2.9}$$

and, from (2.2), we can get a nonparametric estimator of $I(\cdot)$:

$$\hat{I}_{\mathrm{km}}(t) = \frac{N}{1+\psi\exp(-\hat{B}_{\mathrm{km}}(t))} = N\left(\frac{1+\psi\hat{F}(t)}{1+\psi}\right). \tag{2.10}$$

Note that, when dealing with no censorship, the estimator (2.10) is equivalent to the initial estimator $X(t)$, thus $\hat{I}_{\mathrm{km}}(t) = X(t)$.

2.2.3.2 Parametric inference

Let us come back to the parametric NHRS model of Section 2.2.1. Using the nonparametric estimators introduced in Section 2.2.3.1, we are able to introduce new estimators of the parameters β and K and of the functions $I(\cdot)$ and $B(\cdot)$. It is important to note that these new estimators can take into account censored data.

A first approach consists in following the lines of Section 2.2.2 and using the nonparametric estimator of the cumulative contact rate $\hat{B}_{\mathrm{km}}(\cdot)$ given in (2.9) rather than $\hat{I}_{\mathrm{ls}_1}(t)$.

Thus, one can estimate $a_0 = K\ln(\beta)$ and $a_1 = K$ by:

$$(\hat{a}_0, \hat{a}_1) = \underset{(a_0,a_1)}{\arg\min} \sum_{i=1}^{m} \left\{ \ln\left(\ln\left(\frac{1+\psi\hat{F}(t_i)}{1-\hat{F}(t_i)}\right)\right) - (a_0 + a_1\ln(t_i)) \right\}^2$$

where $t_1, \ldots, t_m$ are still the different observed times of infection and $\hat{F}(\cdot)$ is defined in expression (2.8). Hence we obtain new estimators of β and K denoted $\hat{\beta}_{\mathrm{ls}_2}$ and $\hat{K}_{\mathrm{ls}_2}$ respectively. From these estimators we can deduce a parametric estimator of $I(t)$ given by:

$$\hat{I}_{\mathrm{ls}_2}(t) = \frac{N}{1+\psi\exp\left(-\hat{B}_{\mathrm{ls}_2}(t)\right)}$$

where:

$$\hat{B}_{\mathrm{ls}_2}(t) = \hat{\beta}_{\mathrm{ls}_2}^{\hat{K}_{\mathrm{ls}_2}} t^{\hat{K}_{\mathrm{ls}_2}}. \tag{2.11}$$

Recall that when there is no censoring, all these estimators meet the estimators of Section 2.2.2. Another way to estimate the parameters of the NHRS model is to consider the maximum likelihood technic. Indeed, by introducing $f(\cdot)$, the probability density function of the infection time T, one can compute the likelihood of the censored data $(u_i, \delta_i)_{i=1,\ldots,n}$:

$$L(t_1, \ldots, t_{N-I_0}; \beta, K) = \prod_{j=1}^{N-I_0} (f(u_j))^{\delta_j} (1 - F(u_j))^{1-\delta_j}$$

where, thanks to (2.7), the probability distribution function $f(\cdot)$ is given by:

$$f(u) = \frac{\psi\beta(u)e^{-B(u)}}{(1+\psi e^{-B(u)})^2}, \text{ for all } u > 0.$$

Let us denote $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ the estimators of β and K obtained by maximization of the above likelihood. Then we can deduce a new parametric estimator of $I(t)$ given by:

$$\hat{I}_{\mathrm{ml}}(t) = \frac{N}{1+\psi\exp\left(-\hat{B}_{\mathrm{ml}}(t)\right)}$$

where

$$\hat{B}_{\mathrm{ml}}(t) = \hat{\beta}_{\mathrm{ml}}^{\hat{K}_{\mathrm{ml}}} t^{\hat{K}_{\mathrm{ml}}}.$$

Note that all the above estimators based on the maximum likelihood technique are different from the previous one, even when there is no censoring.

2.3 Simulation study of the Euclidian parameters

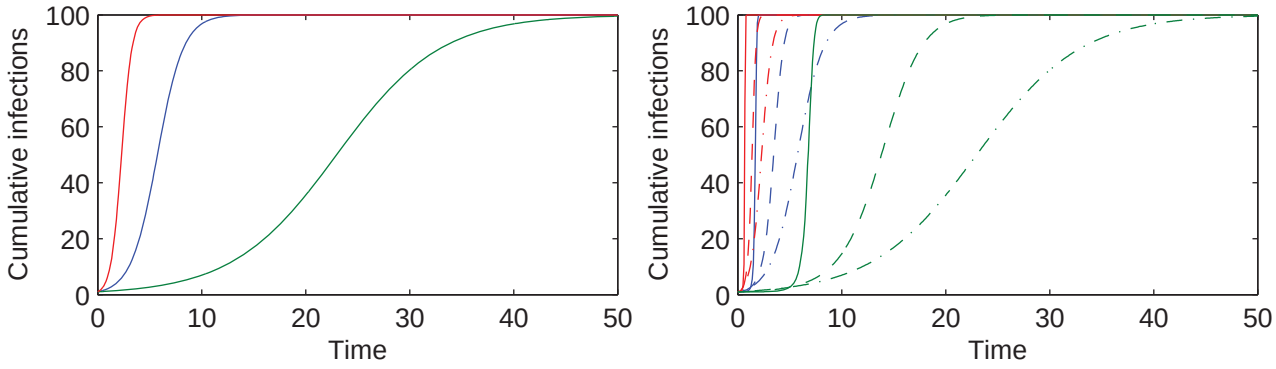
We follow the simulation procedure introduced by Kirmani and Hood (2010). First $N - I_0$ independent random number in $[0, 1]$ are simulated and denoted $\{U_1, \ldots, U_{N-I_0}\}$. Then, $\{T_1, \ldots, T_{N-I_0}\}$ defined by:

$$T_i = \frac{1}{\beta} \ln \left(\frac{1 + \psi U_i}{1 - U_i} \right)^{\frac{1}{K}}, \text{ for } i = 1, \ldots, N - I_0$$

are the times of infection of the $N - I_0$ individuals.

Different values for the parameters are considered, $\beta \in \{0.2, 0.8, 2\}$ for the RCS model and $(\beta, K) \in \{(0.2, 1.5), (0.2, 2), (0.2, 5), (0.8, 1.5), (0.8, 2), (0.8, 5), (2, 1.5), (2, 2), (2, 5)\}$ for the NHRS model.

Figure 2.1 exhibits a sketch of $I(t)$ under both RCS and NHRS models for the different values of β and K.



(a) RCS model **(b)** NHRS model

Figure 2.1: Expected number of hosts infected against time under RCS and NHRS models for different values of β and K. (a) RCS model with $\beta = 0.2$ (green), $\beta = 0.8$ (blue) and $\beta = 2.0$ (red). (b) NHRS model with $\beta = 0.2$ (green), $\beta = 0.8$ (blue), $\beta = 2.0$ (red), $K = 1.0$ (dotted curves), $K = 1.5$ (dashed curves) and $K = 5.0$ (solid curves).

The number $X(t)$ of infected individuals at t is given by:

$$X(t) = I_0 + \sum_{i=1}^{N-I_0} \bullet_{\{T_i \leqslant t\}}.$$

When dealing with estimators which can handle right censored data, the censoring times are simulated according to an exponential distribution with parameter λ chosen in order to obtain a given percentage of censoring. Table 2.1 exhibits the values of λ we found after 1000 simulations which give approximatively the expected percentage of censoring.

Table 2.1: Values of the exponential distribution parameter. The parameter λ approximatively gives 10%, 20% and 50% of censoring.

	Parameter λ	
Percentage of censoring	RCS model	NHRS model
10%	56.0	32.0
20%	26.0	15.0
50%	8.0	4.7

We consider two population sizes $N = 100$ and $N = 1000$ with an initial number of infected individual $I_0 = 1$ and $I_0 = 10$ respectively for both RCS and NHRS models.

Finally, we consider two numbers of observed population: $n = 1$ and $n = 100$. Each of the above simulation scenarios has been replicated 1000 times, allowing us to evaluate empirically the quality criteria of our estimators.

2.3.1 Observation of a single population

In this section, we consider that only one population is observed. Thus, we estimate $I(t)$ by $\hat{I}(t) = X(t)$.

2.3.1.1 Study of the estimators based on an initial estimation of $I(t)$

We consider the least squares estimators $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{K}_{\mathrm{ls}_1}$ of β and K. Tables 2.2 and 2.3 exhibit their empirical relative bias and relative standard deviation under the RCS and the NHRS models. Considering e.g. $\hat{\beta}_{\mathrm{ls}_1}$, the relative bias and relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$ are defined as $\mathbb{E}(\hat{\beta}_{\mathrm{ls}_1} - \beta_{\mathrm{ls}_1})/\beta_{\mathrm{ls}_1}$ and $\sqrt{\mathbb{E}(\hat{\beta}^2_{\mathrm{ls}_1}) - \mathbb{E}(\hat{\beta}_{\mathrm{ls}_1})^2}/\beta_{\mathrm{ls}_1}$ respectively. A more visual information is given by Figures 2.2 and 2.3. First, one can notice that, as expected, the relative bias and relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{K}_{\mathrm{ls}_1}$ decrease in absolute value when the value of N, the population size, increases.

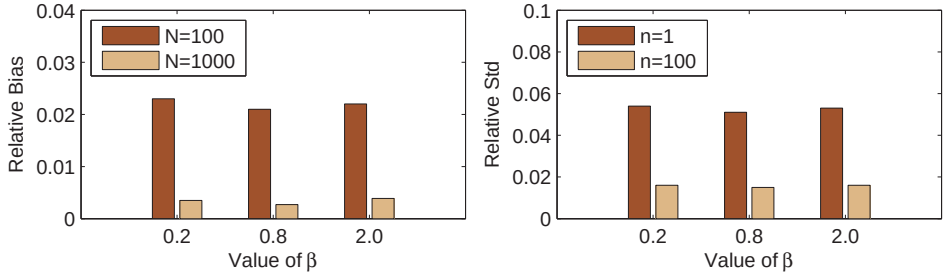
Table 2.2: Simulation results for the RCS model. Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ls}_1}$ for different values of N and β.

Contact rate β	$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
0.2	$2.1 \times 10^{-2}\,(5.4 \times 10^{-2})$	$3.5 \times 10^{-3}\,(1.6 \times 10^{-2})$
0.8	$2.1 \times 10^{-2}\,(5.1 \times 10^{-2})$	$2.7 \times 10^{-3}\,(1.5 \times 10^{-2})$
2.0	$2.2 \times 10^{-2}\,(5.3 \times 10^{-2})$	$3.9 \times 10^{-3}\,(1.6 \times 10^{-2})$

Table 2.3: Simulation results for the NHRS model. Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{K}_{\mathrm{ls}_1}$ for different values of N, β and K.

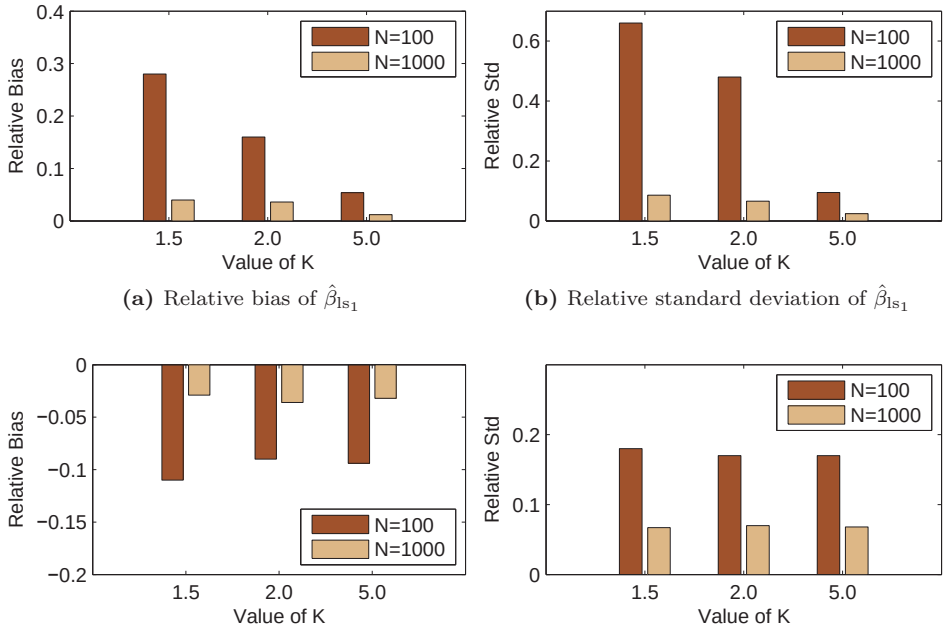
		$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
Contact rate $\beta = 0.2$			
$K = 1.5$	$\hat{\beta}_{\mathrm{ls}_1}$	$2.4 \times 10^{-1}\,(6.2 \times 10^{-1})$	$4.5 \times 10^{-2}\,(8.5 \times 10^{-2})$
	$\hat{K}_{\mathrm{ls}_1}$	$-9.3 \times 10^{-2}\,(1.8 \times 10^{-1})$	$-3.3 \times 10^{-2}\,(6.7 \times 10^{-2})$
$K = 2.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$1.8 \times 10^{-1}\,(4.5 \times 10^{-1})$	$3.0 \times 10^{-2}\,(6.5 \times 10^{-2})$
	$\hat{K}_{\mathrm{ls}_1}$	$-9.9 \times 10^{-2}\,(1.8 \times 10^{-1})$	$-3.0 \times 10^{-2}\,(7.0 \times 10^{-2})$
$K = 5.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$1.1 \times 10^{-1}\,(1.1 \times 10^{-1})$	$1.1 \times 10^{-2}\,(2.4 \times 10^{-2})$
	$\hat{K}_{\mathrm{ls}_1}$	$-1.5 \times 10^{-1}\,(1.6 \times 10^{-1})$	$-3.0 \times 10^{-2}\,(6.6 \times 10^{-2})$
Contact rate $\beta = 0.8$			
$K = 1.5$	$\hat{\beta}_{\mathrm{ls}_1}$	$2.8 \times 10^{-1}\,(6.6 \times 10^{-1})$	$4.0 \times 10^{-2}\,(8.6 \times 10^{-2})$
	$\hat{K}_{\mathrm{ls}_1}$	$-1.1 \times 10^{-1}\,(1.8 \times 10^{-1})$	$-2.9 \times 10^{-2}\,(6.7 \times 10^{-2})$
$K = 2.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$1.6 \times 10^{-1}\,(4.8 \times 10^{-1})$	$3.6 \times 10^{-2}\,(6.6 \times 10^{-2})$
	$\hat{K}_{\mathrm{ls}_1}$	$-9.0 \times 10^{-2}\,(1.7 \times 10^{-1})$	$-3.6 \times 10^{-2}\,(7.0 \times 10^{-2})$
$K = 5.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$5.4 \times 10^{-2}\,(9.5 \times 10^{-2})$	$1.2 \times 10^{-2}\,(2.4 \times 10^{-2})$
	$\hat{K}_{\mathrm{ls}_1}$	$-9.4 \times 10^{-2}\,(1.7 \times 10^{-1})$	$-3.2 \times 10^{-2}\,(6.8 \times 10^{-2})$
Contact rate $\beta = 2.0$			
$K = 1.5$	$\hat{\beta}_{\mathrm{ls}_1}$	$2.3 \times 10^{-1}\,(4.3 \times 10^{-1})$	$4.3 \times 10^{-2}\,(8.5 \times 10^{-2})$
	$\hat{K}_{\mathrm{ls}_1}$	$-9.3 \times 10^{-2}\,(1.8 \times 10^{-1})$	$-3.1 \times 10^{-2}\,(6.7 \times 10^{-2})$
$K = 2.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$1.8 \times 10^{-1}\,(3.2 \times 10^{-1})$	$2.9 \times 10^{-2}\,(6.0 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-1.0 \times 10^{-1}\,(1.9 \times 10^{-1})$	$-2.9 \times 10^{-2}\,(7.0 \times 10^{-2})$
$K = 5.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$5.0 \times 10^{-2}\,(9.0 \times 10^{-2})$	$1.1 \times 10^{-2}\,(2.4 \times 10^{-2})$
	$\hat{K}_{\mathrm{ls}_1}$	$-9.0 \times 10^{-2}\,(1.7 \times 10^{-1})$	$-2.8 \times 10^{-2}\,(6.6 \times 10^{-2})$

One can notice that Table 2.3 exhibits a positive bias for $\hat{\beta}_{\mathrm{ls}_1}$ and a negative bias for $\hat{K}_{\mathrm{ls}_1}$ whatever are the population size N and the original parameters (β, K). One can observe that the relative bias and standard deviation of a parameter aren't impacted by the value of this parameter as displayed in Figures 2.2, 2.3(c) and 2.3(d). Last, Table 2.3 shows that the relative bias and the relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$ decrease when the value of K increases.



(a) Relative bias of $\hat{\beta}_{\mathrm{ls}_1}$ **(b)** Relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$

Figure 2.2: Simulation results under the RCS model.
Histograms of the relative bias (Figure (a)) and the relative standard deviation (Figure (b)) of $\hat{\beta}_{\mathrm{ls}_1}$ based on 1000 replications of the population with $(N, I_0) = (100, 1)$ and $(N, I_0) = (1000, 10)$ for different values of β.



(a) Relative bias of $\hat{\beta}_{\mathrm{ls}_1}$ **(b)** Relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$

(c) Relative bias of $\hat{K}_{\mathrm{ls}_1}$ **(d)** Relative standard deviation of $\hat{K}_{\mathrm{ls}_1}$

Figure 2.3: Simulation results under the NHRS model.
Histograms of the relative bias (Figure (a)), the relative standard deviation (Figure (b)) of $\hat{\beta}_{\mathrm{ls}_1}$, the relative bias (Figure (c)) and the relative standard deviation (Figure (d)) of $\hat{K}_{\mathrm{ls}_1}$ based on 1000 replications of the population with $(N, I_0) = (100, 1)$ and $(N, I_0) = (1000, 10)$ with $\beta = 0.8$ and different values of K.

2.3.1.2 Study of the estimators based on an initial estimation of $F(t)$

We consider the least squares estimators $\hat{\beta}_{\mathrm{ls}_2}$ and $\hat{K}_{\mathrm{ls}_2}$ of β and K. Recall that, unlike the estimators $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{K}_{\mathrm{ls}_1}$ defined in (2.6), the least squares estimators $\hat{\beta}_{\mathrm{ls}_2}$ and $\hat{K}_{\mathrm{ls}_2}$ defined in (2.11) may handle censored observations. However, when there is no censoring, the estimators are the same.

Tables 2.4 and 2.5 give the results of the simulation study of the relative bias and the relative standard deviation of $\hat{\beta}_{\mathrm{ls}_2}$ and $\hat{K}_{\mathrm{ls}_2}$ under both RCS (with $\beta = 0.8$) and NHRS (with $\beta = 0.8$ and $K = 1.5$) models for different sample sizes and percentages of censoring. A more visual information is given by Figures 2.4, 2.5 and 2.6.

Table 2.4: Simulation results for the RCS model ($\beta = 0.8$). Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ls}_2}$ and $\hat{\beta}_{\mathrm{ml}}$ for different percentages of censoring and different values of N.

	Least squares estimator		Maximum likelihood estimator	
Percentage of censoring	$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$	$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
0%	$1.9 \times 10^{-2}\,(5.4 \times 10^{-2})$	$2.5 \times 10^{-3}\,(1.6 \times 10^{-2})$	$1.5 \times 10^{-3}\,(3.3 \times 10^{-2})$	$2.9 \times 10^{-4}\,(1.1 \times 10^{-2})$
10%	$2.5 \times 10^{-2}\,(5.5 \times 10^{-2})$	$2.6 \times 10^{-3}\,(1.8 \times 10^{-2})$	$1.4 \times 10^{-2}\,(3.5 \times 10^{-2})$	$1.2 \times 10^{-2}\,(1.2 \times 10^{-2})$
20%	$2.8 \times 10^{-2}\,(5.9 \times 10^{-2})$	$3.3 \times 10^{-3}\,(1.9 \times 10^{-2})$	$2.9 \times 10^{-2}\,(3.9 \times 10^{-2})$	$2.6 \times 10^{-2}\,(1.3 \times 10^{-2})$
50%	$4.5 \times 10^{-2}\,(7.8 \times 10^{-2})$	$5.9 \times 10^{-3}\,(2.4 \times 10^{-2})$	$9.5 \times 10^{-2}\,(5.6 \times 10^{-2})$	$9.0 \times 10^{-2}\,(1.8 \times 10^{-2})$

Table 2.5: Simulation results for the NHRS model ($\beta = 0.8$ and $K = 1.5$). Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ls}_2}$, $\hat{K}_{\mathrm{ls}_2}$, $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ for different percentages of censored times and different values of N.

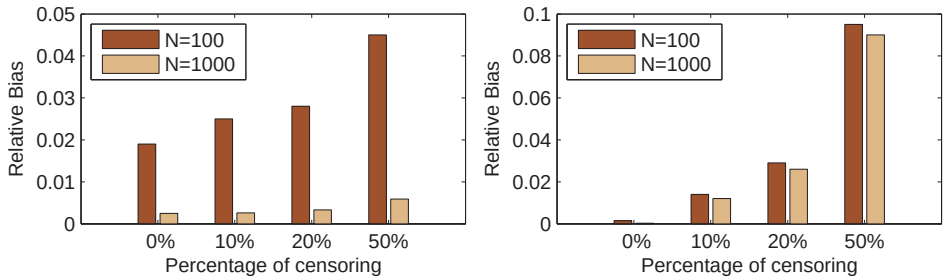
		Least squares estimator		Maximum likelihood estimator	
Percentage of censoring		$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$	$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
0%	$\hat{\beta}$	$2.4 \times 10^{-1}\,(5.0 \times 10^{-1})$	$4.3 \times 10^{-2}\,(8.5 \times 10^{-2})$	$3.6 \times 10^{-3}\,(1.1 \times 10^{-1})$	$3.9 \times 10^{-4}\,(3.3 \times 10^{-2})$
	$\hat{K}$	$-9.3 \times 10^{-2}\,(1.8 \times 10^{-1})$	$-3.1 \times 10^{-2}\,(6.7 \times 10^{-2})$	$1.2 \times 10^{-2}\,(9.3 \times 10^{-2})$	$9.3 \times 10^{-4}\,(2.8 \times 10^{-2})$
10%	$\hat{\beta}$	$2.4 \times 10^{-1}\,(5.6 \times 10^{-1})$	$4.6 \times 10^{-2}\,(9.1 \times 10^{-2})$	$2.0 \times 10^{-2}\,(1.2 \times 10^{-1})$	$1.5 \times 10^{-2}\,(3.6 \times 10^{-2})$
	$\hat{K}$	$-9.3 \times 10^{-2}\,(1.8 \times 10^{-1})$	$-3.4 \times 10^{-2}\,(7.3 \times 10^{-2})$	$2.7 \times 10^{-3}\,(9.3 \times 10^{-2})$	$-6.3 \times 10^{-3}\,(3.0 \times 10^{-2})$
20%	$\hat{\beta}$	$2.9 \times 10^{-1}\,(5.6 \times 10^{-1})$	$4.9 \times 10^{-2}\,(9.4 \times 10^{-2})$	$3.5 \times 10^{-2}\,(1.3 \times 10^{-1})$	$3.4 \times 10^{-2}\,(4.0 \times 10^{-2})$
	$\hat{K}$	$-1.1 \times 10^{-1}\,(1.9 \times 10^{-1})$	$-3.5 \times 10^{-2}\,(7.3 \times 10^{-2})$	$-1.9 \times 10^{-2}\,(1.1 \times 10^{-1})$	$-1.7 \times 10^{-2}\,(3.1 \times 10^{-2})$
50%	$\hat{\beta}$	$4.6 \times 10^{-1}\,(8.5 \times 10^{-1})$	$7.5 \times 10^{-2}\,(1.3 \times 10^{-1})$	$1.3 \times 10^{-1}\,(2.0 \times 10^{-1})$	$1.2 \times 10^{-1}\,(7.5 \times 10^{-2})$
	$\hat{K}$	$-1.5 \times 10^{-1}\,(2.3 \times 10^{-1})$	$-5.5 \times 10^{-2}\,(1.0 \times 10^{-1})$	$-3.3 \times 10^{-2}\,(1.3 \times 10^{-1})$	$-6.1 \times 10^{-2}\,(3.7 \times 10^{-2})$

One can notice that, as expected, the absolute value of the relative bias and the relative standard deviation of $\hat{\beta}_{\mathrm{ls}_2}$ and $\hat{K}_{\mathrm{ls}_2}$ decrease when N increases. It is also not surprising to see that these criterions increase with the percentage of censoring. Finally, we denote that $\hat{K}_{\mathrm{ls}_2}$ has a negative bias disregarding the population size N.

We now consider the maximum likelihood estimators $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ of β and K. Tables 2.4 and 2.5 give the results of the simulation study of the relative bias and the relative standard deviation of $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ under both RCS and NHRS models (with $\beta = 0.8$ and $K = 1.5$). See Figures 2.4, 2.5 and 2.6 for a more visual information.

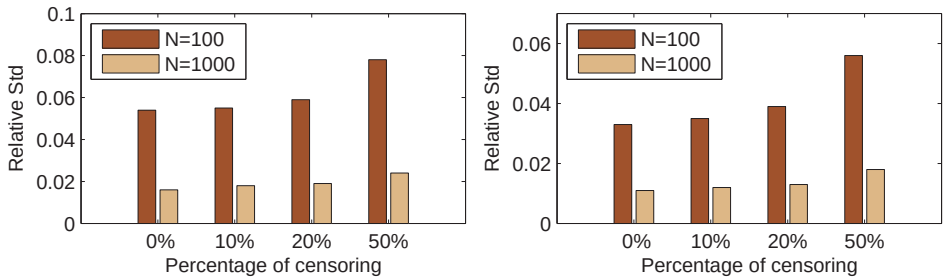
One can notice again that, as expected, the absolute value of the relative bias and the relative standard deviation of $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ increase with the percentage of censoring and decrease with N.

Comparing the least squares estimators $(\hat{\beta}_{\mathrm{ls}_2}, \hat{K}_{\mathrm{ls}_2})$ with the maximum likelihood estimators $(\hat{\beta}_{\mathrm{ml}}, \hat{K}_{\mathrm{ml}})$, one can notice that the values of the relative bias and the relative standard deviation of $(\hat{\beta}_{\mathrm{ml}}, \hat{K}_{\mathrm{ml}})$ are almost always smaller than the values of the relative bias and the relative standard deviation of $(\hat{\beta}_{\mathrm{ls}_2}, \hat{K}_{\mathrm{ls}_2})$.



(a) Relative bias of $\hat{\beta}_{\mathrm{ls}_2}$

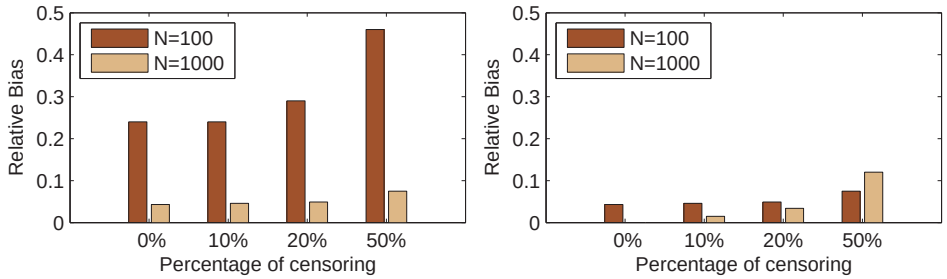
(b) Relative bias of $\hat{\beta}_{\mathrm{ml}}$



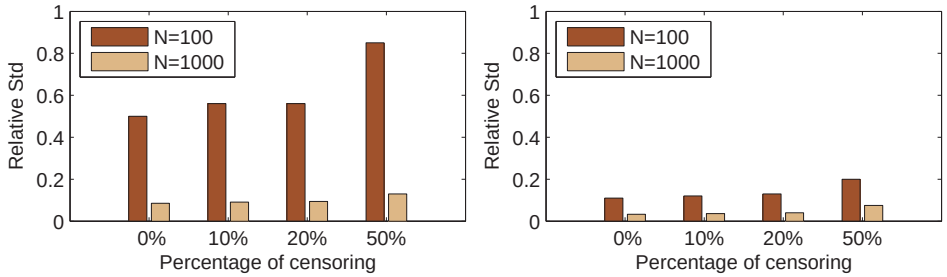
(c) Relative standard deviation of $\hat{\beta}_{\mathrm{ls}_2}$

(d) Relative standard deviation of $\hat{\beta}_{\mathrm{ml}}$

Figure 2.4: Simulation results under the RCS model. Histograms of the relative bias (Figure (a)), the relative standard deviation (Figure (c)) of $\hat{\beta}_{\mathrm{ls}_2}$, the relative bias (Figure (b)) and the relative standard deviation (Figure (d)) of $\hat{\beta}_{\mathrm{ml}}$ based on 1000 replications of the population with $(N, I_0) = (100, 1)$ and $(N, I_0) = (1000, 10)$ with $\beta = 0.8$ for different percentages of censoring.

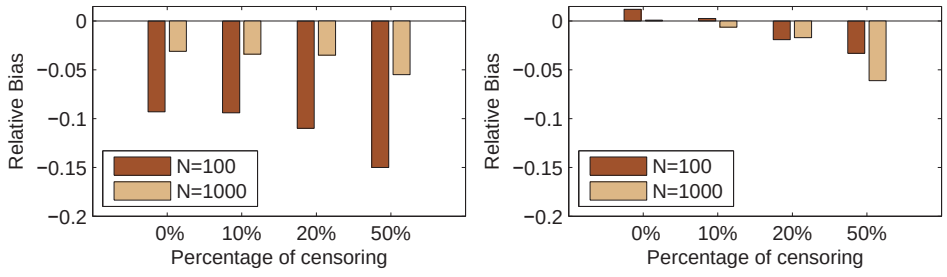


(a) Relative bias of $\hat{\beta}_{\mathrm{ls}_2}$ **(b)** Relative bias of $\hat{\beta}_{\mathrm{ml}}$

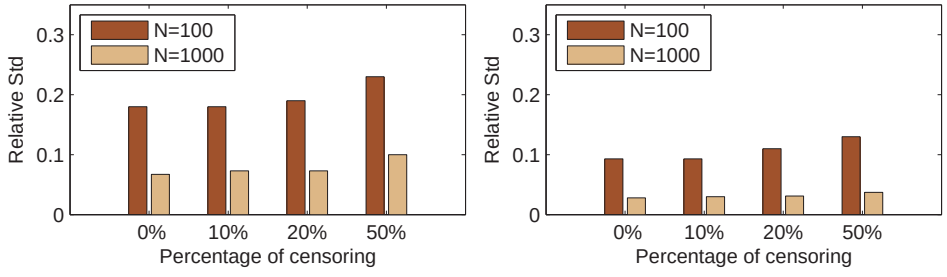


(c) Relative standard deviation of $\hat{\beta}_{\mathrm{ls}_2}$ **(d)** Relative standard deviation of $\hat{\beta}_{\mathrm{ml}}$

Figure 2.5: Simulation results under the NHRS model.
Histograms of the relative bias (Figure (a)), the relative standard deviation (Figure (c)) of $\hat{\beta}_{\mathrm{ls}_2}$, the relative bias (Figure (b)) and the relative standard deviation (Figure (d)) of $\hat{\beta}_{\mathrm{ml}}$ based on 1000 replications of the population with $(N, I_0) = (100, 1)$ and $(N, I_0) = (1000, 10)$ with $\beta = 0.8$ and $K = 1.5$ for different percentages of censoring.



(a) Relative bias of $\hat{K}_{\mathrm{ls}_2}$ **(b)** Bias of $\hat{K}_{\mathrm{ml}}$



(c) Relative standard deviation of $\hat{K}_{\mathrm{ls}_2}$ **(d)** Relative standard deviation of $\hat{K}_{\mathrm{ml}}$

Figure 2.6: Simulation results under the NHRS model.
Histograms of the relative bias (Figure (a)), the relative standard deviation (Figure (c)) of $\hat{K}_{\mathrm{ls}_2}$, the relative bias (Figure (b)) and the relative standard deviation (Figure (d)) of $\hat{K}_{\mathrm{ml}}$ based on 1000 replications of the population with $(N, I_0) = (100, 1)$ and $(N, I_0) = (1000, 10)$ with $\beta = 0.8$ and $K = 1.5$ for different percentages of censoring.

2.3.2 Observation of several independent populations

In this section, we compare the estimation results for $n = 1$ and $n = 100$ independent populations. Recall that, we estimate $I(t)$ by $\hat{I}(t) = \bar{X}(t)$ in the particular case $n = 100$.

Let us first consider the least squares estimators $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{K}_{\mathrm{ls}_1}$ of β and K. Tables 2.6 and 2.7 displays the relative bias and the relative standard deviation of the estimators $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{K}_{\mathrm{ls}_1}$ under the RCS and NHRS models with $\beta = 0.8$, $K = 1.5$ for $(N, I_0) = (100, 1)$ and $(N, I_0) = (1000, 10)$ in the particular case $n = 100$. One can notice that, as expected, the relative bias and relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{K}_{\mathrm{ls}_1}$ decrease when the value of N increases. One can compare those results with the ones displayed in Tables 2.2 and 2.3 for $n = 1$.

In Figure 2.7, we compare the relative bias and standard deviation when $n = 1$ and $n = 100$. We only displayed the case $(N, I_0) = (100, 1)$ since similar trends were observed with $(N, I_0) = (1000, 10)$. One can notice that, as expected, the absolute value of the relative bias and the relative standard deviation greatly decrease as the value of n increases. Figures 2.7(c) and 2.7(d) show again that the relative bias and the relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$ decrease when the value of K increases.

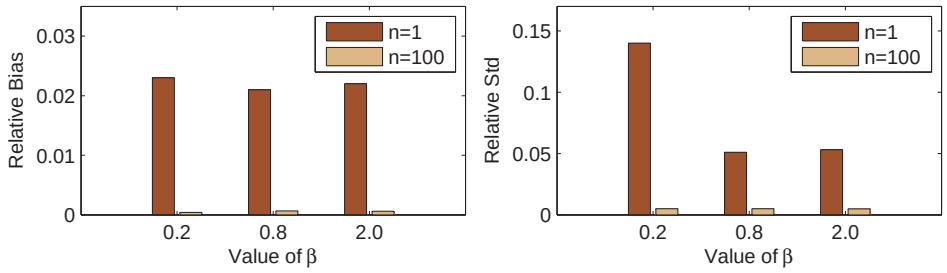
As stated before, we observe a positive bias for $\hat{\beta}_{\mathrm{ls}_1}$ and a negative bias for $\hat{K}_{\mathrm{ls}_1}$ independently of the parameters (β, K). We conclude that the behavior of $\hat{\beta}_{\mathrm{ls}_1}$ (resp. $\hat{K}_{\mathrm{ls}_1}$) when $n = 100$ is similar to the behavior of $\hat{\beta}_{\mathrm{ls}_1}$ (resp. $\hat{K}_{\mathrm{ls}_1}$) when $n = 1$ and that there is a global improvement of the estimations in the case $n = 100$.

Table 2.6: Simulation results for the RCS model. Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ls}_1}$ for different values of N and β with $n = 100$.

Contact rate β	$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
0.2	$4.1 \times 10^{-4}\,(4.9 \times 10^{-3})$	$1.2 \times 10^{-4}\,(1.5 \times 10^{-3})$
0.8	$6.5 \times 10^{-4}\,(5.0 \times 10^{-3})$	$1.8 \times 10^{-4}\,(1.6 \times 10^{-3})$
2.0	$6.0 \times 10^{-4}\,(4.8 \times 10^{-3})$	$1.1 \times 10^{-4}\,(1.6 \times 10^{-3})$

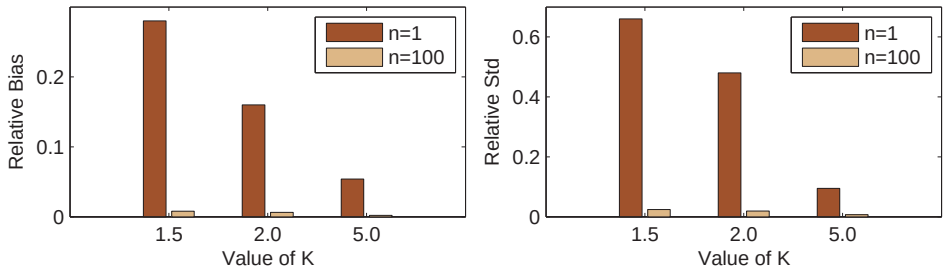
Table 2.7: Simulation results for the NHRS model. Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{K}_{\mathrm{ls}_1}$ for different values of N, β and K with $n = 100$.

		$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
Contact rate $\beta = 0.2$			
$K = 1.5$	$\hat{\beta}_{\mathrm{ls}_1}$	$8.0 \times 10^{-3}\,(2.5 \times 10^{-2})$	$1.5 \times 10^{-3}\,(7.0 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-6.7 \times 10^{-3}\,(2.3 \times 10^{-2})$	$-1.2 \times 10^{-3}\,(6.7 \times 10^{-3})$
$K = 2.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$5.5 \times 10^{-3}\,(1.9 \times 10^{-2})$	$1.0 \times 10^{-3}\,(5.5 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-6.5 \times 10^{-3}\,(2.3 \times 10^{-2})$	$-1.4 \times 10^{-3}\,(7.0 \times 10^{-3})$
$K = 5.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$2.0 \times 10^{-3}\,(7.0 \times 10^{-3})$	$5.0 \times 10^{-4}\,(2.0 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-5.8 \times 10^{-3}\,(2.2 \times 10^{-2})$	$-1.1 \times 10^{-3}\,(7.0 \times 10^{-3})$
Contact rate $\beta = 0.8$			
$K = 1.5$	$\hat{\beta}_{\mathrm{ls}_1}$	$8.1 \times 10^{-3}\,(2.4 \times 10^{-2})$	$1.6 \times 10^{-3}\,(7.2 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-7.3 \times 10^{-3}\,(2.3 \times 10^{-2})$	$-1.5 \times 10^{-3}\,(6.7 \times 10^{-3})$
$K = 2.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$6.2 \times 10^{-3}\,(1.9 \times 10^{-2})$	$6.3 \times 10^{-4}\,(5.7 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-7.5 \times 10^{-3}\,(2.3 \times 10^{-2})$	$-7.5 \times 10^{-4}\,(7.5 \times 10^{-3})$
$K = 5.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$2.1 \times 10^{-3}\,(7.1 \times 10^{-3})$	$3.8 \times 10^{-4}\,(2.2 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-6.4 \times 10^{-3}\,(2.2 \times 10^{-2})$	$-1.3 \times 10^{-3}\,(7.0 \times 10^{-3})$
Contact rate $\beta = 2.0$			
$K = 1.5$	$\hat{\beta}_{\mathrm{ls}_1}$	$8.0 \times 10^{-3}\,(2.4 \times 10^{-2})$	$1.0 \times 10^{-3}\,(7.0 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-7.3 \times 10^{-3}\,(2.1 \times 10^{-2})$	$-1.1 \times 10^{-3}\,(6.4 \times 10^{-3})$
$K = 2.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$6.5 \times 10^{-3}\,(1.9 \times 10^{-2})$	$9.5 \times 10^{-4}\,(5.5 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-7.5 \times 10^{-3}\,(2.3 \times 10^{-2})$	$-1.3 \times 10^{-3}\,(6.5 \times 10^{-3})$
$K = 5.0$	$\hat{\beta}_{\mathrm{ls}_1}$	$2.2 \times 10^{-3}\,(7.5 \times 10^{-3})$	$3.5 \times 10^{-4}\,(2.2 \times 10^{-3})$
	$\hat{K}_{\mathrm{ls}_1}$	$-6.6 \times 10^{-3}\,(2.2 \times 10^{-2})$	$-1.1 \times 10^{-3}\,(7.0 \times 10^{-3})$



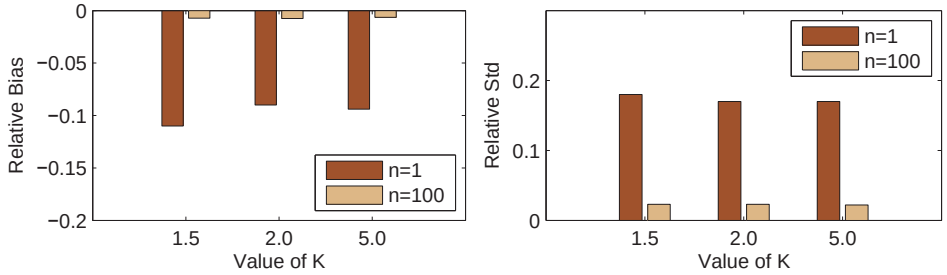
(a) Relative bias of $\hat{\beta}_{\mathrm{ls}_1}$ - RCS

(b) Relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$ - RCS



(c) Relative bias of $\hat{\beta}_{\mathrm{ls}_1}$ - NHRS

(d) Relative standard deviation of $\hat{\beta}_{\mathrm{ls}_1}$ - NHRS



(e) Relative bias of $\hat{K}_{\mathrm{ls}_1}$ - NHRS

(f) Relative standard deviation of $\hat{K}_{\mathrm{ls}_1}$ - NHRS

Figure 2.7: Simulation results for different values of β, K and n.
Histograms of the relative bias (Figure (a)) and the relative standard deviation (Figure (b)) of $\hat{\beta}_{\mathrm{ls}_1}$ based on 1000 replications of population with $(N, I_0) = (100, 1)$ under the RCS model. Histograms of the relative bias (Figure (c)), the relative standard deviation (Figure (d)) of $\hat{\beta}_{\mathrm{ls}_1}$, the relative bias (Figure (e)) and the relative standard deviation (Figure (f)) of $\hat{K}_{\mathrm{ls}_1}$ based on 1000 replications of the population with $(N, I_0) = (100, 1)$ under the NHRS model with $\beta = 0.8$.

We now consider the maximum likelihood estimators $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ of β and K. Tables 2.8 and 2.9 give the results of the simulation study of the relative bias and the relative standard deviation of $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ under, respectively, the RCS and NHRS models with $\beta = 0.8$ and $K = 1.5$ for $(N, I_0) = (100, 1)$ and $(N, I_0) = (1000, 10)$ in the particular case $n = 100$. As expected, we find again that the absolute value of the relative bias and the relative standard deviation of $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ increase with the percentage of censoring. One can compare those results with the ones displayed in Tables 2.4 and 2.5 for $n = 1$. Referring to censored observations in Table 2.9, one can notice that the relative bias of $\hat{K}_{\mathrm{ml}}$ is negative as observed for the previous estimators (see Tables 2.3, 2.5 and 2.7) while the relative bias of $\hat{\beta}_{\mathrm{ml}}$ is positive.

As previously, we only displayed the results of the particular case $N = 100$ in Figure 2.8 where a comparison between $n = 1$ and $n = 100$ is done. One can notice that the value of the standard deviation greatly decreases as the value of n increases. However, only a limited improvement is exhibited for the relative bias of $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$. Considering Figure 2.8(e), one can notice that the relative bias of $\hat{K}_{\mathrm{ml}}$ doesn't decrease when n increases. This phenomenon is a consequence of the strange variations observed in Figure 2.6(b). As observed before, the relative bias of $\hat{K}_{\mathrm{ml}}$ is negative while the relative bias of $\hat{\beta}_{\mathrm{ml}}$ is positive.

Table 2.8: Simulation results for the RCS model ($\beta = 0.8$). Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ml}}$ for different percentages of censoring and different values of N with $n = 100$.

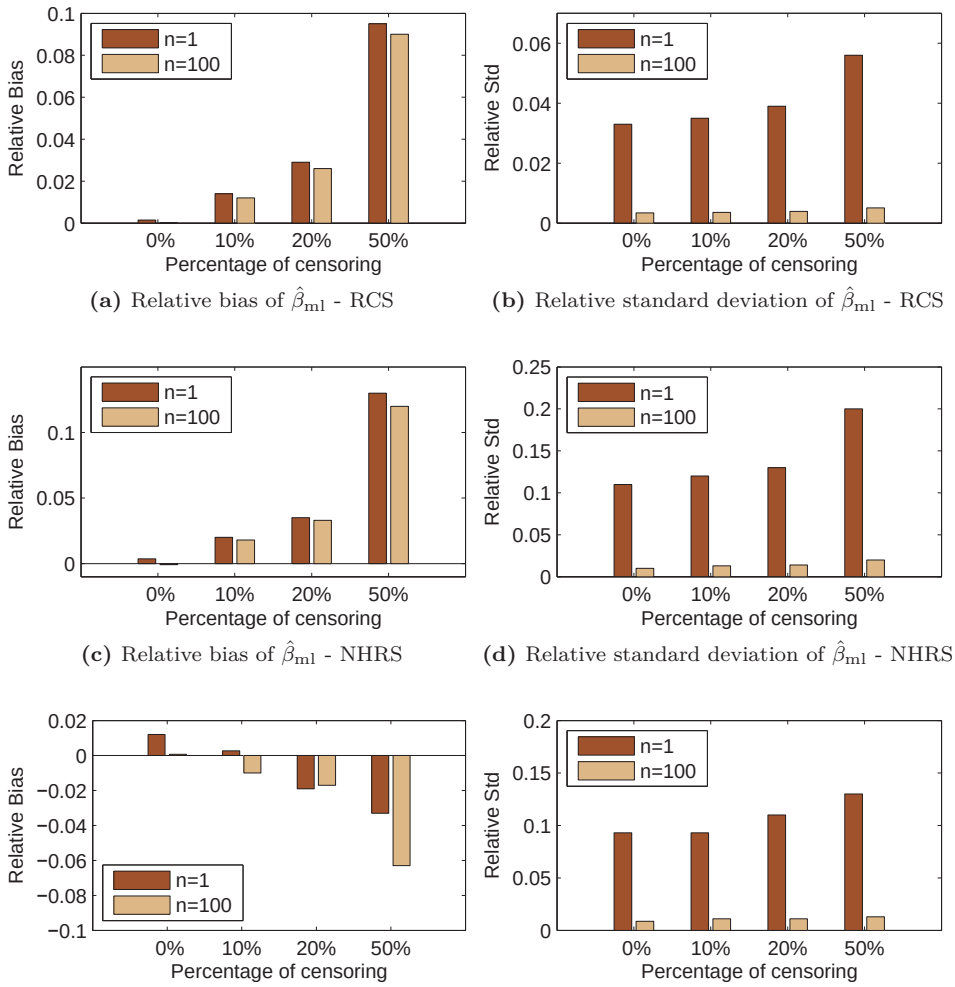
	Maximum likelihood estimator	
Percentage of censoring	$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
0%	$9.1 \times 10^{-5}\,(3.4 \times 10^{-3})$	$7.9 \times 10^{-6}\,(1.1 \times 10^{-3})$
10%	$1.2 \times 10^{-2}\,(3.6 \times 10^{-3})$	$1.2 \times 10^{-2}\,(3.4 \times 10^{-3})$
20%	$2.6 \times 10^{-2}\,(3.9 \times 10^{-3})$	$2.1 \times 10^{-2}\,(3.9 \times 10^{-3})$
50%	$9.0 \times 10^{-2}\,(5.1 \times 10^{-3})$	$8.9 \times 10^{-2}\,(5.0 \times 10^{-3})$

Table 2.9: Simulation results for the NHRS model ($\beta = 0.8$ and $K = 1.5$). Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ for different percentages of censored times and different values of N with $n = 100$.

		Maximum likelihood estimator	
Percentage of censoring		$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
0%	$\hat{\beta}$	$-8.0 \times 10^{-4}\,(1.0 \times 10^{-2})$	$-6.5 \times 10^{-5}\,(3.0 \times 10^{-3})$
	$\hat{K}$	$6.7 \times 10^{-4}\,(8.7 \times 10^{-3})$	$7.3 \times 10^{-5}\,(2.8 \times 10^{-3})$
10%	$\hat{\beta}$	$1.8 \times 10^{-2}\,(1.3 \times 10^{-2})$	$1.3 \times 10^{-2}\,(7.3 \times 10^{-3})$
	$\hat{K}$	$-1.0 \times 10^{-2}\,(1.1 \times 10^{-2})$	$-8.0 \times 10^{-3}\,(2.9 \times 10^{-3})$
20%	$\hat{\beta}$	$3.3 \times 10^{-2}\,(1.4 \times 10^{-2})$	$3.0 \times 10^{-2}\,(1.2 \times 10^{-2})$
	$\hat{K}$	$-1.7 \times 10^{-2}\,(1.1 \times 10^{-2})$	$-1.5 \times 10^{-2}\,(3.0 \times 10^{-3})$
50%	$\hat{\beta}$	$1.2 \times 10^{-1}\,(2.0 \times 10^{-2})$	$-1.2 \times 10^{-1}\,(1.9 \times 10^{-2})$
	$\hat{K}$	$-6.3 \times 10^{-2}\,(1.3 \times 10^{-2})$	$-3.7 \times 10^{-2}\,(6.7 \times 10^{-3})$

Since least squares estimators $(\hat{\beta}_{\mathrm{ls}_1}, \hat{K}_{\mathrm{ls}_1})$ can't handle censored observations, the comparison between the relative bias and the relative standard deviation of $(\hat{\beta}_{\mathrm{ls}_1}, \hat{K}_{\mathrm{ls}_1})$ and $(\hat{\beta}_{\mathrm{ml}}, \hat{K}_{\mathrm{ml}})$ is only driven when there is no censoring. Referring to the estimators' comparison done in Tables 2.4 and 2.5 when $n = 1$, we now compare $(\hat{\beta}_{\mathrm{ls}_1}, \hat{K}_{\mathrm{ls}_1})$ and $(\hat{\beta}_{\mathrm{ml}}, \hat{K}_{\mathrm{ml}})$ when $n = 100$.

Tables 2.10 and 2.11 show that the values of the relative bias and the relative standard deviation of $(\hat{\beta}_{\mathrm{ml}}, \hat{K}_{\mathrm{ml}})$ are smaller than the values of the relative bias and the relative standard deviation of $(\hat{\beta}_{\mathrm{ls}_1}, \hat{K}_{\mathrm{ls}_1})$ under both RCS and NHRS models.



(a) Relative bias of $\hat{\beta}_{\mathrm{ml}}$ - RCS

(b) Relative standard deviation of $\hat{\beta}_{\mathrm{ml}}$ - RCS

(c) Relative bias of $\hat{\beta}_{\mathrm{ml}}$ - NHRS

(d) Relative standard deviation of $\hat{\beta}_{\mathrm{ml}}$ - NHRS

(e) Relative bias of $\hat{K}_{\mathrm{ml}}$ - NHRS

(f) Relative standard deviation of $\hat{K}_{\mathrm{ml}}$ - NHRS

Figure 2.8: Simulation results for different percentages of censoring and different values of n. Histograms of the relative bias (Figure (a)) and the relative standard deviation (Figure (b)) of $\hat{\beta}_{\mathrm{ml}}$ based on 1000 replications of the population with $(N, I_0) = (100, 1)$ under the RCS model with $\beta = 0.8$. Histograms of the relative bias (Figure (c)) and the relative standard deviation (Figure (d)) of $\hat{\beta}_{\mathrm{ml}}$, and the relative bias (Figure (e)) and the relative standard deviation (Figure (f)) of $\hat{K}_{\mathrm{ml}}$ based on 1000 replications of the population with $(N, I_0) = (100, 1)$ under the NHRS model with $\beta = 0.8$ and $K = 1.5$.

Table 2.10: Simulation results for the RCS model ($\beta = 0.8$). Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ls}_1}$ and $\hat{\beta}_{\mathrm{ml}}$ for different values of N with $n = 100$.

	$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
Least squares estimator	$6.5 \times 10^{-4}\,(5.0 \times 10^{-3})$	$1.8 \times 10^{-4}\,(1.6 \times 10^{-3})$
Maximum likelihood estimator	$9.1 \times 10^{-5}\,(3.4 \times 10^{-3})$	$7.9 \times 10^{-6}\,(1.1 \times 10^{-3})$

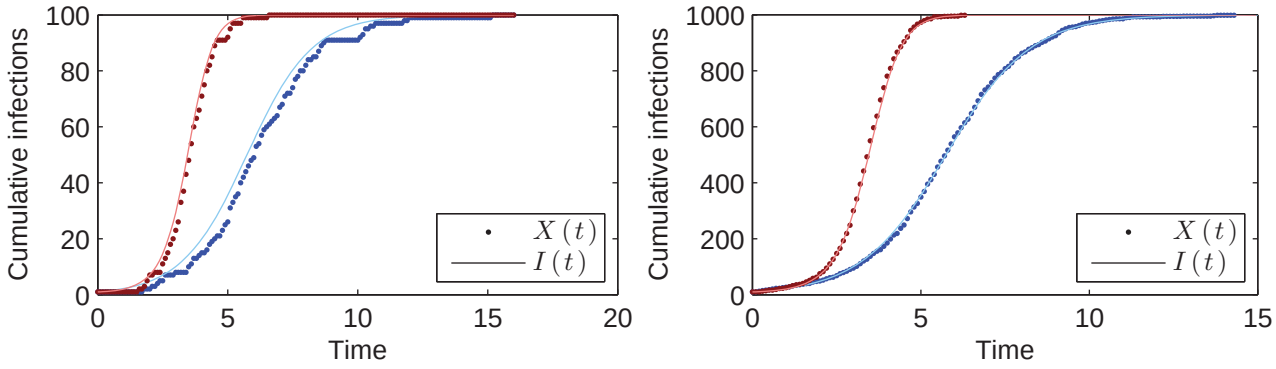
Table 2.11: Simulation results for the NHRS model ($\beta = 0.8$ and $K = 1.5$). Empirical estimation of the relative bias and the relative standard deviation (between parentheses) of $\hat{\beta}_{\mathrm{ls}_1}$, $\hat{K}_{\mathrm{ls}_1}$, $\hat{\beta}_{\mathrm{ml}}$ and $\hat{K}_{\mathrm{ml}}$ for different values of N with $n = 100$.

		$N = 100, I_0 = 1$	$N = 1000, I_0 = 10$
$\hat{\beta}$	Least squares estimator	$8.1 \times 10^{-3}\,(2.4 \times 10^{-2})$	$1.6 \times 10^{-3}\,(7.3 \times 10^{-3})$
	Maximum likelihood estimator	$-4.3 \times 10^{-4}\,(5.3 \times 10^{-3})$	$-3.5 \times 10^{-5}\,(1.6 \times 10^{-3})$
$\hat{K}$	Least squares estimator	$-1.4 \times 10^{-2}\,(4.3 \times 10^{-2})$	$-2.9 \times 10^{-3}\,(1.3 \times 10^{-2})$
	Maximum likelihood estimator	$6.7 \times 10^{-4}\,(8.7 \times 10^{-3})$	$7.3 \times 10^{-5}\,(2.8 \times 10^{-3})$

2.4 Simulation study of the estimators of the function $I(\cdot)$

Contrary to what has been done previously, in this section we will not focus on the estimators of the euclidean parameters β and K anymore but will consider and compare different estimators of the function $I(\cdot)$. Those comparisons will be illustrated via simulated examples for different population sizes, number of populations and percentages of censoring.

First, Figure 2.9 gives a representation of the epidemic model $I(t)$ and the corresponding simulated observations of $X(t)$ under RCS and NHRS models, for $N = 100$ and $N = 1000$ individuals.

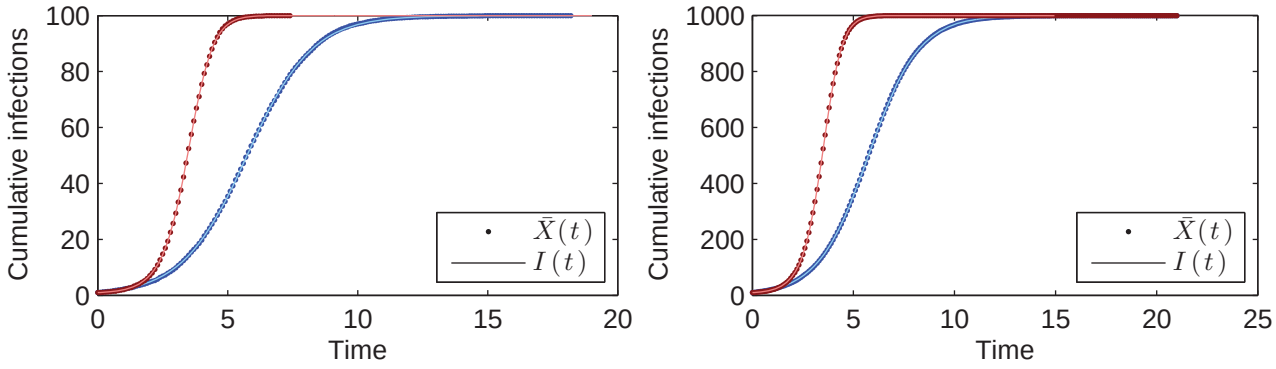


(a) $N = 100, I_0 = 1$ **(b)** $N = 1000, I_0 = 10$

Figure 2.9: Simulation of RCS and NHRS models for $(N = 100, I_0 = 1)$ **and** $(N = 1000, I_0 = 10)$**.** Blue curves correspond to the RCS model ($\beta = 0.8$) and red curves to the NHRS model ($\beta = 0.8$ and $K = 1.5$).

Considering now n independent populations, we estimate $I(t)$ by $\hat{I}(t) = \bar{X}(t)$ where $\bar{X}(t)$ is the mean number of infected individuals at time t among the n independent populations.

Figure 2.10 gives a representation of the epidemic model $I(t)$ and the corresponding observations $\bar{X}(t)$ under RCS and NHRS models for 100 populations of size $N = 100$ and $N = 1000$.

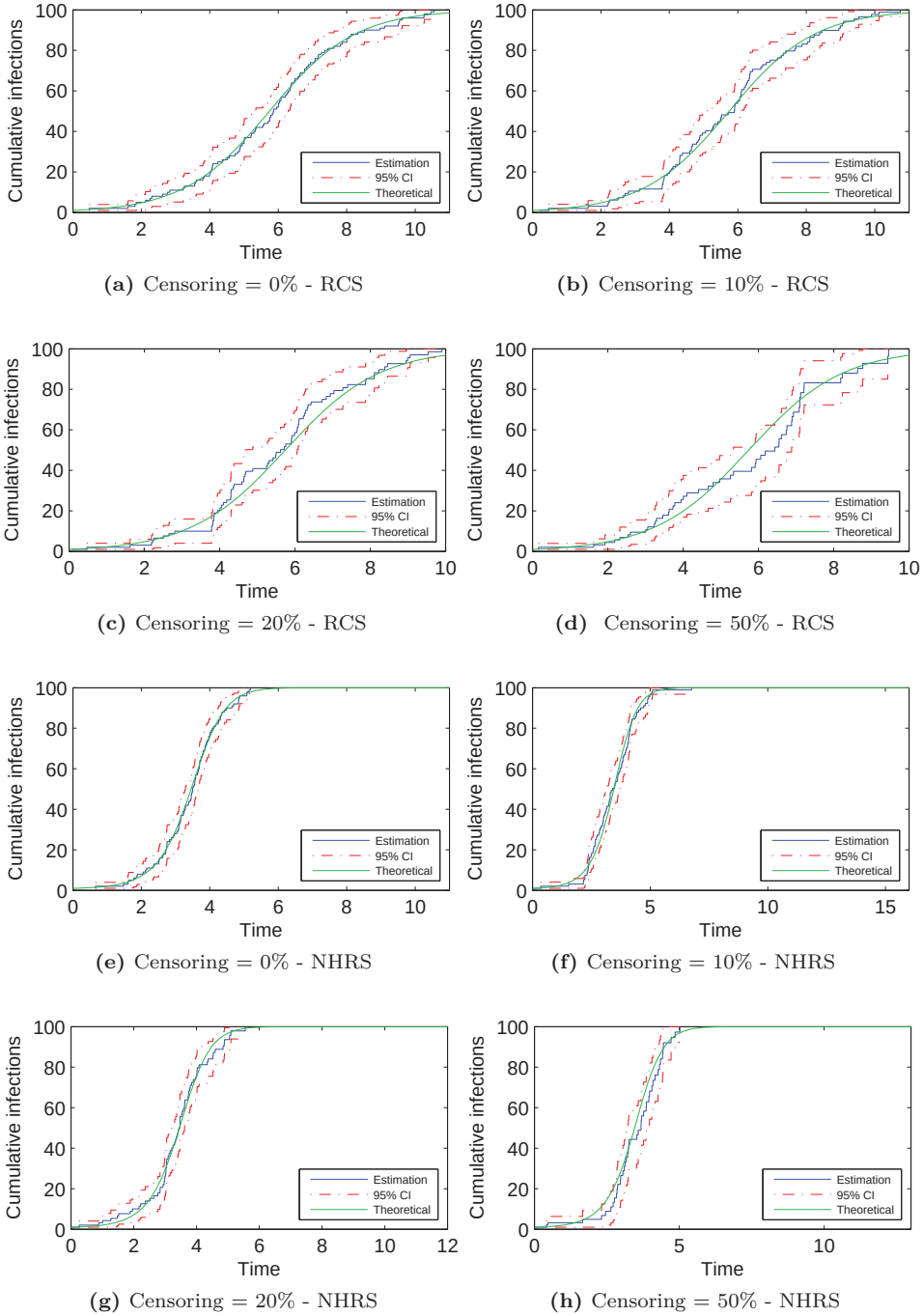


(a) $N = 100, I_0 = 1$ **(b)** $N = 1000, I_0 = 10$

Figure 2.10: Simulation of RCS and NHRS models for $N = 100, I_0 = 1$ **and** $N = 1000, I_0 = 10$**.** Blue curves correspond to the RCS model ($\beta = 0.8$) and red curves to the NHRS model ($\beta = 0.8$ and $K = 1.5$).

Let us consider $\hat{I}_{\mathrm{km}}(t)$. Figure 2.11 gives a representation of the estimator $\hat{I}_{\mathrm{km}}(t)$ of $I(t)$ with different level of censoring compared with the theoretical epidemic model $I(t)$. Recall that this estimation is based on a Kaplan-Meier estimation of the cumulative distribution function $F(t)$.

Figure 2.11 shows the bounds of a 95% confidence interval which are obtained thanks to the Greenwood estimator of the variance. Only the case $N = 100$ was displayed. The particular case $N = 1000$ didn't give relevant information since both the estimation and theoretical curves were mixed up.



(a) Censoring = 0% - RCS **(b)** Censoring = 10% - RCS

(c) Censoring = 20% - RCS **(d)** Censoring = 50% - RCS

(e) Censoring = 0% - NHRS **(f)** Censoring = 10% - NHRS

(g) Censoring = 20% - NHRS **(h)** Censoring = 50% - NHRS

Figure 2.11: Simulated examples. Plot of $I(t)$, $\hat{I}_{\mathrm{km}}(t)$ and the pointwise 95% confidence bounds under the RCS model (with $\beta = 0.8$) and the NHRS model (with $\beta = 0.8$ and $K = 1.5$) with 0%, 10%, 20% and 50% of censoring and $(N, I_0) = (100, 1)$.

As an indicator of the quality of estimation, we consider the mean integrated squared error (MISE) for the different estimation methods of the function $I(\cdot)$. The MISE of an estimator $\hat{I}(\cdot)$ of $I(\cdot)$ is defined by:

$$\mathbb{E}\left[\|\hat{I} - I\|_2^2\right] = \mathbb{E}\left[\int_0^{+\infty} \left(\hat{I}(t) - I(t)\right)^2 \mathrm{d}t\right] \tag{2.12}$$

where $\|\cdot\|_2$ is the L^2 norm. One can give an empirical estimation of (2.12) by computing the arithmetic

mean over m simulations of $\|\hat{I}(\cdot) - I(\cdot)\|_2^2$, i.e:

$$\frac{1}{m}\sum_{i=1}^{m}\left[\int_0^{+\infty}\left(\hat{I}_i(t) - I(t)\right)^2 \mathrm{d}t\right]$$

where $\hat{I}_i(t)$ is the estimator of $I(t)$ obtained on the i^{th} simulation.

First, let us recall the expression of the four estimators of $I(t)$ defined Section 2.2:

$$\begin{aligned}
\hat{I}_{\text{ls}_1}(t) &= \frac{N}{1+\psi\exp\left(-\hat{B}_{\text{ls}_1}(t)\right)} \\
\hat{I}_{\text{km}}(t) &= N\left(\frac{1+\psi\hat{F}(t)}{1+\psi}\right) \\
\hat{I}_{\text{ls}_2}(t) &= \frac{N}{1+\psi\exp\left(-\hat{B}_{\text{ls}_2}(t)\right)} \\
\hat{I}_{\text{ml}}(t) &= \frac{N}{1+\psi\exp\left(-\hat{B}_{\text{ml}}(t)\right)}.
\end{aligned}$$

By following the simulation procedure defined in Section 2.3 for different values of N and n, we can compute the MISE of the different estimators over $m = 1000$ simulations. Hence, Tables 2.12 and 2.13 exhibit the values of the MISE under the RCS ($\beta = 0.8$) and the NHRS ($\beta = 0.8$ and $K = 1.5$) models respectively.

Recall that only the estimators $\hat{I}_{\text{ls}_1}(t)$ and $\hat{I}_{\text{ml}}(t)$ of $I(t)$ have been introduced when dealing with 100 independent populations, thus values of the MISE are missing for $\hat{I}_{\text{km}}(t)$ and $\hat{I}_{\text{ls}_2}(t)$ in Tables 2.12 and 2.13 when $n = 100$.

Since $I(t) \in [0, N]$, one can notice that the value of the MISE increases with the value of N. As expected, the MISE of each estimator increases with the percentage of censoring and decreases when the value n increases.

Considering Table 2.12, one can notice that the MISE of the maximum likelihood estimator $\hat{I}_{\text{ml}}(\cdot)$ is the best under the RCS model when dealing with no censored observations and whatever the values of N and n are. However, when censored data are involved, the MISE of the least squares estimator $\hat{I}_{\text{ls}_2}(\cdot)$ is the lowest.

Similar comments can be made when considering Table 2.13 which deals with the NHRS model. As long as no censored data are considered, the MISE of $\hat{I}_{\text{ml}}(\cdot)$ is the best but $\hat{I}_{\text{ls}_2}(\cdot)$ is doing better when censored data are involved.

Table 2.12: Simulation results for the RCS model ($\beta = 0.8$). Empirical estimation of the MISE of the different estimations of $I(t)$ for different values of N, n and different percentages of censoring.

	$N = 100, I_0 = 1$		$N = 1000, I_0 = 10$		
	$n = 1$	$n = 100$	$n = 1$	$n = 100$	Censoring (%)
$\hat{I}_{\mathrm{ls}_1}(\cdot)$					
	127.4	1.2	1 116	11.8	0%
$\hat{I}_{\mathrm{km}}(\cdot)$					
	127.9	–	1 205	–	0%
	240.4	–	6 022	–	10%
	519.1	–	25 374	–	20%
	4 208	–	380 480	–	50%
$\hat{I}_{\mathrm{ls}_2}(\cdot)$					
	135.7	–	1 243	–	0%
	151.0	–	1 393	–	10%
	170.5	–	1 611	–	20%
	305.4	–	2 647	–	50%
$\hat{I}_{\mathrm{ml}}(\cdot)$					
	51.9	0.6	559.7	5.5	0%
	62.3	1.1	1 294	9.7	10%
	100.1	27.1	3 839	306.9	20%
	460.9	68.2	33 694	1 256	50%

Table 2.13: Simulation results for the NHRS model ($\beta = 0.8$ and $K = 1.5$). Empirical estimation of the MISE of the different estimations of $I(t)$ for different values of N, n and different percentages of censoring.

	$N = 100, I_0 = 1$		$N = 1000, I_0 = 10$		
	$n = 1$	$n = 100$	$n = 1$	$n = 100$	Censoring (%)
$\hat{I}_{\mathrm{ls}_1}(\cdot)$					
	104.5	0.8	934.4	7.5	0%
$\hat{I}_{\mathrm{km}}(\cdot)$					
	54.8	–	651.8	–	0%
	95.8	–	1 602	–	10%
	168.0	–	7 714	–	20%
	1 383	–	117 750	–	50%
$\hat{I}_{\mathrm{ls}_2}(\cdot)$					
	96.0	–	941.6	–	0%
	99.2	–	1 075	–	10%
	123.1	–	1 117	–	20%
	265.5	–	3 124	–	50%
$\hat{I}_{\mathrm{ml}}(\cdot)$					
	31.4	0.3	305.5	3.0	0%
	37.8	0.42	510	6.1	10%
	46.1	1.2	1 140	18.4	20%
	149	2.4	8 658	102.6	50%

2.5 Applications on real data sets

We study two datasets to give a concrete illustration of the different estimators. First, we use the dataset of the Code Red v2 (CRv2) epidemic used by Kirmani et al. (2010). Then we use a dataset of the reported AIDS (acquired immune deficiency syndrome) cases in France.

2.5.1 Application to the Code Red v2 worm propagation

Let us consider in this section the Code Red v2 (CRv2) dataset used by Moore et al. (2002). This dataset, published on the website of Caida[1], presents the propagation of the CRv2 worm during approximately 30 hours. Code Red v2 is a computer worm attacking computers running Microsoft's ISS web server which infected more than 359000 computers mainly in the United States, Korea, China and Taiwan during the 19^{th} and the 20^{th} of July, 2001.

The dataset provided on the website of Caida is merged from three different sources which recorded the spread of CRv2 starting from the 19^{th} of July at midnight UTC to the 20^{th} July at 7 UTC in the morning. That dataset consists of a count of each single one infection with the corresponding relative time (in seconds) at infection. Figure 2.12 displays a sketch of the number of distinct IP addresses from $t = 0$ to $t = 30$ where t is in hours.

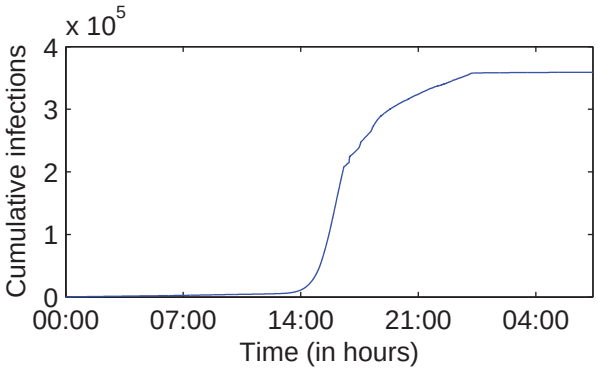


Figure 2.12: Cumulative number of unique IP addresses infected by the Code Red v2 worm.

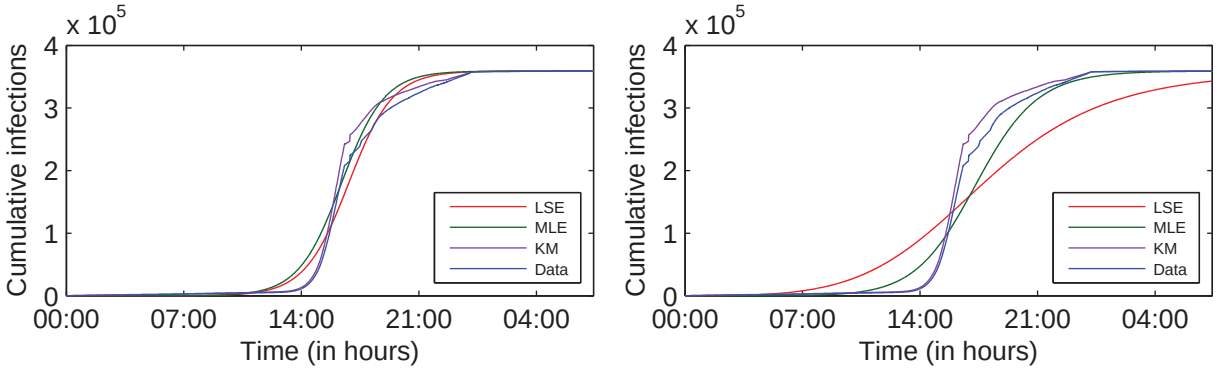
We now estimate the parameters of the RCS and the NHRS models by using the estimation methods introduced earlier. From Table 2.14, one can notice that the least squares estimators and the maximum likelihood estimators give similar values of $\hat{\beta}$ for the RCS model . However, we denote some differences, especially for $\hat{\beta}$, between the two methods when dealing with the NHRS model. Those differences are illustrated in Figure 2.13.

Table 2.14: Estimation results for the RCS model and the NHRS model. Empirical estimation of the parameters β for the RCS model and (β, K) for the NHRS model.

		NHRS model	
	RCS model	β	K
Least squares estimator	2.1×10^{-4}	1.4×10^{-2}	3.8×10^{-1}
Maximum likelihood estimator	2.2×10^{-4}	5.0×10^{-4}	7.4×10^{-1}

Figure 2.13 gives an illustration of the estimation of the RCS and NHRS models for the different estimators. One can notice that the three estimation methods give a quite similar estimation of the RCS model. In contrast, we denote that the least squares technic differs from the two others methods when considering the NHRS model.

[1] http://www.caida.org/home/



(a) RCS model

(b) NHRS model

Figure 2.13: Code Red v2 worm propagation dataset. Plot of $X(t)$, $\hat{I}_{\mathrm{ls}_1}(t)$, $\hat{I}_{\mathrm{ml}}(t)$ and $\hat{I}_{\mathrm{km}}(t)$ under the RCS model (a) and the NHRS model (b).

2.5.2 Application to the human immunodeficiency virus (HIV) transmission

We now focus on the spread of HIV in France. In order to do this, we refer to the dataset of yearly reported AIDS cases from 1982 to 2011 published on the website of Eco-Santé[2]. Since AIDS is an advanced stage of HIV infection and by assuming a constant AIDS incubation period, we may consider that the evolution of AIDS incidence reflects quite well the number of new HIV infections per year.

Since the dataset only gives the incidence of AIDS cases, we reconstruct the prevalence curve by summing the values from previous years. Thus, Figure 2.14 displays the prevalence curve of AIDS cases from an initial number of $I_0 = 31$ cases in 1982 to a final number of $N = 67508$ cases in 2011.

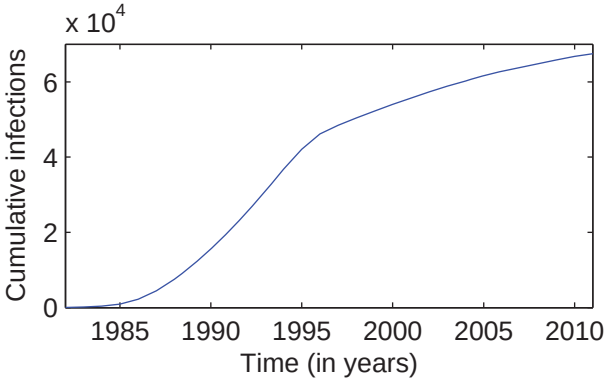


Figure 2.14: Cumulative number of AIDS cases in France.

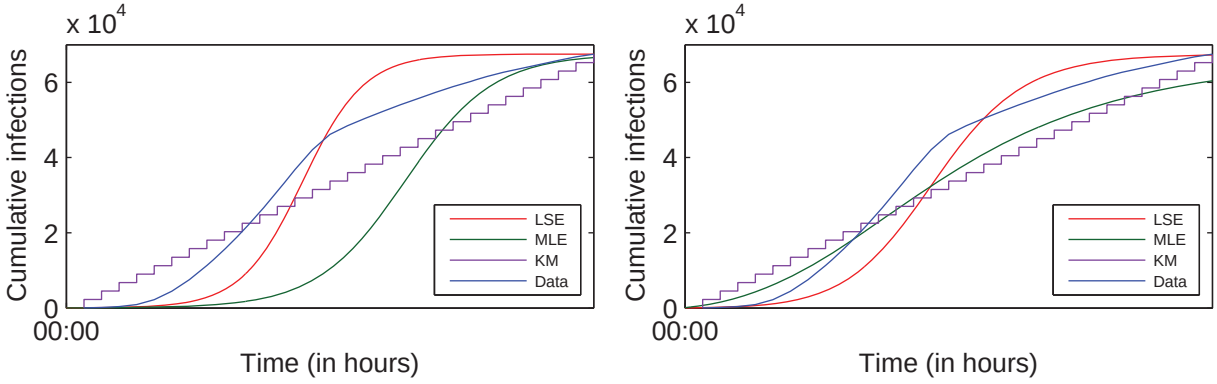
Next, we estimate the parameters of the RCS model and the NHRS model by computing the same estimation methods as used in Section 2.5.1. From Table 2.15, one can denote that the least squares estimator and the maximum likelihood estimator give different estimations for both RCS and NHRS models. As shown in Table 2.14, the higher difference still comes from $\hat{\beta}$ in the particular case of the NHRS model.

Figure 2.15 gives an illustration of the estimation of the RCS and NHRS models for the different estimators. One can notice that for both RCS and NHRS model, estimations are quite difficult whatever the estimation method. It is especially true for the Kaplan-Meier estimator. We may conclude that both RCS and NHRS models aren't suitable to fit that dataset.

[2]http://www.ecosante.fr

Table 2.15: Estimation results for the RCS model and the NHRS model. Empirical estimation of the parameters β for the RCS model and (β, K) for the NHRS model.

		NHRS model	
	RCS model	β	K
Least squares estimator	5.7×10^{-1}	1.2	7.3×10^{-1}
Maximum likelihood estimator	4.0×10^{-1}	3.0×10	3.4×10^{-1}



(a) RCS model **(b)** NHRS model

Figure 2.15: Reported AIDS cases dataset. Plot of $X(t)$, $\hat{I}_{\mathrm{ls}_1}(t)$, $\hat{I}_{\mathrm{ml}}(t)$ and $\hat{I}_{\mathrm{km}}(t)$ under the RCS model (a) and the NHRS model (b).

2.6 Conclusions

We investigated several estimation methods for the well known logistic model of worm propagation. First, we studied the relative bias and the relative standard deviation of the euclidean parameters in Section 2.3. Under the RCS and the NHRS models, the maximum likelihood estimator provided almost always better estimations than the least squares estimator for both the relative bias and the relative standard deviation. The least squares estimators were doing better only in the particular case of a high amount of censoring. Then, by referring to the study of the mean integrated squared error (MISE) in Section 2.4, our study demonstrated again that the maximum likelihood estimator was the best estimator as long as a relatively low amount of censoring is involved. However, when the amount of censoring increases, the least squares estimator tends to be more reliable. Finally, we applied those estimation methods to a real dataset of the Code Red v2 worm propagation. In the particular case of the RCS model, the different estimation methods gave similar results. Considering the NHRS model, results were more contrasted with a fairly poor fit for the least squares estimator. However, considering the HIV dataset, all the methods demonstrated a poor fit to the epidemic curve exhibiting the fact that both models were not suitable for this dataset.

The studied estimators have proved to be quite efficient tools to deal with the estimation of epidemic models' parameters and could be applied to much wider problems such as models of the propagation of communicable diseases in populations to support health care decision makers. However, some issues are still to be addressed. Indeed, the development of statistical methods to assess whether the model is in line with the nature of the studied data should be considered. Moreover, one could find of interest to theoretically study the asymptotic properties of the different estimators.

3 Some models of back-calculation for the hepatitis C virus infection incidence in France

Français

L'hépatite C étant une maladie asymptomatique, l'estimation de sa prévalence reste aujourd'hui un défi pour les responsables des politiques de santé. En faisant l'hypothèse que le taux d'incidence annuel du virus de l'hépatite C (VHC) suit une distribution de Poisson, nous traitons dans cette partie deux méthodes de rétro-calcul basées sur l'estimation par maximum de vraisemblance et sur l'algorithme EM. En ce basant sur un jeu de données du nombre de décès dus au carcinome hépatocellulaire (CHC) suivi d'un travail de nettoyage de données pour s'assurer de la pertinence de celles-ci, les deux méthodes de rétro-calcul sont comparées avec l'approche proposée par Deuffic et al. Nos résultats montrent que la méthode basée sur l'algorithme EM démontre de meilleures performances par rapport aux autres et offre des résultats en accord avec la prévalence estimée du VHC en France.

English

Since hepatitis C is an asymptomatic disease, its prevalence estimation remains a challenge for health care policy makers. Assuming that hepatitis C virus (HCV) annual incidence rates follow a Poisson distribution, we consider in this part two back-calculation methods based on the maximum likelihood approach and on the EM-algorithm. Using a CépiDc's database on the number of deaths related to hepatocellular carcinoma (HCC) and after a large data cleansing activity to ensure for the data's relevance, both back-calculation methods are compared with the approach proposed by Deuffic et al. It is found that the method based on the EM-algorithm demonstrates better performances than other approaches and offers consistent results with the estimated HCV prevalence in France.

Contents

Résumé **47**
3.1 Introduction **47**
3.2 Models and estimation methods **48**
3.2.1 Model 48
3.2.2 Maximum likelihood estimation 49
3.2.3 Expectation maximization smoothing algorithm 50
3.2.4 Precision 51
3.3 The CépiDc dataset of death by HCC in France between 1979 and 2012 . **52**
3.3.1 Data 52
3.3.2 Natural history model 54
3.3.3 Lifetime distribution of an individual who develop and die of HCC 56
3.3.4 Distribution of the time to death from HCC in the overall population 57
3.3.5 Simplified Deuffic et al. model 58
3.3.6 Quality of the models 59
3.4 Results **60**
3.5 Discussion **65**

Résumé

Ce troisième chapitre se propose d'étudier plusieurs modèles de rétro-calcul pour l'estimation de l'incidence du virus de l'hépatite C (VHC) en France. Les résultats de ces modèles sont notamment comparés avec l'approche de référence en France développée par Deuffic-Burban et al. (1999) et basée sur un modèle logistique.

L'hépatite C est une maladie asymptomatique dont l'une des conséquences est la survenue, au bout de quelques décennies, d'un carcinome hépatocellulaire (CHC) dont le pronostic médical est souvent fatal. En se basant d'une part sur les données des décès liés au CHC et d'autre part sur le temps entre l'infection par le VHC et le décès lié au CHC, nous nous proposons de développer des modèles de rétro-calcul afin de remonter à l'incidence passée du VHC.

L'hypothèse générale autour de laquelle s'articule ce chapitre suppose que le nombre d'infections du au VHC peut être modélisé via des distributions de Poisson. Ainsi, en notant Y_j le nombre de décès dus au CHC l'année j et X_{ij} le nombre de contamination par le VHC l'année i dont le décès par CHC s'est produit l'année j, on écrit :

$$\mathbb{E}(Y_j) = \sum_i \theta_i f_{j-i}$$

où θ_i est le nombre moyen de nouvelles infections l'année i et f_{j-i} la probabilité qu'un individu infecté par le VHC l'année i décède d'un CHC l'année j.

Dans le but d'estimer θ_i, nous nous intéressons dans un premier temps à deux modèles paramétriques (modèle logistique et modèle exponentiel gaussien) et à la méthode du maximum de vraisemblance.

Puis, afin de s'affranchir du biais généré par ces modèles paramétriques, nous nous orientons vers une approche basée sur l'algorithme d'espérance maximisation lissé (EMS pour expectation maximization smoothing, Silverman and Nychka (1990)) extension directe de l'algorithme d'espérance maximisation (EM) développé par Dempster et al. (1977).

Dans ce chapitre, le traitement des données prend une place prépondérante. En effet, suite à un changement de CIM (classification internationale des maladies) en 1999, les données de décès liés au CHC ont été reportées différemment avant et après 1999 entraînant un décalage conséquent entre les données de décès pré-1999 et post-1999. Nous parvenons ainsi à obtenir une série chronologique des décès par CHC lié au VHC cohérente et supposons que ce nombre de décès est nul en France avant 1920.

Nous nous proposons ensuite de donner une estimation de f_{j-i} à partir d'un modèle d'histoire naturelle de la maladie. Ce modèle de type chaîne de Markov à temps discrets est construit à partir de données de la littérature, Deuffic-Burban et al. (1999) principalement, et permet de donner une estimation de f_{j-i} en calculant le temps moyen passé par un individu dans la chaîne de Markov.

Nous nous focalisons ensuite sur le développement d'une version simplifiée du modèle de Deuffic-Burban et al. (1999). Les performances de ces modèles sont directement mesurées par l'intermédiaire du critère des moindres carrés.

Les résultats obtenus à partir de nos trois modèles sont satisfaisants avec un avantage net pour le modèle basé sur l'algorithme d'espérance maximisation lissé. En effet, ces résultats sont à la fois en accord avec les résultats obtenus par le modèle simplifié de Deuffic-Burban et al. (1999) en ce qui concerne l'estimation de l'incidence passée du VHC en France ainsi qu'avec l'estimation de la prévalence réalisée par l'INVS pour l'année 2004.

3.1 Introduction

In the last decades, hepatitis C virus (HCV) has become a major threat to public health with an estimated 80 million chronic carriers worldwide (see Gower et al. (2014)). Leading to severe end-stage liver complications such as liver cirrhosis, hepatic failure and hepatocellular carcinoma (HCC), HCV is responsible for hundreds of thousands of deaths each year. In France, the prevalence of anti-HCV antibodies has been estimated at 0.84% (95%CI:0.65-1.10) in 2004 with a proportion of individuals aware of their sero-status estimated at 57% (95%CI:43-71) (see Institut de veille sanitaire (2007)). As new therapeutics

emerge, establishing successful strategies to eliminate the virus requires to estimate the total number of infected individuals and to identify the stage of fibrosis.

To address this issue, one might use mathematical technics, such as back-calculation, to reconstruct the past incidence and model the future burden of the disease (see Deuffic-Burban et al. (1999); Griffiths and Nix (2002); McEwan et al. (2014); Sweeting et al. (2007)). Basically, back-calculation is based on the use of actual observed data (i.e death certificates, diagnoses, etc) and on the assessment of the time between infection and those data to estimate the number of infections that took place in the past. Considering the particular case of France and assuming that the past incidence of HCV is a logistic function, Deuffic-Burban et al. (1999) proposed an efficient back-calculation approach based on the weighted least squares criterion and reported HCC deaths to estimate that past incidence.

In 2014, to estimate the past incidence of HCV in Taiwan, McEwan et al. (2014) used an approach based on the human immunodeficiency virus (HIV) past incidence reconstruction (see Bacchetti et al. (1993); Becker and Marschner (1993); Brookmeyer and Gail (1994)) assuming all HCV infections were independent random variables following Poisson distributions. This approach makes use of the expectation-maximization-smoothing (EMS) algorithm by Silverman and Nychka (1990), based on the original expectation maximization (EM) algorithm by Dempster et al. (1977), and does not make assumptions on the shape of the past incidence curve.

The objective of this chapter is to propose an adaptation of the approach developed by McEwan et al. (2014) to the French national data of the observed number of deaths due to HCC, to estimate the past incidence of the chronic hepatitis C virus infection in France. Along with this, we introduce some parametric models based on well-known epidemic models from the literature to give an alternative approach of the work of McEwan et al. (2014). After a large work of formatting the data, to ensure that we only focus on HCC deaths, comparisons will be made with the model developed by Deuffic-Burban et al. (1999).

This chapter is organized as follows: Section 2 provides an overview of the model studied and the different estimation methods considered. The presentation of the dataset and of the lifetime distribution are given in Section 3. Then, Section 4 exhibits our results and compares the different models. Finally, Section 5 presents our conclusions.

3.2 Models and estimation methods

In this part, we propose to model the past incidence of HCV via Poisson distributions. Several estimation methods will be introduced to make some comparisons with the model developed by Deuffic-Burban et al. (1999) on the estimation of the past infections of HCV.

3.2.1 Model

Let us denote $\{X_{ij}\}$ the unobserved number of people infected in year i who died of HCC related to HCV in year j, for $i \leqslant j$. For the sake of practicality, we assume that $j - i \leqslant 100$. We denote $\{X_i\}$ the number of people infected in year i which leads to $X_i = \sum_{j=i}^{i+100} X_{ij}$.

We assume that the random variables $\{X_{ij}\}$ are realizations of independent Poisson variables with parameters $\theta_i f_{ij}$ where θ_i is the mean number of new infections in year i and f_{ij} the probability that an individual infected by HCV in year i dies of HCC related to HCV in year j.

We assume that f_{ij} remains the same for each individual and does not vary over time. For more convenience, we change the notation f_{ij} to f_{j-i} where f_{j-i} is the probability that the duration between the infection of a given individual and his death by HCC is $j - i$. In the following pages, we refer to f as lifetime distribution.

However, as $\{X_i\}$ and $\{X_{ij}\}$ are unobserved data, we introduce the observed data $\{Y_j\}$ which is the number of people who died from HCC related to HCV in year j, $-\infty < j \leqslant T$ where T corresponds to the last year of observation. In practice, we have $t_0 \leqslant j \leqslant T$ where t_0 is the time of the beginning of the epidemic. Thus, one can write:

$$Y_j = \sum_{i=t_{0j}}^{j} X_{ij}.$$

where $t_{0j} = \max(t_0, j - 100)$. Since we assumed that each infected individual is assigned a lifetime independently and that the number of new infections is independent from one year to another, the random variable $\{Y_j\}$ are independent Poisson random variables with mean:

$$\mathbb{E}(Y_j) = \sum_{i=t_{0j}}^{j} \theta_i f_{j-i}. \tag{3.1}$$

In the next section, we propose to set several parametric models over the θ_i to give an estimation of the past incidence of HCV based on well-known epidemic models of the literature.

3.2.2 Maximum likelihood estimation

To give an estimation of θ_i using the maximum likelihood estimation method, we compute the likelihood of $\{Y_j\}$ with parameter $\theta = (\theta_{t_0}, \ldots, \theta_T)$:

$$L(Y;\theta) = \prod_{j=t_0}^{T} \left[\frac{1}{y_j!} \left(\sum_{i=t_{0j}}^{j} \theta_i f_{j-i} \right)^{y_j} \exp\left(- \sum_{i=t_{0j}}^{j} \theta_i f_{j-i} \right) \right].$$

It follows that the log-likelihood of $\{Y_j\}$ with parameter θ is:

$$l(Y;\theta) = \sum_{j=t_0}^{T} \left[y_j \times \ln\left(\sum_{i=t_{0j}}^{j} \theta_i f_{j-i} \right) - \sum_{i=t_{0j}}^{j} \theta_i f_{j-i} - \ln(y_j!) \right]. \tag{3.2}$$

We will now introduce two different parametric models for θ. Let us first consider the logistic model for population dynamics introduced by Verhulst (1838) with equation:

$$f_L(t; K, \alpha, \beta) = \frac{K}{1 + e^{-\alpha(t-\beta)}}, \tag{3.3}$$

where K is the plateau value as $t \to \infty$, the parameter α is the growth rate and β a translation parameter. To allow for a general decrease of new infected cases corresponding to the elimination of transfusion-associated HCV infections after 1989, we modify (3.3) by assuming a reduction of 40% of new cases, exactly as Deuffic-Burban et al. (1999) did. Thus, we have:

$$f_{\mathcal{L}}(t; K, \alpha, \beta) = \begin{cases} \dfrac{K}{1 + e^{-\alpha(t-\beta)}}, & \text{for } t < 1990 \\ 0.6 \times \dfrac{K}{1 + e^{-\alpha(t-\beta)}}, & \text{otherwise.} \end{cases}$$

Next, by assuming a declining number of infected cases after a peak of infection, we introduce a model based on an exponentially modified Gaussian (EMG) function (see Grushka (1972)):

$$f_E(t; H, \lambda, \mu, \sigma) = \frac{H\lambda}{2} e^{\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2t)} \operatorname{ercf}\left(\frac{\mu + \lambda\sigma^2 - t}{\sqrt{2}\sigma} \right),$$

where H is a scale parameter and ercf the complementary error function defined as:

$$\operatorname{ercf}(t) = \frac{2}{\sqrt{\pi}} \int_t^{\infty} e^{-x^2} \mathrm{d}x.$$

By replacing the expression of θ_i in equation (3.2) by $f_{\mathcal{L}}(i;\cdot)$ or $f_E(i;\cdot)$, one can estimate the model parameters by maximizing the following likelihoods:

$$\begin{aligned} l_{\mathcal{L}}(Y; K, \alpha, \beta) = \sum_{j=t_0}^{T} \Bigg[& y_j \times \ln\left(\sum_{i=t_{0j}}^{j} \left(\frac{K f_{j-i}}{1 + e^{-\alpha(i-\beta)}} \cdot^{\bullet}_{\{i<1990\}} + \frac{0.6 \times K f_{j-i}}{1 + e^{-\alpha(i-\beta)}} \cdot^{\bullet}_{\{i \geqslant 1990\}} \right) \right) \\ & - \sum_{i=t_{0j}}^{j} \left(\frac{K f_{j-i}}{1 + e^{-\alpha(i-\beta)}} \cdot^{\bullet}_{\{i<1990\}} + \frac{0.6 \times K f_{j-i}}{1 + e^{-\alpha(i-\beta)}} \cdot^{\bullet}_{\{i \geqslant 1990\}} \right) - \ln(y_j!) \Bigg], \end{aligned}$$

$$l_E(Y;H,\lambda,\mu,\sigma) = \sum_{j=t_0}^{T}\left[y_j \times \ln\left(\sum_{i=t_{0j}}^{j}\frac{H\lambda}{2}e^{\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2i)}\,\mathrm{ercf}\left(\frac{\mu+\lambda\sigma^2-i}{\sqrt{2}\sigma}\right)f_{j-i}\right) - \sum_{i=t_{0j}}^{j}\frac{H\lambda}{2}e^{\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2i)}\,\mathrm{ercf}\left(\frac{\mu+\lambda\sigma^2-i}{\sqrt{2}\sigma}\right)f_{j-i} - \ln(y_j!)\right].$$

Let us denote $\hat{K}$, $\hat{\alpha}$, $\hat{\beta}$, $\hat{H}$, $\hat{\lambda}$, $\hat{\mu}$ and $\hat{\sigma}$ the maximum likelihood estimators of K, α, β, H, λ, μ and σ obtained by maximization of the above likelihoods. Then we can deduce maximum likelihood estimators of θ given by:

$$\hat{\theta}_{i\mathcal{L}} = \frac{\hat{K}}{1+e^{-\hat{\alpha}(i-\hat{\beta})}}\cdot{}_{\bullet\,\{i<1990\}} + \frac{0.6\times\hat{K}}{1+e^{-\hat{\alpha}(i-\hat{\beta})}}\cdot{}_{\bullet\,\{i\geqslant 1990\}}$$

and

$$\hat{\theta}_{iE} = \frac{\hat{H}\hat{\lambda}}{2}e^{\frac{\hat{\lambda}}{2}(2\hat{\mu}+\hat{\lambda}\hat{\sigma}^2-2i)}\,\mathrm{ercf}\left(\frac{\hat{\mu}+\hat{\lambda}\hat{\sigma}^2-i}{\sqrt{2}\hat{\sigma}}\right).$$

We will compare the models introduced previously with a simplified version of the model developed by Deuffic-Burban et al. (1999) (see Section 3.3.5).

However, setting parametric models on θ_i, induce some bias inherent to the models. Indeed, by choosing a logistic or an exponentially modified Gaussian model, we constraint θ_i to a particular shape. Thus, to address this issue, we propose to use an estimation procedure based on the expectation maximization smoothing (EMS) algorithm developed by Silverman and Nychka (1990).

3.2.3 Expectation maximization smoothing algorithm

To give an estimation of θ, one can apply the EMS algorithm on the unobserved complete data $\{X_{ij}\}$, for $t_0 \leqslant i \leqslant j \leqslant T$. As $\{X_{ij}\}$ follows a Poisson distribution with mean $\theta_i f_{j-i}$, one can compute the log-likelihood of $\{X_{ij}\}$ given θ:

$$l(X;\theta) \propto \sum_{i=t_0}^{T}\sum_{j=i}^{T_{\max,i}}\left[x_{ij}\times\ln\left(\theta_i f_{j-i}\right)-\theta_i f_{j-i}\right], \tag{3.4}$$

where $T_{\max,i} = \min(i+100, T)$. Then, we compute the conditional expectation of (3.4) given $\{Y_j\}$, for $t_0 \leqslant j \leqslant T$, and $\tilde{\theta}^{(p)}$:

$$\mathbb{E}\left(l(X;\theta)|\{Y_j=y_j; j=t_0,\ldots,T\},\tilde{\theta}^{(p)}\right) \propto \sum_{i=t_0}^{T}\sum_{j=i}^{T_{\max,i}}\left(\mathbb{E}\left(X_{ij}|\{Y_j=y_j; j=t_0,\ldots,T\},\tilde{\theta}^{(p)}\right) \times\ln\left(\theta_i f_{j-i}\right)-\theta_i f_{j-i}\right), \tag{3.5}$$

where $\tilde{\theta}^{(p)}$ is the estimate of θ at the p-th iteration. From (3.5) we deduce that the E-step of the EM algorithm is reduced to computing $\mathbb{E}\left(X_{ij}|\{Y_j=y_j; j=t_0,\ldots,T\},\tilde{\theta}^{(p)}\right)$.

Since for independent Poisson variables X and Y with parameters θ_X and θ_Y, we have:

$$\mathbb{E}(X|X+Y) = (X+Y)\frac{\theta_X}{\theta_X+\theta_Y},$$

we easily deduce that:

$$\mathbb{E}\left(X_{ij}|\{Y_j=y_j; j=t_0,\ldots,T\},\tilde{\theta}^{(p)}\right) = y_j \times \frac{\tilde{\theta}_i^{(p)} f_{j-i}}{\sum_{k=t_{0j}}^{j}\tilde{\theta}_k^{(p)} f_{j-k}}.$$

Next, we compute the M-step by deriving (3.5) with respect to θ_i and setting the derivative equal to 0.

$$\sum_{j=i}^{T_{\max,i}} \left[y_j \times \frac{\tilde{\theta}_i^{(p)} f_{j-i}}{\sum_{k=t_{0j}}^{j} \tilde{\theta}_k^{(p)} f_{j-k}} \times \frac{f_{j-i}}{\tilde{\theta}_i f_{j-i}} - f_{j-i} \right] = 0, \text{ for } i = t_0, \ldots, T.$$

Thus, the estimate of θ_i at the $(p+1)^{\text{th}}$ iteration is given by:

$$\tilde{\theta}_i^{(p+1)} = \frac{1}{\sum_{j=i}^{T_{\max,i}} f_{j-i}} \sum_{j=i}^{T_{\max,i}} \left[y_j \times \frac{\tilde{\theta}_i^{(p)} f_{j-i}}{\sum_{k=t_{0j}}^{j} \tilde{\theta}_k^{(p)} f_{j-k}} \right], \text{ for } i = t_0, \ldots, T. \tag{3.6}$$

However, such an estimation technic leads to a very unstable estimation of θ_i contrasting with the relative smoothness of the curve of incident cases per year. Indeed, as highlighted by Becker and Marschner (1993), the unsmoothed estimate tends to have some erratic pattern. Hence, we add an extra step to the EM algorithm by applying a moving average on the $\tilde{\theta}_i$ provided by equation (3.6). We write:

$$\hat{\theta}_i^{(p+1)} = \sum_{l=0}^{L} \omega_l \times \tilde{\theta}_{i+l-L/2}^{(p+1)} \tag{3.7}$$

where L is the order of the moving average and the ω_l are symmetric weights such as $\sum_l \omega_l = 1$. Referring to Bacchetti et al. (1993); Becker and Marschner (1993); Brookmeyer and Gail (1994); Silverman and Nychka (1990), we rely on a weighted moving average by setting ω_l proportional to the binomial coefficients. Hence, we have:

$$\omega_l = \frac{1}{2^L} \times \binom{L}{l}, \quad l = \{0, 1, \ldots, L\},$$

which is known as binomial filter with bandwidth L. The value obtained in equation (3.7) is then reused in a new iteration of the EMS algorithm.

Finally, we define a criterion for convergence for the EMS algorithm. We choose ε as small as possible such that:

$$\frac{\left\| \hat{\theta}^{(p+1)} - \hat{\theta}^{(p)} \right\|_2}{\left\| \hat{\theta}^{(p)} \right\|_2} < \varepsilon$$

where $\|\cdot\|_2$ is the L^2 norm in $\mathbb{R}^T$. In the following pages, we choose $\varepsilon = 10^{-4}$ and denote $\hat{\theta}$ the resulting estimator.

3.2.4 Precision

Dealing with a large number of parameters and a smoothing step which leads to additional uncertainties, one can compute a confidence interval to get additional information on the precision of the estimation. Following the procedure proposed by Becker and Marschner (1993), we generate a point-wise bootstrap confidence interval for $\hat{\theta}$.

Firstly, for $t_0 \leqslant j \leqslant T$, we compute N realizations $Y_j^{(1)}, \ldots, Y_j^{(N)}$ from a Poisson distribution with mean

$$\sum_{i=t_{0j}}^{j} \tilde{\theta}_i f_{j-i}$$

where $\tilde{\theta}_i$ is the unsmoothed estimator of θ_i obtained in equation (3.6). We use the unsmoothed estimator of θ_i to compensate for the uncertainty brought by the density estimation obtained via the kernel-smoothing method (see Becker and Marschner (1993)).

Secondly, we apply the EMS algorithm to the N realizations $Y_j^{(1)}, \ldots, Y_j^{(N)}$, for $t_0 \leqslant j \leqslant T$. We obtain N new smoothed estimators of θ_i denoted $\hat{\theta}_i^{(1)}, \ldots, \hat{\theta}_i^{(N)}$, for $t_0 \leqslant i \leqslant T$.

Finally, we sort $\hat{\theta}_i - \hat{\theta}_i^{(1)}, \ldots, \hat{\theta}_i - \hat{\theta}_i^{(N)}$ in ascending order and compute $\mu_{i,1-\alpha/2}$ (resp. $\mu_{i,\alpha/2}$) the $1-\alpha/2$ (resp. $\alpha/2$) percentile of the bootstrap distribution of $\hat{\theta}_i$. Thus, $\left[\hat{\theta}_i - \mu_{i,1-\alpha/2}; \hat{\theta}_i - \mu_{i,\alpha/2}\right]$ are pointwise bootstrapped $1-\alpha$ confidence intervals of $\hat{\theta}_i$, for $i = t_0, \ldots, T$. The same procedure is applied to the models described in Section 3.2.2 to get pointwise bootstrapped $1-\alpha$ confidence intervals of $\hat{\theta}_{i\mathcal{L}}$ and $\hat{\theta}_{iE}$.

3.3 The CépiDc dataset of death by HCC in France between 1979 and 2012

3.3.1 Data

We refer to the database of the number of reported deaths by HCC in France from 1979 to 2012 published on the website of CépiDc[1]. As diseases classification changed in 2000, we reported in Table 3.1 the codes used for the data extraction where ICD-9 and ICD-10 stand for International Classification of Diseases, Ninth and Tenth Revisions (see World Health Organization (1977, 1992)). Since we consider HCC, we only focus on the codes 155.0 and C22.0. However, due to some misclassified codes, the codes 155.2 and C22.9 were also used to estimate the data.

Table 3.1: Causes of death extracted from CépiDc.

ICD version	Code	Description
ICD-9	155.0	Malignant neoplasm of liver, primary
	155.2	Malignant neoplasm of liver, not specified as primary or secondary
ICD-10	C22.0	Liver cell carcinoma
	C22.9	Malignant neoplasm of liver, not specified as primary or secondary

Only HCC deaths related to compensated cirrhosis were considered since HCC deaths related to decompensated cirrhosis were reported as death related to cirrhosis and not HCC (see Mourad (2014)).

By computing the mean proportion of people of each gender in the database of HCC deaths, one can find that the proportion of men is 81.4%. Since this proportion does not reflect the proportion of men in the overall population infected by HCV, we decide to study each gender apart and adjust the results to estimate the overall infected population. Based on the cohort studied by Deuffic-Burban et al. (2008), we assume that the proportion of men is 57%. Then by multiplying the number of men who died of HCC related to HCV by $0.57/0.814 = 0.7002$ and the number of women who died of HCC related to HCV by $0.43/0.186 = 2.3118$, one can compute the number of HCC related to HCV deaths for the overall population.

As a consequence of the changement of classification, one can notice large gaps between values in year 1999 and values in year 2000 for both men and women in Figures 3.1(a) and 3.1(b). This lack of correspondence between data before and after 2000 has already been raised by Duberg and Hultcrantz (2008) in a correspondance with Bosetti et al. (2008).

However, considering the two gaps, one can notice a relative symmetry which allows us to think that some death certificates might have been misclassified. Qualified experts of CépiDC confirmed that hypothesis by affirming that some death certificates could have been coded 155.0 instead of 155.2.

Hence, for the sake of simplicity, we propose to multiply the number of deaths coded 155.0 for men (resp. women) by some constant c_m (resp. c_w) assuming a general increase of the trend in mortality

[1] Centre d'épidémiologie sur les causes médicales de décès - http://www.cepidc.inserm.fr/site4/

of HCC. As the number of deaths coded C22.0 after 2000 is nearly constant until 2012, we assume that the number of deaths coded 155.0 in 1999 for men (resp. for women) equals the number of deaths coded C22.0 in 2000 divided by c_m (resp. c_w). Thus, by denoting Y^m_{1999} (resp. Y^w_{1999}) the number of death certificates coded 155.0 in 1999 for men (resp. women) and Y^m_{2000} (resp. Y^w_{2000}) the number of death certificates coded C22.0 in 2000 for men (resp. women), we write:

$$c_m = \frac{Y^m_{2000}}{Y^m_{1999}} = 0.8354 \quad \text{and} \quad c_w = \frac{Y^w_{2000}}{Y^w_{1999}} = 0.6897. \tag{3.8}$$

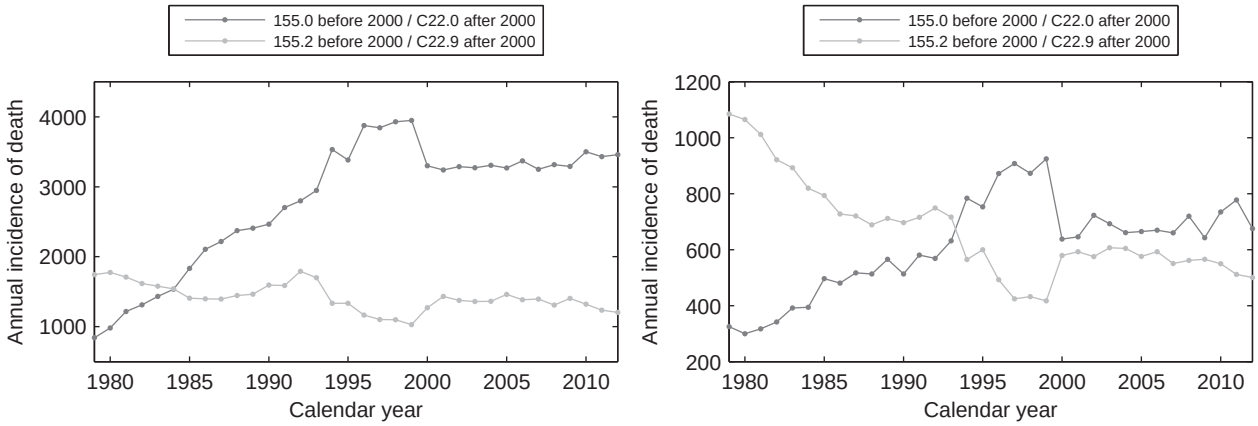
Since we are studying HCV, we need to estimate the proportion of HCC deaths related to HCV as various causes lead to HCC. Many studies have reported so far some contrasted values for the proportion of HCC attributable to HCV in different countries (see Deuffic-Burban et al. (1999); El-Serag (2002); El-Zayadi et al. (2005); Yoshizawa (2002)). Globally, Perez et al. (2006) reported a worldwide proportion p of HCV-related HCC of 25% for the year 2002 while Marcellin et al. (2008) reported a proportion of 22.5% for the year 2001 in France. Let us take $p = 0.225$ as it corresponds to the most recent estimation based on French data.

On the basis of what precedes, Figure 3.2 displays the adjusted dataset for the overall population and for both men and women between 1979 and 2012.

Finally, to face some computational issues due to the lack of data before 1979, we extrapolate the number of HCC deaths related to HCV in the past. To give a realistic shape to the curve of the past deaths and assuming that there is no death from HCC related to HCV before 1920, we fit a gaussian function

$$f(x) = \alpha \exp\left(-\left(\frac{(x-\beta)}{\gamma}\right)^2\right)$$

with a non-linear least squares procedure which gives us the following values $\alpha = 912.8$, $\beta = 2008$ and $\gamma = 23.51$. Values of $f(x)$ were rounded up. Extrapolated data before 1979 are displayed in Figure 3.2 for $t_0 = 1920$ and $T = 2012$.



(a) Annual number of reported death for men

(b) Annual number of reported death for women

Figure 3.1: Number of reported deaths due to a liver tumor in France extracted from CépiDC.
The mid grey dots represent the number of annual reported deaths coded 155.0 then C22.0 and the light grey dots represent the number of annual reported deaths coded 155.2 then C22.9. (a) Annual number of deaths for men. (b) Annual number of deaths for women.

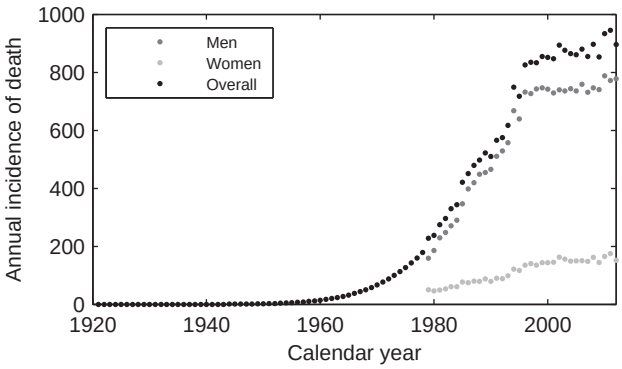


Figure 3.2: Adjusted number of HCC deaths related to HCV in France.
The mid grey dots represent the adjusted data for men, the light grey dots represent the adjusted data for women and the black dots represent the adjusted data for the overall population.

3.3.2 Natural history model

Since HCV has a long asymptomatic period, it is quite difficult to know the period duration between the infection and the deaths by HCC. To estimate this lifetime distribution, we model the natural history of the disease. Once the infection occurs, the subject enters in the so-called acute phase of infection denoted state A. This phase is followed by the chronic phase if no spontaneous clearance happens within the six months after infection. The chronic phase corresponds to the slow decay of the liver in a 5-step process (from F_0 to F_4) named fibrosis which leads to the development of cirrhosis. As the cirrhosis progresses, the subject may develop a hepatocellular carcinoma (state HCC) or a decompensated cirrhosis (state DC) which could lead to death.

Let us consider a discrete time Markov chain as a model of the natural history of HCV, with graph displayed in Figure 3.3. We denote by $p_{S_i \to S_j}$ the annual transition probabilities between two given stages S_i and S_j of the Markov chain. For simplification purposes, death unrelated to HCV was not displayed on the graph.

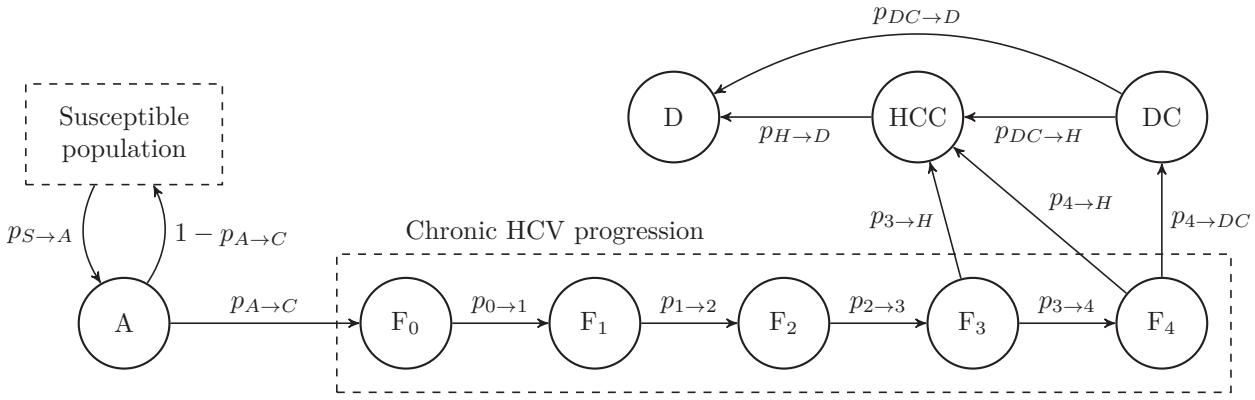


Figure 3.3: Markov model of the natural history of HCV.
A, acute infection; F_0-F_4, stages of the fibrosis progression based on the Metavir scoring system indicating the degree of inflammation; HCC, hepatocellular carcinoma; DC, decompensated cirrhosis; D, death due to HCV.

As the total prevalence was estimated for chronic carriers in 2004 in France (see Institut de veille sanitaire (2007)), we choose to focus on chronic carriers of HCV. Moreover, since no data is available for the number of death from HCC related to a decompensated cirrhosis nor from decompensated cirrhosis related to HCV, the state DC is discarded of the model. Thus, as a first step, we will only study the population who died from HCC related to HCV and simplify the previous Markov chain to focus on the

chain with graph displayed in Figure 3.4.

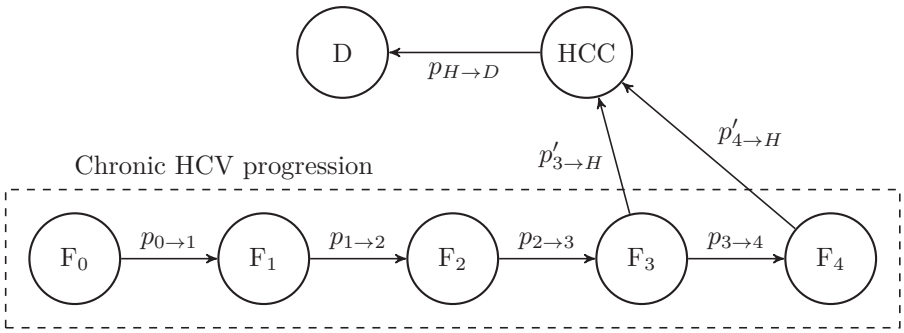


Figure 3.4: Simplified Markov model of the natural history of HCV.
F_0-F_4, stages of the fibrosis progression based on the Metavir scoring system indicating the degree of inflammation; HCC, hepatocellular carcinoma; D, death due to HCV.

Especially, one can notice that transition probabilities from F_3 and F_4 to HCC are not the same between the two chains. The reason of that modification is that $p_{3\to H}$ and $p_{4\to H}$ apply to the whole infected population while $p'_{3\to H}$ and $p'_{4\to H}$ apply to the infected population who will develop HCC.

As all infected patients do not necessarily develop HCC, we have:

$$p'_{3\to H} = p_{3\to H}/h \times \rho \quad \text{and} \quad p'_{4\to H} = p_{4\to H}/h \times \phi \tag{3.9}$$

where h is the proportion of patients who will develop HCC and ρ and ϕ the proportions of patients coming from F_3 and F_4 respectively.

To determine $p'_{3\to H}$ and $p'_{4\to H}$, we need to determine the proportion h of patients who actually develop HCC. Since Lok et al. (2009) reported a proportion of 4.9% among a population of 1,005 patients in 2009, we choose $h = 0.049$. In the same time, they estimated a cumulative 5-year HCC incidence of 4.1% and 7.0% for patients with bridging fibrosis (F_3) and for patients with cirrhosis (F_4) respectively. From this, we may compute:

$$\rho = \frac{0.041}{(0.07 + 0.041)} = 0.37 \quad \text{and} \quad \phi = \frac{0.07}{(0.07 + 0.041)} = 0.63.$$

We can also compute the transition probabilities $p_{3\to H}$ and $p_{4\to H}$ by applying the following transformation:

$$-\frac{\ln(1 - CIr)}{d}$$

where CIr is the cumulative incidence rate and d the period of time during which CIr is measured. With $d = 5$ and $CIr = 0.041$ for patients with bridging fibrosis and $CIr = 0.07$ for patients with cirrhosis given by Lok et al. (2009), we obtain $p_{3\to H} = 0.0084$ and $p_{4\to H} = 0.0145$ respectively. From (3.9), we obtain $p'_{3\to H} = 0.0634$ and $p'_{4\to H} = 0.1864$.

The remaining transition probabilities reported in Table 3.2 were found in the literature. Those were mainly found in the work of Deuffic-Burban et al. (2008) who considered gender, age and alcohol consumption (slight vs heavy drinkers) as covariates in HIV-negative patients. Hence, to compute our transition probabilities, we assume that gender, age and alcohol consumption distributions have the same distributions as the cohort studied by Deuffic-Burban et al. (2008) since no data on those covariates were available in the database of CépiDc.

We assume a spontaneous clearance of 25% over 6 months, as reported by Micallef et al. (2006). Thus, we obtain $p_{A\to C} = 0.5625$ as an annual rate .

Table 3.2: Annual transition probabilities of the natural history model.
†Transition during the first year. ‡Transition during the second year and after.

Parameters	Values used in the model	Sources
$p_{A\to C}$	0.5625	Micallef et al. (2006)
$p_{0\to 1}$	0.1401	Deuffic-Burban et al. (2008)
$p_{1\to 2}$	0.1401	Deuffic-Burban et al. (2008)
$p_{2\to 3}$	0.1401	Deuffic-Burban et al. (2008)
$p_{3\to 4}$	0.1581	Deuffic-Burban et al. (2008)
$p_{3\to H}$ / $p'_{3\to H}$	0.0084 / 0.0634	Lok et al. (2009)
$p_{4\to H}$ / $p'_{4\to H}$	0.0145 / 0.1864	Lok et al. (2009)
$p_{4\to DC}$	0.0711	Deuffic-Burban et al. (2008)
$p_{DC\to H}$	0.0534	Planas et al. (2004)
$p_{DC\to D}$	$0.3900^{\dagger}$ / $0.1100^{\ddagger}$	D'Amico et al. (2006)
$p_{H\to D}$	$0.7456^{\dagger}$ / $0.2726^{\ddagger}$	El-Serag et al. (2001)

3.3.3 Lifetime distribution of an individual who develop and die of HCC

From the Markov chain established in the previous section, we want to determine the lifetime distribution f defined in Section 3.2.1. The survival can be characterized as the time spent by a given individual in the Markov chain. We denote by S the random variable describing the time spent by an individual in the Markov chain:

$$\{\mathrm{F}_0, \mathrm{F}_1, \mathrm{F}_2, \mathrm{F}_3, \mathrm{F}_4, \mathrm{HCC}, \mathrm{D}\}.$$

Let us denote P the transition matrix of the Markov chain described in Figure 3.4. Hence, we have:

$$P = \begin{pmatrix} 1-p_{0\to 1} & p_{0\to 1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1-p_{1\to 2} & p_{1\to 2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1-p_{2\to 3} & p_{2\to 3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1-p_{3\to 4}-p'_{3\to H} & p_{3\to 4} & p'_{3\to H} & 0 \\ 0 & 0 & 0 & 0 & 1-p'_{4\to H} & p'_{4\to H} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1-p_{H\to D} & p_{H\to D} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

We may simplify P as the following block matrix:

$$P = \begin{pmatrix} Q & Q_0 \\ \mathbf{0} & 1 \end{pmatrix}$$

where $Q_0 = (0,0,0,0,0,p_{H\to D})^{\intercal}$, $\mathbf{0} = (0,0,0,0,0,0)$ and

$$Q = \begin{pmatrix} 1-p_{0\to 1} & p_{0\to 1} & 0 & 0 & 0 & 0 \\ 0 & 1-p_{1\to 2} & p_{1\to 2} & 0 & 0 & 0 \\ 0 & 0 & 1-p_{2\to 3} & p_{2\to 3} & 0 & 0 \\ 0 & 0 & 0 & 1-p_{3\to 4}-p'_{3\to H} & p_{3\to 4} & p'_{3\to H} \\ 0 & 0 & 0 & 0 & 1-p'_{4\to H} & p'_{4\to H} \\ 0 & 0 & 0 & 0 & 0 & 1-p_{H\to D} \end{pmatrix}.$$

Thus, by denoting $\tau = (1,0,0,0,0,0)$ the initial distribution, we obtain:

$$\mathbb{P}(S=k) = \tau Q^{k-1} Q_0,$$

where k is the number of cycles spent in the Markov chain. By writing $f(k) = \mathbb{P}(S=k)$, the random variable S is said to be phase-type distributed with probability distribution function f. One can consult Asmussen (2003) for a full account of the theory of phase-type distributions. With our previous notation, $f_{j-i} = \tau Q^{(j-i)-1} Q_0$.

The main disadvantage of this method is the fact that by referring to a discrete time Markov chain, we are limited to $f_{j-i} = 0$, for $j - i = 1, \ldots, 6$, since 6 years is considered as the minimum absorbing time of the Markov chain. To solve this issue, we decide to fit the phase-type distribution f with a given distribution by estimating its parameters with respect to the least squares criterion.

Since the phase-type distribution considered is discrete, the following distributions were tested:

$$\{\text{Binomial, Poisson, Negative Binomial, Geometric}\}.$$

It turned out that the negative binomial distribution was the best fit with estimators $(\hat{r}, \hat{p}) = (7.2634, 0.8122)$ with graph displayed in Figure 3.5. We conclude that:

$$f_{j-i} = \binom{j - i + \hat{r} - 1}{j - i} \cdot \hat{p}^{j-i}(1 - \hat{p})^{\hat{r}}.$$

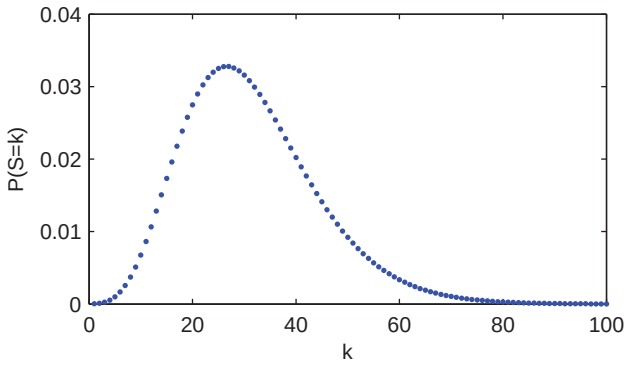


Figure 3.5: Lifetime distribution. Lifetime distribution adjusted by a negative binomial distribution with parameters $(\hat{r}, \hat{p}) = (7.2634, 0.8122)$.

By computing the mean of the negative binomial distribution with parameters estimated above, we obtain 31.4 years. As occurrence of HCC after HCV infection and the mean lifetime of untreated patients with HCC have been estimated to 25-30 years (see Castells et al. (1995); Tong et al. (1995)) and to 0-3 years (see Cabibbo et al. (2010); Giannini et al. (2015)) respectively, our estimation is comparable to these.

3.3.4 Distribution of the time to death from HCC in the overall population

Until now, we focused on estimating the population who will actually develop and die of HCC. To estimate the overall population, we denote by Θ_i the mean number of people infected by HCV in year i. We assume that Θ_i follows the same trend as θ_i, more precisely that $\Theta_i = \lambda \cdot \theta_i$, where $\lambda > 0$. Thus, we will estimate Θ_i by $\hat{\lambda} \cdot \hat{\theta}_i$ where the estimator $\hat{\lambda}$ is given below. Similar argument holds for $\hat{\theta}_{i\mathcal{L}}$ and $\hat{\theta}_{iE}$ introduced in Section 3.2.2.

To give an estimation of λ, let us modify the Markov chain described in Figure 3.4 by including a new state OCD (Other Cause of Death) corresponding to the patient flows leaving the Markov chain, i.e the patients leaving due to natural mortality or development of a decompensated cirrhosis. The states of this new Markov chain are $\{F_0, F_1, F_2, F_3, F_4, \text{HCC}, \text{D}, \text{OCD}\}$ with graph described in Figure 3.6 where p_m is the French natural mortality rate extracted from the database of Insee[2].

[2]Institut national de la statistique et des études économiques - http://www.insee.fr/

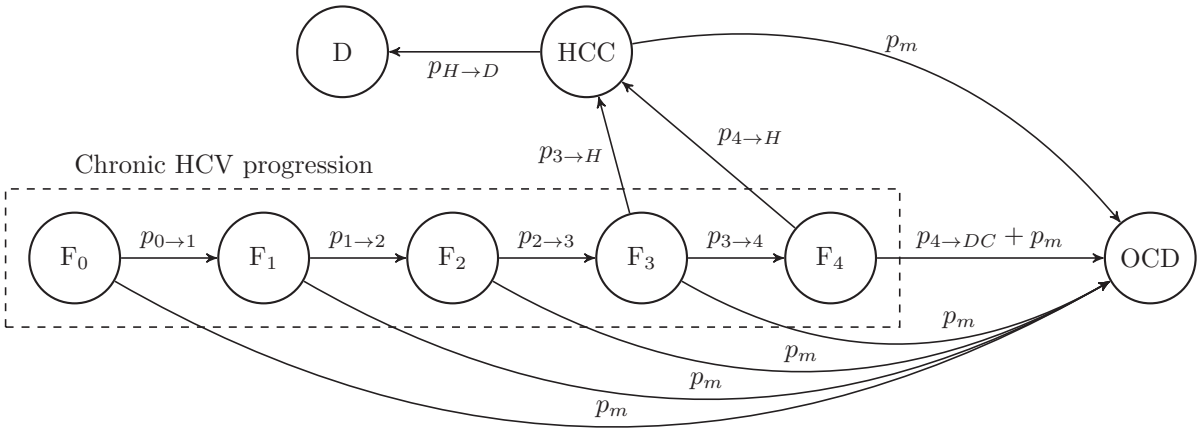


Figure 3.6: Simplified Markov model of the natural history of HCV with extra state OCD. F_0-F_4, stages of the fibrosis progression based on the Metavir scoring system indicating the degree of inflammation; HCC, hepatocellular carcinoma; D, death due to HCV; OCD, patients who left due to natural mortality or decompensated cirrhosis.

By denoting P' the transition matrix of the Markov chain with graph displayed in Figure 3.6, we have:

$$P' = \begin{pmatrix} p_0 & p_{0\to 1} & 0 & 0 & 0 & 0 & 0 & p_m \\ 0 & p_1 & p_{1\to 2} & 0 & 0 & 0 & 0 & p_m \\ 0 & 0 & p_2 & p_{2\to 3} & 0 & 0 & 0 & p_m \\ 0 & 0 & 0 & p_3 & p_{3\to 4} & p_{3\to H} & 0 & p_m \\ 0 & 0 & 0 & 0 & p_4 & p_{4\to H} & 0 & p_m + p_{4\to DC} \\ 0 & 0 & 0 & 0 & 0 & 1 - p_{H\to D} - p_m & p_{H\to D} & p_m \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}. \tag{3.10}$$

To simplify the expression of (3.10), we used the following notations:

$$\begin{cases} p_0 = 1 - p_{0\to 1} - p_m \\ p_1 = 1 - p_{1\to 2} - p_m \\ p_2 = 1 - p_{2\to 3} - p_m \\ p_3 = 1 - p_{3\to 4} - p_{3\to H} - p_m \\ p_4 = 1 - p_{4\to H} - p_{4\to DC} - p_m. \end{cases}$$

Let $\mathbf{x}_k$ be a row vector of length 8 describing the estimated mean number of individuals in each state of the Markov chain at the end of year k. In particular, assuming that the value of λ is known, we have $\mathbf{x}_{t_0}(1) = \lambda \cdot \hat{\theta}_{t_0}$ and $\mathbf{x}_{t_0}(6) = 0$ which are the estimated mean number of people in the states "F_0" and "HCC" the first year respectively. One can calculate the mean number of people in each state at year $k+1$ by computing:

$$\mathbf{x}_{k+1} = \mathbf{x}_k \times P' + \begin{bmatrix} \lambda \cdot \hat{\theta}_{k+1} & 0 & \cdots & 0 \end{bmatrix}.$$

Thus, we estimate λ with respect to the least squares criterion:

$$\hat{\lambda} = \arg\min_{\lambda} \sum_{k=t_0}^{T} \left(Y_k - \mathbf{x}_{k-1}(6) \times p_{H\to D}\right)^2$$

where T is the year of last observation. Similarly, based on $\hat{\theta}_{i\mathcal{L}}$ and $\hat{\theta}_{iE}$, we obtain $\hat{\lambda}_{\mathcal{L}}$ and $\hat{\lambda}_E$.

3.3.5 Simplified Deuffic et al. model

In this section, we describe a simplified version of the model initially developed by Deuffic-Burban et al. (1999). As age of the patients was not taken into consideration in the previous models, we had to simplify their model to make the comparisons. Furthermore, as the prevalence of HCV in France has been

estimated at 221,386 (95%CI:158,909-283,862) cases for the year 2004, Deuffic-Burban et al. (1999) developed their model in the presence of an equality constraint taking into account the estimated prevalence of HCV in 2004. However, as this estimation admits a relatively wide confidence interval, we decided to dispense with that constraint in our simplified Deuffic model (SDM).

Finally, in their original model, they estimated the transition probabilities of their Markov chain while we just took their own estimations to build ours as a complement to published probabilities from the literature. Hence, some computation differences may arise but the general theory stays the same. We consider the same Markov chain as the one with graph displayed in Figure 3.6.

The models described in Sections 3.2.2 and 3.2.3 estimate the past number of HCV-infected individuals who died from HCC. With the estimation of λ in Section 3.3.4, one can reconstruct the past number of all HCV infections. On the contrary, the model developed by Deuffic-Burban et al. (1999) directly estimates the past number of all HCV infections and does not need any additional estimation step. Hence, in this section we focus on Θ_i the mean number of people infected by HCV in year i.

The simplified model of Deuffic-Burban et al. (1999) supposes a parametric model for Θ in the form of the following equation:

$$f_D(t;\alpha,\beta,\gamma) = \begin{cases} \dfrac{\exp(a(t-t_0))}{b+\beta^{-1}\exp(a(t-t_0))}, & \text{for } t < 1990 \\ 0.6 \times \dfrac{\exp(a(t-t_0))}{b+\beta^{-1}\exp(a(t-t_0))}, & \text{otherwise} \end{cases}$$

with $a = 4 \times \alpha \times \beta^{-1}$ and $b = \beta^{-1} \times \exp(a(\gamma - t_0))$ where α is the slope of the curve, β the asymptotic plateau value and γ the year of mid-epidemic.

From the notations introduced in the previous section, let us write:

$$\mathbf{x}_{k+1} = \mathbf{x}_k \times P' + \begin{bmatrix} f_D(t;\alpha,\beta,\gamma) & 0 & \cdots & 0 \end{bmatrix}.$$

Deuffic-Burban et al. (1999) proposed to estimate the parameters (α,β,γ) with respect to the weighted least squares criterion:

$$(\hat{\alpha},\hat{\beta},\hat{\gamma}) = \underset{(\alpha,\beta,\gamma)}{\arg\min} \sum_{k=t_0}^{T} \frac{\left(Y_k - \mathbf{x}_{k-1}(6) \times p_{H\to D}\right)^2}{\sigma^2_{Y_k}}$$

where $\sigma^2_{Y_k}$ is the estimated variance of Y_k. Assuming that Y_k follows a binomial distribution, we have:

$$Y_k \sim \mathcal{B}\Big(\mathbf{x}_{k-1}(6), p_{H\to D}\Big),$$

where n_k is the number of individuals in the Markov chain in year k and p_k the probability that an individual die from HCC related to HCV in year k. Thus, one can write:

$$\sigma^2_{Y_k} = \mathbf{x}_{k-1}(6) \times p_{H\to D} \times (1 - p_{H\to D}).$$

We obtain an estimator of Θ given by:

$$\hat{\Theta}_{iD} = \frac{\exp(\hat{a}(t-t_0))}{\hat{b}+\hat{\beta}^{-1}\exp(\hat{a}(t-t_0))} \cdot \bullet_{\{i<1990\}} + \frac{0.6 \times \exp(\hat{a}(t-t_0))}{\hat{b}+\hat{\beta}^{-1}\exp(\hat{a}(t-t_0))} \cdot \bullet_{\{i\geqslant 1990\}}$$

where $\hat{a}$ and $\hat{b}$ are the estimators of a, b. Assuming the parameters $(\hat{\alpha},\hat{\beta},\hat{\gamma})$ admit a normal asymptotic distribution, the corresponding 95% confidence intervals are computed by estimating the variance-covariance matrix of the parameters.

3.3.6 Quality of the models

To determine which model fits best the data, we will use the mean squared error (MSE) criterion:

$$\text{MSE} = \frac{1}{T}\sum_{k=1}^{T} \left(Y_k - \mathbf{x}_{k-1}(6) \times p_{H\to D}\right)^2$$

where $\mathbf{x}_{k+1} = \mathbf{x}_k \times P' + \begin{bmatrix} \hat{\Theta}_{k+1} & 0 & \cdots & 0 \end{bmatrix}$ and P' is the transition matrix defined in Section 3.3.4.

3.4 Results

First, let us recall the walkthrough of this study. Our estimations are based on observed data Y_j, for $j = t_0, \dots, T$, from death certificates of French patients who died of hepatocellular carcinoma (HCC) related to hepatitis C virus (HCV). Several epidemic models and statistical procedures, introduced in Sections 3.2.2 and 3.2.3, are computed to determine $\hat{\theta}_i$, $\hat{\theta}_{i\mathcal{L}}$ and $\hat{\theta}_{iE}$ the estimators of the mean past incidence of patients who got infected in year i and who died of HCC.

From this, models are scaled to the overall population thanks to a parameter λ which is estimated with respect to the least squares criterion and given in Table 3.3. One can notice that those coefficients admit similar values for the three models.

Table 3.3: Scale coefficients estimation.

Models	Coefficients
EMS	7.3187
Logistic	7.2465
EMG	7.3100

Considering the smoothing step of the EMS algorithm, several values for the order L of the binomial filter were tested and it turned out that $L = 4$ was approximatively reflecting the expected trend of HCV incidence in France. In fact, since the precision of $\hat{\theta}_i$ declines as i increases, a solution would be to increase the bandwidth of the binomial filter with i. However, we noticed that by increasing the bandwidth L, the distant values of $\hat{\theta}_i$ did not admit strong fluctuations while recent values did. This motivated our choice for $L = 4$.

Graphics on Figure 3.7 exhibit the past incidence of HCV in France ($\hat{\Theta}_i$, $\hat{\Theta}_{i\mathcal{L}}$ and $\hat{\Theta}_{iE}$) estimated by the three methods introduced earlier and the simplified Deuffic model ($\hat{\Theta}_{iD}$). One can notice that the four models follow a similar trend and admit a maximum expected number of infections around 7,000 – 9,000 cases. The model estimated by the EMS algorithm and the parametric model based on the EMG function admit a peak of infection during the late 70s'. One can observe that the SDM and logistic model admit the same trend since the plateau value is reached in the early 80s'. While no assumption are made on the general shape of the model estimated by the EMS algorithm, one can see that the model still reflects the general decrease of infections past the infection peak.

One can also notice on Figure 3.7 that the different confidence intervals are not symmetric. This phenomenon is due to the fact that the confidence intervals are generated via a Poisson distribution which is not a symmetric distribution. Particularly, a peak in the confidence interval of the EMS model can be observed on Figure 3.7(a). This anomaly is a consequence of the lack of data available for the estimation of Θ_i when i gets closer to T. The same anomaly is exhibited on Figure 3.10(a). Estimators and 95% confidence intervals of the different parameters are given in Table 3.4.

Graphics on Figure 3.8 plot for each models the estimated number of death from HCC due to HCV as well as the number of observed HCC deaths. Confidence intervals for our three models were constructed by bootstrap as described in Section 3.2.4. One can notice that all the models seem to fit correctly the adjusted data from CépiDC.

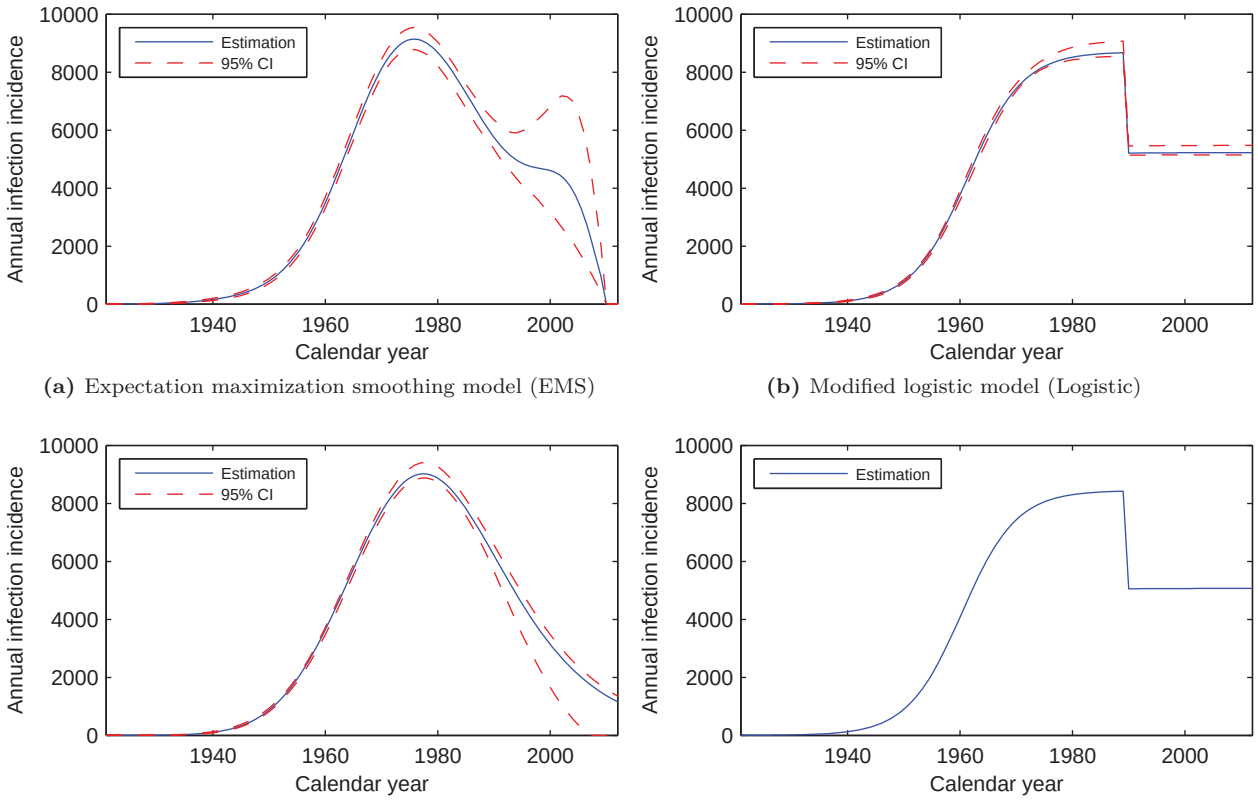
Table 3.5 exhibits the mean squared error computed for the different models. One can see that the EMS-based model seems to have the lower MSE among the models. It is followed by the EMG and SDM models which admit similar values.

In Table 3.6, we compare our results with the prevalence of chronic carriers of HCV estimated by InVS[3] (see Institut de veille sanitaire (2007)) in 2004 in France. One can notice that our estimates seem lower than what was estimated by InVS for the year 2004. However, confidence intervals are overlapping. Such a difference could be explained by the fact that in Section 3.3.4, no consideration is made about patients exiting the model when getting cured (standard treatment or liver transplant) and how much

[3] "Institut de veille sanitaire", the French institute for public health surveillance.

the treatment decelerates the disease's progression. Moreover, no patient was assumed to develop and eventually die from a fulminant hepatitis neither. However, one can observe that the models based on a logistic function (logistic model and SDM) admit the highest values and, then, are closer to the value estimated by InVS. As exhibited in Figure 3.9, one can see that the confidence intervals obtained for our models are significantly smaller than the one obtained by InVS.

Finally, Figure 3.10 shows the estimated prevalence of HCV in France. As said above, one can observe that the four models follow a similar trend. There is a general increase from the end of 1940 to reach a relative stagnation during 1990 decade. From there, one can notice a decrease in the prevalence of HCV in France.



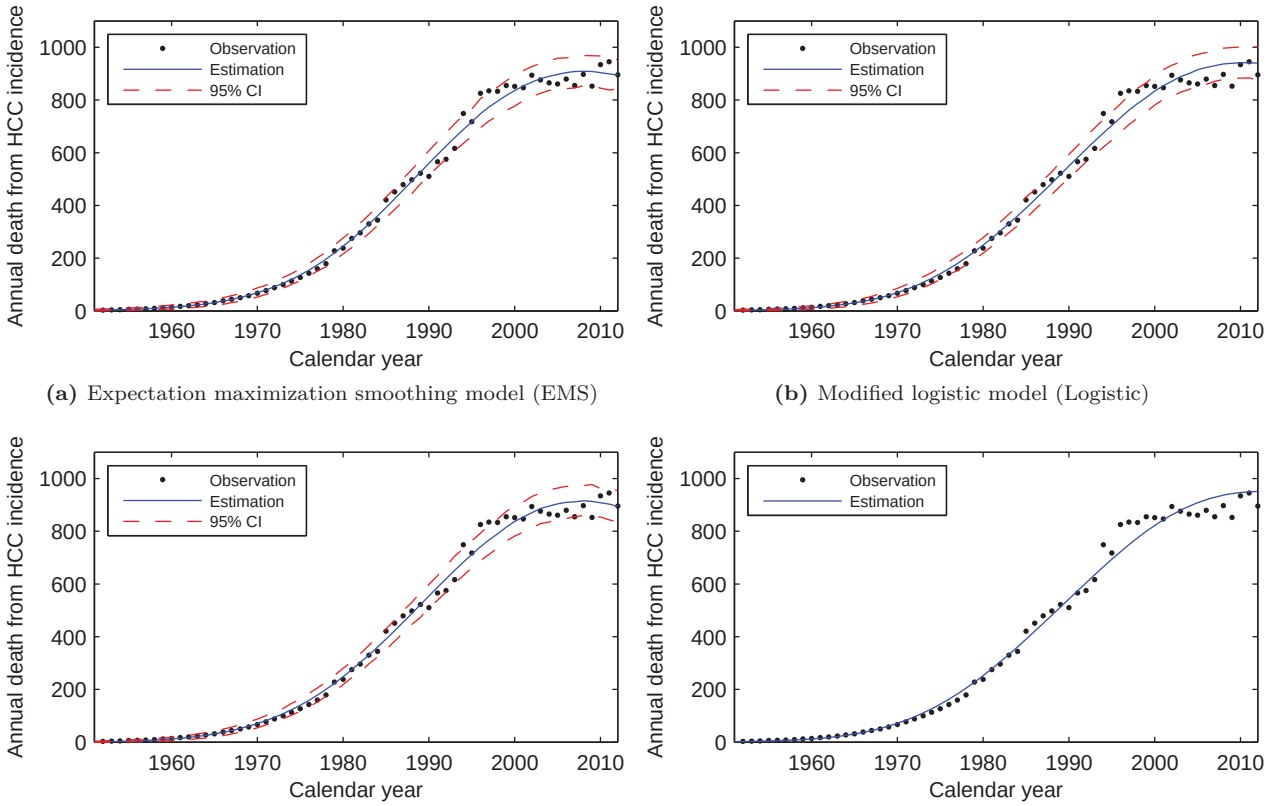
(a) Expectation maximization smoothing model (EMS)

(b) Modified logistic model (Logistic)

(c) Exponential modified Gaussian model (EMG)

(d) Simplified Deuffic et al. model (SDM)

Figure 3.7: Estimation of the expected number of HCV infections.
Solid curves represent the estimation of the expected number of annual infections while dashed curves are the bootstrapped 95% confidence intervals.



(a) Expectation maximization smoothing model (EMS)

(b) Modified logistic model (Logistic)

(c) Exponential modified Gaussian model (EMG)

(d) Simplified Deuffic et al. model (SDM)

Figure 3.8: Estimation of the expected number of deaths from HCC related to HCV. Solid curves represent the estimation of the expected number of annual deaths while dashed curves are the bootstrapped 95% confidence intervals. Dotted curve are the number of observed HCC due to HCV each year.

Table 3.4: Parameter estimation and confidence interval for the different models.

	Parameters	
Models	Estimates	CI 95%
Logistic		
K	1,200.9	[1,152.7 – 1,237.8]
α	0.2039	[0.1997 – 0.2209]
β	1,961.31	[1,960.41 – 1,961.77]
EMG		
K	4.25×10^6	$[1.45 \times 10^6 - 8.48 \times 10^6]$
λ	1.3942	[1.3276 – 2.303]
μ	0.622	[-119.03 – 1.1858]
σ	11.3638	[11.0478 – 12.9392]
SDM		
α	435.65	[416.91 – 454.38]
β	8,443.69	[8,192.02 – 8,695.35]
θ	1,960.32	[1,959.66 – 1,960.98]

Table 3.5: Mean squared error result for the different models.

Models	Mean squared error
EMS	467.16
Logistic	723.35
EMG	522.08
SDM	585.27

Table 3.6: Estimation and 95% confidence interval of total HCV chronic carriers in 2004.

	Prevalence in 2004	
Models	Estimates	CI 95%
InVS	221,386	[158,909 – 283,862]
EMS	187,038	[166,989 – 213,656]
Logistic	197,604	[194,450 – 205,609]
EMG	179,864	[160,812 – 189,432]
SDM	194,014	–

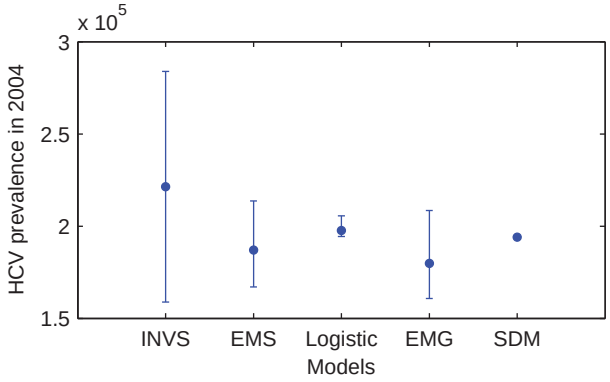
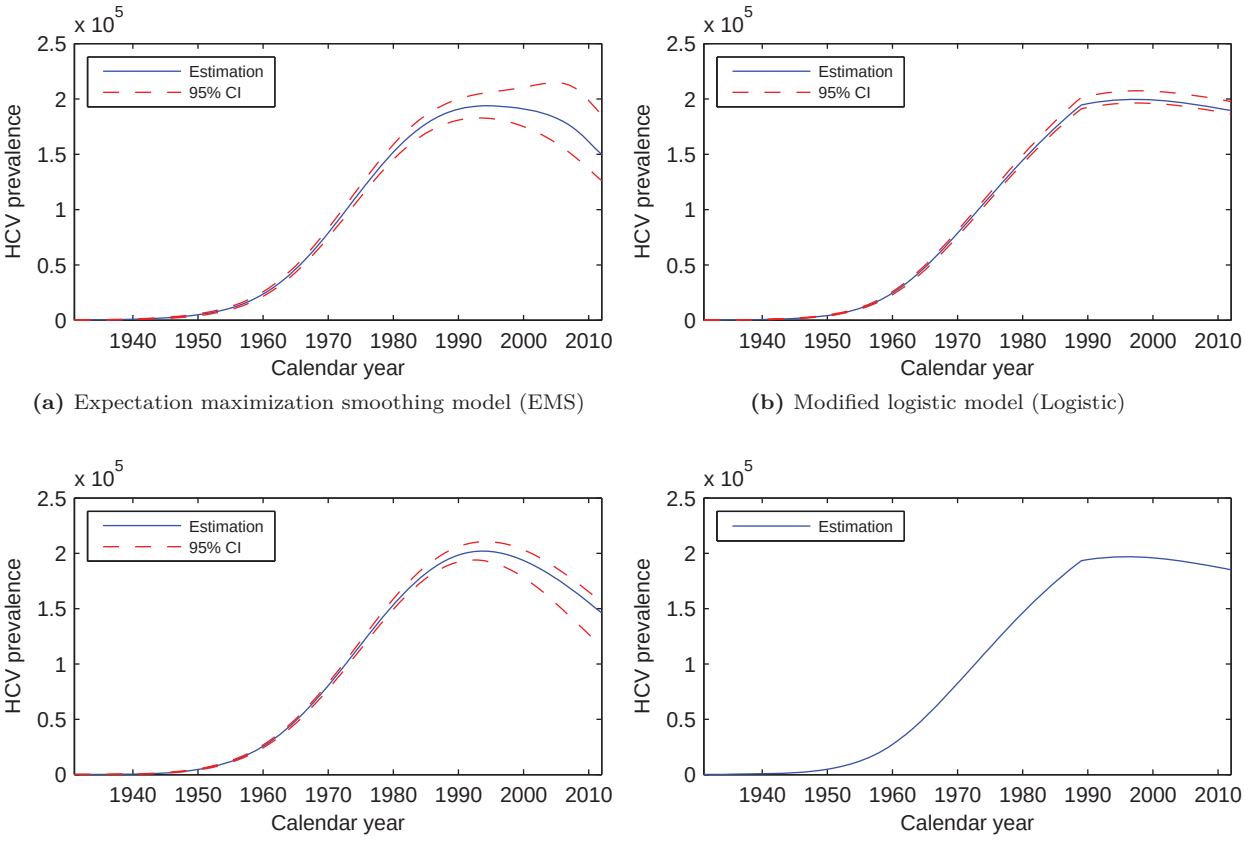


Figure 3.9: Estimation and 95% confidence interval of total HCV chronic carriers in 2004.



(a) Expectation maximization smoothing model (EMS)

(b) Modified logistic model (Logistic)

(c) Exponential modified Gaussian model (EMG)

(d) Simplified Deuffic et al. model (SDM)

Figure 3.10: Estimated prevalence of total chronic HCV in France.
Solid curves represent the estimation of the prevalence of the chronic carries of HCV while dashed curves are the bootstrapped 95% confidence intervals.

3.5 Discussion

In this chapter, we have presented several models based on the natural history of HCV to estimate the past incidence of the virus in France from 1930 to 2012. Those estimations were conducted via a back-calculation approach computed on actual French data of death from HCC. Our results have been and were compared with the model developed by Deuffic-Burban et al. (1999). The different models exhibited similar trends on the evolution of incidence and prevalence of HCV in France with an incidence's peak value reached by the models at around 8,000 cases during the late 70'. Similarly, the different models estimated the prevalence of HCV at a little less than 200,000 cases in 2004 which is consistent with the estimation reported by InVS at 221,386 (95%CI:159,000-284,000) cases (see Institut de veille sanitaire (2007)). The difference between the prevalence values could be explained by the fact that no treatment has been incorporated in the model. As the addition of therapeutics would strongly alter the lifetime distribution, the total number of infected individuals must have been underestimated.

The main advantage of the model developed with the expectation maximization smoothing algorithm is the fact that no particular assumption was made on the form of the $\theta_i(\cdot)$ function. This helped us to limit the bias induced by an arbitrary choice of a given parametric function. Indeed, the poor performance of the parametric models (particularly the logistic model and, to a lesser extent, the SDM and the EMG model) on the mean squared error computation, precisely highlights this advantage. Then, an advantage inherent to the use of the bootstrap method and Poisson distributions, is the possibility to easily compute confidence intervals of the different functions (incidence and prevalence), while this might be harder to achieve with the model developed by Deuffic-Burban et al. (1999) without knowing the distribution of the estimated parameters.

However, our approach presents some limitations. First, as HCV progression is quite long, such a back-calculation approach cannot provide accurate estimations of recent HCV incidence. The consideration of HCV diagnoses in addition to HCC certificate deaths could lead to a better approximation of the recent incidence (see Cui and Becker (2000) or Bellocco and Marschner (2000) for the particular case of HIV). Second, contrary to what has been done by Deuffic-Burban et al. (1999), we only focus on the population who died of HCC related to HCV. To estimate the overall population, we need an extra step which induce some additional bias. Third, in terms of model construction, some important parameters have been set aside. Indeed, we would like to emphasize the fact that no consideration was made regarding the use of covariates such as age or gender of the patients. No consideration was made about treated patients or HIV-coinfected patients while these could have a significant impact on the fibrosis' progression. However, as the development of therapeutics against HCV is relatively recent, these should have a limited impact on our back-calculation approach.

Last, we would like to highlight the fact that this study is influenced by the quality of the data. Since no accurate data exists on neither the number of death from HCC related to HCV nor the number of death from HCV, the results reported in this study could be improved. However, we believe that we have done a substantial work in the processing of data of CépiDC and one could expect a reliable approximation of the reality from those data.

In conclusion, the models addressed in this study have shown consistent results with the expected incidence of HCV in France. However, those should be completed in consideration with the prognostic factors of HCV for the sake of efficiency. The limitations formulated above could be subject to further research.

4 Network model for the propagation of hepatitis C virus in France

Français

Nous étudions un modèle de propagation épidémique pour un virus de type hépatite C sur un réseau à deux niveaux, à savoir le réseau des aires urbaines françaises. Le premier niveau de ce réseau modélise les interactions entre individus au sein de chaque aire urbaine. Le second niveau est celui des aires urbaines elles-mêmes, reliées entre elles par les individus voyageant entre celles-ci et propageant par la même occasion l'épidémie de ville en ville. Nous encodons le second niveau du réseau en tant que noeud supplémentaire à l'intérieur du premier niveau et observons que cette méthode donne des résultats tangibles en termes d'ampleur et de vitesse de propagation d'une épidémie. Nous appliquons ensuite cette méthodologie à un modèle de propagation épidémique plus complexe afin de mesurer l'impact des nouvelles thérapeutiques anti-hépatite C sur la population infectée en France.

English

We examine a model of epidemic propagation for a virus such as hepatitis C on a multiplex network, namely the network of French urban areas. One network level is that of the individual interactions inside each urban area. The second level is that of the areas themselves, linked by individuals travelling between these areas and potentially helping the epidemic spread from one city to another. We choose to encode the second level of the network as extra, special nodes in the first level. We observe that such an encoding leads to sensible results in terms of the extent and speed of propagation of an epidemic, depending on its source point. We then applied this methodology to a more complex model of epidemic propagation in order to measure the impact of new anti-hepatitis C therapeutics on the French infected population.

Contents

Résumé **67**
4.1 Introduction **68**
4.2 Epidemic models **68**
4.2.1 Initial model of the spread of hepatitis C virus 68
4.2.2 Incorporating contact with HIV-infected individuals 71
4.3 Epidemic propagation on a metapopulation model **72**
4.4 Simulations **74**
4.4.1 Parameter values 74
4.4.2 Results on an abstract three-city network 76
4.4.3 Results on the network of France's 100 largest urban areas 79
4.5 Enhanced HCV model **83**
4.5.1 Model structure 83
4.5.2 Parameter values 88
4.5.3 Migration crisis 93
4.5.4 Prevalence of HCV and initial conditions of the model 93
4.6 Results **94**
4.7 Discussion **98**

Résumé

Ce dernier chapitre de thèse se propose d'étudier un modèle dynamique de propagation du virus de l'hépatite C (VHC) en France afin de quantifier l'impact des nouvelles thérapeutiques anti-VHC. Ce travail se décompose en deux parties au cours desquelles sont successivement présentés :

- un modèle de métapopulation dont chaque sous-population est décrite par un réseau de contact sur lequel est appliqué un modèle simpliste d'histoire naturelle du VHC;
- une extension de ce modèle simpliste incluant des fonctionnalités plus avancées telles que le génotype de la maladie ou la progression à travers les différents états de fibrose.

A l'heure actuelle, la grande majorité des nouvelles contaminations par le VHC dans les pays développés a lieu au sein de la population toxicomane. En effet, l'hépatite C se transmettant très principalement par le sang, les usagers de drogues injectables sont une population d'autant plus vulnérable que la plupart ignore leur statut sérologique.

Aussi, afin de tenter de prendre en compte l'hétérogénéité de cette population en termes d'usage de produits illicites, nous nous intéressons à une modélisation par graphe de contact où chaque contact correspond à un échange de matériel (seringues, aiguilles, etc). Cette modélisation par graphe de contact souffre néanmoins d'une limitation importante. En effet, très peu de données sont disponibles sur la structure d'un tel graphe et, quand bien même ces données seraient disponibles, on pourrait se poser la question de la pertinence d'un tel graphe réalisé à l'échelle nationale.

Dans cette optique, la première partie de ce chapitre s'articule donc autour du concept de métapopulation. Ce concept, formulé par le mathématicien Richard Levins en 1969, considère les dynamiques de plusieurs sous-populations (ou patchs) interagissant les unes avec les autres par un phénomène appelé dispersion.

Dans le cas qui nous intéresse, nous modélisons donc plusieurs sous-populations toxicomanes, par des graphes de contact générés avec des distributions données, interagissant entre elles par un phénomène de migration s'appuyant sur les données de déménagement publiées par l'INSEE.

Concrètement la structure d'un graphe donné reste inchangée (nombre de noeuds, distribution des degrés) mais chaque graphe voisin influe sur la propagation de l'épidémie au sein dudit graphe à hauteur de la proportion d'individus infectés au sein de chaque graphe d'une part et des migrations entre ledit graphe et les autres d'autre part. Nous modélisons ensuite les dynamiques de ces sous-populations à l'aide d'un modèle type SIS (susceptible-infected-susceptible) adapté au VHC.

Les premiers résultats obtenus sont encourageants et offre une illustration de l'importance de considérer les interactions entre les différentes sous-populations dans la modélisation de la propagation du virus.

Dans la seconde partie de ce chapitre, nous nous proposons de garder la structure de métapopulation introduite en première partie et de se concentrer sur un modèle dynamique de la propagation du VHC plus poussé.

En se basant cette fois-ci sur la prévalence actuelle du virus en France, nous déterminons l'impact des nouvelles thérapies en prenant en compte les différents états de fibrose du patient, le génotype par lequel il est infecté mais aussi la possibilité pour un individu de rejoindre ou de quitter la population toxicomane. Afin de mesurer au mieux l'impact des nouvelles thérapeutiques anti-VHC, l'évolution des individus quittant la population toxicomane est capturée par un second modèle.

Ce modèle alternatif est un simple modèle d'histoire naturelle similaire en tout point au modèle original mais ne prenant pas en compte la possibilité pour les individus de se faire infecter. En effet, tout au long de ce chapitre, nous nous focalisons sur la population toxicomane comme principal foyer de nouvelles infections.

Les résultats obtenus dans cette seconde partie permettent de mettre en évidence un impact marqué des nouvelles thérapeutiques anti-VHC ainsi que l'importance des comportements individuels en termes de partage de seringues sur la prévalence de la maladie.

4.1 Introduction

Modeling the propagation of epidemics is a long-standing endeavour in the medical and mathematical sciences. In the past decade or so, multiple developments in the science of complex systems in general, and networks in particular, has led to new models, along with the possibility to test their predictions on datasets, the size of which keeps increasing (see for instance Colizza et al. (2007); Hufnagel et al. (2004), and references therein).

Particularly challenging are the study and modeling of epidemics involving specific, sometimes marginal, segments of the general population. For instance, viruses and diseases that are particularly prevalent among people who inject drugs (PWID). Things become even more involved if one tries to take into account correlations between interrelated epidemics, such as hepatitis C virus (HCV) and AIDS (HIV) among PWID populations.

In this chapter, we introduce a simple model of propagation for HCV on a network of networks, representing PWID populations in distinct but interconnected urban areas. We then applied this methodology to a more complex epidemic model in order to measure the impact of new anti-hepatitis C therapeutics on the French infected population.

In our models, the PWID population of each urban area is a network of individuals, and the fact that people may travel or even move houses from one city to another makes for a second level of links, between cities themselves. We choose to represent this second level via the introduction of special nodes at the first level, that will represent at the individual level in each city, contacts with individuals from other cities. This means that if cities A and B are linked (in the sense that a number of individuals travel from one of them to the other), then we add a node in the network representing individuals in city A. The degree and other characteristics of this special node in city A will depend on the number of individuals from city B travelling to city A and on the size of the infected population in city B. This makes for a simple treatment of the links between cities. We also adapt the standard susceptible-infected-recovered (SIR) equations to reflect the specifics of the HCV epidemics, as well as to take into account the special type of network we are working with. The construction of the simple model is detailed in Section 2 and 3.

In Section 4, we examine, via numerical simulations, how our model works, first on a toy-country with just three cities and then on the network of France's 100 largest urban areas. We explore part of the phase space arising when the values chosen for certain crucial parameters (such as the transmission rate) are allowed to vary across ranges away from their current reported or observed value. We find that simulations run for different values of the parameters lead to sensible stylized facts in terms of the extent and speed of propagation of an HCV epidemics in France.

In Section 5, we develop our second model for HCV propagation by considering some important features such as the fibrosis stages or the different genotypes and apply it on the network of networks methodology. Corresponding results are presented in Section 6. Section 7 presents our conclusions.

4.2 Epidemic models

4.2.1 Initial model of the spread of hepatitis C virus

Let us consider the compartmental model describing the infection dynamics of hepatitis C virus (HCV) displayed in Figure 4.1. This model is based on a susceptible-infected-susceptible (SIS) model which is a derivative of the classic susceptible-infected-recovered (SIR) model introduced by Kermack and McKendrick (1927).

We denote by S the number of susceptible individuals, by A the number of individuals in the acute phase of the infection and by C the number of individuals in the chronic phase of the infection. Assuming that the number of individuals in the total population is constant over time, which is not a strong assumption for large populations, we write $S(t) + A(t) + C(t) = N$. In particular, we assume that the number of people who died from HCV or from some other cause equals the number of people joining the population as new susceptibles.

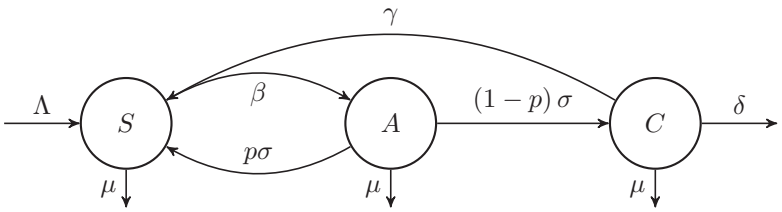


Figure 4.1: Compartmental model of HCV's natural history.
S, susceptible; A, acute HCV infection; C, chronic HCV infection; D, death.

From now on, we choose to work in terms of proportions rather than absolute numbers. The proportion of susceptible, acute-HCV infected and chronic-HCV infected at time t are denoted $s(t), a(t)$ and $c(t)$ respectively. By denoting β the transmission rate, γ the recovery rate, δ the death rate, μ the natural mortality rate, Λ the joining rate, $1/\sigma$ the duration of the acute phase and p the spontaneous clearance, i.e. the probability that a given individual clears the infection, one can write the following set of differential equations:

$$\frac{\mathrm{d}s(t)}{\mathrm{d}t} = \Lambda - \beta s(t)\big(a(t)+c(t)\big) + p\sigma a(t) + \gamma c(t) - \mu s(t), \tag{4.1a}$$

$$\frac{\mathrm{d}a(t)}{\mathrm{d}t} = \beta s(t)\big(a(t)+c(t)\big) - (\sigma+\mu)a(t), \tag{4.1b}$$

$$\frac{\mathrm{d}c(t)}{\mathrm{d}t} = (1-p)\sigma a(t) - (\gamma+\delta+\mu)c(t). \tag{4.1c}$$

Since HCV is an asymptomatic disease, very limited individuals are diagnosed during the acute phase of the infection. Thus we assume that only individuals in the chronic phase have access to a treatment, i.e only people in the chronic phase can recover.

We want to emphasize the fact that one can rewrite the transmission rate β as:

$$\beta = -\ln\Big(1-(1-(1-\pi)^{\omega})\Big)$$

where ω is the contact rate (the number of contacts per unit time) and π the transmission risk (the risk of infection for a given contact). Moreover, as not every patient is actually treated, one can rewrite γ as $\gamma = \nu \times q$ where ν is the recovery rate per patient and q the proportion of treated patients.

The model defined in the set of equations (4.1), also called mass action model, offers a simple and convenient manner of describing the dynamics of an epidemic in a large population. However, it fails at dealing with population heterogeneities and individual interactions which can greatly affect the rate of transmission of the virus.

Considering HCV, it has been stated that the primary route of transmission in the developed world is intravenous drug use (IDU) (see Maheshwari et al. (2010)). As people who inject drugs (PWID) tend to share drug injecting equipment with a limited number of partners (see for example the social network described by Rolls et al. (2012b)), the use of the model described by (4.1) might by questionable.

Hence, assuming that individuals with the same number of sharing partners present the same risk in the infectious process, we choose to focus on the heterogeneous mean field approach, first introduced by Pastor-Satorras and Vespignani (2001). In this approach, each individual is assigned a degree (a number of partners) k, where $1 \leqslant k \leqslant N-1$, and constitutes an element of the network (also referred to as node in graph theory). As individuals with 0 partner do not contribute to the spread of the epidemic, only individuals with at least one partner will be considered. Figure 4.2 offers a visual representation of a static network of sharing partners infected by HCV. For every individual dying, it is assumed that a new susceptible takes its place.

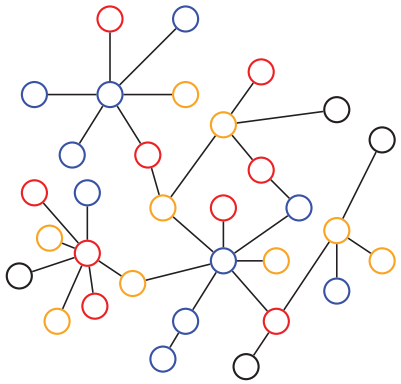


Figure 4.2: Network of PWID infected by HCV and their sharing partners.
Blue nodes represent susceptible individuals, yellow nodes represent individuals in the acute phase of HCV, red nodes represent individuals in the chronic phase of HCV and black nodes are individuals who died from HCV.

Let us denote by ρ the degree distribution, i.e. $\rho(k)$ is the proportion of individuals with degree k. Be denoting $s_k(t)$ (resp. $a_k(t)$ and $c_k(t)$) the proportion of susceptible (resp. acute-HCV infected and chronic-HCV infected) individuals with degree k, one can write:

$$s(t) = \sum_{k=1}^{N-1} s_k(t),\ a(t) = \sum_{k=1}^{N-1} a_k(t) \text{ and } c(t) = \sum_{k=1}^{N-1} c_k(t).$$

Referring now to Pastor-Satorras and Vespignani (2001), the proportion of infected partners for an individual with degree k is given by:

$$\theta_k(t) = \sum_{k'=1}^{N-1} \rho_N(k'|k)\Big(a_{k'}(t) + c_{k'}(t)\Big)$$

where $\rho_N(k'|k)$ is the proportion of partners with degree k' among all partners of an individual with degree k. It can be shown that:

$$\rho_N(k'|k) = \frac{k'\rho(k')}{\langle k\rangle}$$

where $\langle k\rangle$ is the mean degree of the network. As $\rho_N(k'|k)$ does not depend on k, we rewrite $\rho_N(k'|k)$ as $\rho_N(k')$. From what precedes, one can rewrite (4.1) as:

$$\frac{\mathrm{d}s_k(t)}{\mathrm{d}t} = \Lambda - k\beta s_k(t)\theta_k(t) + p\sigma a_k(t) + \gamma c_k(t) - \mu s_k(t), \tag{4.2a}$$

$$\frac{\mathrm{d}a_k(t)}{\mathrm{d}t} = k\beta s_k(t)\theta_k(t) - (\sigma+\mu)a_k(t), \tag{4.2b}$$

$$\frac{\mathrm{d}c_k(t)}{\mathrm{d}t} = (1-p)\sigma a_k(t) - (\gamma+\delta+\mu)c_k(t). \tag{4.2c}$$

To determine which degree distribution should be used to model the drug equipment sharing pattern between PWID, we refer to the work of Dombrowski et al. (2013) who studied the social network of PWID in the Bushwick neighborhood of Brooklyn, New York, by analyzing its degree distribution (see also Pelude (2007)).

This analysis showed that the resulting degree distribution could be assimilated to a power law distribution. Being the main characteristic of scale-free networks (see Barabási and Réka (1999)), power law distributions reflect a preferential attachment dynamics which is consistent with the fact that new PWID network members tend to join a particular injector network via someone who is already part of it.

Henceforth, we assume that the social network of PWID admits a scale-free topology. In particular, one can write $\rho(k) \sim k^{-\alpha}$ where α is called scale parameter. However, according to Clauset et al. (2009), it appears that power laws with exponential cutoff tend to fit natural networks even better than pure

power laws. Indeed, the power law is a long tailed distribution, hence, by applying an exponential cutoff, one can shrink the tail of the distribution faster as individuals with a huge amount of partners is highly improbable. Thus, based on this result and for an arbitrary cutoff κ, one can write:

$$\rho(k) = \frac{k^{-\alpha} e^{-k/\kappa}}{\mathrm{Li}_\alpha\left(e^{-1/\kappa}\right)}$$

where Li_n is the polylogarithm of order n.

4.2.2 Incorporating contact with HIV-infected individuals

The model developed in the previous section should lead to a reasonable approximation of the spread of HCV on a network. However, one may want to add some details by allowing individuals to get infected by human immunodeficiency virus (HIV). Indeed, by dealing with PWID, people infected by HCV are often co-infected by HIV which greatly increases the risk of death.

We now consider the possibility for each individual to get infected by HIV. Hence, we introduce three new compartments representing individuals infected only by HIV (S^*), individuals in the acute phase of the HCV infection and infected by HIV (A^*) and individuals in the chronic phase of the HCV infection and infected by HIV (C^*). Particularly, we have $S(t) + A(t) + C(t) + S^*(t) + A^*(t) + C^*(t) = N$.

We also introduce γ^* and δ^* the recovery and the death rates of HCV for people already infected by HIV. As well, we introduce $\beta^* = 1 - (1 - \pi^*)^\omega$ the transmission rate of HIV, $\beta^\dagger = 1 - (1 - \pi^\dagger)^\omega$ the transmission rate of both HCV and HIV and p^* the spontaneous clearance of HCV for people infected by HIV. Finally, we denote by μ^* the mortality rate of PWID infected by HIV. A visual representation of the model is given in Figure 4.3.

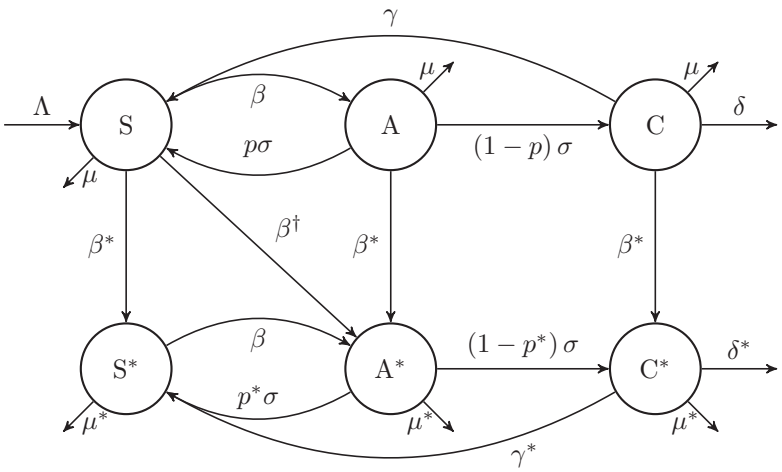


Figure 4.3: Compartmental model of HCV's natural history with HIV coinfection.
S, susceptible; A, acute HCV infection; C, chronic HCV infection. Starred compartments represent HIV infected individuals.

As done in Section 4.2.1, proportions of the different quantities are denoted with small letters. We now introduce the probabilities that an individual with degree k has an infected partner whichever the infection. Hence, we have the probability that an individual with degree k has an HCV infected partner:

$$\theta_k(t) = \sum_{k'=1}^{N-1} \rho_N(k') \Big(a_{k'}(t) + c_{k'}(t) + a^*_{k'}(t) + c^*_{k'}(t) \Big), \tag{4.3}$$

the probability that an individual with degree k has an HIV infected partner:

$$\theta^*_k(t) = \sum_{k'=1}^{N-1} \rho_N(k') \Big(s^*_{k'}(t) + a^*_{k'}(t) + c^*_{k'}(t) \Big), \tag{4.4}$$

and the probability that an individual with degree k has an HCV/HIV co-infected partner:

$$\theta_k^{**}(t) = \sum_{k'=1}^{N-1} \rho_N(k') \Big(a_{k'}^*(t) + c_{k'}^*(t) \Big). \tag{4.5}$$

From those equations and based on the model of Figure 4.3, one can easily rewrite (4.2) as:

$$\frac{\mathrm{d}s_k(t)}{\mathrm{d}t} = \Lambda - k\beta s_k(t)\theta_k(t) - k\beta^* s_k(t)\theta_k^*(t) - k\beta^\dagger s_k(t)\theta_k^{**}(t) + p\sigma a_k(t) + \gamma c_k(t) - \mu s_k(t), \tag{4.6a}$$

$$\frac{\mathrm{d}s_k^*(t)}{\mathrm{d}t} = -k\beta s_k^*(t)\theta_k(t) + k\beta^* s_k(t)\theta_k^*(t) + p^*\sigma a_k^*(t) + \gamma^* c_k^*(t) - \mu^* s_k^*(t), \tag{4.6b}$$

$$\frac{\mathrm{d}a_k(t)}{\mathrm{d}t} = k\beta s_k(t)\theta_k(t) - (\sigma+\mu)a_k(t) - k\beta^* a_k(t)\theta_k^*(t), \tag{4.6c}$$

$$\frac{\mathrm{d}a_k^*(t)}{\mathrm{d}t} = k\beta s_k^*(t)\theta_k(t) + k\beta^\dagger s_k(t)\theta_k^{**}(t) + k\beta^* a_k(t)\theta_k^*(t) - (\sigma+\mu^*)a_k^*(t), \tag{4.6d}$$

$$\frac{\mathrm{d}c_k(t)}{\mathrm{d}t} = (1-p)\sigma a_k(t) - k\beta^* c_k(t)\theta_k^*(t) - (\gamma+\delta+\mu)c_k(t), \tag{4.6e}$$

$$\frac{\mathrm{d}c_k^*(t)}{\mathrm{d}t} = (1-p^*)\sigma a_k^*(t) + k\beta^* c_k(t)\theta_k^*(t) - (\gamma^*+\delta^*+\mu^*)c_k^*(t). \tag{4.6f}$$

Such a framework allows for the study of an epidemic on a given network. However, when dealing with a large population, it might be wiser to deal with several subnetworks connected with each other. Indeed, when dealing with an epidemic, say at a country scale, computing a subnetwork for each city and allowing human movements between them might be more accurate than computing a single giant network. Based on this hypothesis, we propose to study, in the next section, the spread of the epidemic of HCV defined by (4.6) across several subnetworks.

4.3 Epidemic propagation on a metapopulation model

Metapopulation models, first described by Levins (1969), consist in modeling the interactions of several spatially separated populations. Those models have been extensively used in the field of epidemiology to describe the spread of infectious diseases at a large scale by dividing the population into subpopulations corresponding to different households, cities, etc. Depending on human movements, a subpopulation, or patch, containing infected individuals is allowed to interact with other patches spreading the infection to neighboring subpopulations. We refer the reader for instance to Hufnagel et al. (2004) and Colizza et al. (2007) for an illustration of the role of the global aviation network in the spread of the SARS infection.

Here we adapt this type of model to study the spread of HCV on a network of cities, assuming that the population of each city can be considered as a patch of a global metapopulation model. Each subpopulation admits a scale-free structure as defined in Section 4.2.1 and the spread of the virus is governed by the set of differential equations (4.6). To keep things simple while making them concrete, we work in this section on networks of two and three cities.

Two-city network. Let us consider two cities A and B with injecting drug populations N_A and N_B respectively. Assuming that these populations are organized in scale-free networks, we denote respectively by $\langle k_A \rangle$ and $\langle k_B \rangle$ their mean degree and by $\rho_A(k)$ and $\rho_B(k)$ their degree distributions. We update the notations of the quantities of (4.6) by adding a subscript of the corresponding city, e.g. $s_A(t)$ and $s_B(t)$ denote the proportions of susceptible individuals in cities A and B respectively.

To model the effect of B on the spread of the disease in A, we propose to add a node in the network of A which is interacting with every node of A. In the same way, we add an extra node into the network of B to model the effect of A on B. These extra nodes allow the population of A (resp. B) to connect with the new individuals moving from B to A (resp. from A to B). A visual representation of the process is given in Figure 4.4.

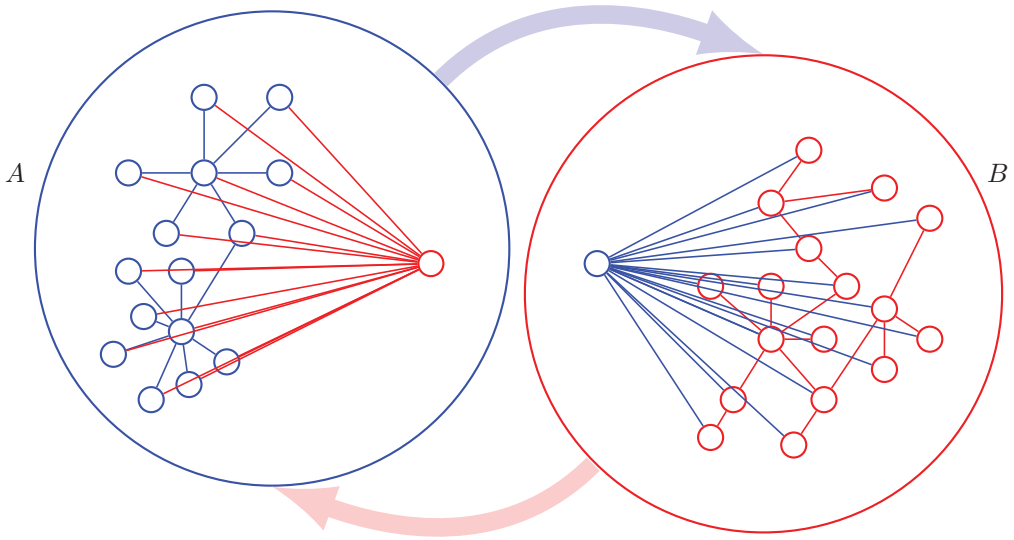


Figure 4.4: Networks of A and B with the additional nodes representing people from A in B and vice versa.
Network of the city A (resp. B) in blue (resp. red) with the additional node of B (resp. A) in red (resp. in blue) interacting with every node. The blue light arrow represents the population travelling from A to B while the red light arrow represents the population from B to A.

The extra node in the network of A contains the main characteristics of B, i.e. the proportion of infected individuals in B and $\tau_{B,A}$ the number of people travelling from B to A. Similarly, the extra node in the network of B contains the main characteristics of A.

Hence, by focusing on the first equation of (4.6a), one can write:

$$\begin{aligned}\frac{\mathrm{d}s_{k,A}(t)}{\mathrm{d}t} &= -k\beta s_{k,A}(t)\theta_{k,A}(t) - k\beta^* s_{k,A}(t)\theta^*_{k,A}(t) - k\beta^\dagger s_{k,A}(t)\theta^{**}_{k,A}(t) - \mu s_{k,A}(t) \\ &\quad -s_{k,A}(t)\left(\frac{\tau_{B,A}}{n_A+\tau_{B,A}}\right)\cdot\Big[\beta\Big(a_B(t)+c_B(t)+a^*_B(t)+c^*_B(t)\Big) \\ &\quad +\beta^*\Big(s^*_B(t)+a^*_B(t)+c^*_B(t)\Big)+\beta^\dagger\Big(a^*_B(t)+c^*_B(t)\Big)\Big] + p\sigma a_{k,A}(t)+\gamma c_{k,A}(t)\end{aligned} \tag{4.7}$$

where n_A is the population still alive in city A. In particular, n_A is defined as:

$$n_A = s_A(t) + a_A(t) + c_A(t) + s^*_A(t) + a^*_A(t) + c^*_A(t).$$

Three-city network. Let us introduce a third city C with population N_C. As above, we assume that C is organized in a scale-free network with mean degree $\langle k_C\rangle$ and degree distribution $\rho_C(k)$. To model the interactions between A and C (resp. B and C), we add another node in the graph of A (resp. B) containing the main characteristics of C. See Figure 4.5 for a graphical visualization of the process in the particular case of A.

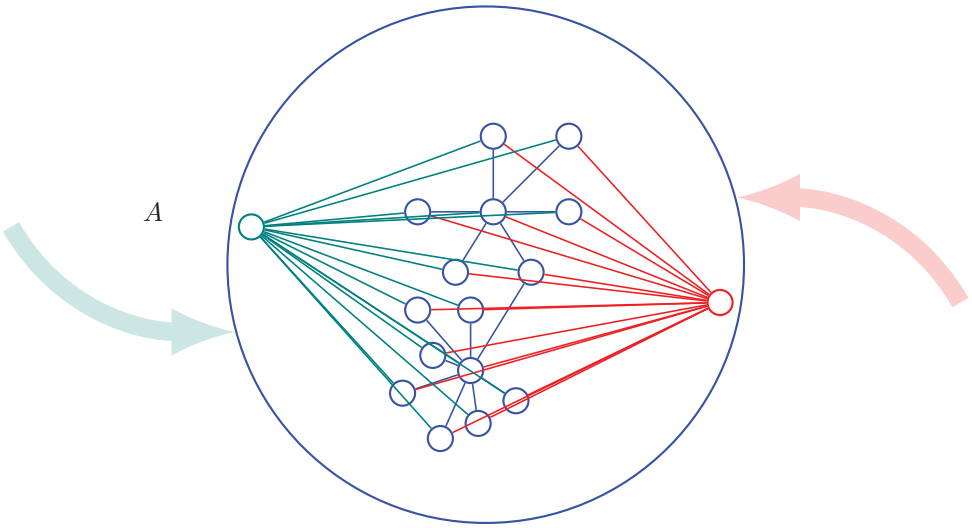


Figure 4.5: Network of A with the additional nodes corresponding to cities B and C. The network of city A is represented in blue. The additional nodes corresponding to cities B and C are represented in red and green respectively. The red light arrow represents the population travelling from B to A while the green light arrow represents the population travelling from C to A.

Adding the new node corresponding to city C, equation (4.7) becomes:

$$\begin{aligned}\frac{\mathrm{d}s_{k,A}(t)}{\mathrm{d}t} &= -k\beta s_{k,A}(t)\theta_{k,A}(t) - k\beta^* s_{k,A}(t)\theta^*_{k,A}(t) - k\beta^\dagger s_{k,A}(t)\theta^{**}_{k,A}(t) - \mu s_{k,A}(t) \\ &\quad - s_{k,A}(t)\sum_{\Phi\in\{B,C\}}\left(\frac{\tau_{\Phi,A}}{n_A+\sum_{\Phi\in\{B,C\}}\tau_{\Phi,A}}\right)\cdot\Big[\beta\Big(a_\Phi(t)+c_\Phi(t)+a^*_\Phi(t)+c^*_\Phi(t)\Big) \\ &\quad + \beta^*\Big(s^*_\Phi(t)+a^*_\Phi(t)+c^*_\Phi(t)\Big)+\beta^\dagger\Big(a^*_\Phi(t)+c^*_\Phi(t)\Big)\Big] + p\sigma a_{k,A}(t)+\gamma c_{k,A}(t).\end{aligned} \tag{4.8}$$

Full details of the complete set of differential equations are available in Appendix A.

4.4 Simulations

4.4.1 Parameter values

Let us first consider three abstract cities A, B and C which PWID communities of 10000, 8000 and 5000 injecting drug users respectively. Cities with larger populations tend to be more attractive, so we use a simple gravitational model to determine the proportion of a city's travelling population driven to a particular city among all other cities:

$$\tau_{A,B} = \frac{m_p n_A n_B}{n_B + n_C}$$

where m_p is the moving population percentage. Values of the different population movements were rounded up and reported in Table 4.1. Those percentages were kept low to consider that fact the population of a given city is far greater than the moving population.

As far as network parameters are concerned, very limited data is available for their estimation. Hence, we use the values obtained by Dombrowski et al. (2013), which were $\alpha = 1.8$ for the scale parameter and $\langle k\rangle = 3.0$ for the mean degree. Based on this, we initiate 1000 simulations of the degree distribution with $\alpha = 1.8$ and different values of the exponential cut-off until reaching the value of 3.0 for the mean degree. The resulting cutoff was $\kappa = 40$. Figure 4.6 illustrates the degree distribution of a network of N=10,000.

One can see on Figure 4.6 that degrees can reach high values ($> 10^2$). Since the model described in Appendix A is degree based, a high degree can lead to a very large amount of differential equations. Hence, for the sake of practicality, we decide to split the population into four arbitrary groups: people with degree 1, people whose degree lies between 2 and 5, people whose degree lies between 6 and 10 and

	Destination					
	$m_p = 0.1\%$			$m_p = 1\%$		
	A	B	C	A	B	C
A	0	6	4	0	62	39
B	5	0	3	53	0	27
C	3	2	0	28	22	0

Table 4.1: Annual origin-destination matrix.

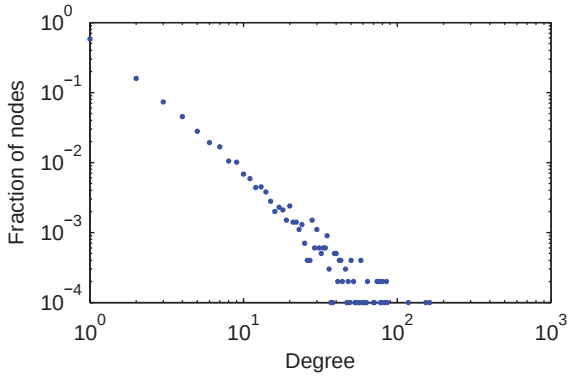


Figure 4.6: Degree distribution of a scale-free network constituted of 10,000 nodes. The degree distribution was computed for a network of 10,000 nodes based on a power law distribution with exponential cutoff which parameters are $\alpha = 1.8$ and $\kappa = 40$.

people whose degree is greater than 10.

The values of the transmission risks of HCV and HIV are difficult to estimate. Several studies reported different values for the per contact probability of HCV transmission for PWID ranging from 0.5% to 10% (see Bayoumi and Zaric (2008); Boelen et al. (2014); O'Leary and Green (2003); Rolls et al. (2012b); Vickerman et al. (2007)). As for the transmission of HCV, we refer to Rolls et al. (2012b) who estimated a transmission risk of 1% in a network of PWID in Australia. Considering the transmission of HIV, we refer to Patel et al. (2014) who reported 63 transmissions for 10,000 exposures for the US epidemic. These transmission risks represent the probabilities of getting infected per act of sharing needles. Assuming that these probabilities are independent, the transmission risk $\pi^\dagger$ of both HCV and HIV is given by $\pi^\dagger = \pi \times \pi^*$.

As for the frequency of needle/syringe sharing, difficulties arise when one tries to obtain accurate estimations. Referring to Vickerman et al. (2007) we choose the value of 16 sharing acts per month. However, as this frequency is shared among the different partners of a given individual, we assume a per-partner sharing frequency such as $16/3 = 5.33$ where 3 is the mean number of partners.

The values of the transmission risks introduced above reflect somehow the true HCV and HIV transmission risks for PWID. However, for illustrative purposes, we propose to temporary choose greater values for those risks by multiplying the transmission risks π, π^* and $\pi^\dagger$ by some constant λ to observe the impact of migrating people on the spread of the epidemic.

Considering the treatment of HCV, we decided to focus on the recent direct-acting antivirals (DAAs) which leads to high recovery rates. Indeed, based on a recent ANRS press release, the recovery rate of the new DAAs has been estimated up to 93% for HCV/HIV co-infected patients (see ANRS (2015)). Similar results were observed for HCV mono-infected patients. Due to the fact that HCV is an asymptomatic disease, a large part of HCV infected patients are not aware of their serostatus. Globally, this involves a very low proportion of infected patients accessing to health care. Indeed, although France has one of the highest treatment rates in Europe, this proportion hardly reaches 5.2% (see Razavi et al. (2014)).

As for the mortality rates of PWID infected by HCV, HIV or HCV and HIV, we refer to the work of van Santen et al. (2014) who calculated all-cause and cause-specific crude mortality rates (per 1000 person-years) and standardized mortality ratios for PWID in the Netherlands. A summary of the parameters used in the model are given in Table 4.2.

Since we want to quantify the impact of the interactions between cities, we assume that the epidemic starts in one of the three cities and observe the impact on the two others. For the epidemic to begin, we set a relatively low amount of individual in city A such as $s_A(0) = 0.985$, $c_A(0) = 0.01$ (HCV mono-infected individuals in chronic phase) and $c_A^*(0) = 0.005$ (HIV/HCV co-infected individuals in chronic phase). The values chosen for the initiation of the model are purely arbitrary.

Due to the random nature of the network framework, we will focus on the mean of each quantity over 1000 simulations.

Table 4.2: Parameters used in the model to simulate the spread of HCV.

Parameter	Notation	Value	Source
Risk of transmission per contact			
HCV, %	π	1	Rolls et al. (2012b)
HIV, %	π^*	0.63	Patel et al. (2014)
HCV and HIV, %	$\pi^\dagger$	6.3×10^{-5}	–
Semi-annual HCV spontaneous clearance			
HCV mono-infected, %	p	26	Micallef et al. (2006)
HCV/HIV co-infected, %	p^*	10	Hernandez and Sherman (2011)
Annual risk of death			
HCV/HIV uninfected, ‰	μ	10.4	van Santen et al. (2014)
HCV mono-infected, ‰	δ	22.7	van Santen et al. (2014)
HIV mono-infected, ‰	μ^*	44.3	van Santen et al. (2014)
HCV/HIV co-infected, ‰	δ^*	54.9	van Santen et al. (2014)
Network characteristics			
Mean degree	$\langle k \rangle$	3.0	Dombrowski et al. (2013)
Scale parameter	α	1.8	Dombrowski et al. (2013)
Exponential cutoff	κ	40	–
Anti-HCV treatment characteristics			
Recovery rate, %	ν	93	ANRS (2015)
Proportion of treated patients, %	q	5.2	Razavi et al. (2014)
Other parameters			
Per-partner sharing frequency, per month	ω	5.33	Vickerman et al. (2007)
Duration of the acute phase, years	$1/\sigma$	1/2	Micallef et al. (2006)
Unit time, years	–	1	–

4.4.2 Results on an abstract three-city network

Figures 4.7, 4.8 and 4.9 display the evolution of the susceptible, the HCV mono-infected and HIV/HCV co-infected populations in the whole system (cities A, B and C). The different curves of the three figures represent the population dynamics when no migration is assumed (blue markers), 0.1% of the population is migrating (red markers) and 1% of the population is migrating (green markers). Each quantity displayed in Figures 4.7, 4.8 and 4.9 is the sum of the corresponding quantity over the four degree groups.

Considering Figure 4.7, one can see that the susceptible population gradually depletes with the percentage of migrating population. As more and more people move from city A, the epidemic spreads faster

and has a stronger impact on the cities B and C. As expected, the spread of the epidemic is even stronger when the coefficient α increases. One can see that the blue curves reach a minimum value between 60% and 70%. This phenomenon is due to the fact that the susceptible populations of the cities B and C stay intact (no initial infected individual nor migration from infected cities).

As for Figure 4.8, one can observe the same dynamics as for Figure 4.7. The proportion of HCV mono-infected individuals in the population is greatly impacted by the percentage of migrating people and by the coefficient α. Similarly, the blue curves here correspond to the proportion of HCV mono-infected individuals in the city A. This explains why that proportion is never higher than 33.3%.

Lastly, Figure 4.9 exhibits the same evolutions as the two previous figures. One can particularly observe on Figure 4.9 that the evolution of the HIV/HCV co-infected population is greater than the evolution of the HCV mono-infected population of Figure 4.8. This is due to the fact that, contrary to HCV, no treatment for HIV is considered in the model. Thus, each HCV mono-infected individual tends to get also infected by HIV as time goes by.

In the next section we propose to study the spread of the infection assuming the epidemic starts from a particular patch. To highlight the impact of people migration on the spread of an infection, two different patches are studied.

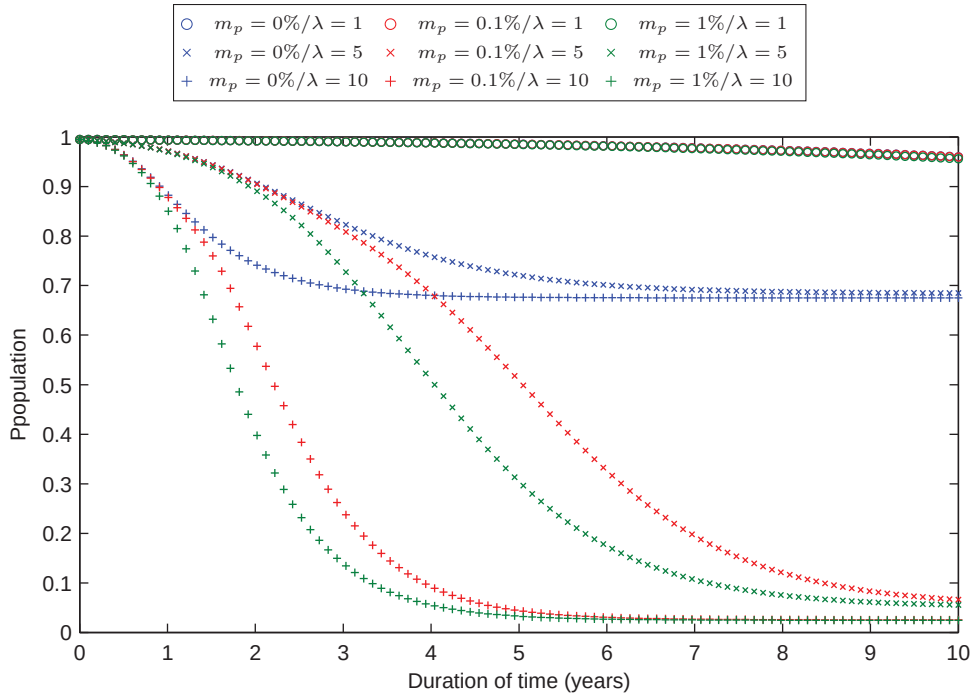


Figure 4.7: Mean evolution of the susceptible population in the whole system.
The different curves represent the mean quantities over 1000 simulations. Blue curves represent the evolution of the susceptible population when no interactions between the cities are allowed ($m_p = 0\%$), red curves represent the evolution of the susceptible population assuming interactions up to 0.1% of the populations ($m_p = 0.1\%$) and green curves represent the evolution of the susceptible population assuming interactions up to 1% of the populations ($m_p = 1\%$).

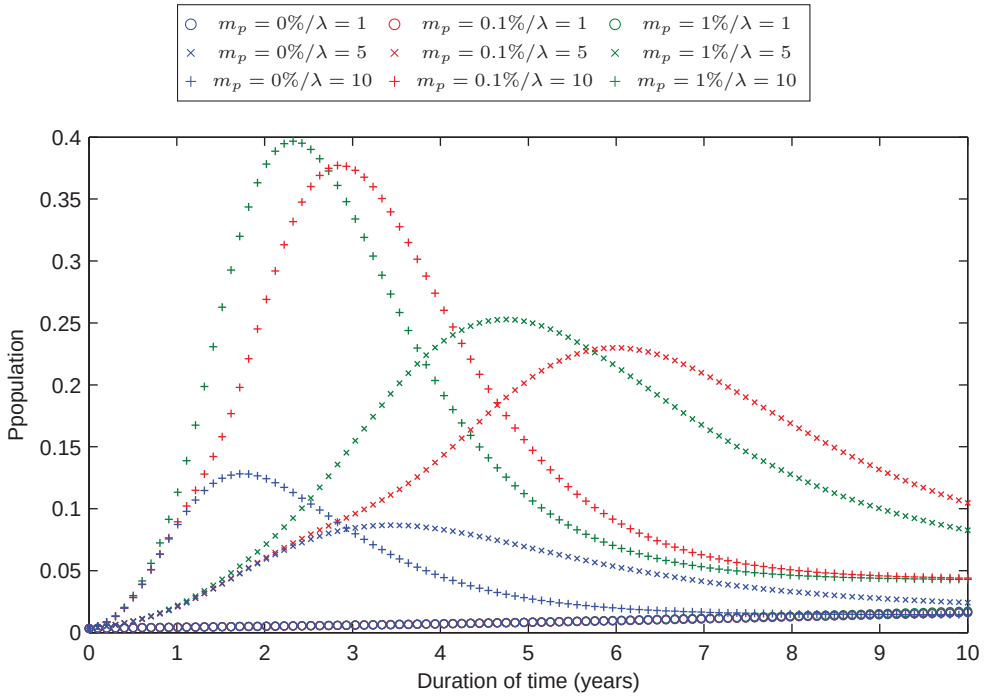


Figure 4.8: Mean evolution of the HCV mono-infected population in the whole system. The different curves represent the mean quantities over 1000 simulations. Blue curves represent the evolution of the HCV mono-infected population when no interactions between the cities are allowed ($m_p = 0\%$), red curves represent the evolution of the HCV mono-infected population assuming interactions up to 0.1% of the populations ($m_p = 0.1\%$) and green curves represent the evolution of the HCV mono-infected population assuming interactions up to 1% of the populations ($m_p = 1\%$).

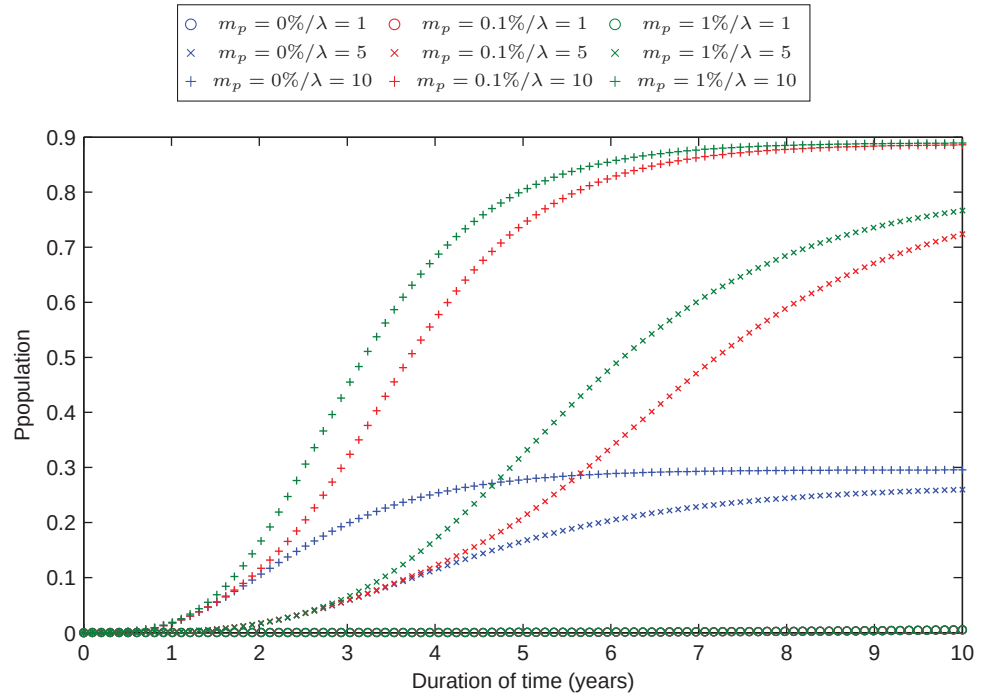


Figure 4.9: Mean evolution of the HIV/HCV co-infected population in the whole system. The different curves represent the mean quantities over 1000 simulations. Blue curves represent the evolution of the HIV/HCV co-infected population when no interactions between the cities are allowed ($m_p = 0\%$), red curves represent the evolution of the HIV/HCV co-infected population assuming interactions up to 0.1% of the populations ($m_p = 0.1\%$) and green curves represent the evolution of the HIV/HCV co-infected population assuming interactions up to 1% of the populations ($m_p = 1\%$).

4.4.3 Results on the network of France's 100 largest urban areas

We generalize the metapopulation model of Section 4.4.1 by studying the spread of HCV across a whole country: France. As we want our predictions to be relatively accurate without considering every single city, we choose to focus on the first hundred biggest urban areas in France (metropolitan France only) for a population of 43,187,838 individuals in 2012. Indeed, by referring to the work of Rachlis et al. (2007) who identified legal problems, entering drug treatment program and drug tourism as PWID's main drivers of mobility, we assume that the great majority of PWID tend to move towards and between the main urban areas.

Then, the population of these urban areas is multiplied by the prevalence of PWID in France estimated at 5.9 ‰ by the OFDT[1] in 2006 (see Costes et al. (2009)). Since this prevalence has been estimated for the population aged 15 to 64 (63,8% of the total French population in 2012, i.e 40,433,870 individuals), we multiply the population of the French urban areas by $40433870/43187838 = 0.94$ for both populations to match. The original populations retained for the simulations are reported in Appendix B.

Considering the origin-destination matrix of such a model, no data is available for the annual migration of PWID. Hence, we refer to a French database of residential mobility during a five-year period published on the website of INSEE[2]. Values are divided by 5 and rounded to obtain an annual origin-destination matrix. To only consider PWID, values are multiplied by 5.9 ‰.

As we want to study the impact of migrating populations on the spread of the infection, we shall consider two different starting points. We first choose Paris as a starting point since this city has the highest level of interactions with nearly all other French cities. To select the second starting point, we need a city which possesses a far lower level of interactions with all other cities than Paris. However, to prevent the snowball effect, it is better to also focus on a city which has very limited interactions with Paris. Hence, we choose Forbach, a small urban area in the North-Eastern part of France, as the starting point in a second set of simulations.

Referring to Section 4.4.1, we set the proportion of susceptible individuals to 98.5% in the city where the epidemic starts. The proportions of HCV mono-infected (chronic phase) and HIV mono-infected patients are set to 1% and 0.5% respectively in that same city. All other cities are initiated at 100% of susceptible individuals. Similarly, we assume that each urban area is organized according to a scale-free network which degree distribution follows a power law with exponential cutoffs and assume a multiplicative constant λ of 10. Parameters considered for this model are reported in Table 4.2.

Results. Evolution of the different mean quantities are represented in Figure 4.10 where two different scenarios are exhibited in plain and dashed curves. One notices that, as expected, the epidemic spreads much more easily to the whole country when starting from Paris rather than Forbach. Indeed, as Paris has a lot more interactions with other urban areas than Forbach, the epidemic spreads faster. Particularly, one can see that the proportion of susceptible individuals depletes rapidly between the third and the eighth year when the epidemic starts in Paris while this proportion depletes between the seventh and the fourteenth year when the epidemic starts in Forbach. Similarly, one can observe that the proportion of HCV mono-infected patients increases during a few years then decreases due to the superinfection with HIV.

Figures 4.11 and 4.12 give a comparison of the spread of the infection in the country. Generally, one can see that the epidemic spreads further and faster when starting in Paris. Indeed, one can observe that, ten years after the beginning of the epidemic, the high majority of urban areas admits a proportion of HCV infected individuals (mono-infection or co-infection) superior to 40% when the epidemic outbreak takes place in Paris. As for Forbach's epidemic outbreak, not even all urban areas are affected by the infection after ten years. Last, one can denote that the last epidemic-affected city seems to be Cluses in both Figures 4.11 and 4.12. This can be explained by the fact that this urban area presents the lowest level of interaction with other French cities (it only interacts significantly with Genève-Annemasse).

[1] *Observatoire français des drogues et des toxicomanies*, the French monitoring centre for drugs & drug addiction.
[2] *Institut national de la statistique et des études économiques*, the French office for national statistics.

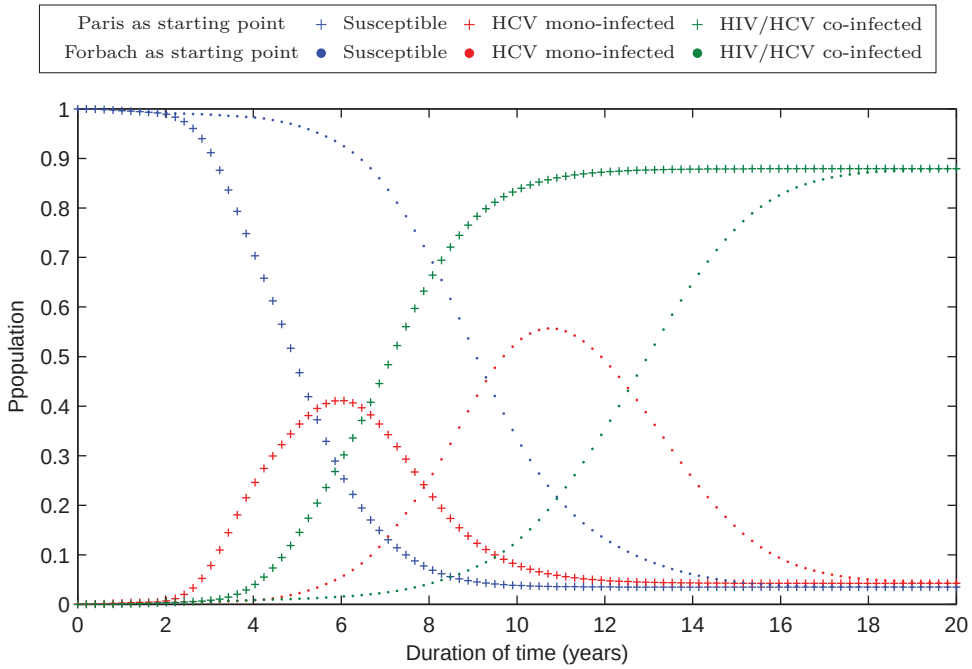
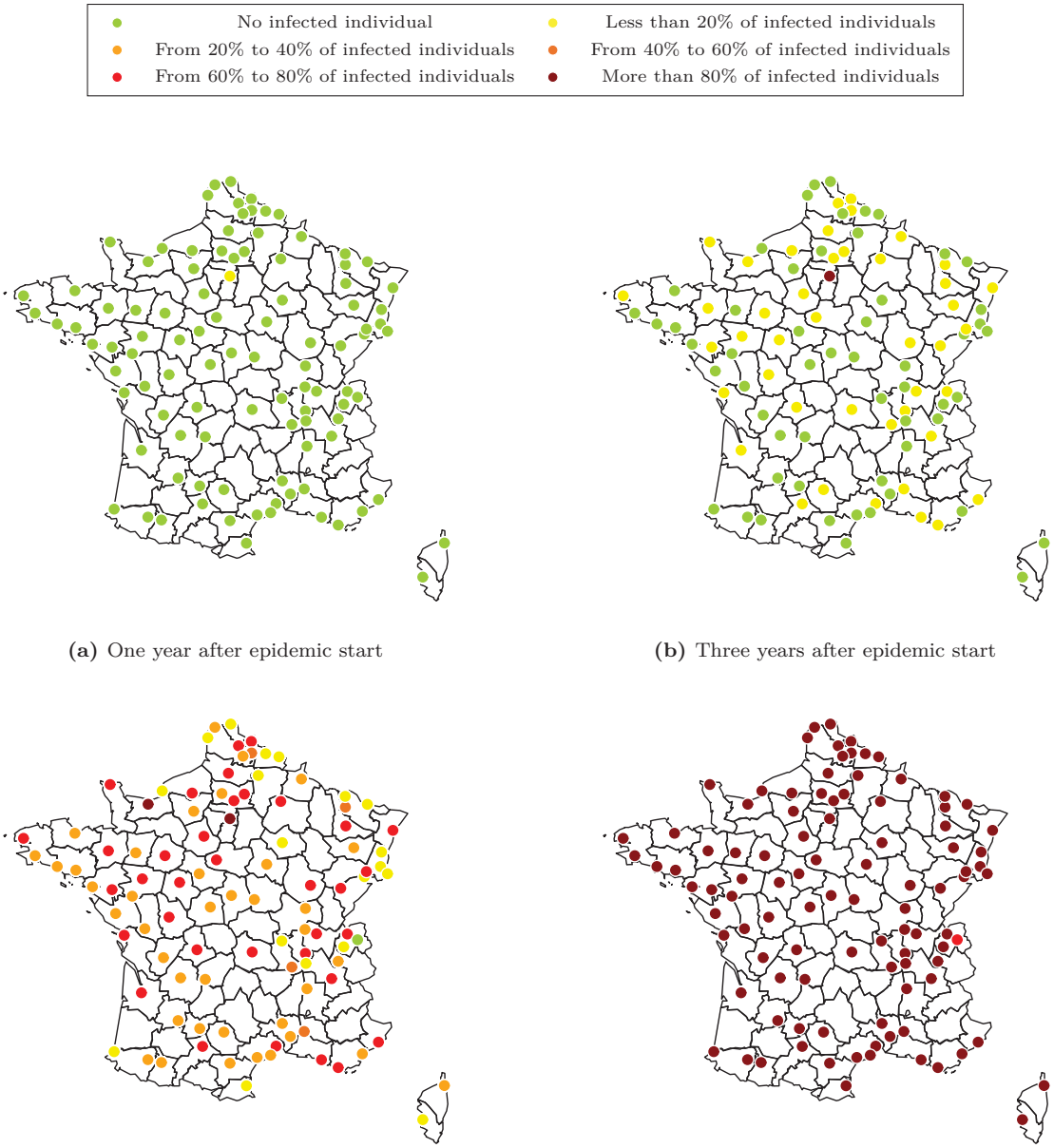


Figure 4.10: Mean evolution of the population dynamics in France.
The different curves represent the mean quantities over 1000 simulations. Dotted curves represent the evolution of the French population when Forbach is chosen as the epidemic starting point while (+) markers represent the evolution of the French population when Paris is chosen as the epidemic starting point. The blue markers represent the susceptible population while the red and the green markers represent the HCV mono-infected and the HIV/HCV co-infected populations respectively



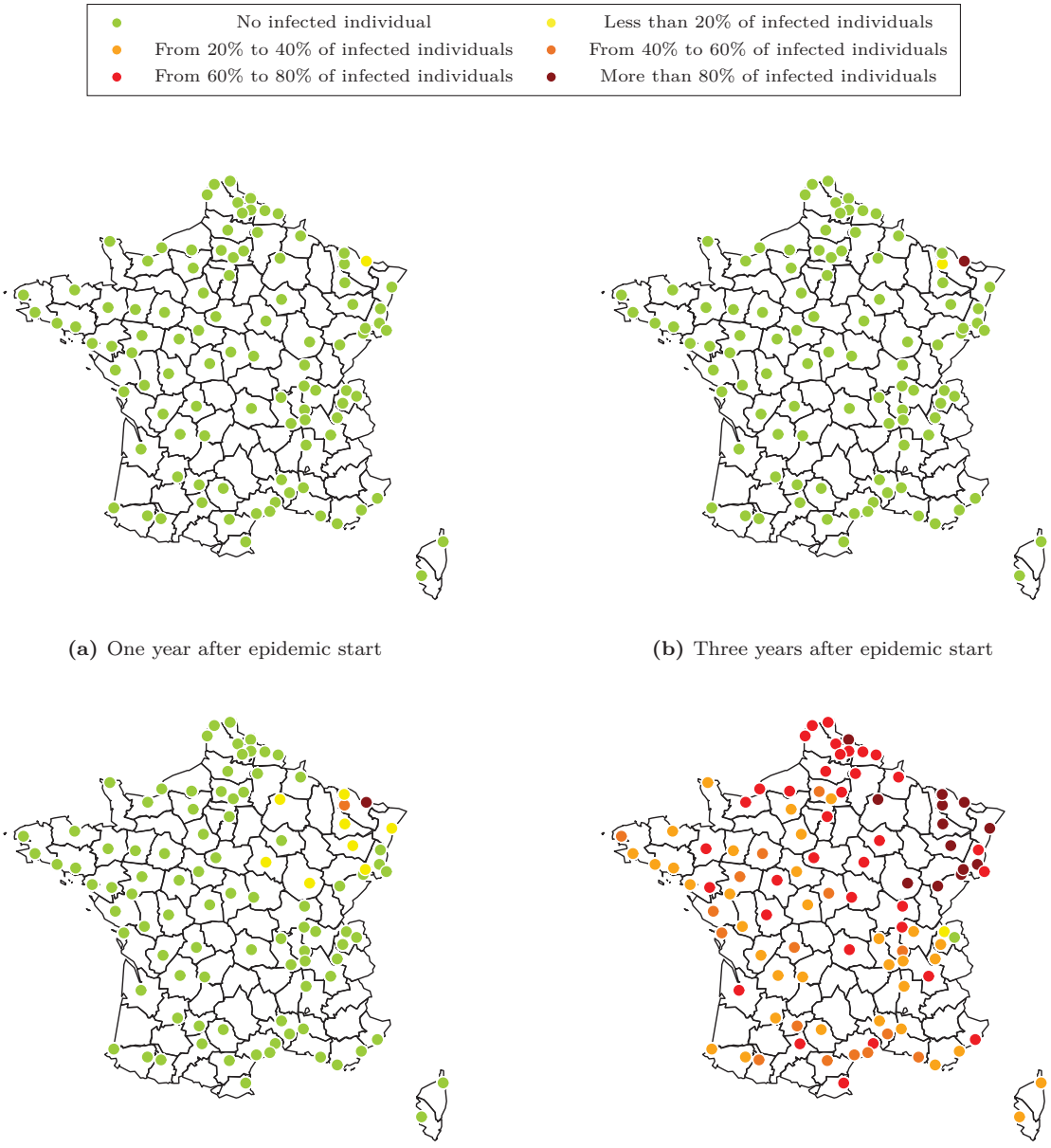
(a) One year after epidemic start

(b) Three years after epidemic start

(c) Five years after epidemic start

(d) Ten years after epidemic start

Figure 4.11: Evolution of the spread of HCV in France with an epidemic start in Paris. Estimations of the proportions of HCV infected individuals in France over 1000 simulations. Each colored dot represents a French urban area. Starting point was 98.5% of susceptible, 1% of HCV infected individuals and 0.5% of HIV infected individuals in Paris. Every other city started at 100% of susceptible individuals.



(a) One year after epidemic start

(b) Three years after epidemic start

(c) Five years after epidemic start

(d) Ten years after epidemic start

Figure 4.12: Evolution of the spread of HCV in France with an epidemic start in Forbach. Estimations of the proportions of HCV infected individuals in France over 1000 simulations. Each colored dot represents a French urban area. Starting point was 98.5% of susceptible, 1% of HCV infected individuals and 0.5% of HIV infected individuals in Forbach. Every other city started at 100% of susceptible individuals.

4.5 Enhanced HCV model

In the following sections, we enhance the original model developed in Section 4.2.2 by adding supplementary details such as the progression of fibrosis, the development of end-stage liver diseases or the differentiation between the main HCV genotypes. We also examine an alternative model for non-injecting drug users to track HCV infected patients once they leave the injecting-drug population. In this alternative model, transmission of HCV or HIV is not considered as non-IDU do not participate to the force of infection within the injecting-drug population anymore.

4.5.1 Model structure

4.5.1.1 Fibrosis progression and end-stage liver diseases

The natural history of HCV is mainly characterized by the destruction of the liver in a process called fibrosis. The progression of fibrosis along the years can be divided in five different stages associated with the gravity of the fibrosis. See the Metavir scores in Table 4.3 for a description of the different stages.

Table 4.3: Metavir fibrosis scoring

Score	Description
F_0	No fibrosis
F_1	Portal fibrosis without septa
F_2	Portal fibrosis with few septa
F_3	Numerous septa without cirrhosis
F_4	Cirrhosis

Once an individual reaches the late fibrosis stages (typically F_3 or F_4), one may develop a decompensated cirrhosis (DC) and/or a hepatocellular carcinoma (HCC) which lead to a short-term death if no appropriate treatment is undertaken.

Let us denote by $F_0(t)$ (resp. $F_1(t)$, $F_2(t)$, $F_3(t)$ and $F_4(t)$) the number of individuals in stage 0 (resp. 1, 2, 3 and 4) of fibrosis at time t, by $HCC(t)$ the number of individuals with a hepatocellular carcinoma at t and by $DC(t)$ the number of individuals with a decompensated cirrhosis at t. From what precedes, one can modify state C of the model in Figure 4.3 by including the stages corresponding to the fibrosis progression and to end-stage liver diseases (ESLD). Such a model is displayed in Figure 4.13 in the specific case of HCV mono-infection where λ_1, λ_2, λ_3, λ_4, λ_h, λ_H, λ_{D_H}, λ_D, λ_{H_D} and λ_{D_D} are the different progression rates to the stages of fibrosis, HCC, decompensated cirrhosis and death due to HCV. The natural mortality rate μ is not displayed for better clarity.

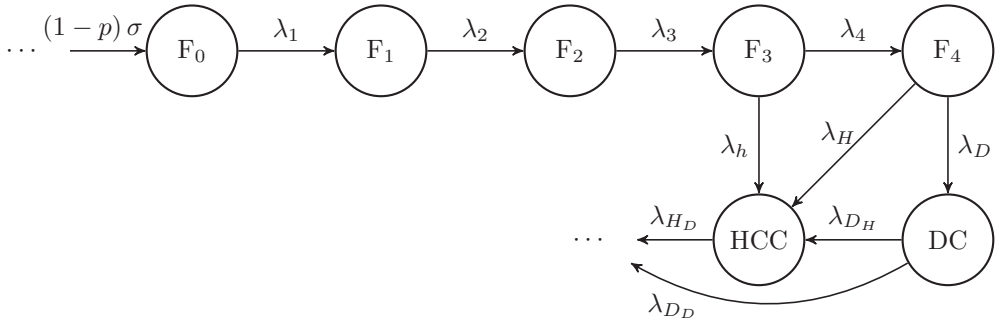


Figure 4.13: Compartmental model of fibrosis progression and end-stage liver diseases. F_0-F_4, stages of the fibrosis progression based on the Metavir scoring system indicating the degree of inflammation; HCC, hepatocellular carcinoma; DC, decompensated cirrhosis.

To compute the differential equation of the proportion of susceptible individuals in that particular

framework, we denote by:

$$\phi(t) = a(t) + f_0(t) + f_1(t) + f_2(t) + f_3(t) + f_4(t) + dc(t) + hcc(t)$$

the proportion of individuals infected by HCV at time t. Similarly, let us denote by:

$$\phi^*(t) = a^*(t) + f_0^*(t) + f_1^*(t) + f_2^*(t) + f_3^*(t) + f_4^*(t) + dc^*(t) + hcc^*(t)$$

the proportion of individuals infected by HCV and HIV at time t. Hence, one can respectively rewrite equations (4.3), (4.4) and (4.5) as:

$$\theta_k(t) = \sum_{k'=1}^{N-1} \rho_N(k') \Big(\phi_{k'}(t) + \phi_{k'}^*(t) \Big), \tag{4.9a}$$

$$\theta_k^*(t) = \sum_{k'=1}^{N-1} \rho_N(k') \Big(s_{k'}^*(t) + \phi_{k'}^*(t) \Big) \tag{4.9b}$$

$$\text{and} \quad \theta_k^{**}(t) = \sum_{k'=1}^{N-1} \rho_N(k') \times \phi_{k'}^*(t). \tag{4.9c}$$

Based on these notations, one can rewrite (4.8) as:

$$\begin{aligned} \frac{\mathrm{d}s_{k,A}(t)}{\mathrm{d}t} \quad = \quad & -k\beta s_{k,A}(t)\theta_{k,A}(t) - k\beta^* s_{k,A}(t)\theta_{k,A}^*(t) - k\beta^\dagger s_{k,A}(t)\theta_{k,A}^{**}(t) - \mu s_{k,A}(t) \\ & - s_{k,A}(t) \sum_{\Phi} \left(\frac{\tau_{\Phi,A}}{n_A + \sum_\Phi \tau_{\Phi,A}} \right) \cdot \Big[\beta \Big(\phi_\Phi(t) + \phi_\Phi^*(t) \Big) \\ & + \beta^* \Big(s_\Phi^*(t) + \phi_\Phi^*(t) \Big) + \beta^\dagger \phi_\Phi^*(t) \Big] + p\sigma a_{k,A}(t). \end{aligned} \tag{4.10}$$

4.5.1.2 Main HCV genotypes

In the previous section, no particular attention was paid to the genotype of HCV. Yet, it has been clearly established that HCV genotypes entice a significant difference in the progression of fibrosis, the occurence of ESLD and the response to treatment. Indeed, recent studies have demonstrated an acceleration of the fibrosis progression and an increase of hepatocellular carcinoma risk incidence in genotype 3 HCV infected patients compared to genotype 1 HCV infected patients (see Asselah et al. (2006); Bochud et al. (2009); Kanwal et al. (2014); Nkontchou et al. (2011); Tapper and Afdhal (2013)). Those results are particularly important since genotypes 1 and 3 are the most common genotypes in France (see Payan et al. (2005)).

Henceforth, we choose to focus on genotypes 1 and 3. Let us assume that the notations introduced in Section 4.5.1.1 refer now to individuals infected by genotype 1 HCV and let us introduce equivalent notations for genotype 3 HCV infected individuals. By denoting $\check{\mathrm{A}}$ the number of patients in the acute phase of genotype 3 HCV infection, each individual can either get infected by genotype 1 HCV or genotype 3 HCV as displayed in Figure 4.14. As above, the natural mortality rate μ is not displayed for better clarity.

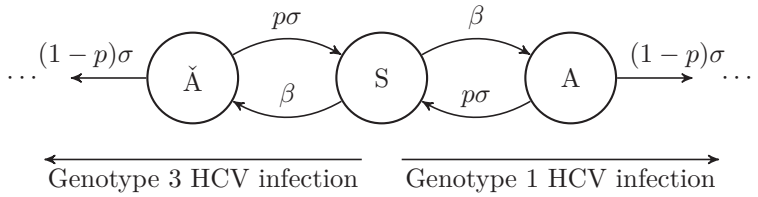


Figure 4.14: Compartmental model of the infection by the different HCV genotypes.

Next, we denote by $\check{\mathrm{F}}_0(t)$ (resp. $\check{\mathrm{F}}_1(t)$, $\check{\mathrm{F}}_2(t)$, $\check{\mathrm{F}}_3(t)$ and $\check{\mathrm{F}}_4(t)$) the number of genotype 3 HCV infected individuals in the stage 0 (resp. 1, 2, 3 and 4) of fibrosis at t, by $\mathrm{H\check{C}C}(t)$ the number of genotype 3 HCV infected individuals with a hepatocellular carcinoma at t and by $\mathrm{D\check{C}}(t)$ the number of genotype 3 HCV

infected individuals with a decompensated cirrhosis at t. Based on these notations, we consider the model displayed in Figure 4.15. As done in Figure 4.13, HIV co-infected individuals were not displayed in Figure 4.15 for the sake of simplicity. The progression rates κ_1, κ_2, κ_3, κ_4, κ_h, κ_H, κ_{D_H}, κ_D, κ_{H_D} and κ_{D_D} are equivalent to the rates introduced in Section 4.5.1.1 for genotype 3 HCV infected individuals.

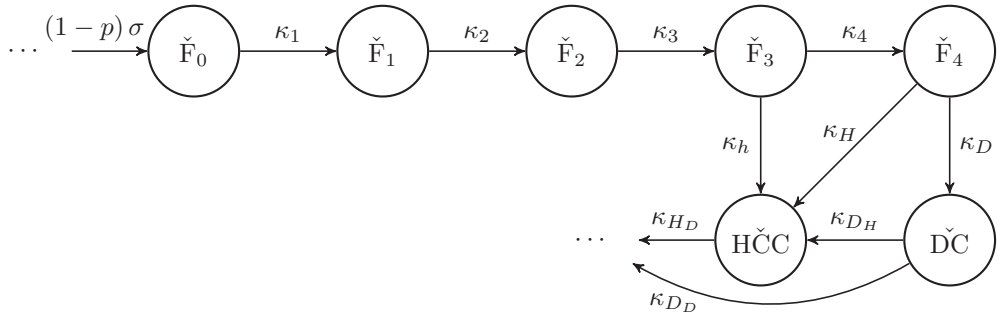


Figure 4.15: Compartmental model of fibrosis progression and end-stage liver diseases for patients infected by genotype 3 HCV.
$\check{\mathrm{F}}_0$-$\check{\mathrm{F}}_4$, stages of the fibrosis progression based on the Metavir scoring system indicating the degree of inflammation; HČC, hepatocellular carcinoma; DČ, decompensated cirrhosis.

Based on the model displayed in Figure 4.15, we denote by:

$$\psi(t) = \check{a}(t) + \check{f}_0(t) + \check{f}_1(t) + \check{f}_2(t) + \check{f}_3(t) + \check{f}_4(t) + \check{dc}(t) + \check{hcc}(t)$$

the proportion of individuals with genotype 3 HCV infection at t. Similarly, we denote by:

$$\psi^*(t) = \check{a}^*(t) + \check{f}_0^*(t) + \check{f}_1^*(t) + \check{f}_2^*(t) + \check{f}_3^*(t) + \check{f}_4^*(t) + \check{dc}^*(t) + \check{hcc}^*(t)$$

the proportion of individuals with genotype 3 HCV and HIV co-infection at time t. Hence, one can modify the expression of (4.9) as follows:

$$\begin{aligned}
\theta_k(t) &= \sum_{k'=1}^{N-1} \rho_N(k')\Big(\phi_{k'}(t) + \phi^*_{k'}(t) + \psi_{k'}(t) + \psi^*_{k'}(t)\Big),\\
\theta^*_k(t) &= \sum_{k'=1}^{N-1} \rho_N(k')\Big(s^*_{k'}(t) + \phi^*_{k'}(t) + \psi^*_{k'}(t)\Big)\\
\text{and}\quad \theta^{**}_k(t) &= \sum_{k'=1}^{N-1} \rho_N(k')\Big(\phi^*_{k'}(t) + \psi^*_{k'}(t)\Big).
\end{aligned}$$

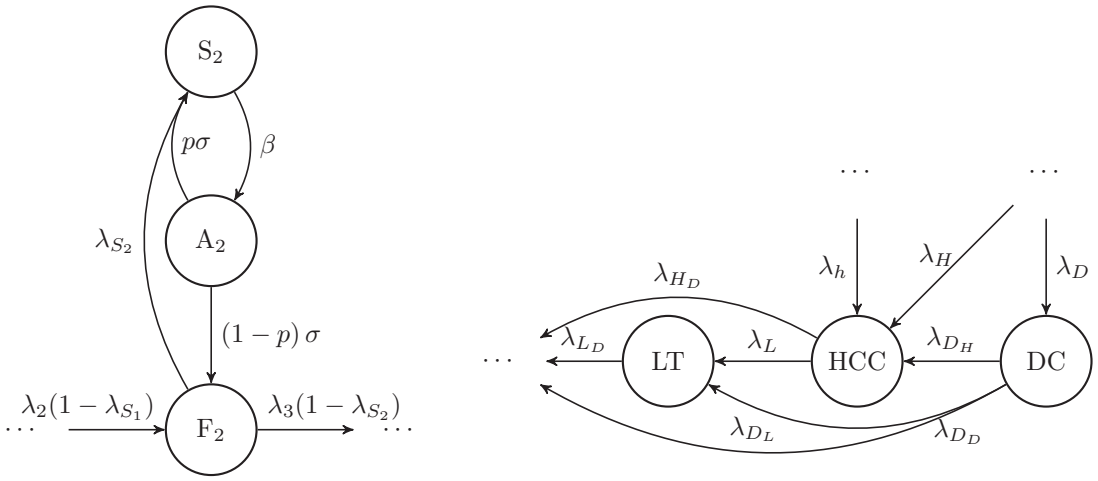
Based on what precedes, one can write:

$$\begin{aligned}
\frac{\mathrm{d}s_{k,A}(t)}{\mathrm{d}t} &= -k\beta s_{k,A}(t)\theta_{k,A}(t) - k\beta^* s_{k,A}(t)\theta^*_{k,A}(t) - k\beta^\dagger s_{k,A}(t)\theta^{**}_{k,A}(t) - \mu s_{k,A}(t)\\
&\quad -s_{k,A}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,A}}{n_A + \sum_\Phi \tau_{\Phi,A}}\right)\cdot\Big[\beta\Big(\phi_\Phi(t) + \phi^*_\Phi(t) + \psi_\Phi(t) + \psi^*_\Phi(t)\Big)\\
&\quad +\beta^*\Big(s^*_\Phi(t) + \phi^*_\Phi(t) + \psi^*_\Phi(t)\Big) + \beta^\dagger\Big(\phi^*_\Phi(t) + \psi^*_\Phi(t)\Big)\Big] + p\sigma(a_{k,A}(t) + \check{a}_{k,A}(t)).
\end{aligned} \tag{4.11}$$

In the next section, we add another feature by integrating the possibility for patients to receive treatment or transplant during the course of HCV.

4.5.1.3 Anti-HCV treatment and liver transplant

Treatment of HCV has always been a challenge due to the low success rates of the different treatments and to the occurrence of numerous side effects. In the past few years, new treatments have been developed offering a greater improvement in terms of success rates and a lower adverse events incidence. Defined as the achievement of sustained virology response (undetectable viral load), the success rates of those new therapies, referred as second-generation direct-acting antivirals (2^{nd}-gen DAAs), can exceed 95% after



(a) Direct-acting antivirals treatment **(b)** Liver transplant

Figure 4.16: Compartmental models of HCV treatment and liver transplant.
HCC, hepatocellular carcinoma; DC, decompensated cirrhosis; LT, liver transplant. (a) Compartmental model for HIV uninfected patients in fibrosis stage F2 treated with direct-acting antivirals. (b) Compartmental model for HIV uninfected patients transplanted after reaching hepatocellular carcinoma and/or decompensated cirrhosis.

12 or 24 weeks of therapy (see Hézode (2016); Pol et al. (2016); Wyles et al. (2015)). Moreover, 2nd-gen DAAs showed high sustained virology response (SVR) rates for every genotype and for patients in any fibrosis stage.

Considering treatment indication, patients at greater risk initiate 2nd-gen DAAs sooner to limit the liver destruction caused by the fibrosis progression. Indeed, in France, patients with HCV mono-infection do not initiate 2nd-gen DAAs before reaching the fibrosis stage F2 while HCV/HIV co-infected patients can initiate the treatment at any fibrosis stage (see Haute Autorité de Santé (2014)). However, the French National Authority for Health recently recommended to extend the indication of those 2nd-gen DAAs to F0-F1 patients without comorbidities. Despite the efficacy of those new treatments, 2nd-gen DAAs initiation might not be the most viable option for patients who reach ESLD. Indeed, with survival rates constantly improving, liver transplant might represent a better alternative for eligible patients.

In this section, we add to our model more specific patient care by assigning a probability to reach SVR depending on the fibrosis stage. Considering liver transplant, this feature is only available in the alternative model since PWID are not eligible (see Dhumeaux (2014)).

For the sake of simplicity, we assume that the fibrosis process cannot regress even after SVR achievement. Hence, we introduce new compartments corresponding to individuals who cured the virus but are still affected by either fibrosis, decompensated cirrhosis and/or HIV. Similarly, some new compartments are introduced for patients who enters an acute phase after reinfection. By denoting S_2 (resp. A_2) the number of susceptible (resp. acutely infected) individuals in fibrosis stage F2, one can observe on Figure 4.16(a) an illustration of HIV uninfected patients who cured the virus in fibrosis stage F_2 and got reinfected. Figure 4.16(b) displays the compartmental model of patients accessing liver transplantation where λ_L (resp. λ_{D_L}) is the transition rate from hepatocellular carcinoma (resp. decompensated cirrhosis) to liver transplant.

Denoting $s_{2,k,A}(t)$ the proportion of susceptible individual with stage F2 fibrosis and degree k in sub-population A, one can deduce from Figure 4.16(a) and Equation (4.11) the following differential

equation:

$$\begin{aligned}\frac{\mathrm{d}s_{2,k,A}(t)}{\mathrm{d}t} &= -k\beta s_{2,k,A}(t)\theta_{k,A}(t) - k\beta^* s_{2,k,A}(t)\theta^*_{k,A}(t) - k\beta^\dagger s_{2,k,A}(t)\theta^{**}_{k,A}(t)\\ &\quad -s_{2,k,A}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,A}}{n_A+\sum_\Phi \tau_{\Phi,A}}\right)\cdot\Big[\beta^\dagger\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\\ &\quad +\beta^*\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big)\Big]\\ &\quad +p\sigma\Big(a_{2,k,A}(t)+\breve{a}_{2,k,A}(t)\Big)+\lambda_{S_2}f_{2,k,A}(t)+\kappa_{S_2}\breve{f}_{2,k,A}(t)-\mu s_{2,k,A}(t).\end{aligned}$$

where $\lambda_{S_2}=\nu_2\times q$ and $\kappa_{S_2}=\breve{\nu}_2\times q$ correspond to the proportions of patients recovering from stage F2 fibrosis either infected by genotype 1 or genotype 3 respectively. Note that q is the proportion of treated patients introduced in Section 4.2.1 and ν_2 (resp. $\breve{\nu}_2$) is the recovery rate of a patient in fibrosis stage F2 infected by the genotype 1 (resp. genotype 3). In all the following, we denote by ν_i (resp. $\breve{\nu}_i$) the recovery rate of a patient in fibrosis stage Fi, $i=0,\dots,4$, either infected by genotype 1 or genotype 3 respectively. Recovery rate for patients with decompensated cirrhosis and genotype 1 (resp. genotype 3) infection is denoted by ν_D (resp. $\breve{\nu}_D$). Similarly, referring to Figure 4.16(b), one can write:

$$\begin{aligned}\frac{\mathrm{d}lt_{k,A}(t)}{\mathrm{d}t} &= \lambda_{D_L}dc_{k,A}(t)+\lambda_L hcc_{k,A}(t)-k\beta^* lt_{k,A}(t)\theta^*_{k,A}(t)-(\lambda_{L_D}+\mu)lt_{k,A}(t)\\ &\quad -lt_{k,A}(t)\sum_{\Phi}\beta^*\left(\frac{\tau_{\Phi,A}}{n_A+\sum_\Phi\tau_{\Phi,A}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big).\end{aligned}$$

4.5.1.4 Initiation and cessation rates

Let us now consider the possibility for an individual to enter or leave the population. Indeed, by focusing on IDU, one should consider the possibility for an individual to initiate the use of injectable drugs and join the network of IDU or to discontinue the use of drugs. In that optic, let us denote by τ the cessation rate for each individual.

Considering the number of individuals entering the population each year, we assume that the total number of people in each sub-population stays constant over time. Indeed, as difficulties emerge when one try to obtain reliable data not only on the incidence of IDU in France but also on the prevalence itself, assuming a constant-size population might be the best option. Hence, we denote by $\Lambda_A(t)$ the proportion of people entering sub-population A at time t which is assumed to be the sum of the following proportions:

- people of A discontinuing the use of injectable drugs at t,
- people of A dying from HCV at t,
- and people of A dying from a cause unrelated to HCV at t.

For simplicity purposes, we assume that every entering individual is a susceptible individual. Thus, from what precedes, one can rewrite (4.11) as follows:

$$\begin{aligned}\frac{\mathrm{d}s_{k,A}(t)}{\mathrm{d}t} &= \Lambda_{k,A}(t)-k\beta s_{k,A}(t)\theta_{k,A}(t)-k\beta^* s_{k,A}(t)\theta^*_{k,A}(t)-k\beta^\dagger s_{k,A}(t)\theta^{**}_{k,A}(t)\\ &\quad -s_{k,A}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,A}}{n_A+\sum_\Phi\tau_{\Phi,A}}\right)\cdot\Big[\beta\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big)\\ &\quad +\beta^*\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta^\dagger\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big]+p\sigma(a_{k,A}(t)+\breve{a}_{k,A}(t))\\ &\quad +\lambda_{S_0}f_{0,k,A}(t)+\kappa_{S_0}\breve{f}_{0,k,A}(t)-(\tau+\mu)s_{k,A}(t).\end{aligned}\tag{4.12}$$

4.5.2 Parameter values

Let us consider the values of the different parameters addressed in the previous sections so far. Since limited data are available for genotype 3 HCV infected patients, we will assume that there is no significant difference when appropriate. A summary of all the parameters value is available in Tables 4.4 and 4.6.

First, we refer to Bochud et al. (2009) who estimated fibrosis progression rates for both individuals infected by genotype 1 or genotype 3 HCV from the Swiss Hepatitis C Cohort database. It was reported a mean fibrosis progression acceleration by 1.39 for genotype 3 HCV infected patients. We also refer to Thein et al. (2008) who estimated fibrosis progression rates based on a meta-analysis of 111 studies of chronic HCV infected patients (0.117, 0.085, 0.120 and 0.116 for the progression to F_1, F_2, F_3 and F_4 respectively). As for HIV co-infection, Soto et al. (1997) estimated a mean time interval from HCV infection to cirrhosis of 23.2 versus 6.9 years for HCV mono-infected patients and HCV/HIV co-infected patients respectively. Considering the work of Benhamou et al. (1999), an estimation of the median fibrosis progression rates for HCV/HIV co-infected and HCV mono-infected patients resulted to 0.153 and 0.106 respectively. Last, Mohsen et al. (2003) estimated a fibrosis progression acceleration of 1.4 and an acceleration of advanced fibrosis (stages 3 and 4) development of 3 for co-infected patients. Based on those results, we assume that HIV co-infection contributes to the acceleration of fibrosis progression by a mean factor of 2 for the progression to F_1 and F_2 and by a mean factor of 2.6 for the progression to F_3 and F_4.

Considering the progression rates from cirrhosis to decompensated cirrhosis, McGarry et al. (2012) used the value of 0.03 for their economic model and Re et al. (2014) reported a risk for decompensation of 9.5 events per 1000 person-years vs 5.7 events per 1000 person-years for HCV/HIV co-infected and HCV mono-infected patients respectively, i.e a relative risk of 1.67. As the majority of patients were infected by genotype 1 HCV, we assume no significant impact of genotype 3 infection on the incidence of decompensated cirrhosis.

As for the proportion of HCV-infected patients who develop HCC, Lok et al. (2009) reported a proportion of 4.9% among a population of 1,005 patients. Moreover, they estimated a cumulative 5-year HCC incidence of 4.1% and 7.0% for patients with bridging fibrosis (F3) and for patients with cirrhosis (F4) respectively. These results give an annual progression rate of 0.84% and 1.45% for patients in fibrosis stage F_3 and F_4 respectively. As 93% of patients were infected by genotype 1, no impact on HCC incidence had been measured. However, Nkontchou et al. (2011) and Kanwal et al. (2014) estimated a hazard ratio of 3.54 and 1.80 for the impact of genotype 3 infection on the development of HCC respectively. Based on this, we choose a mean rate of HCC progression acceleration of 2.67. Considering progression to HCC in co-infected patients, only scarce data was available, however Kramer et al. (2015) estimated an incidence rate 1.48 times higher among HIV-positive patients than among HIV-negative patients.

Let us consider the progression rate from decompensated cirrhosis to HCC. In their economic model, McGarry et al. (2012) reported an annual rate of 0.079. As data on the impact of HIV is not available, we assume an incidence rate of HCC 1.48 times higher as above (0.1169).

As for the spontaneous clearance, we refer to Micallef et al. (2006) and Hernandez and Sherman (2011) who respectively reported rates of 26% and 10% for HCV mono-infected and HIV/HCV co-infected individuals. No association between viral clearance and HCV genotype has been highlighted. However, results were contradictory. Indeed, Lehmann et al. (2004) demonstrated a better spontaneous clearance for genotype 3 HCV infected individuals while Grebely et al. (2014) demonstrated better viral clearance for genotype 1 HCV infected individuals. In the following, we assume that there is no significant impact of genotype 3 on the clearance.

For the mortality rate related to decompensated cirrhosis, Pineda et al. (2005) estimated a median survival time for HCV/HIV co-infected and HCV mono-infected patients of 16 and 48 months, respectively. Hence, one can compute the annual mortality rates:

$$\left(1 - 0.5^{(12/16)}\right) = 0.4054 \quad \text{and} \quad \left(1 - 0.5^{(12/48)}\right) = 0.1591. \tag{4.13}$$

HCV genotype was not associated with survival of decompensated cirrhosis. Considering the mortality rate related to HCC, Lewin et al. (2015) estimated a median survival time for HCV/HIV co-infected and HCV mono-infected patients of 17.2 and 54.7 months, respectively. From (4.13), we obtain the following annual mortality rates 0.3834 and 0.1411 respectively. No information was available for genotype 3.

We now focus on the different values of the treatment-related and transplantation-related parameters. First, by referring to the economic model of McGarry et al. (2012), the values of 0.0310 and 0.1033 were chosen for the annual transition rates of liver transplant initiation. Considering the mortality after liver

transplantation, Terrault et al. (2012) reported a 3-year survival rate of 60% for HCV/HIV co-infected patients and 79% for HCV mono-infected respectively. From (4.13), we easily deduce a mortality rate of 0.1566 for HCV/HIV co-infected patients and 0.0756 for HCV mono-infected respectively. As no information on genotype 3 is available, we assume that the mortality rate is higher for genotype 3 HCV infected patients by a factor of 1.39 which corresponds to the mean fibrosis progression acceleration factor (see Section 4.5.1.2).

As for the SVR rates, Wyles et al. (2015) estimated those rates for HCV/HIV co-infected patients in each fibrosis stage. Based on a Daclatasvir + Sofosbuvir therapy during 12 weeks, SVR rates obtained were 93.8%, 100%, 100%, 96.6% and 97.5% for patients in fibrosis stage F_0, F_1, F_2, F_3 and F_4 respectively. Since it has been established that HIV co-infection has no statistically significant impact on treatment effectiveness (odds ratio 1.07) (see Highleyman (2016)), we assume the exact same SVR rates for HCV mono-infected patients. For patients in decompensated cirrhosis, Pol et al. (2016) reported a SVR rate of 86% for an indication of Ledipasvir + Sofosbuvir during 12 weeks in genotype 1 and 4 HCV infected patients.

Next, considering patients infected by genotype 3 HCV, we refer to Hézode (2016) who reported SVR rates for different indications. Indeed, an indication of Sofosbuvir + Daclatasvir during 12 weeks demonstrated a SVR rate of 95% for patients without cirrhosis, an indication of Sofosbuvir + Daclatasvir + Ribavirin during 12 weeks demonstrated a SVR rate of 90% for cirrhotic patients and an indication of Sofosbuvir + Daclatasvir + Ribavirin during 24 weeks demonstrated a SVR rate of 78.5% for patients with decompensated cirrhosis.

Let us now consider the annual cessation rate. We refer to Angelis et al. (2004) who used the values 5%, 9% and 12% in a back-calculation model of the incidence of opiate use/injecting drug use. Similarly, Martin et al. (2011) used a cessation rate of 7.75% for their modeling study of the impact of antiviral therapy on the prevalence of HCV in IDU. Hence, we choose the mean annual cessation rate of 8.44%.

As stated earlier, people leaving the population of IDU because of the cessation rate are tracked via an alternative model with respect to their compartment (health state) of origin. Indeed, if one leaves the original model in a given health state, say fibrosis state F_2, one will enter the alternative model in that exact same health state. An illustration of this alternative model for genotype 1 HCV infected individuals is displayed in Appendix D. Additionally, the annual number of HCV infections occurring in the non-IDU population is integrated in the alternative model. Since HCV infections occurring in the population of IDU represent 76.5% of the total number of new infections, the number of HCV infections occurring in the non-IDU population will be computed such as:

$$\begin{array}{c}\#\text{ of HCV infections in}\\ \text{the population of non-IDU}\end{array} = \begin{array}{c}\#\text{ of HCV infections in}\\ \text{the population of IDU}\end{array} \times \frac{1-0.765}{0.765}.$$

Finally, let us focus on the needle/syringe sharing frequency. Referring to Section 4.4.1, we previously used the value of 5.33 for our simple HCV compartmental model. In order to give a more precise estimation of the transmission risk of HCV, we now refer to Vickerman et al. (2007, 2008) where several values have been estimated for the frequency of needle/syringe sharing in different cities in England. Reported frequencies were 2.85, 5.6 and 16 per month for Teesside, Bristol and London respectively.

However, the majority of PWID in Bristol and Teesside (49.7% and 76.2% respectively) were on opiate substitution therapy (OST) and PWID on OST tend to lower their frequency of injecting and sharing. In France, it has been estimated by the OEDT[3] that around 50% of PWID were on OST (see Costes et al. (2010)). Taking this into account, we retained the frequency of needle/syringe sharing observed in Bristol, i.e 5.6 sharing events per month. To obtain a per-partner sharing frequency, we compute 5.6/3=1.87 where 3 corresponds to the mean number of partners.

Parameters are reported in Table 4.6 and Table 4.4. Table 4.5 gives the sources of the different indices used in Table 4.6. All other parameters were previously reported in Table 4.2. An illustration of the final model, for the particular case of genotype 1, is given in Figure 4.17. Each starred compartment or rate is the HIV-infected counterpart of the non-starred quantities. For a better clarity, natural mortality, cessation and initiation rates were not displayed. Full details of the complete set of differential equations for the genotype 1 are available in Appendix C. We only focused on genotype 1 in Appendix C since the differential equations of genotype 3 admit a similar formulation. the same reason holds for the alternative model.

[3] *Observatoire européen des drogues et des toxicomanies*, the European monitoring centre for drugs & drug addiction.

Table 4.4: Annual rates used in the model to simulate the spread of HCV

Parameter	Notation	Value	Source
Annual risk of death			
HCV mono-infected			
With decompensated cirrhosis	λ_h	0.1591	Pineda et al. (2005)
With hepatocellular carcinoma	λ_H	0.1411	Lewin et al. (2015)
HCV/HIV co-infected			
With decompensated cirrhosis	λ_h	0.4054	Pineda et al. (2005)
With hepatocellular carcinoma	λ_H	0.3834	Lewin et al. (2015)
Other parameters			
Per-partner sharing frequency, per month	ω	1.87	Vickerman et al. (2007, 2008)
Cessation of injecting drugs, %	–	8.44	Angelis et al. (2004); Martin et al. (2011)

Table 4.5: Sources of the annual rates

Source	Index
Bochud et al. (2009)	[1]
Thein et al. (2008)	[2]
Soto et al. (1997)	[3]
Mohsen et al. (2003)	[4]
Benhamou et al. (1999)	[5]
Nkontchou et al. (2011)	[6]
Kanwal et al. (2014)	[7]
Lok et al. (2009)	[8]
McGarry et al. (2012)	[9]
Re et al. (2014)	[10]
Kramer et al. (2015)	[11]
Wyles et al. (2015)	[12]
Hézode (2016)	[13]
Pol et al. (2016)	[14]
Terrault et al. (2012)	[15]

Table 4.6: Annual rates used in the model to simulate the spread of HCV (by genotype)

	Genotype 1		Genotype 3		
Parameter	Notation	Value	Notation	Value	Source
Annual fibrosis progression					
HCV mono-infected					
From F0 to F1	λ_1	0.1170	κ_1	0.1626	[1,2]
From F1 to F2	λ_2	0.0850	κ_2	0.1182	[1,2]
From F2 to F3	λ_3	0.1200	κ_3	0.1668	[1,2]
From F3 to F4	λ_4	0.1160	κ_4	0.1612	[1,2]
HCV/HIV co-infected					
From F0 to F1	λ_1^*	0.2340	κ_1^*	0.3252	[1,2,3,4,5]
From F1 to F2	λ_2^*	0.1700	κ_2^*	0.2364	[1,2,3,4,5]
From F2 to F3	λ_3^*	0.3120	κ_3^*	0.4337	[1,2,3,4,5]
From F3 to F4	λ_4^*	0.3016	κ_4^*	0.4191	[1,2,3,4,5]
Annual progression to ESLD					
HCV mono-infected					
From F3 to HCC	λ_h	0.0084	κ_h	0.0224	[6,7,8]
From F4 to HCC	λ_H	0.0145	κ_H	0.0387	[6,7,8]
From F4 to DC	λ_D	0.0300	–	–	[9]
From DC to HCC	λ_{D_H}	0.0790	–	–	[9]
From HCC to LT	λ_L	0.1033	–	–	[9]
From DC to LT	λ_{D_L}	0.0310	–	–	[9]
HCV/HIV co-infected					
From F3 to HCC	λ_h^*	0.0124	κ_h^*	0.0332	[6,7,8,11]
From F4 to HCC	λ_H^*	0.0215	κ_H^*	0.0573	[6,7,8,11]
From F4 to DC	λ_D^*	0.0501	–	–	[9,10]
From DC to HCC	$\lambda_{D_H}^*$	0.1169	–	–	[9,11]
From HCC to LT	λ_L^*	0.1033	–	–	[9]
From DC to LT	$\lambda_{D_L}^*$	0.0310	–	–	[9]
Treatment of HCV†					
With fibrosis state F0	ν_0	0.9380	$\check{\nu}_0$	0.9500	[12,13]
With fibrosis state F1	ν_1	1.0000	$\check{\nu}_1$	0.9500	[12,13]
With fibrosis state F2	ν_2	1.0000	$\check{\nu}_2$	0.9500	[12,13]
With fibrosis state F3	ν_3	0.9660	$\check{\nu}_3$	0.9500	[12,13]
With fibrosis state F4	ν_4	0.9750	$\check{\nu}_4$	0.9000	[12,13]
With DC	ν_D	0.8650	$\check{\nu}_D$	0.7850	[13,14]
Mortality after liver transplant					
HCV mono-infected	λ_{L_D}	0.0756	κ_{L_D}	0.1051	[1,15]
HCV/HIV co-infected	$\lambda_{L_D}^*$	0.1566	$\kappa_{L_D}^*$	0.2177	[1,15]

† Rates based on a treatment duration of 12 or 24 weeks. ESLD, end-stage liver disease; S, susceptible; A, acute HCV infection; DC, decompensated cirrhosis; HCC, hepatocellular carcinoma; LT, liver transplant. Starred compartments represent HIV infected individuals. Sources are given in Table 4.5.

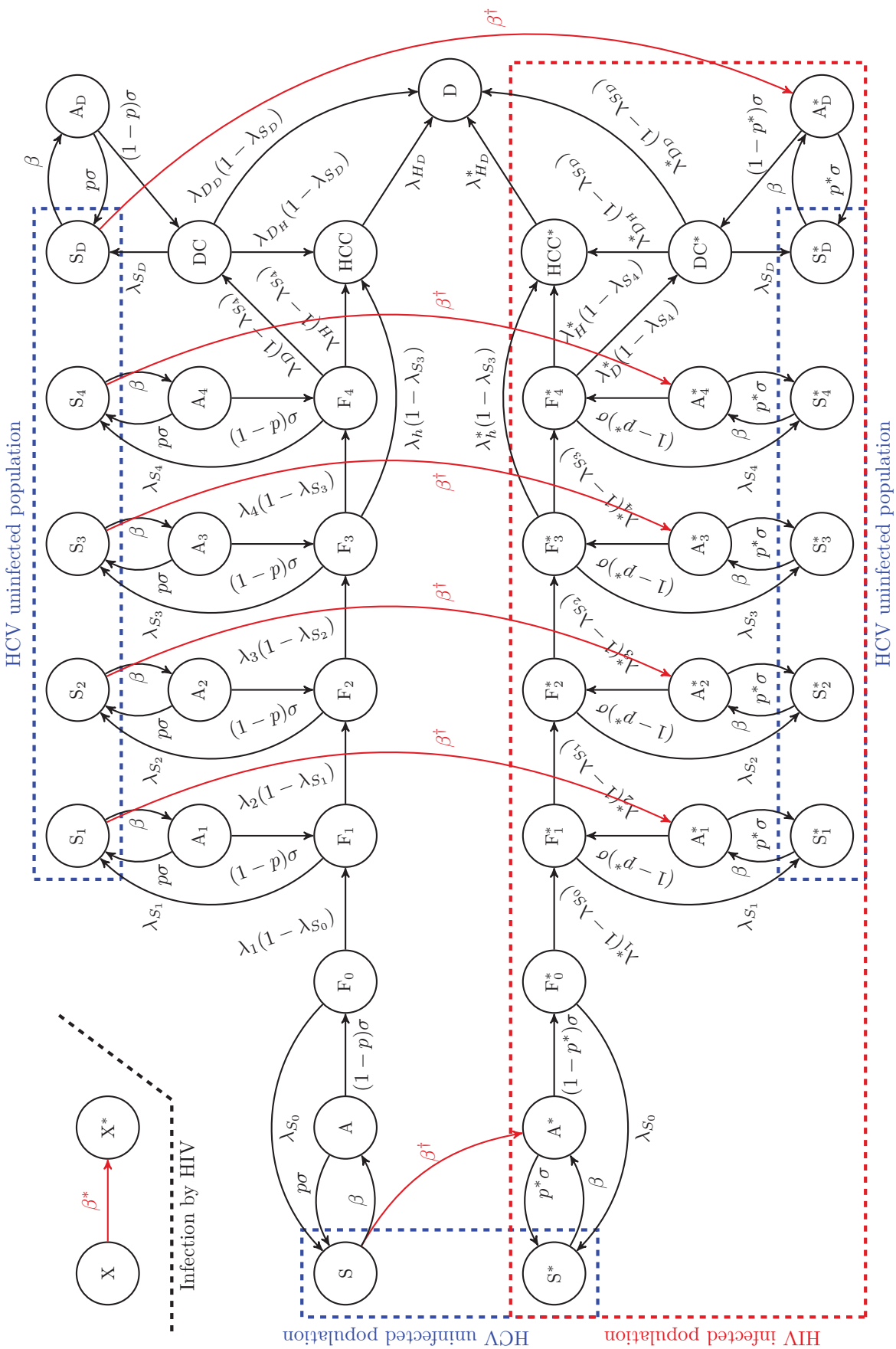


Figure 4.17: Dynamic model of patients infected by HCV (Genotype 1).
S, susceptible population; A, acute infection; F, fibrosis stage; HCC, hepatocellular carcinoma; DC, decompensated cirrhosis; LT, liver transplant; D, death due to HCV. The subscript i, for $i \in \{0, 1, 2, 3, 4\}$, corresponds to the stage of the fibrosis progression based on the Metavir scoring system indicating the degree of inflammation. Starred compartments represent HIV infected individuals. Any patient in state X, where X$\in$ {S, A, S_i, A_i, F_i, HCC, DC, LT} for $i \in \{2, 3, 4, D\}$, can enter state X^* with probability β^*.

4.5.3 Migration crisis

Migrations, and more specifically human migrations, play a major role in the dissemination of diseases all over the world. When the host country and the country of origin admit different prevalence rates, migrations flows can interfere with the host country infected population and impact patient care. Moreover, as immigrants tend to be often more vulnerable to exposure to drugs than non-immigrant people with a similar socio-economic background (see Carballo et al. (2011)), strong immigration flows can have a direct impact on the spread of HCV.

Currently, European countries are facing a major challenge in welcoming thousands of immigrants fleeing the Syrian civil war. Indeed, with an estimated prevalence of HCV reaching 1% (see Daw and Dau (2012)) (with a genotype 3 HCV infection proportion of 1.6% (see Chemaitelly et al. (2015)) for the Syrian population) and an estimated prevalence of HIV of 0.01% according to the Central Intelligence Agency, the arrival of Syrian immigrants should be considered in the model with a great attention.

Today, the French government is planning to receive 30,000 immigrants in the next two years from Syria and the neighboring regions. Thus, we will add 15,000 individuals per year for the first two years in the model distributed into 6 different groups depending on their status (uninfected, HCV infected, HCV/HIV co-infected, etc). Since obtaining reliable data is difficult in the particular case of war refugees, we will successively assume different initiation-to-drugs rates for our simulations. Moreover, as no data is available on the health state of refugees (fibrosis stage, HCC, etc), we assume that every HCV infected individual enters the model in stage fibrosis F_0. Every migrating individual is considered in the expression $\Lambda_{k,A}(t)$ of Equation 4.12.

4.5.4 Prevalence of HCV and initial conditions of the model

Let us now consider the proportions of patients infected by HCV and/or HIV in the French IDU population. Those proportions will allow us to initiate the model.

As for the proportions of HCV infected patients and of HCV/HIV co-infected patients, we refer to the French survey Coquelicot on the prevalence of HCV and HIV in the French population of IDU. Indeed, in 2011 it was estimated that the proportion of HCV infected patients in the French IDU population was 44% while the proportion of HIV infected patients was 10% (see Jauffret-Roustide et al. (2013)). No indication was available on the proportion of HCV/HIV co-infected patients. However, in 2004, Larsen et al. (2005) reported a value of 92.8% for the proportion of HCV/HIV co-infected individuals in the French IDU population. Moreover, in the same French survey Coquelicot (see Jauffret-Roustide et al. (2013)), values were available for the prevalence of HCV and HIV in some urban areas. Those values are reported in Table 4.7. For all others urban areas, it is assumed a prevalence of HCV of 44% and a prevalence of HIV of 10%. For the absolute number of PWID in each urban area, we refer to Section 4.4.3.

Table 4.7: HCV and HIV prevalence in IDU population of some urban areas in 2011

Urban area	HCV prevalence	HIV prevalence
Bordeaux	24%	4%
Lille	28%	0%
Marseille	56%	17%
Paris	44%	10%
Strasbourg	47%	3%

To determine the proportions of the different genotypes in the French population, we refer to Laperche et al. (2012) who reported in 2010 the approximated values 56%, 9%, 25%, 5%, 3% and 2% for genotypes 1, 2, 3, 4, 5, 6 respectively. However, since non-3 genotypes admit relatively similar SVR rates, we decide to group them and stick to a proportion of non-3 genotype (75%) and a proportion of genotype 3 (25%).

Last, to consider the proportion of individuals in the different health states, we refer to Razavi et al. (2014) who proposed a model for the estimation of the present and future disease burden of HCV in several countries. The model offered an evolution of the different quantities (F_0-F_4, DC, HCC, LT) from 1950 to 2030. Estimated values obtained for the year 2013 are reported in Table 4.8. As liver transplant is not considered in our model for IDU, the proportion of individuals in the 'Liver transplant' state has been split up to decompensated cirrhosis and HCC according to their relative proportions.

Thus, the proportion of individuals with decompensated cirrhosis used in our model is $0.52 + 0.60 \times 0.52 = 0.832$ while the proportion of individuals with HCC used is $0.48 + 0.60 \times 0.48 = 0.768$. Proportions used in the alternative model are the same as displayed in Table 4.8. Referring to Table 4.8, the total number of HCV infected patients in the non-IDU population is estimated to $201,627 - 112,116 = 89,511$. As no data was available for the acute phase of the disease, we assumed that no patient was in acute phase at the initiation of the model.

Table 4.8: HCV prevalence by health state in France for the year 2013

Health state	Number (%)
F_0	51,387 (25.49)
F_1	71,098 (35.26)
F_2	31,586 (15.67)
F_3	27,925 (13.85)
F_4	16,408 (8.14)
DC	1,053 (0.52)
HCC	967 (0.48)
LT	1,203 (0.60)
Total	201,627 (100)

4.6 Results

The evolution of the HCV prevalence in France is presented in the following figures for different assumptions. Those assumptions are directly related to the proportion of teated patients, the sharing frequency between PWID and the rate of immigration. Each assumption is compared to the base case scenario which is based on the values introduced in the previous sections, i.e a proportion of treated patients of 5.2%, a sharing frequency of 1.87 per partner per month and a rate of immigration of 15,000 people per year for two successive years.

First, Figure 4.18 exhibits the evolution of both genotypes 3 and non-3 starting from the year 2013 to the year 2040. One can notice that the evolution of the proportions of both genotypes stay relatively constant over time. Since genotype 3 tends to be more aggressive than other genotypes, we could have expected a lower genotype 3 proportion over time due to a higher mortality. However, as 2^{nd}-gen DAAs admit very high SVR rates whatever the genotype, this could explain the fact that no sustained decline in terms of proportion is observed.

Next, we focus on the proportion of treated patients in France. Since the proportion of treated patients in France hardly reaches the value of 5.2%, we made some assumptions on higher proportions to observe the potential impact of 2^{nd}-gen DAAs on the prevalence of HCV in France. By referring to Figure 4.19, one can observe that, as expected, the prevalence of HCV decreases with the proportion of treated patients. Indeed, by assuming a proportion of treated patients of 50%, the prevalence of HCV decreases below 50,000 infected individuals just in a few years. However, it has to be recalled that since HCV is an asymptomatic disease, a large proportion of HCV infected individuals are still undiagnosed and that 50% of treated patients might be unrealistic for now.

In Figure 4.20, we propose to study the impact of sharing frequencies in the IDU population. Indeed, as sharing frequencies are directly related to the transmission of the virus, being able to limit the frequency of syringe/needle sharing could have a positive impact on the total prevalence of HCV in France. Since we calculated the number of transmissions occurring outside of the IDU population based on the number of transmissions occurring in the IDU population (see Section 4.5.2), we made a modification as follows:

$$\begin{array}{c}\#\text{ of HCV infections in}\\ \text{the population of non-IDU}\end{array} = \begin{array}{c}\#\text{ of HCV infections in}\\ \text{the population of IDU}\end{array} \times \frac{1 - 0.765/s_f}{0.765/s_f}$$

where s_f is the sharing frequency reduction. With this modification the reduction of sharing frequency does not impact the number of new infections occurring outside the IDU population. One can observe on

Figure 4.20 that the evolution of the prevalence slightly declines when the sharing frequency decreases. This impact might not be as strong as the impact of 2nd-gen DAAs but the sharing frequency only affect the number of new infections and not the current prevalence.

Last, we study the potential effect of immigration on the prevalence of HCV. By referring to Figure 4.21, one can see that immigration has a limited impact on the HCV prevalence and that immigration rates have to last for several years to be significant. However, those immigration rates are based on governmental sources which do not reflect the the numbers of the clandestine immigration. Hence, despite of the results exhibited by Figure 4.21, one can expect a greater impact of immigration on the HCV prevalence.

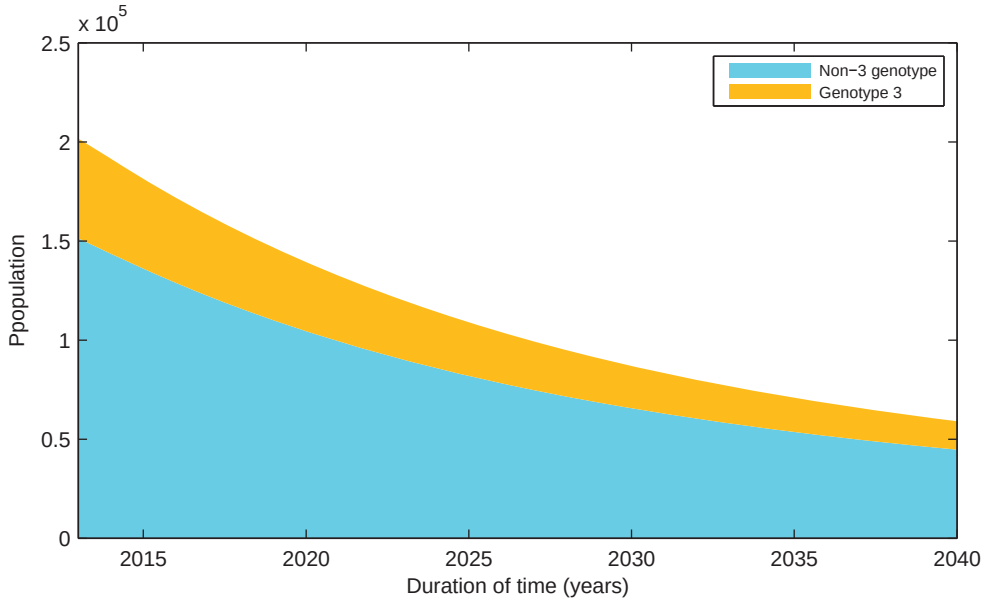


Figure 4.18: Mean evolution of the HCV prevalence in France by genotype.
The different curves represent the mean quantities over 100 simulations. The blue area represents the evolution of the prevalence of non-3 genotype while the orange area represents the evolution of the prevalence of genotype 3.

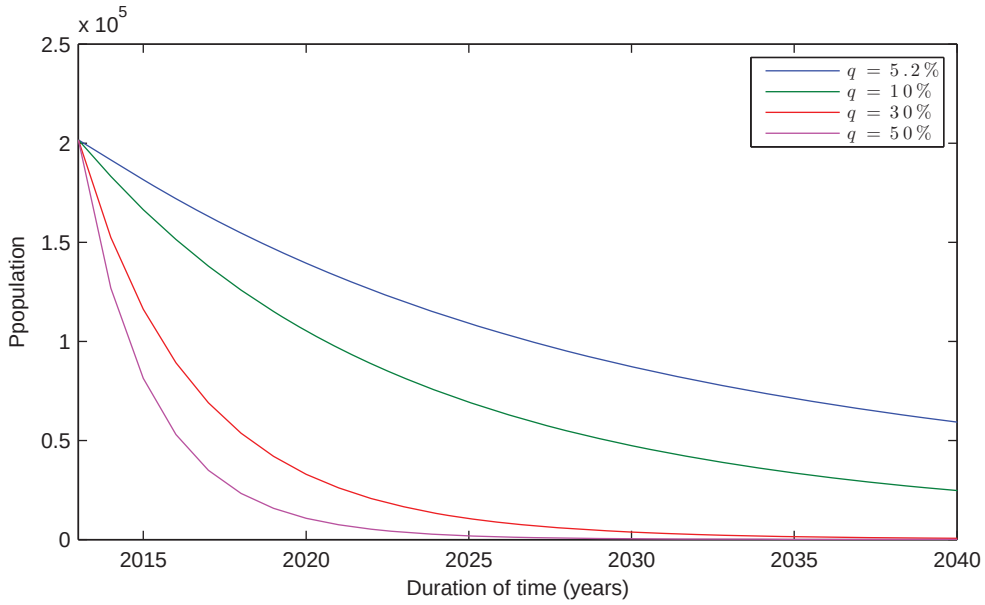


Figure 4.19: Mean evolution of the HCV prevalence in France for different proportion of treated patients.
The different curves represent the mean quantities over 100 simulations. The blue curve represents the base case with a proportion of treated patients of $q = 5.2\%$. The three other curves (in green, red and pink) represent the evolution of the prevalence of HCV in France for three different values of q ($10\%, 30\%$ and 50% respectively).

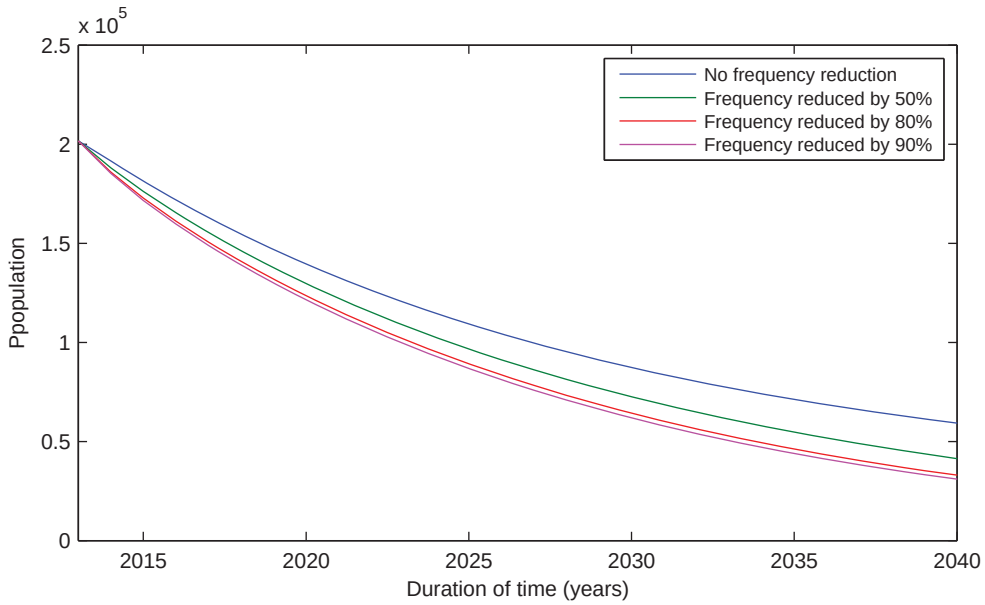


Figure 4.20: Mean evolution of the HCV prevalence in France for different sharing frequencies.
The different curves represent the mean quantities over 100 simulations. The blue curve represents the base case with a sharing frequency of 1.87×12 per partner per month. The three other curves (in green, red and pink) represent the evolution of the prevalence of HCV in France for a frequency reduced by $50\%, 80\%$ and 90% respectively.

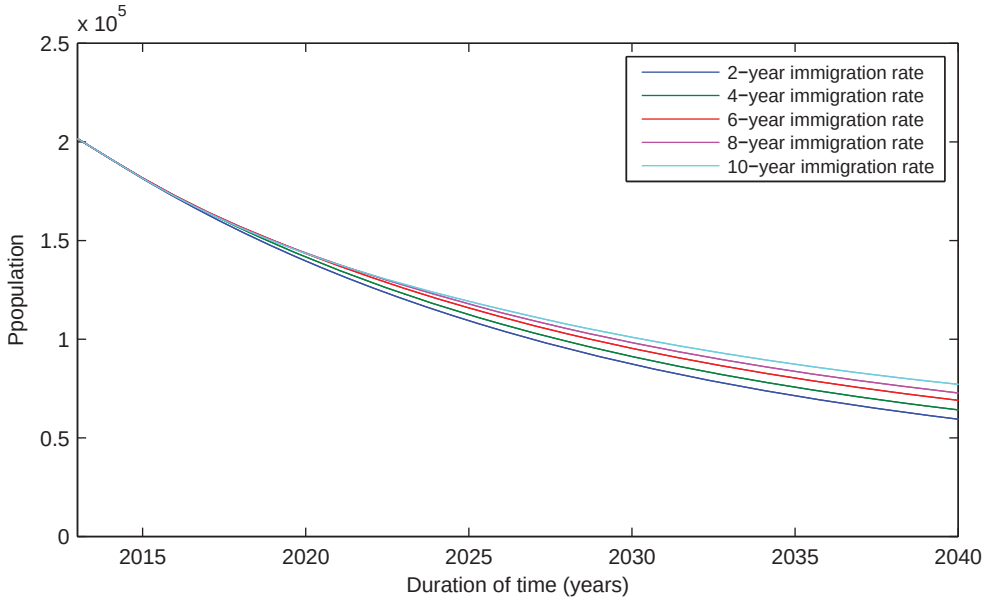


Figure 4.21: Mean evolution of the HCV prevalence in France for different immigration rates.
The different curves represent the mean quantities over 100 simulations. The blue curve represents the base case with 15,000 people immigrating in France each year for two successive years. The four other curves (in green, red, pink and light blue) represent the evolution of the prevalence of HCV in France with 15,000 people immigrating in France each year for four, six, eight and ten successive years respectively.

4.7 Discussion

In this chapter, we have introduced two compartmental models of the propagation of HCV in France to estimate the impact of the new 2nd-gen DAAs on the size of the infected population. The first part of the chapter was meant to develop a simple compartmental model of the propagation of HCV on a network of networks representing PWID populations in distinct but interconnected urban areas (metapopulation model). The second part was entirely dedicated to the development of a more complex model taking into account more in-depth mechanisms of the infection on the same network of interconnected urban areas.

The simple model allowed us to give a representation of the basic mechanisms of the HCV infection by considering the heterogeneities of the PWID population. It was applied on a three-city network and demonstrated, as expected, that the epidemic spread was correlated with the amount of interactions between the cities in this three-city system. Indeed, it has been showed that the epidemic was growing faster and stronger when the amount of interactions between cities was high. Then, this model corroborated this result on a network of 100 interconnected urban areas where two epidemic starting points were set in two different urban areas, mainly differing from each other by the amount of interactions they have with other urban areas, and where very different outcomes in terms of prevalence and propagation speed were observed.

In the second part of this chapter, we still considered our network of urban areas but largely focused our attention on developing a much more complex HCV compartmental model. By considering the different fibrosis stages and the main HCV genotypes, this second model was meant to describe at best the future evolution of the HCV prevalence in France with the arrival on the market of new therapeutics. Our results have shown that the new 2nd-gen DAAs have a strong impact on the prevalence of HCV assuming that a sufficient number of patients has access to the treatment. Indeed, with the current proportion of treated patients, the infection will not disappear before several decades. Needle sharing frequencies and migrations rates seem to have a limited impact on the prevalence.

The main advantages of our approach come from the consideration of heterogeneities in the PWID population to model the transmission process on the one hand, and on the other, the use of a metapopulation model to model the infectious dynamics at a country scale. Indeed, the use of a metapopluation model allowed us to divide the general PWID population into several PWID subpopulations in order to better handle population heterogeneities with a contact network.

However, our approach presents some limitations. First, when dealing with the PWID population, difficulties arise when trying to obtain reliable data. Indeed, whether it be the needle sharing behavior for the estimation of HCV transmission or the degree of each individual for the establishment of a pertinent contact network, the lack of data forced us to make strong assumptions. Moreover, no cessation of treatment was assumed despite the fact that the rate of noncompliance among PWID might be significant. Second, as the number of annual HCV infections was difficult to estimate, the number of HCV infections occurring outside the PWID population was simply computed as a proportion of the number of HCV infections occurring in the PWID population. Third, for the sake of simplicity, we only focused on the new 2nd-gen DAAs without considering other treatments. Fourth, as for immigration, more reliable numbers should be considered.

To conclude, the methodology addressed in this chapter allows for an interesting way to deal with the spread of an epidemic at a national scale. Moreover, by considering contact networks, one can relieve the homogeneous mixing assumption and use classic compartmental models with population heterogeneities. However, our model still needs to be applied on real datasets for validation. The different limitations stated above could be subject to further research.

Conclusions

English

The work presented in this doctoral thesis answered to three different epidemiology-related problematics using mathematical modeling. Despite the fact that the main goal of this work was the assessment of the new therapeutics' impact on the HCV infected population, we explored several ways in the application of mathematical models to epidemics. In a series of three papers, we investigated several statistical estimation methods, some back-calculation models for the reconstruction of past incidence and the association of two dynamic compartmental models on a contact network with a metapopulation model.

In the first paper (Chapter 2), we studied different statistical estimation technics for two parametric models (RCS and NHRS). The estimation of the relative bias and the relative standard deviation showed that the maximum likelihood estimator provided almost always better estimations than the least squares estimator.

However, in presence of a high amount of censoring, it turned out that the least squares estimator was performing better. The computation of the mean integrated squared error (MISE) corroborated those results for both low and high amount of censoring. We finally applied those estimation methods to two real datasets of the Code Red v2 worm propagation and of the spread of HIV in France. Considering the Code Red v2 dataset, the different estimation methods gave similar results for the RCS model. As for the NHRS model, results were more contrasted with a fairly poor fit for the least squares estimator. As for the HIV dataset, both RCS and NHRS models were not suitable to this dataset since the results were not as good as they were for the Code Red v2 dataset. Estimation methods studied in this part turned out to be very efficient tools to deal with the estimation of epidemic models parameters and could be applied to much wider problems such as models of the propagation of communicable diseases in populations to support health care decision makers.

However, some issues are still to be addressed. Indeed, the development of statistical methods to assess whether the model is in line with the nature of the studied data should be considered. Moreover, one could find of interest to theoretically study the asymptotic properties of the different estimators.

In the second paper (Chapter 3), in order to estimate the past incidence of HCV, we compared several back-calculation methods based on either the maximum likelihood method (logistic and exponentially modified Gaussian models) or the EMS algorithm with the approach of reference in France which is based on the weighted least squares method.

Similar trends were exhibited on the evolution of incidence and prevalence of HCV for the different models and results were consistent with the prevalence estimation reported by INVS for the year 2004. In particular, the method based on the EMS algorithm offered better performances in terms of mean squared error compared with the other methods and allowed us to reduce the bias induced by an arbitrary choice of a given parametric model. This latter advantage was precisely highlighted by the poor performances exhibited by the parametric models (particularly the logistic model). Moreover, by assuming that the random variables X_{ij} (the unobserved number of people infected in year i who died of HCC in year j) were realizations of independent Poisson variables, we were able to easily compute a confidence interval for our results thanks to the bootstrap method. However, our approach has some limitations. Indeed, due to the long term progression of HCV, our approach is not able to provide reliable estimation for the recent HCV incidence. Regarding the population, only sex was consider as a covariate while other covariates such as age or alcohol consumption should be considered. Last, no treatment was considered in this study.

As for further research, our recommendations would be to focus on the points raised in the limitations and to obtain more reliable date even if a large work was undertaken about this aspect in this study.

In the last paper (Chapter 4), we focused our attention on two compartmental models of HCV propagation and on a metapopulation model representing the network of the 100 largest French urban areas. The main objective was to estimate the impact of new anti-HCV therapeutics by focusing on the PWID population and its role in the spread of the infection.

Hence, a two-level network was built to deal with the interactions between the urban areas on the one hand, and on the other, to model the interactions between individuals inside each sub-population (urban area). Then, the two compartmental models were applied on these sub-populations. The first compartmental model, the simplest one, was mainly set as an example to test the suitability of the metapopulation model to deal with the spread of an epidemic at a country scale. Numerical simulations exhibited sensible stylized facts in terms of the extent and speed of propagation of an epidemic in France. The second compartmental model presented more in-depth features related to the infection such as fibrosis, cirrhosis or genotype. Our main results showed that the new 2^{nd}-gen DAAs had a significant impact on the prevalence of HCV even if a greater proportion of patients should be treated in order to achieve an eradication of HCV in France in the future decades. The main advantage of our work is the fact that it presents an easy way to deal with social and spatial heterogeneities. Moreover, we built an advanced model of HCV propagation which considers several essential components of the disease mechanisms. However, with more advanced models comes more complexity and more assumptions. Indeed, by focusing on such a marginal population (PWID), some data were not available and numerous assumptions had to be made. Mathematically, we witnessed a general increase of the complexity as each new compartment leads to a new differential equation. Moreover, as no data was available to validate the model on such a country scale, it makes sense to remain cautious about the interpretation of the results. Last, we did not consider any treatments other than the new 2^{nd}-gen DAAs in our models.

To conclude, the methodology addressed in this chapter allows for an interesting way to deal with the spread of an epidemic at a national scale. However, our model still needs to be applied on real datasets for validation. The different limitations stated above could be subject to further research.

Français

Le travail présenté dans cette thèse de doctorat répond à trois problématiques épidémiologiques différentes, toutes liées à la modélisation mathématique. Bien que l'objectif principal de cette thèse soit d'évaluer l'impact des nouvelles thérapeutiques anti-VHC sur la population infectée, nous avons exploré plusieurs manières d'utiliser des modèles mathématiques pour l'étude des épidémies. Dans cette série de trois papiers, nous avons abordé quelques méthodes d'estimation statistique, plusieurs modèles de rétrocalcul dans le cadre de l'estimation de l'incidence passée d'une maladie et l'association de deux modèles dynamiques compartimentaux basés sur un réseau de contacts à un modèle de métapopulation.

Dans le premier papier (Chapitre 2), nous avons étudié différentes méthodes d'estimation statistique appliquées à deux modèles paramétriques (RCS et NHRS). Les estimations du biais relatif et de l'écart-type relatif ont montré que l'estimateur du maximum de vraisemblanceprésentait presque toujours de meilleures estimations que l'estimateur des moindres carrés.

Cependant, en présence d'un grand nombre de données censurées, il s'est avéré que c'est l'estimateur des moindres carrés qui démontrait de meilleures performances. Le calcul du "mean integrated squared error (MISE) a d'ailleurs permis de corroborer ces résultats pour différents niveaux de censure. Enfin, nous avons appliqué ces méthodes d'estimation à deux jeux de données sur la propagation du ver Code Red v2 et sur la propagation du VIH en France. En ce qui concerne le jeu de données sur le ver Code Red v2, les différentes méthodes d'estimation ont donné des résultats similaires pour le modèle RCS. Pour le modèle NHRS, les résultats ont été plus contrastés puisque l'estimateur des moindres carrés a démontré de moins bonnes performances. Concernant le jeu de données sur le VIH, les deux modèles RCS et NHRS se sont révélés peu adaptés puisque les résultats n'étaient clairement pas aussi bons que pour le jeu de données sur le ver Code Red v2. Les méthodes d'estimation étudiées dans cette partie se sont avérées être des outils efficaces pour l'estimation des paramètres de modèles épidémiques et pourraient être appliquées à des problématiques plus larges telles que la propagation d'une épidémie au sein d'une population afin de venir en aide aux preneurs de décision du milieu de la santé.

Cependant, quelques problèmes nécessitent toujours d'être résolus comme par exemple le développement de méthodes statistiques permettant d'évaluer si le modèle est en adéquation avec les données étudiées. De plus, d'un point de vue théorique, il pourrait être intéressant de s'intéresser aux propriétés

asymptotiques des différents estimateurs.

Dans ce deuxième papier (Chapitre 3), et dans le but d'estimer l'incidence passée du VHC, nous avons comparé avec l'approche de référence en France, basée sur la méthode des moindres carrés pondérés, plusieurs méthodes de rétro-calcul basées, d'une part, sur la méthode du maximum de vraisemblance (modèle logistique et exponentiel gaussien) et, d'autre part, sur l'algorithme EMS.

Des tendances similaires ont été observées pour les différents modèles en ce qui concerne l'évolution de l'incidence et de la prévalence du VHC. Les résultats se sont d'ailleurs avérés cohérents avec l'estimation de la prévalence effectuée par l'INVS pour l'année 2004. En particulier, la méthode basée sur l'algorithme EMS a démontré de meilleures performances en termes d'erreur des moindres carrés, comparativement aux autres méthodes. Cette méthode a également permis de réduire le biais inhérent au choix d'un modèle paramétrique. Ce second avantage a notamment été souligné par les moins bonnes performances obtenues avec les modèles paramétriques (particulièrement le modèle logistique). De plus, en se basant sur l'hypothèse selon laquelle les X_{ij} (le nombre non-observé de personnes infectées l'année i qui meurt de CHC l'année j) étaient des réalisations indépendantes de loi de Poisson, nous avons été facilement capable de générer des intervalles de confiance grâce à la méthode du bootstrap. Cependant, notre approche présente des limitations. En effet, le VHC étant une maladie à progression lente, notre approche n'est pas capable de fournir des estimations suffisamment précises sur l'incidence récente du VHC. Concernant la population, seul le sexe a été retenu en tant que covariable alors que d'autres covariables toutes aussi pertinentes telles que l'âge ou la consommation d'alcool aurait pu l'être également. Enfin, aucun traitement n'a été pris en compte dans ce travail.

Afin de prolonger ces travaux au cours de futures recherches, nos recommandations se focaliseraient sur les points soulevés plus haut et sur l'obtention de données plus fiables même si un grand travail a déjà été fait dans ce cadre-là au sein de ce papier.

Dans le dernier papier (Chapitre 4), nous avons focalisé notre attention sur deux modèles compartimentaux de propagation du VHC et sur un modèle de métapopulation représentant les 100 plus grandes aires urbaines françaises. L'objectif principal a été d'estimer l'impact des nouvelles thérapeutiques anti-VHC en se concentrant sur la population toxicomane et son rôle dans la propagation de l'infection.

Ainsi, un réseau à deux niveaux a été construit pour prendre en compte les interactions entre les aires urbaines d'une part, et d'autre part, pour modéliser les interactions entre individus au sein de chaque sous-population (aire urbaine). Puis les deux modèles compartimentaux ont été appliqué sur ces sous-populations. Le premier modèle compartimental, le plus simple des deux, a principalement été développé pour tester l'adéquation du modèle de métapopulation à une problématique de type propagation d'une épidémie à l'échelle d'un pays. Des simulations numériques ont permis de mettre en avant des résultats tangibles en termes d'ampleur et de vitesse de propagation d'une épidémie en France. Le second modèle compartimental qui a été développé, s'est principalement axé sur une description plus en profondeur du VHC prenant en compte les états de fibrose, la cirrhose ou le génotype. Nos principaux résultats ont montré que les antiviraux à action directe de seconde génération avaient un impact significatif sur la prévalence du VHC même si une proportion plus grande de patients traités serait nécessaire pour éradiquer le VHC en France dans les prochaines décennies. Le principal avantage de notre travail est qu'il constitue une manière simple de modéliser les hétérogénéités sociales et spatiales. De plus, nous avons construit un modèle de propagation du VHC plus avancé, prenant en compte bien plus de mécanismes essentiels de la maladie. Cependant, ce modèle plus avancé entraîne un accroissement de la complexité et du nombre d'hypothèses. En effet, en se focalisant sur une population aussi marginale que la population toxicomane, on se retrouve confronté à une quantité de données disponibles limitée nécessitant de faire de nombreuses hypothèses. Mathématiquement, nous témoignons d'un accroissement général de la complexité puisque chaque nouveau compartiment conduit à une nouvelle équation différentielle. De plus, aucune donnée n'étant disponible pour la validation du modèle à cette échelle, il convient de rester prudent sur l'interprétation des résultats. Enfin, aucun autre traitement n'a été pris en compte dans nos modèles.

En conclusion, la méthodologie abordée dans ce chapitre permet de modéliser de façon originale la propagation d'une épidémie à l'échelle d'un pays. Cependant, notre modèle nécessite d'être validé sur des jeux de données réelles. Enfin, les différentes limitations abordées plus haut peuvent être sujettes à de futurs travaux de recherche.

Bibliography

R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002. 11

E. Allen. Modeling with Itô Stochastic Differential Equations. *Mathematical Modelling: Theory and Applications*, 22, 2007. 17

L. J. S. Allen. An Introduction to Stochastic Epidemic Models. *Mathematical Epidemiology*, 1945:81–130, 2008. 13, 17

R. M. Anderson. Vaccination against rubella and measles: quantitative investigation of different policies. *Journal of Hygiene*, 90:259–325, 1983. 12

R. M. Anderson. The epidemiology of HIV infection: variable incubation plus infectious periods and heterogeneity in sexual activity. *Journal of the Royal Statistical Society. Serie A (Statistics in Society)*, 151:66–98, 1988. 17

R. M. Anderson and R. M. May. Infectious Diseases of Humans: Dynamics and Control. *Oxford Science Publications*, 1992. 12

D. De Angelis, M. Hckman, and S. Yang. Estimating Long-term Trends in the Incidence and Prevalence of Opiate Use/Injecting Drug Use and the Number of Former Users: Back-Calculation Methods and Opiate Overdose Deaths. *American Journal of Epidemiology*, 160:994–1004, 2004. 89, 90

ANRS. Direct-acting antivirals for hepatitis C. First real-life efficacy data in HIV/HCV co-infected patients. *Press release*, 2015. 75, 76

S. Asmussen. Applied Probability and Queues. *Stochastic Modelling and Applied Probability*, 51, 2003. 56

T. Asselah, L. Rubbia-Brandt, et al. Steatosis in chronic hepatitis C: why does it really matter? *Gut*, 55:123–130, 2006. 84

H. R. Babad, D. J. Nokes, et al. Predacting the impact of measles vaccination in Egland and Wales: model validation and analysis of policy options. *Epidemiology and Infection*, 114:319–344, 1995. 12

P. Bacchetti, M. R. Segal, and N. P. Jewell. Backcalculation of HIV Infection Rates. *Statistical Sciences*, 8:82–101, 1993. 48, 51

F. Ball and P. O'Neill. A Modification of the General Stochastic Epidemic Motivated by AIDS Modelling. *Advances in Applied Probability*, 25:39–62, 1993. 15

A. L. Barabási and A. Réka. Emergence of scaling in random networks. *Science*, 286:509–512, 1999. 70

M. S. Bartlett. Some Evolutionary Stochastic Processes. *Journal of the Royal Statistical Society. Serie B (Methodological)*, 11:211–229, 1949. 13

A. M. Bayoumi and G. S. Zaric. The cost-effectiveness of Vancouver's supervised injection facility. *Canadian Medical Association Journal*, 179:1143–1151, 2008. 75

N. G. Becker and I. C. Marschner. A method for estimating the age-specific relative risk of HIV infection from AIDS incidence data. *Biometrika*, 80:165–178, 1993. 2, 48, 51

R. Bellocco and I. C. Marschner. Joint analysis of HIV and AIDS surveillance data in back-calculation. *Statistics in Medicine*, 19:297–311, 2000. 65

Y. Benhamou, M. Bochet, et al. Liver Fibrosis Progression in Human Immunodeficiency Virus and Hepatitis C Virus Coinfected Patients. *Hepatology*, 30:1054–1058, 1999. 88, 90

A. Bhadra, E. L. Ionides, and K. Laneri. Malaria in Northwest India: Data Analysis via Partially Observed Stochastic Differential Equation Models Driven by Levy Noise. *Journal of the American Statistical Association*, 106:440–451, 2011. 16

D. Bishai, B. Johns, et al. Measles Eradication versus Measles Control:†An Economic Analysis. *Journal of Vaccines & Vaccination*, 2012. 13

P. Y. Bochud, T. Cai, et al. Genotype 3 is associated with accelerated fibrosis progression in chronic hepatitis C. *Journal of Hepatology*, 51:655–666, 2009. 84, 88, 90

L. Boelen, S. Teutsch, et al. Per-Event Probability of Hepatitis C Infection during Sharing of Injecting Equipment. *PLoS One*, 9, 2014. 75

J. A. Bondy and U. S. R. Murty. *Graph Theory with Applications*. North Holland, 1976. 9

C. Bosetti, F. Levi, et al. Trends in Mortality from Hepatocellular Carcinoma in Europe, 1980-2004. *Hepatology*, 48:137–145, 2008. 52

R. Brookmeyer and M. H. Gail. AIDS epidemiology: a quantitative approach. *Oxford University Press*, page 376, 1994. 48, 51

G. Cabibbo, M. Enea, et al. A meta-analysis of survival rates of untreated patients in randomized clinical trials of hepatocellular carcinoma. *Hepatology*, 51:1274–1283, 2010. 57

M. Carballo, R. Cody, and E. OReilly. *Migration, hepatitis B and hepatitis C*. International Centre for Migration, Health and Development, 2011. 93

L. Castells, V. Vargas, et al. Long interval between HCV infection and development of hepatocellular carcinoma. *Liver International*, 15:159–163, 1995. 57

S. Cauchemez and N. M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *Journal of the Royal Society*, 5: 885–897, 2008. 15

H. Chemaitelly, K. Chaabna, and L. J. Abu-Raddad. The Epidemiology of Hepatitis C Virus in the Fertile Crescent: Systematic Review and Meta-Analysis. *PLOS ONE*, 10, 2015. 93

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51:661–703, 2009. 70

V. Colizza, A. Barrat, et al. Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. *BMC Medicine*, 5, 2007. 68, 72

J. M. Costes, L. Vaissade, et al. Prévalence de l'usage problématique de drogues en France. *Observatoire français des drogues et des toxicomanies*, 2009. 79

J. M. Costes, L. Vaissade, et al. Rapport annuel 2010: état du phénomène de la drogue en Europe. *Observatoire européen des drogues et des toxicomanies*, 2010. 89

J. Cui and N. G. Becker. Estimating HIV incidence date of both HIV and AIDS diagnoses. *Statistics in Medecine*, 19:1165–1177, 2000. 65

N. Dalal, D. Greenhalgh, and X. Mao. A stochastic model of AIDS and condom use. *Journal of Mathematical Analysis and Applications*, 325:36–53, 2007. 16

G. D'Amico, G. Garcia-Tsao, and L. Pagliaro. Natural history and prognostic indicators of survival in cirrhosis: a systematic review of 118 studies. *Journal of Hepatology*, 44:217–231, 2006. 56

C. Dargatz. A Diffusion Approximation for an Epidemic Model. *EconStor*, 8, 2007. 17

M. A. Daw and A. A. Dau. Hepatitis C virus in Arab world: a state of concern. *The Scientific World Journal*, 2012, 2012. 93

A. P. Dempster, N. M . Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977. 7, 47, 48

S. Deuffic-Burban, L. Buffat, et al. Modeling the Hepatitis C Virus Epidemic in France. *Hepatology*, 29: 1596–1601, 1999. 2, 47, 48, 49, 50, 53, 58, 59, 65

S. Deuffic-Burban, P. Deltenre, et al. Impact of viral eradication on mortality related to hepatitis C: A modeling approach in France. *Journal of Hepatology*, 49:175–183, 2008. 52, 55, 56

D. Dhumeaux. *Prise en charge des personnes infectées par les virus de lhépatite B ou de lhépatite C*. Agence nationale de recherches sur le sida et les hépatites virales & Association française pour l'étude du foie, 2014. 86

K. Dombrowski, R. Curtis, et al. Topological and Historical Considerations for Infectious Disease Transmission among Injecting Drug Users in Bushwick, Brooklyn (USA). *World Journal of AIDS*, 1:1–9, 2013. 70, 74, 76

A. S. Duberg and R. Hultcrantz. Misleading Figures on Trends in Mortality from Hepatocellular Carcinoma in Europe. *Hepatology*, 49:336, 2008. 52

J. Dushoff, J. B. Plotkin, et al. Dynamical resonance can account for seasonality of influenza epidemics. *Proceeding of the National Academy of Sciences*, 101:16915–16916, 2004. 15

H. B. El-Serag. Heptocellular Carcinoma and Hepatitis C in the United States. *Hepatology*, 36:74–83, 2002. 53

H. B. El-Serag, A. C. Mason, and C. Key. Trends in survival of patients with hepatocellular carcinoma between 1977 and 1996 in the United States. *Hepatology*, 33:62–65, 2001. 56

A. R. El-Zayadi, H. M. Badran, et al. Heptocellular carcinoma in Egypt: a single center study over a decade. *World Journal of Gastroenterology*, 11:5193–5198, 2005. 53

P. Erdős and A. Rényi. On Random Graphs. *Publicationes Mathematicae*, 6:290–297, 1959. 10

European Centre for Disease Prevention and Control. *Hepatitis B and C surveillance in Europe 2012*, 2012. 1

N. M. Ferguson, C. A. Donnelly, and R. M. Anderson. The Foot-and-Mouth Eepidemic in Great Britain: Pattern of Spread and Impact of Interventions. *Science*, 292:1155–1160, 2001. 12

R. Fisher. On an Absolute Criterion for Fitting Frequency Curves. *Messenger of Mathematics*, 1912. 6

R. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, 222:309–368, 1922. 6

J. Geng, D. Xu, et al. Assessing hepatitis A virus epidemic stochastic process in eight cities in China in 1990. *International Journal of Epidemiology*, 27:320–322, 1998. 13

A. C. Ghani, J. Swinton, and G. P. Garnett. The role of sexual partnership networks in the epidemiology of gonorrhea. *Sexually transmitted diseases*, 24:45–56, 1997. 17

E. G. Giannini, F. Farinati, et al. Prognosis of untreated hepatocellular carcinoma. *Hepatology*, 61: 184–190, 2015. 57

E. Gower, C. Estes, et al. Global epidemiology and genotype distribution of the hepatitis C virus infection. *Journal of Hepatology*, 61:45–57, 2014. 47

J. Grebely, K. Page, et al. The effects of female sex, viral genotype, and IL28B genotype on spontaneous clearance of acute hepatitis C virus infection. *Hepatology*, 59:109–120, 2014. 88

M. Greenwood. The natural duration of cancer. *Reports of Public Health and Related Subjects*, 33:1–26, 1926. 7

J. Griffiths and B. Nix. Modeling the Hepatitis C Virus Epidemic in France Using Temporal Pattern of Hepatocellular Carcinoma Deaths. *Hepatology*, 35:709–715, 2002. 48

E. Grushka. Characterization of Exponentially Modified Gaussian Peaks in Chromatography. *Analytical Chemistry*, 44:1733–1738, 1972. 49

Haute Autorité de Santé. *Prise en charge de l'hépatite C par les médicaments anti-viraux à action directe (AAD)*. Haute Autorité de Santé (HAS), 2014. 86

M. D. Hernandez and K. E. Sherman. HIV/HCV Coinfection Natural History and Disease Progression, A Review of The Most Recent Literature. *Current Opinion in HIV and AIDS*, 6:478–482, 2011. 76, 88

L. Highleyman. *Does having HIV affect response to hepatitis C treatment?* infohep, 2016. 89

L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences*, 101:15124–15129, 2004. 68, 72

C. Hézode. Anti-NS5A associé au sofosbuvir : place dans la prise en charge des patients de génotype 3. *Hépato-Gastro et Oncologie Digestive*, 23:20–26, 2016. 86, 89, 90

Institut de veille sanitaire. *Prevalence des hepatites B et C en France en 2004*, 2007. 47, 54, 60, 65

M. Jauffret-Roustide, J. Pillonel, et al. Estimation de la séroprévalence du VIH et de l'hépatite C chez les usagers de drogues en France. Premiers résultats de l'enquête ANRS-Coquelicot 2011. *Bulletin épidémiologique hebdomadaire*, 39-40:504–509, 2013. 93

J. L. W. V. Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30:175–193, 1906. 8

F. Kanwal, J. R. Kramer, et al. HCV Genotype 3 is Associated With an Increased Risk of Cirrhosis and Hepatocellular Cancer in a National Sample of U.S. Veterans with HCV. *Hepatology*, 60:98–105, 2014. 84, 88, 90

E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of American Statistical Association*, 63:457–481, 1958. 6, 25

T. G. Kassem and J. N. Ndam. A stochastic modeling of recurrent measles epidemics. *Science World Journal*, 3:29–32, 2008. 16

W. O. Kermack and A. G. A. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society*, 115:700–721, 1927. 1, 2, 12, 23, 68

E. Kirmani and C. S. Hood. Analysis of a scanning model of worm propagation. *Journal of Computational Virology*, 6:31–42, 2010. 2, 22, 23, 24, 25, 27

J. P. Klein and M. L. Moeschberger. Survival Analysis: Techniques for Censored and Truncated Data. *Statistics for Biology and Health*, 16:538, 2003. 25

J. R. Kramer, M. A. Kowalkowski, et al. The effect of HIV viral control on the incidence of hepatocellular carcinoma in veterans with hepatitis C and HIV coinfection. *Journal of Acquired Immune Deficiency Syndromes*, 68:456–462, 2015. 88, 90

S. Laperche, A. Servant-Delmas, et al. La surveillance de la diversité des virus VIH, VHB et VHC chez les donneurs de sang français entre 2000 et 2010. *Bulletin épidémiologique hebdomadaire*, 39-40:447–452, 2012. 93

C. Larsen, G. Pialoux, et al. Prévalence des co-infections par les virus des hépatites B et C dans la population VIH+, France, juin 2004. *Bulletin épidémiologique hebdomadaire*, 23:109–116, 2005. 93

A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Didot, 1805. 5

M. Lehmann, M. F. Meyer, et al. High rate of spontaneous clearance of acute hepatitis C virus genotype 3 infection. *Journal of Medical Virology*, 73:387–391, 2004. 88

R. Levins. Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the Entomological Society of America*, 15:237–240, 1969. 72

M. Lewin, M. Gelu-Simeon, et al. Imaging Features and Prognosis of Hepatocellular Carcinoma in Patients with Cirrhosis Who Are Coinfected with Human Immunodeficiency Virus and Hepatitis C Virus. *Radiology*, 277:443–453, 2015. 88, 90

M. Liljenstam, D. M. Nicol, et al. Simulating Realistic Network Worm Traffic for Worm Warning System Design and Testing. *Proceedings of the 2003 ACM workshop on Rapid malcode*, pages 24–33, 2003. 23

A. S. Lok, L. B. Seeff, et al. Incidence of hepatocellular carcinoma and associated risk factors in hepatitis C-related advanced liver disease. *Gastroenterology*, 136:136–148, 2009. 55, 56, 88, 90

A. Maheshwari, P. J. Thuluvath, et al. Management of acute hepatitis C. *Clinics in liver disease*, 14: 169–176, 2010. 69

P. Marcellin, F. Pequignot, et al. Mortality related to chronic hepatitis B and chronic hepatitis C in Frace: Evidence for the role of HIV confection and alcohol consumption. *Journal of Hepatology*, 48: 200–207, 2008. 53

N. K. Martin, P. Vickerman, et al. Can antiviral therapy for hepatitis C reduce the prevalence of HCV among injecting drug user populations? A modeling analysis of its prevention. *Journal of Hepatology*, 54:1137–1144, 2011. 89, 90

P. McEwan, T. Ward, et al. Estimating the Incidence and Prevalence of Chronic Hepatitis C Infection in Taiwan Using Back Projection. *Value In Health Regional Issues*, 3C:5–11, 2014. 2, 48

L. J. McGarry, V. S. Pawar, et al. Economic model of a birth cohort screening program for hepatitis C virus. *Hepatology*, 55:1344–1355, 2012. 88, 90

A. G. McKendrick. Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 44:98–130, 1926. 13

J. M. Micallef, J. M. Kaldor, and G. J. Dore. Spontaneous Viral Clearance Following Acute Hepatitis C Infection: A Systematic Review of Longitudinal Studies. *Journal of Viral Hepatitis*, 13:34–41, 2006. 55, 56, 76, 88

S. Milgram. The Small World Problem. *Psychology Today*, 1:60–67, 1967. 11

A. H. Mohsen, P. J. Easterbrook, et al. Impact of human immunodeficiency virus (HIV) infection on the progression of liver fibrosis in hepatitis C virus infected patients. *Gut*, 52:1035–1040, 2003. 88, 90

D. Moore, C. Shannon, and J. Brown. Code-red: a case study on the spread and victims of an internet worm. *Proceedings of the 2nd Internet Measurement Workshop*, pages 273–284, 2002. 43

H. C. Moore, P. Jacoby, et al. Modelling the Seasonal Epidemics of Respiratory Syncytial Virus in Young Children. *PLoS ONE*, 9, 2014. 12

A. Mourad. Modelisation de la morbi-mortalite du carcinome hepatocellulaire en France par stade de gravite : Evaluation de differentes strategies en fonction du depistage et des ressources therapeutiques. PhD thesis. Universite Lille Nord de France. page 165, 2014. 52

G. Nkontchou, M. Ziol, et al. HCV genotype 3 is associated with a higher hepatocellular carcinoma incidence in patients with ongoing viral C cirrhosis. *Journal of Viral Hepatitis*, 18:516–522, 2011. 84, 88, 90

A. O. Ogunmola. On the Use of Discrete-Time Markov Process for HIV/AIDS Epidemic Modeling. *International Journal of Modern Mathematical Sciences*, 2:82–95, 2014. 13

F. M. O'Leary and T. C. Green. Community acquired needlestick injuries in non-health care workers presenting to an urban emergency department. *Emergency Medicine*, 15:434–440, 2003. 75

R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86, 2001. 69, 70

R. Pastor-Satorras and A. Vespignani. Epidemics and immunization in scale-free networks. *Cornell University Library*, 2002. 18

P. Patel, C. B. Borkowf, et al. Estimating per-act HIV transmission risk: a systematic review. *AIDS*, 28: 1509–1519, 2014. 75, 76

C. Payan, F. Roudot-Thoraval, et al. Changing of hepatitis C virus genotype patterns in France at the beginning of the third millenium: The GEMHEP GenoCII Study. *Journal of Viral Hepatitis*, 12: 405–413, 2005. 84

R. Pearl and L. J. Reed. On the rate of growth of the population of united states since1790 and its mathematical representation. *Proceedings of the National Academy of Sciences*, 6:275–288, 1920. 23

L. Pelude. Networks among Injection Drug Users: Random or Scale-Free? 2007. 70

J. F. Perez, G. L. Armstrong, et al. The contribution of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide. *Journal of Hepatology*, 45:529–538, 2006. 53

J. A. Pineda, M. Romero-Gómez, et al. HIV Coinfection Shortens the Survival of Patients With Hepatitis C Virus-Related Decompensated Cirrhosis. *Hepatology*, 41:779–789, 2005. 88, 90

R. Planas, B. Ballesté, et al. Natural history of decompensated hepatitis C virus-related cirrhosis. A study of 200 patients. *Journal of Hepatology*, 40:823–830, 2004. 56

S. Pol, M. Corouge, and A. Vallet-Pichard. Daclatasvir-sofosbuvir combination therapy with or without ribavirin for hepatitis C virus infection: from the clinical trials to real life. *Hepatic Medicine: Evidence and Research*, 8:21–26, 2016. 86, 89, 90

B. Rachlis, K. C. Brouwer, et al. Migration and transmission of blood-borne infections among injection drug users: Understanding the epidemiologic bridge. *Drug and Alcohol Dependence*, 90:107–119, 2007. 79

G. F. Ragget. Modeling the Eyam plague. *The Institute of Mathematics and its Applications*, 18:221–226, 1982. 15

H. Razavi, I. Waked, et al. The present and future disease burden of hepatitis c virus (HCV) infection with todays treatment paradigm. *Journal of Viral Hepatitis*, 21:34–59, 2014. 75, 76, 93

V. Lo Re, M. J. Kallan, et al. Co-Infection with HIV Increases Risk for Decompensation in Patients with HCV. *Annals of Internal Medicine*, 160:369–379, 2014. 88, 90

D. A. Rolls, G. Daraganova, and al. Modelling hepatitis c transmission over a social network of injecting drug users. *Journal of Theoretical Biology*, 297:73–87, 2012a. 17

D. A. Rolls, G. Daraganova, et al. Modelling hepatitis C transmission over a social network of injecting drug users. *Journal of Theoretical Biology*, 297:73–87, 2012b. 69, 75, 76

B. W. Silverman and D. W. Nychka. A Smoothed EM Approach to Indirect Estimation Problems, with Particular Reference to Stereology and Emission Tomography. *Journal of the Royal Statistical Society. Series B (Methodological)*, 52:271–324, 1990. 2, 8, 47, 48, 50, 51

B. Soto, A. Sánchez-Quijano, et al. Human immunodeficiency virus infection modifies the natural history of chronic parenterally-acquired hepatitis C with an unusually rapid progression to cirrhosis. *Journal of Hepatology*, 26:1–5, 1997. 88, 90

S. Staniford, V. Paxson, and N. Weaver. How to 0wn the Internet in Your Spare Time. *Proceedings of the 11th USENIX Security Symposium*, pages 149–167, 2002. 2, 22, 23

Y. Sun, C. Liu, and al. Epidemic spreading on weighted complex networks. *Physics Letters A*, 378: 635–640, 2014. 18

M. J. Sweeting, D. De Angelis, et al. The burden of hepatitis C in England. *Journal of Viral Hepatitis*, 14:570–576, 2007. 48

E. B. Tapper and N. H. Afdhal. Is 3 the new 1: perspectives on virology, natural history and treatment for hepatitis C genotype 3. *Journal of Viral Hepatitis*, 20:669–677, 2013. 84

N. A. Terrault, M. E. Roland, et al. Outcomes of Liver Transplantation in HCV-HIV Coinfected Recipients. *Liver Transplantation*, 18:716–726, 2012. 89, 90

H. H. Thein, Q. Yi, et al. Estimation of stage-specific fibrosis progression rates in chronic hepatitis C virus infection: A meta-analysis and meta-regression. *Hepatology*, 48:418–432, 2008. 88, 90

M. J. Tong, N. S. El-Farra, et al. Clinical outcomes after transfusion-associated hepatitis C. *The New England Journal of Medecine*, 22:1463–1466, 1995. 57

A. Traulsen, J. C. Claussen, and C. Hauert. Stochastic differential equations for evolutionary dynamics with demographic noise and mutations. *Cornell University Library*, 2012. 16

D. K. van Santen, J. J. van der Helm, et al. Temporal trends in mortality among people who use drugs compared with the general Dutch population differ by hepatitis C virus and HIV infection status. *AIDS*, 28:25892599, 2014. 76

P. F. Verhulst. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance mathématique et physique*, 10:113–121, 1838. 22, 23, 49

P. Vickerman, M. Hickman, and A. Judd. Modeling the impact of Hepatitis C transmission of reducing string sharing: London case study. *International Journal of Epidemiology*, 36:396–405, 2007. 75, 76, 89, 90

P. Vickerman, A. Miners, and J. Williams. Assessing the cost-effectiveness of interventions linked to needle and syringe programmes for injecting drug users: An economic modelling report. *National Institute for Health and Care Excellence*, 2008. 89, 90

M. Vojnović and A. Ganesh. On the Race of Worms, Alerts and Patches. *IEEE/ACM Transactions on Networking*, 16:1066–1079, 2008. 23

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998. 10

World Health Organization. *International Classification of Diseases: 9th revision*. Geneva, Switzerland: World Health Organization, 1977. 52

World Health Organization. *International Statistical Classification of Diseases and related Health Problems: 10th revision*. Geneva, Switzerland: World Health Organization, 1992. 52

D. L. Wyles, P. J. Ruane, et al. Daclatasvir plus Sofosbuvir for HCV in Patients Coinfected with HIV-1. *The New England Journal of Medicine*, 373:714–725, 2015. 86, 89, 90

M. Xu, Y. Ding, and L. Hu. A Stochastic Model for Prevention and Control of HIV/AIDS Transmission Dynamics. *Lecture Notes in Computer Science*, 4689:28–37, 2007. 16

H. Yoshizawa. Hepatocellular carcinoma associated with hepatitis C virus infection in Japan: projection to other counties in the foreseeable future. *Oncology*, 62:8–17, 2002. 53

M. Youssef and C. Scoglio. An individual-based approach to SIR epidemics in contact networks. *Journal of Theoretical Biology*, 283:136–144, 2011. 18

C. C. Zou, W. Gong, and D. Towsley. Code Red Worm Propagation Modeling and Analysis. *Proceedings of the 9th ACM conference on Computer and communications security*, pages 138–147, 2002. 1, 23

C. C. Zou, D. Towsley, and W. Gong. On the performance of internet worm scanning strategies. *Performance Evaluation*, 63:700–723, 2006. 24

Appendices

A Differential equations

$$
\begin{aligned}
\frac{\mathrm{d}s_{k,\cdot}(t)}{\mathrm{d}t} &= \Lambda - k\beta s_{k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* s_{k,\cdot}(t)\theta^*_{k,\cdot}(t) - k\beta^\dagger s_{k,\cdot}(t)\theta^{**}_{k,\cdot}(t) - \mu s_{k,\cdot}(t) \\
&\quad -s_{k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta\Big(a_\Phi(t) + c_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big) \\
&\quad +\beta^*\Big(s^*_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big) + \beta^\dagger\Big(a^*_\Phi(t) + c^*_\Phi(t)\Big)\Big] + p\sigma a_{k,\cdot}(t) + \gamma c_{k,\cdot}(t), \\
\frac{\mathrm{d}s^*_{k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s^*_{k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* s_{k,\cdot}(t)\theta^*_{k,\cdot}(t) + p^*\sigma a^*_{k,\cdot}(t) + \gamma^* c^*_{k,\cdot}(t) - \mu^* s^*_{k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta^* s_{k,\cdot}(t)\Big(s^*_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big) \\
&\quad -\beta s^*_{k,\cdot}(t)\Big(a_\Phi(t) + c_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big)\Big], \\
\frac{\mathrm{d}a_{k,\cdot}(t)}{\mathrm{d}t} &= k\beta s_{k,\cdot}(t)\theta_{k,\cdot}(t) - (\sigma+\mu)a_{k,\cdot}(t) - k\beta^* a_{k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&\quad +\beta s_{k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(a_\Phi(t) + c_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big), \\
\frac{\mathrm{d}a^*_{k,\cdot}(t)}{\mathrm{d}t} &= k\beta s^*_{k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^\dagger s_{k,\cdot}(t)\theta^{**}_{k,\cdot}(t) + k\beta^* a_{k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\sigma+\mu^*)a^*_{k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s^*_{k,\cdot}(t)\Big(a_\Phi(t) + c_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big) \\
&\quad +\beta^* a_{k,\cdot}(t)\Big(s^*_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big) + \beta^\dagger s_{k,\cdot}(t)\Big(a^*_\Phi(t) + c^*_\Phi(t)\Big)\Big], \\
\frac{\mathrm{d}c_{k,\cdot}(t)}{\mathrm{d}t} &= (1-p)\sigma a_{k,\cdot}(t) - k\beta^* c_{k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\gamma+\delta+\mu)c_{k,\cdot}(t) \\
&\quad -\beta^*(a_{k,\cdot}(t) + c_{k,\cdot}(t))\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big), \\
\frac{\mathrm{d}c^*_{k,\cdot}(t)}{\mathrm{d}t} &= (1-p^*)\sigma a^*_{k,\cdot}(t) + k\beta^* c_{k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\gamma^*+\delta^*+\mu^*)c^*_{k,\cdot}(t) \\
&\quad +\beta^* c_{k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t) + a^*_\Phi(t) + c^*_\Phi(t)\Big).
\end{aligned}
$$

B The main French urban areas

Urban area	Population	Urban area	Population
Paris	12,341,418	Saint-Brieuc	171,721
Lyon	2,214,068	Béziers	165,498
Marseille - Aix-en-Provence	1,727,070	Montbéliard	162,326
Toulouse	1,270,760	Niort	152,721
Lille†	1,166,452	Vannes	150,860
Bordeaux	1,158,431	Chartres	145,735
Nice	1,004,914	Bourges	139,968
Nantes	897,713	Thionville	135,627
Strasbourg†	768,868	Chalon-sur-Saône	133,557
Rennes	690,467	Boulogne-sur-Mer	133,062
Grenoble	679,863	Maubeuge	129,931
Rouen	658,285	Arras	128,784
Toulon	611,237	Colmar	127,625
Montpellier	569,956	Blois	127,053
Douai - Lens	540,981	Calais	126,266
Avignon	515,536	Quimper	125,487
Saint-Etienne	512,830	Beauvais	125,095
Tours	483,744	Bourg-en-Bresse	122,806
Clermont-Ferrand	469,922	Laval	121,399
Nancy	434,479	La Roche-sur-Yon	117,965
Orléans	423,123	Creil	117,654
Angers	403,633	Cherbourg-Octeville	116,517
Caen	403,765	Tarbes	115,557
Metz	389,700	Belfort	114,077
Dijon	377,590	Alès	113,769
Béthune	368,633	Vienne	112,334
Valenciennes†	367,094	Agen	111,663
Le Mans	344,893	Saint-Quentin	111,474
Reims	317,611	Evreux	111,449
Brest	314,844	Roanne	107,209
Perpignan	309,962	Charleville-Mézières	106,835
Amiens	293,671	Montauban	105,654
Genève - Annemasse†	292,180	Cholet	104,917
Le Havre	290,890	Périgueux	102,417
Bayonne†	288,359	Sarrebruck - Forbach†	101,806
Mulhouse	284,739	Nevers	101,586
Limoges	282,971	Brive-la-Gaillarde	101,435
Nîmes	259,348	Ajaccio	100,643
Dunkerque	257,773	Mâcon	99,873
Poitiers	255,831	Carcassonne	97,801
Besançon	246,841	Albi	97,667
Pau	240,857	Compiègne	97,502
Annecy	221,111	Bastia	93,971
Chambéry	217,356	Epinal	93,891
Lorient	215,591	Fréjus	93,562
Saint-Nazaire	213,083	Bâle - Saint-Louis†	93,018
La Rochelle	207,211	Châteauroux	92,723
Troyes	191,505	Auxerre	92,307
Angoulême	180,593	Sète	91,101
Valence	175,636	Cluses	90,872

Table B.1: Population of the French urban areas based on census data for the year 2012.
† Only the French part is considered.

C

Differential equations - Enhanced model - Genotype 1

$$
\begin{aligned}
\frac{\mathrm{d}s_{k,\cdot}(t)}{\mathrm{d}t} &= \Lambda_{k,\cdot}(t) - k\beta s_{k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* s_{k,\cdot}(t)\theta^*_{k,\cdot}(t) - k\beta^\dagger s_{k,\cdot}(t)\theta^{**}_{k,\cdot}(t) \\
&\quad - s_{k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta\Big(\phi_\Phi(t) + \phi^*_\Phi(t) + \psi_\Phi(t) + \psi^*_\Phi(t)\Big) \\
&\quad + \beta^*\Big(s^*_\Phi(t) + \phi^*_\Phi(t) + \psi^*_\Phi(t)\Big) + \beta^\dagger\Big(\phi^*_\Phi(t) + \psi^*_\Phi(t)\Big)\Big] + p\sigma(a_{k,\cdot}(t) + \breve{a}_{k,\cdot}(t)) \\
&\quad + \lambda_{S_0} f_{0,k,\cdot}(t) + \kappa_{S_0}\breve{f}_{0,k,\cdot}(t) - (\tau+\mu)s_{k,\cdot}(t) \\
\frac{\mathrm{d}s^*_{k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s^*_{k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* s_{k,\cdot}(t)\theta^*_{k,\cdot}(t) + p^*\sigma(a^*_{k,\cdot}(t) + \breve{a}^*_{k,\cdot}(t)) - (\tau+\mu^*)s^*_{k,\cdot}(t) \\
&\quad + \sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta^* s_{k,\cdot}(t)\Big(s^*_\Phi(t) + \phi^*_\Phi(t) + \psi^*_\Phi(t)\Big) \\
&\quad - \beta s^*_{k,\cdot}(t)\Big(\phi_\Phi(t) + \phi^*_\Phi(t) + \psi_\Phi(t) + \psi^*_\Phi(t)\Big)\Big] + \lambda_{S_0} f^*_{0,k,\cdot}(t) + \kappa_{S_0}\breve{f}^*_{0,k,\cdot}(t) \\
\frac{\mathrm{d}a_{k,\cdot}(t)}{\mathrm{d}t} &= k\beta s_{k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* a_{k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\sigma+\tau+\mu)a_{k,\cdot}(t) \\
&\quad + \sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s_{k,\cdot}(t)\Big(\phi_\Phi(t) + \phi^*_\Phi(t) + \psi_\Phi(t) + \psi^*_\Phi(t)\Big) \\
&\quad - \beta^* a_{k,\cdot}(t)\Big(s^*_\Phi(t) + \phi^*_\Phi(t) + \psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}a^*_{k,\cdot}(t)}{\mathrm{d}t} &= k\beta s^*_{k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* a_{k,\cdot}(t)\theta^*_{k,\cdot}(t) + k\beta^\dagger s_{k,\cdot}(t)\theta^{**}_{k,\cdot}(t) - (\sigma+\tau+\mu^*)a^*_{k,\cdot}(t) \\
&\quad + \sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s^*_{k,\cdot}(t)\Big(\phi_\Phi(t) + \phi^*_\Phi(t) + \psi_\Phi(t) + \psi^*_\Phi(t)\Big) \\
&\quad + \beta^* a_{k,\cdot}(t)\Big(s^*_\Phi(t) + \phi^*_\Phi(t) + \psi^*_\Phi(t)\Big) + \beta^\dagger s_{k,\cdot}(t)\Big(\phi^*_\Phi(t) + \psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}f_{0,k,\cdot}(t)}{\mathrm{d}t} &= (1-p)\sigma a_{k,\cdot}(t) - \lambda_1(1-\lambda_{S_0})f_{0,k,\cdot}(t) - \lambda_{S_0} f_{0,k,\cdot}(t) - k\beta^* f_{0,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&\quad - \beta^* f_{0,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t) + \phi^*_\Phi(t) + \psi^*_\Phi(t)\Big) - (\tau+\mu)f_{0,k,\cdot}(t) \\
\frac{\mathrm{d}f^*_{0,k,\cdot}(t)}{\mathrm{d}t} &= (1-p^*)\sigma a^*_{k,\cdot}(t) - \lambda^*_1(1-\lambda_{S_0})f^*_{0,k,\cdot}(t) - \lambda_{S_0} f^*_{0,k,\cdot}(t) + k\beta^* f_{0,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&\quad + \beta^* f_{0,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t) + \phi^*_\Phi(t) + \psi^*_\Phi(t)\Big) - (\tau+\mu^*)f^*_{0,k,\cdot}(t)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}s_{1,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s_{1,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* s_{1,k,\cdot}(t)\theta^*_{k,\cdot}(t) - k\beta^\dagger s_{1,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) \\
&\quad -s_{1,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad +\beta^*\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta^\dagger\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + p\sigma(a_{1,k,\cdot}(t)+\breve{a}_{1,k,\cdot}(t)) \\
&\quad +\lambda_{S_1} f_{1,k,\cdot}(t) + \kappa_{S_1}\breve{f}_{1,k,\cdot}(t) - (\tau+\mu)s_{1,k,\cdot}(t) \\
\frac{\mathrm{d}s^*_{1,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s^*_{1,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* s_{1,k,\cdot}(t)\theta^*_{k,\cdot}(t) + p^*\sigma(a^*_{1,k,\cdot}(t)+\breve{a}^*_{1,k,\cdot}(t)) - (\tau+\mu^*)s^*_{1,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta^* s_{1,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad -\beta s^*_{1,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + \lambda_{S_1} f^*_{1,k,\cdot}(t) + \kappa_{S_1}\breve{f}^*_{1,k,\cdot}(t) \\
\frac{\mathrm{d}a_{1,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s_{1,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* a_{1,k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\sigma+\tau+\mu)a_{1,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s_{1,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad -\beta^* a_{1,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}a^*_{1,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s^*_{1,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* a_{1,k,\cdot}(t)\theta^*_{k,\cdot}(t) + k\beta^\dagger s_{1,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) - (\sigma+\tau+\mu^*)a^*_{1,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s^*_{1,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad +\beta^* a_{1,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta^\dagger s_{1,k,\cdot}(t)\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}f_{1,k,\cdot}(t)}{\mathrm{d}t} &= (1-p)\sigma a_{1,k,\cdot}(t) + \lambda_1(1-\lambda_{S_0})f_{0,k,\cdot}(t) - \lambda_2(1-\lambda_{S_1})f_{1,k,\cdot}(t) - \lambda_{S_1} f_{1,k,\cdot}(t) - k\beta^* f_{1,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&\quad -\beta^* f_{1,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu)f_{1,k,\cdot}(t) \\
\frac{\mathrm{d}f^*_{1,k,\cdot}(t)}{\mathrm{d}t} &= (1-p^*)\sigma a^*_{1,k,\cdot}(t) + \lambda^*_1(1-\lambda_{S_0})f^*_{0,k,\cdot}(t) - \lambda^*_2(1-\lambda_{S_1})f^*_{1,k,\cdot}(t) - \lambda_{S_1} f^*_{1,k,\cdot}(t) + k\beta^* f_{1,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&\quad +\beta^* f_{1,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu^*)f^*_{1,k,\cdot}(t) \\
\frac{\mathrm{d}s_{2,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s_{2,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* s_{2,k,\cdot}(t)\theta^*_{k,\cdot}(t) - k\beta^\dagger s_{2,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) \\
&\quad -s_{2,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad +\beta^*\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta^\dagger\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + p\sigma(a_{2,k,\cdot}(t)+\breve{a}_{2,k,\cdot}(t)) \\
&\quad +\lambda_{S_2} f_{2,k,\cdot}(t) + \kappa_{S_2}\breve{f}_{2,k,\cdot}(t) - (\tau+\mu)s_{2,k,\cdot}(t) \\
\frac{\mathrm{d}s^*_{2,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s^*_{2,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* s_{2,k,\cdot}(t)\theta^*_{k,\cdot}(t) + p^*\sigma(a^*_{2,k,\cdot}(t)+\breve{a}^*_{2,k,\cdot}(t)) - (\tau+\mu^*)s^*_{2,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta^* s_{2,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad -\beta s^*_{2,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + \lambda_{S_2} f^*_{2,k,\cdot}(t) + \kappa_{S_2}\breve{f}^*_{2,k,\cdot}(t)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}a_{2,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s_{2,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* a_{2,k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\sigma+\tau+\mu)a_{2,k,\cdot}(t) \\
&+ \sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s_{2,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&- \beta^* a_{2,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}a^*_{2,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s^*_{2,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* a_{2,k,\cdot}(t)\theta^*_{k,\cdot}(t) + k\beta^\dagger s_{2,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) - (\sigma+\tau+\mu^*)a^*_{2,k,\cdot}(t) \\
&+ \sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s^*_{2,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&+ \beta^* a_{2,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) + \beta^\dagger s_{2,k,\cdot}(t)\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}f_{2,k,\cdot}(t)}{\mathrm{d}t} &= (1-p)\sigma a_{2,k,\cdot}(t) + \lambda_2(1-\lambda_{S_1})f_{1,k,\cdot}(t) - \lambda_3(1-\lambda_{S_2})f_{2,k,\cdot}(t) - \lambda_{S_2}f_{2,k,\cdot}(t) - k\beta^* f_{2,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&-\beta^* f_{2,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu)f_{2,k,\cdot}(t) \\
\frac{\mathrm{d}f^*_{2,k,\cdot}(t)}{\mathrm{d}t} &= (1-p^*)\sigma a^*_{2,k,\cdot}(t) + \lambda^*_2(1-\lambda_{S_1})f^*_{1,k,\cdot}(t) - \lambda^*_3(1-\lambda_{S_2})f^*_{2,k,\cdot}(t) - \lambda_{S_2}f^*_{2,k,\cdot}(t) + k\beta^* f_{2,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&+\beta^* f_{2,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu^*)f^*_{2,k,\cdot}(t) \\
\frac{\mathrm{d}s_{3,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s_{3,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* s_{3,k,\cdot}(t)\theta^*_{k,\cdot}(t) - k\beta^\dagger s_{3,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) \\
&-s_{3,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&+\beta^*\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) + \beta^\dagger\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + p\sigma(a_{3,k,\cdot}(t)+\breve{a}_{3,k,\cdot}(t)) \\
&+\lambda_{S_3}f_{3,k,\cdot}(t) + \kappa_{S_3}\breve{f}_{3,k,\cdot}(t) - (\tau+\mu)s_{3,k,\cdot}(t) \\
\frac{\mathrm{d}s^*_{3,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s^*_{3,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* s_{3,k,\cdot}(t)\theta^*_{k,\cdot}(t) + p^*\sigma(a^*_{3,k,\cdot}(t)+\breve{a}^*_{3,k,\cdot}(t)) - (\tau+\mu^*)s^*_{3,k,\cdot}(t) \\
&+\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta^* s_{3,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&-\beta s^*_{3,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + \lambda_{S_3}f^*_{3,k,\cdot}(t) + \kappa_{S_3}\breve{f}^*_{3,k,\cdot}(t) \\
\frac{\mathrm{d}a_{3,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s_{3,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* a_{3,k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\sigma+\tau+\mu)a_{3,k,\cdot}(t) \\
&+\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s_{3,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&-\beta^* a_{3,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}a^*_{3,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s^*_{3,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* a_{3,k,\cdot}(t)\theta^*_{k,\cdot}(t) + k\beta^\dagger s_{3,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) - (\sigma+\tau+\mu^*)a^*_{3,k,\cdot}(t) \\
&+\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s^*_{3,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&+\beta^* a_{3,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) + \beta^\dagger s_{3,k,\cdot}(t)\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}f_{3,k,\cdot}(t)}{\mathrm{d}t} &= (1-p)\sigma a_{3,k,\cdot}(t) + \lambda_3(1-\lambda_{S_2})f_{2,k,\cdot}(t) - \lambda_4(1-\lambda_{S_3})f_{3,k,\cdot}(t) - \lambda_{S_3}f_{3,k,\cdot}(t) - \lambda_h(1-\lambda_{S_3})f_{3,k,\cdot}(t) \\
&\quad -\beta^* f_{3,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu)f_{3,k,\cdot}(t) - k\beta^* f_{3,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
\frac{\mathrm{d}f^*_{3,k,\cdot}(t)}{\mathrm{d}t} &= (1-p^*)\sigma a^*_{3,k,\cdot}(t) + \lambda^*_3(1-\lambda_{S_2})f^*_{2,k,\cdot}(t) - \lambda^*_4(1-\lambda_{S_3})f^*_{3,k,\cdot}(t) - \lambda_{S_3}f^*_{3,k,\cdot}(t) - \lambda^*_h(1-\lambda_{S_3})f^*_{3,k,\cdot}(t) \\
&\quad +\beta^* f_{3,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu^*)f^*_{3,k,\cdot}(t) + k\beta^* f_{3,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
\frac{\mathrm{d}s_{4,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s_{4,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* s_{4,k,\cdot}(t)\theta^*_{k,\cdot}(t) - k\beta^\dagger s_{4,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) \\
&\quad -s_{4,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad +\beta^*\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta^\dagger\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + p\sigma(a_{4,k,\cdot}(t)+\breve{a}_{4,k,\cdot}(t)) \\
&\quad +\lambda_{S_4}f_{4,k,\cdot}(t) + \kappa_{S_4}\breve{f}_{4,k,\cdot}(t) - (\tau+\mu)s_{4,k,\cdot}(t) \\
\frac{\mathrm{d}s^*_{4,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s^*_{4,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* s_{4,k,\cdot}(t)\theta^*_{k,\cdot}(t) + p^*\sigma(a^*_{4,k,\cdot}(t)+\breve{a}^*_{4,k,\cdot}(t)) - (\tau+\mu^*)s^*_{4,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta^* s_{4,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad -\beta s^*_{4,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + \lambda_{S_4}f^*_{4,k,\cdot}(t) + \kappa_{S_4}\breve{f}^*_{4,k,\cdot}(t) \\
\frac{\mathrm{d}a_{4,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s_{4,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* a_{4,k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\sigma+\tau+\mu)a_{4,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s_{4,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad -\beta^* a_{4,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}a^*_{4,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s^*_{4,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* a_{4,k,\cdot}(t)\theta^*_{k,\cdot}(t) + k\beta^\dagger s_{4,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) - (\sigma+\tau+\mu^*)a^*_{4,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s^*_{4,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad +\beta^* a_{4,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta^\dagger s_{4,k,\cdot}(t)\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}f_{4,k,\cdot}(t)}{\mathrm{d}t} &= (1-p)\sigma a_{4,k,\cdot}(t) + \lambda_4(1-\lambda_{S_3})f_{3,k,\cdot}(t) - \lambda_H(1-\lambda_{S_4})f_{4,k,\cdot}(t) - \lambda_{S_4}f_{4,k,\cdot}(t) - \lambda_D(1-\lambda_{S_4})f_{4,k,\cdot}(t) \\
&\quad -\beta^* f_{4,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu)f_{4,k,\cdot}(t) - k\beta^* f_{4,k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
\frac{\mathrm{d}f^*_{4,k,\cdot}(t)}{\mathrm{d}t} &= (1-p^*)\sigma a^*_{4,k,\cdot}(t) + \lambda^*_4(1-\lambda_{S_3})f^*_{3,k,\cdot}(t) - \lambda^*_H(1-\lambda_{S_4})f^*_{4,k,\cdot}(t) - \lambda_{S_4}f^*_{4,k,\cdot}(t) - \lambda^*_D(1-\lambda_{S_4})f^*_{4,k,\cdot}(t) \\
&\quad +\beta^* f_{4,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu^*)f^*_{4,k,\cdot}(t) + k\beta^* f_{4,k,\cdot}(t)\theta^*_{k,\cdot}(t)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\mathrm{d}s_{D,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s_{D,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* s_{D,k,\cdot}(t)\theta^*_{k,\cdot}(t) - k\beta^\dagger s_{D,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) \\
&\quad -s_{D,k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad +\beta^*\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta^\dagger\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + p\sigma(a_{D,k,\cdot}(t)+\breve{a}_{D,k,\cdot}(t)) \\
&\quad +\lambda_{S_D} dc_{k,\cdot}(t) + \kappa_{S_D}\breve{dc}_{k,\cdot}(t) - (\tau+\mu)s_{D,k,\cdot}(t) \\
\frac{\mathrm{d}s^*_{D,k,\cdot}(t)}{\mathrm{d}t} &= -k\beta s^*_{D,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* s_{D,k,\cdot}(t)\theta^*_{k,\cdot}(t) + p^*\sigma(a^*_{D,k,\cdot}(t)+\breve{a}^*_{D,k,\cdot}(t)) - (\tau+\mu^*)s^*_{D,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta^* s_{D,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad -\beta s^*_{D,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] + \lambda_{S_D} dc^*_{k,\cdot}(t) + \kappa_{S_D}\breve{dc}^*_{k,\cdot}(t) \\
\frac{\mathrm{d}a_{D,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s_{D,k,\cdot}(t)\theta_{k,\cdot}(t) - k\beta^* a_{D,k,\cdot}(t)\theta^*_{k,\cdot}(t) - (\sigma+\tau+\mu)a_{D,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s_{D,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad -\beta^* a_{D,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}a^*_{D,k,\cdot}(t)}{\mathrm{d}t} &= k\beta s^*_{D,k,\cdot}(t)\theta_{k,\cdot}(t) + k\beta^* a_{D,k,\cdot}(t)\theta^*_{k,\cdot}(t) + k\beta^\dagger s_{D,k,\cdot}(t)\theta^{**}_{k,\cdot}(t) - (\sigma+\tau+\mu^*)a^*_{D,k,\cdot}(t) \\
&\quad +\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big[\beta s^*_{D,k,\cdot}(t)\Big(\phi_\Phi(t)+\phi^*_\Phi(t)+\psi_\Phi(t)+\psi^*_\Phi(t)\Big) \\
&\quad +\beta^* a_{D,k,\cdot}(t)\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)+\beta^\dagger s_{D,k,\cdot}(t)\Big(\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big)\Big] \\
\frac{\mathrm{d}dc_{k,\cdot}(t)}{\mathrm{d}t} &= (1-p)\sigma a_{D,k,\cdot}(t) + \lambda_D(1-\lambda_{S_4})f_{4,k,\cdot}(t) - \lambda_{D_D}(1-\lambda_{S_D})dc_{k,\cdot}(t) - \lambda_{S_D} dc_{k,\cdot}(t) - \lambda_{D_H}(1-\lambda_{S_D})dc_{k,\cdot}(t) \\
&\quad -\beta^* dc_{k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu)dc_{k,\cdot}(t) - k\beta^* dc_{k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
\frac{\mathrm{d}dc^*_{k,\cdot}(t)}{\mathrm{d}t} &= (1-p^*)\sigma a^*_{D,k,\cdot}(t) + \lambda^*_D(1-\lambda_{S_4})f^*_{4,k,\cdot}(t) - \lambda^*_{D_D}(1-\lambda_{S_D})dc^*_{k,\cdot}(t) - \lambda_{S_D} dc^*_{k,\cdot}(t) - \lambda^*_{D_H}(1-\lambda_{S_D})dc^*_{k,\cdot}(t) \\
&\quad +\beta^* dc_{k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu^*)dc^*_{k,\cdot}(t) + k\beta^* dc_{k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
\frac{\mathrm{d}hcc_{k,\cdot}(t)}{\mathrm{d}t} &= \lambda_H(1-\lambda_{S_4})f_{4,k,\cdot}(t) + \lambda_{D_H}(1-\lambda_{S_D})dc_{k,\cdot}(t) - \lambda_{H_D} hcc_{k,\cdot}(t) - k\beta^* hcc_{k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&\quad -\beta^* hcc_{k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu)hcc_{k,\cdot}(t) \\
\frac{\mathrm{d}hcc^*_{k,\cdot}(t)}{\mathrm{d}t} &= \lambda^*_H(1-\lambda_{S_4})f^*_{4,k,\cdot}(t) + \lambda^*_{D_H}(1-\lambda_{S_D})dc^*_{k,\cdot}(t) - \lambda^*_{H_D} hcc^*_{k,\cdot}(t) + k\beta^* hcc_{k,\cdot}(t)\theta^*_{k,\cdot}(t) \\
&\quad +\beta^* hcc_{k,\cdot}(t)\sum_{\Phi}\left(\frac{\tau_{\Phi,\cdot}}{n_\cdot + \sum_\Phi \tau_{\Phi,\cdot}}\right)\cdot\Big(s^*_\Phi(t)+\phi^*_\Phi(t)+\psi^*_\Phi(t)\Big) - (\tau+\mu^*)hcc^*_{k,\cdot}(t)
\end{aligned}
$$

D

Alternative model

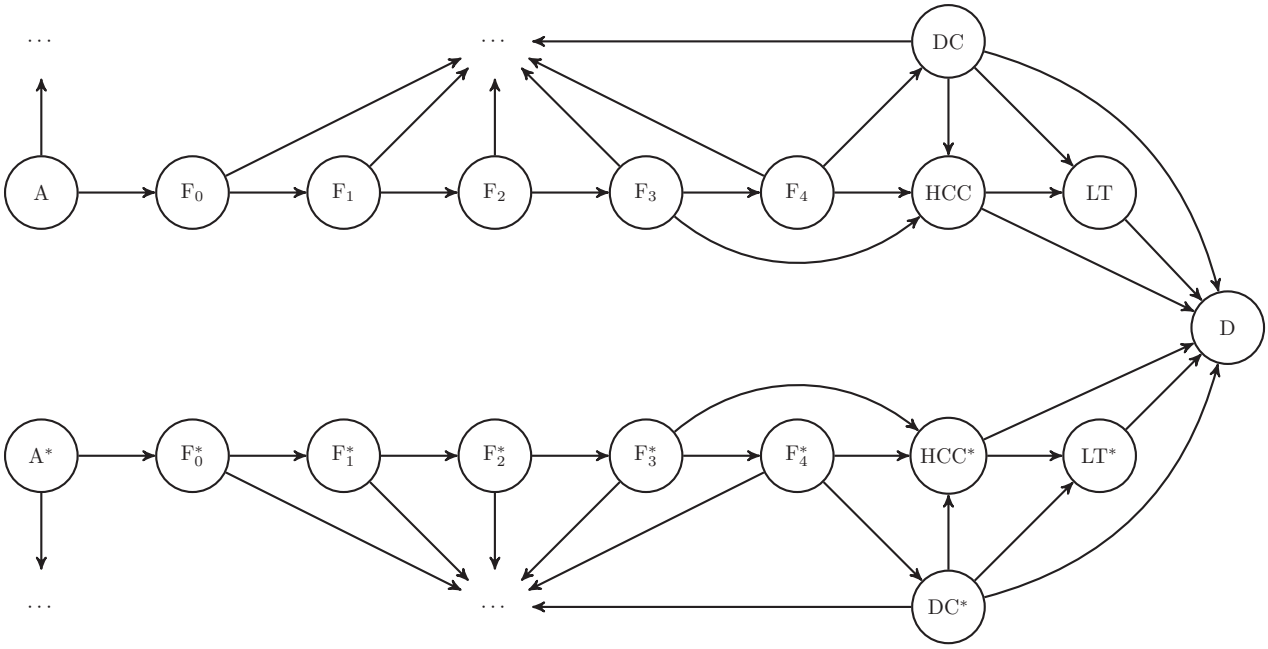


Figure D.1: Alternative model of the patients infected by HCV (Genotype 1).
The different transmission rates do not appear for the sake of clarity. For full details on the transmission rates, one can refer to Figure 4.17. A, acute infection; F, stage of the fibrosis; HCC, hepatocellular carcinoma; DC, decompensated cirrhosis; LT, liver transplant; D, death due to HCV. The subscript i, for $i \in \{0, 1, 2, 3, 4\}$, corresponds to the stage of the fibrosis progression. Starred compartments represent HIV infected individuals.

Développement d'une approche basée sur les modèles dynamiques compartimentaux pour évaluer le bénéfice et l'impact des nouveaux médicaments en population générale : application au cas de l'hépatite C

Auteur : Arnaud Nucit
Directeur de thèse : Jean-Yves Dauxois
Encadrement industriel : Nathalie Schmidely
Date et lieu de soutenance : 16/12/2016 – Conservatoire national des arts et métiers
Discipline : Mathématiques appliquées

Français

Ce travail de thèse s'articule autour de trois parties distinctes abordant chacune un thème précis lié à l'épidémiologie. La première partie de ces travaux s'inscrit dans le cadre de la propagation de virus via l'utilisation de modèles épidémiques. Dans cette partie, sont analysés différentes méthodes d'estimations paramétriques et y sont étudiés la qualité de ces estimateurs. Une application à des virus informatiques est proposée. La deuxième partie de cette thèse propose une méthode d'estimation de la prévalence actuelle du virus de l'hépatite C en France par l'intermédiaire d'un modèle de rétro-calcul associé à un modèle de Markov modélisant l'histoire naturelle de la maladie. Cette méthode et les résultats qui en découlent sont comparés avec les résultats obtenus via l'approche de référence en France. Enfin, la dernière partie s'intéresse à l'étude de l'impact des nouvelles thérapeutiques anti-hépatite C susceptible d'éradiquer le virus à moyen terme. En assimilant la population d'intérêt à un groupement de graphes aléatoires, la propagation du virus est modélisée à partir d'un modèle de métapopulation construit sur la base de données migratoires où les dynamiques de chaque sous-population sont régies par un ensemble d'équations différentielles déterministes. Ce travail doctoral a été réalisé dans le cadre dune convention CIFRE avec les laboratoires Bristol-Myers Squibb.

Mots clé : modélisation, modèles compartimentaux, statistique, hépatite C, modèles dynamiques, épidémiologie.

English

The works undertaken in this doctoral thesis are conducted in three parts, each one dealing with a specific epidemiology-related domain. The first part of this work deals with the propagation of viruses by using well-known epidemic models. It is mainly focused on the analyze of different estimation methods and on their performance. An application on computer virus is proposed. The second part of this thesis gives an estimation method of the hepatitis C virus prevalence in France based on a back-calculation model in association with a Markov model of the disease's natural history. This method and its results are compared with those generated by the reference approach in France. The last part is focused on the study of the recent anti-hepatitis C therapeutics impact on the population since is has been stated that those could eradicate the virus at middle term. In that optic, based on published migration data and assuming that the population of interest is organized into a set of specific contact networks, a metapopulation is computed in which the dynamics of each sub-population is governed by a set of deterministic differential equations. This doctoral research has been conducted through a CIFRE industrial research agreement with the Bristol-Myers Squibb pharmaceutical company.

Keywords: modeling, compartmental models, statistics, hepatitis C, dynamic models, epidemiology.