

Mémoire présenté le :

**pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA
et l'admission à l'Institut des Actuaires**

Par : Elie MADRANGES

Titre Proposition de développements de la modélisation épidémiologique au
service de la quantification actuarielle des risques

Confidentialité : NON OUI (Durée : 1 an 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

*Membre présents du jury de l'Institut
des Actuaires*

David MARIUZZA

signature

Entreprise :

Nom : Optimind

Signature : 

Membres présents du jury de l'ISFA

Pierre RIBEREAU

Directeur de mémoire en entreprise :

Nom : Yann DISERBEAU

Signature : 


Invité :

Nom :


Signature :

**Autorisation de publication et de mise
en ligne sur un site de diffusion de
documents actuariels (après expiration
de l'éventuel délai de confidentialité)**

Signature du responsable entreprise



Signature du candidat





Remerciements

Je tiens à remercier dans un premier temps Christophe EBERLÉ, président fondateur d'Optimind, qui a permis la réalisation de mon alternance au sein de cette entreprise. Merci également aux *partners* Gildas ROBERT et Valérie DEPPE, qui ont participé à un cadre de travail propice à l'élaboration de ce mémoire au sein de la *practice Actuarial & Financial Services*.

Je souhaite adresser un très grand merci et témoigner de toute ma gratitude à mon tuteur Yann DISERBEAU pour son encadrement de qualité, ses enseignements enrichissants, ses encouragements, sa bienveillance constante et son soutien infaillible tout au long de ce mémoire.

Je tiens à remercier l'ensemble des collaborateurs d'Optimind qui m'ont permis de monter en compétences et d'apprendre à leur côté tout au long de mon alternance, et notamment Bisous RAZAFINDRABARY pour sa disponibilité et sa bonne humeur. Je remercie particulièrement Emmanuel BERTELÉ et Hugo DIJOURD pour leurs relectures ainsi que l'ensemble de l'*Expertise Center Santé & Prévoyance* pour leurs conseils avisés.

Je souhaiterais remercier l'ensemble des professeurs de l'ISFA, et notamment Pierre RIBEREAU, pour son accessibilité et ses conseils.

J'adresserais également mes remerciements à ma famille pour leur soutien inconditionnel.

Enfin, je souhaiterais remercier Fanie GARCIA et Jurnerio SYZIU pour les moments passés ensemble, et Clara FORT d'avoir su m'aider et me soutenir mieux que quiconque.

Résumé

Les épidémies, ou pandémies, sont indéniablement, depuis la crise du covid-19, au centre des préoccupations. Leur impact sanitaire et économique les placent sans difficulté parmi les catastrophes naturelles les plus surveillées du *XXI^{ème}* siècle. Dans le domaine de l'assurance, nombreuses sont les conséquences des épidémies sur les activités des acteurs du secteur : chômage, arrêts de travail, perte d'activité, décès, perturbation des marchés financiers, etc.

Toutefois, anticiper les risques associés à de tels phénomènes se révèle à ce jour complexe pour les acteurs du secteur de l'assurance. Les modèles épidémiologiques existants, comme le modèle *SIR* (Susceptibles-Infectés-Rétablis), bien qu'utiles dans certaines situations, se révèlent trop imprécis pour une exploitation actuarielle satisfaisante. De plus, l'anticipation souvent déterministe des trajectoires épidémiques au sein de ces modèles ne permet que l'étude d'un scénario moyen, peu significatif au regard de la volatilité importante des phénomènes sous-jacents.

Dans ce contexte, ce mémoire propose plusieurs améliorations notables du modèle *SIR*. Une refonte probabiliste des dynamiques de transition inter-états permet une modélisation plus fine des trajectoires épidémiques, autorisant ainsi l'exploitation du modèle dans le cadre de prédictions sur le long-terme. La mise en place de méthodes de calibrations adaptées à des problèmes d'optimisation en haute dimension, analogues à celles utilisées pour l'ajustement de réseaux de neurones, permet de garantir une adéquation satisfaisante du modèle par *backtesting* à diverses observations historiques des dynamiques épidémiques. Enfin, l'introduction d'une composante stochastique répliquant la volatilité historique des dynamiques épidémiques observées permet une quantification pertinente des risques épidémiologiques auxquels les institutionnels du secteur de l'assurance sont exposés.

Les méthodologies proposées sont illustrées à travers le cas pratique de l'estimation de la distribution de charge associée à une garantie arrêt de travail pour cause d'infection par le virus de la grippe.

Abstract

Epidemics or pandemics are undeniably, since the covid-19 crisis, at the center of concerns. Their health and economic impact make them one of the most closely watched natural disasters of the 21st century. In the field of insurance, epidemics have many impacts on the sector's activities : unemployment, work stoppages, loss of business, death, disruption of financial markets, etc.

However, anticipating the risks associated with such phenomena has proven to be a complex task for players in the insurance sector. Existing epidemiological models, such as the SIR model (Susceptible-Infected-Recovered), although useful in certain situations, are too imprecise for a satisfactory actuarial exploitation. Moreover, the often deterministic anticipation of epidemic trajectories in these models only allows the study of an average scenario, which is not very significant in view of the significant volatility of epidemic phenomena.

In this context, this thesis proposes several notable improvements to the SIR model. A probabilistic redesign of the inter-state transition dynamics allows a more accurate modeling of epidemic trajectories, thus allowing the model to be used for long-term predictions. The implementation of calibration methods adapted to high-dimensional optimization problems, similar to those used for neural network fitting, allows to guarantee a satisfactory adequacy of the model by backtesting to various historical observations of epidemic dynamics. Finally, the introduction of a stochastic component replicating the historical volatility of the observed epidemic dynamics allows a relevant quantification of the epidemiological risks to which insurance institutions are exposed.

The proposed methodologies are illustrated through the practical case of the estimation of the cost distribution associated with a work stoppage guarantee due to an influenza infection.

Glossaire

ADAM	<i>Adaptive Moment Estimation</i>
ALD	<i>Affectation Longue Durée</i>
CCN	<i>Convention Collective Nationale</i>
IJ	<i>Indemnité journalière</i>
MSE	<i>Mean Squared Error</i>
PIB	<i>Produit Intérieur Brut</i>
SARS	<i>Severe Acute Respiratory Syndrom</i>
SEIR	<i>Susceptibles-Exposés-Infecteds-Rétablis</i>
SIDA	<i>Syndrôme d'Immunodéficience Acquis</i>
SIR	<i>Susceptibles-Infecteds-Rétablis</i>
SIRD	<i>Susceptibles-Infecteds-Rétablis-Décédés</i>
SJB	<i>Salaire Journalier de Base</i>
SMIC	<i>Salaire Minimum de Croissance</i>
SS	<i>Sécurité Sociale</i>
TN-ILMS	<i>Transmission Network for Individual Level Models</i>
TNS	<i>Travailleur Non Salarié</i>
VIH	<i>Virus d'Immunodéficience Humaine</i>

Table des matières

Remerciements	i
Résumé	ii
Abstract	iii
Glossaire	iv
Introduction	ix
Synthèse	xi
Summary	xviii
I Mise en contexte	xxv
1 Caractérisation du risque épidémiologique	1
1.1 Nomenclature et principaux concepts	1
1.1.1 Définition de l'épidémiologie	1
1.1.2 Notions d'endémie, d'épidémie et de pandémie	1
1.2 Impacts historiques des pandémies et épidémies depuis le XX ^{ème} siècle	2
1.3 Focus sur les virus de la grippe	4
1.3.1 Agent pathogène	4
1.3.2 Symptômes et transmission	4
1.3.3 Statistiques en France et dans le monde	4
1.3.4 Mesures de prévention mises en place	5
1.4 Impacts possibles de ce type de catastrophes et enjeux assurantiels afférents	6
1.4.1 Surmortalité engendrée	6
1.4.2 Impacts d'ordres économiques et sociaux	7
1.4.3 Le cas des épidémies de grippe saisonnière	8

2	Introduction au cas pratique : enjeux assurantiels associés aux épidémies de grippe saisonnière sur une garantie « arrêt de travail »	10
2.1	Modalités de la garantie arrêt de travail	10
2.1.1	Modalités pour un salarié	11
2.1.2	Loi de Mensualisation du 19 janvier 1978	11
2.1.3	Modalités pour un indépendant	11
2.1.4	Garanties arrêt de travail	11
II	Vers une évolution du modèle épidémiologique SIR : valeur ajoutée du SIR^+ déterministe	13
3	État de l'art	15
3.1	Le modèle SI	15
3.2	Le modèle SIR	17
3.2.1	Le modèle SIR dans la littérature	17
3.2.2	Une discrétisation temporelle du modèle SIR plus cohérente avec la maille de surveillance de la grippe	18
4	Le modèle SIR^+, amélioration de l'existant SIR	20
4.1	Refonte des dynamiques déterministes de transition inter-état	20
4.1.1	Nécessité de correction des erreurs induites par les modèles « constant-rate »	20
4.1.2	Modélisation plus fine du nombre de nouveaux infectés par pas de temps	21
4.1.3	Modélisation plus fine du nombre de nouveaux rétablis par pas de temps	22
4.2	Données utilisées pour l'ajustement du modèle : historique de l'incidence grippale en France	26
4.2.1	Le réseau Sentinelles	27
4.2.2	Retraitement des données	29
5	Adaptation d'une technique d'optimisation de l'ajustement issue du <i>Machine Learning</i> : <i>Adaptive Moment Estimation</i>	35
5.1	Spécificités de l'algorithme de descente de gradient <i>ADAM</i>	36
5.2	Démonstration de l'existence du gradient de l'écart quadratique	39
5.2.1	Initialisation de la récurrence	40
5.2.2	Dérivabilité de la fonction de répartition mesurant la durée d'infection	41
5.2.3	Finalisation de la récurrence	46
5.3	Approximation du gradient de la <i>MSE</i>	47
5.3.1	Estimation des dérivées partielles par quotient différentiel	47
5.3.2	Introduction d'une métrique sur les quotients différentiels, critère de précision	50

5.4	Implémentation et mise en application de l'algorithme d'ADAM	51
5.4.1	Détermination des paramètres initiaux d'ADAM	52
5.4.2	Détermination de l'hyperparamètre λ , le <i>learning-rate</i> initial	52
5.4.3	Introduction d'un retard dans le début de simulation	53
5.5	Résultats de la calibration du modèle SIR^+	54
III	Construction du modèle SIR^+ stochastique	61
6	Tentative d'obtention d'un modèle stochastique par ajustement d'une loi sur la distribution des paramètres	62
7	Tentative d'obtention d'un modèle stochastique par refonte des dynamiques de transition déterministes	65
7.1	Proposition de refonte de la dynamique d'incidence	65
7.1.1	Rappel de la dynamique déterministe	65
7.1.2	Introduction d'un processus stochastique	66
7.2	Adéquation entre la dynamique déterministe et l'espérance de la dynamique stochastique : incompatibilité avec la structure du modèle	67
7.3	Difficultés à obtenir une relation de récurrence entre les espérances stochastiques successives	69
8	Transformation du déterministe SIR_{det}^+ par le biais d'un mouvement brownien géométrique	71
8.1	Rappel des propriétés des mouvements browniens géométriques	71
8.2	Transformation appliquée à l'incidence déterministe	72
9	Calibration du taux d'erreur du modèle SIR_{sto}^+	75
9.1	Minimisation de l'écart quadratique entre les taux d'erreur théoriques et les taux d'erreur empiriques	75
9.2	Résultats de la calibration et trajectoires stochastiques des épidémies simulées	77
IV	Cas pratique : distribution de provisions en arrêts de travail engendrés par des infections par virus grippal	80
10	Processus de simulation d'épidémies et distribution des provisions	81
10.1	Hypothèses retenues	82
10.1.1	Portefeuille	82
10.1.2	Garantie arrêt de travail	83

10.2	Création du portefeuille fictif	84
10.3	Simulation macroscopique d'une épidémie de grippe	85
10.4	Simulation individuelle : processus de répartition des incidences et des nouveaux rétablis	86
10.5	Calcul de la charge totale des infectés dans le portefeuille	87
10.6	Distributions des durées d'infection et distributions des provisions	89
Conclusion		93

Introduction

La crise du Covid-19 ayant exacerbé les impacts potentiels en termes de surmortalité et de pertes économiques associés aux phénomènes épidémiologiques, l'intérêt pour l'étude de ces derniers a crû de manière importante, bien que la surveillance de ces catastrophes ne soit pas apparu de concert avec ce virus et ses variants. De fait, l'apparition de virus particulièrement virulents et dangereux pour l'Homme s'est fortement accélérée ces dernières décennies. De nombreux infectologues, virologues, médecins et bien d'autres professionnels de la santé ont participé, sous la houlette de la journaliste Marie-Monique Robin, à la rédaction de « La Fabrique des Pandémies », enquête dans laquelle ils affirment que l'Humanité est entrée dans « une ère d'épidémies, de pandémies et de confinements chroniques ». L'apparition de maladies infectieuses, quantifiée à une tous les 15 ans, est désormais de l'ordre d'une à cinq tous les ans.

D'un point de vue assurantiel, la survenance d'une épidémie ou d'une pandémie a des conséquences importantes, d'autant plus si cette catastrophe est précédée d'une impréparation des autorités. Le risque le plus largement relayé est généralement la surmortalité engendrée par les contaminations. Cependant les conséquences économiques peuvent être bien plus préoccupantes, et ne sont parfois visibles que longtemps après l'épisode épidémique : chômage, perte d'activité, arrêts de travail, perturbation des marchés financiers, ralentissement de l'économie, etc.

À cet effet, la modélisation épidémiologique peut venir grandement aider les acteurs du secteur de l'assurance à quantifier ces risques, et à pouvoir plus précisément les appréhender selon leurs propres besoins. Un des modèles les plus classiques en la matière est le modèle *SIR* (*S* pour « Susceptibles », *I* pour « Infectés » et *R* pour « Rétablis »), qui segmente la population en trois compartiments correspondant chacun à un état sanitaire. Cependant ce modèle aboutit vraisemblablement à certaines imprécisions le rendant inadapté à la prévision de dynamiques épidémiques sur le long terme. De plus, étant déterministe, il ne permet pas de prendre en compte la nature aléatoire de ces dernières. Le présent mémoire part donc d'un constat qui est la nécessité d'une amélioration des modèles épidémiologiques.

Dès lors une problématique se dégage de ce constat : comment est-il possible d'améliorer ces modèles

pour faire en sorte qu'ils reflètent de manière plus adéquate les dynamiques épidémiques sous-jacentes, et ainsi rendre pertinent leur utilisation dans le cadre de la quantification actuarielle des risques ?

Dans ce contexte, ce mémoire propose plusieurs améliorations notables du modèle *SIR*. Une refonte probabiliste des dynamiques de transition inter-états permettra une modélisation plus fine des trajectoires épidémiques, autorisant ainsi l'exploitation du modèle dans le cadre de prédictions sur le long-terme. La mise en place de méthodes de calibrations adaptées à des problèmes d'optimisation en haute dimension, analogues à celles utilisées pour l'ajustement de réseaux de neurones, permettra de garantir une adéquation satisfaisante du modèle par *backtesting* à diverses observations historiques des dynamiques épidémiques. Enfin, l'introduction d'une composante stochastique répliquant la volatilité historique des dynamiques épidémiques observées permettra une quantification pertinente des risques épidémiologiques auxquels les institutionnels du secteur de l'assurance sont exposés.

Les méthodologies proposées seront illustrées à travers le cas pratique de l'estimation de la distribution de charge associée à une garantie arrêt de travail pour cause d'infection par le virus de la grippe.

Synthèse

État de l'art et enjeux assurantiels de la modélisation épidémiologique

La modélisation épidémiologique n'a cessé au fil des années de capter l'attention des acteurs du secteur de l'assurance, et cet intérêt connaît, à la lumière des récents événements, une croissance décuplée. Les conséquences qui se dégagent à la suite de pandémies ou d'épidémies sont multiples : perte d'activité, chômage, arrêts de travail, perturbation importante des marchés financiers etc.

De ce constat émerge donc une certaine nécessité : développer des modèles robustes et fiables capables d'aider les assureurs à mieux appréhender le risque que ces phénomènes véhiculent, pour mieux s'en prémunir.

Les modèles épidémiologiques utilisés jusqu'à présent se répartissent en deux catégories : les modèles individuels et les modèles compartimentaux. Les modèles compartimentaux modélisent la population en séparant les individus par « classes », assimilables à des états sanitaires, homogènes. C'est le cas du modèle *SIR*, qui segmente la population en « Susceptibles », « Infectés », « Rétablis ». Ces trois états sont liés entre eux par des équations de passage inter-état qui représentent les mouvements de population ou plus précisément les changements d'état sanitaires dans la population considérée. Elles sont caractérisées par deux paramètres : le premier, β , correspond au nombre moyen de personnes qu'un individu infecté contamine pendant une unité de temps et régit le passage de l'état *S* à l'état *I*, alors que le deuxième, α , représente la proportion d'infectés qui guérissent par unité de temps et est lui fondamental dans le passage de l'état *I* à l'état *R*.

Refonte des dynamiques d'un des modèles déterministes les plus connus : construction du *SIR*⁺

Il apparaît de ses équations que le modèle *SIR* en l'état constitue un modèle dont la simplicité et les hypothèses sur lesquelles il s'appuie, comme par exemple la non prise en compte de la durée passée dans l'état *I* pour ce qu'il s'agit de la transition de l'état *I* à l'état *R*, le rendent sensiblement perfectible. La nécessité d'opérer une révision des dynamiques de ce modèle s'est alors imposée, pour réussir

à mieux appréhender les dynamiques d'entrée et de sortie de chaque état et permettre d'accroître la précision d'un modèle qui semble jusqu'alors trop grossier. En adoptant une approche probabiliste, les équations du nouveau modèle SIR^+ ont été établies avec une structure similaire à celles du SIR , et ont présenté une complexification notable et souhaitable de celui-ci.

Le modèle SIR^+ développé possède alors 3 paramètres : le premier, noté θ , est la probabilité qu'un contact entre un individu susceptible et un individu infecté engendre la contamination de l'individu susceptible sachant qu'un tel contact a eu lieu. En notant m le nombre moyen de contacts qui ont lieu par unité de temps, il ressort que le paramètre β du modèle SIR n'est ni plus ni moins que $m \times \theta$ dans le SIR^+ .

Enfin le deuxième et le troisième paramètre du modèle SIR^+ , notés respectivement a et b , sont les paramètres d'une loi *Gamma* : cette loi représente la durée que passe un individu infecté dans l'état I , et influera sur les sorties de l'état I , et donc sur les entrées dans l'état R .

Introduction du cas pratique illustratif, calibration du SIR^+ et premier résultat

Le modèle SIR^+ ainsi construit semble pallier les défauts du SIR évoqués. Il peut donc être calibré selon les besoins de modélisation. L'utilisation du modèle ainsi que la méthodologie de calibration sont illustrés à travers un cas pratique : l'estimation de la distribution de charge associée à une garantie arrêt de travail pour cause d'infection par le virus de la grippe. Les données utilisées sont des données d'incidences de la grippe en France de 1984 à nos jours, documentées par le réseau Sentinelles. Ces données sont en *open-data* et sont utilisables pour une calibration du modèle SIR^+ . Elles nécessitent cependant un retraitement : la maille d'observation de Sentinelles est la semaine, or la grippe dure en général moins d'une semaine : afin de modéliser le plus finement possible il a été nécessaire de passer d'une maille semaine à une maille jour. Pour ce faire, une répartition uniforme suivie d'un lissage par moyenne mobile a été opéré.

La calibration, qui a pour but d'optimiser les paramètres en minimisant l'écart quadratique entre les données réelles et les prédictions du SIR^+ , présente ici la particularité d'exploiter un algorithme qui trouve sa place plus habituellement pour répondre à des problématiques de *deep-learning* : l'algorithme *ADAM* (*Adaptive Moment Estimation*). Cet algorithme présente plusieurs avantages intéressants dans les travaux de ce mémoire. Le premier est sa vitesse accrue de convergence vers le minimum global tout en évitant les minima locaux, et la seconde est sa capacité à calibrer un modèle muni de très nombreux paramètres. N'étant qu'au nombre de 3 ici, cet aspect de l'algorithme *ADAM* ne présente pas un avantage conséquent, mais son utilisation ouvre la voie à une potentielle complexification du modèle par ajout d'état et/ou de paramètres. Ainsi l'utilisation de l'algorithme *ADAM* permet une

certaine généralisation de cette méthode de calibration sur ce modèle. Classiquement, la fonction de perte minimisée *via* l'algorithme *ADAM* est la *MSE* (*Mean Squared Error*), qui représente l'erreur quadratique moyenne entre les données d'incidence fournies par Sentinelles et les incidences modélisées par le *SIR*⁺. L'algorithme *ADAM* utilise, à chaque itération, le calcul du gradient de la fonction de perte, ici le gradient de l'erreur quadratique c'est à dire ses dérivées partielles selon chacun de ses paramètres. Afin de pouvoir utiliser cet algorithme, il a donc fallu démontrer que ce gradient existe, et donc, faute de documentation sur ce sujet précis, proposer une démonstration intégrale de la dérivabilité de la fonction de perte selon chacun de ses paramètres.

Une fois cette démonstration effectuée, un obstacle demeure : l'impossibilité de déterminer une formule fermée pour les dérivées partielles de la fonction de perte par rapport aux trois paramètres. Il a donc été nécessaire d'approcher numériquement la dérivée par le quotient différentiel (ou taux d'accroissement) de la *MSE* selon ces trois paramètres, quotient différentiel qui dépend d'une précision (aussi appelé le « pas ») choisie au préalable. Le choix de ce pas a été effectué conformément à une métrique introduite permettant d'assurer la pertinence de cette approximation numérique. Une fois la méthodologie d'approximation numérique mise en place, la calibration *via* l'algorithme *ADAM* est possible et permet d'obtenir des paramètres « optimaux ».

Les résultats de la calibration dans le cadre du cas pratique illustratif donnent, pour chaque saison épidémique, un jeu de paramètres θ , a et b optimaux au regard de la *MSE*. Ils se présentent comme suit pour quelques saisons :

Saison	θ optimal	a optimal	b optimal
1996	0,0190	3,6678	0,8880
1998	0,0221	0,7466	4,3034
2008	0,0265	3,8223	0,6407

TABLE 1 – Jeux de paramètres optimaux du modèle *SIR*⁺ pour quelques saisons épidémiques

Les résultats de la calibration peuvent être visualisés en comparaison des courbes d'incidence historiques :

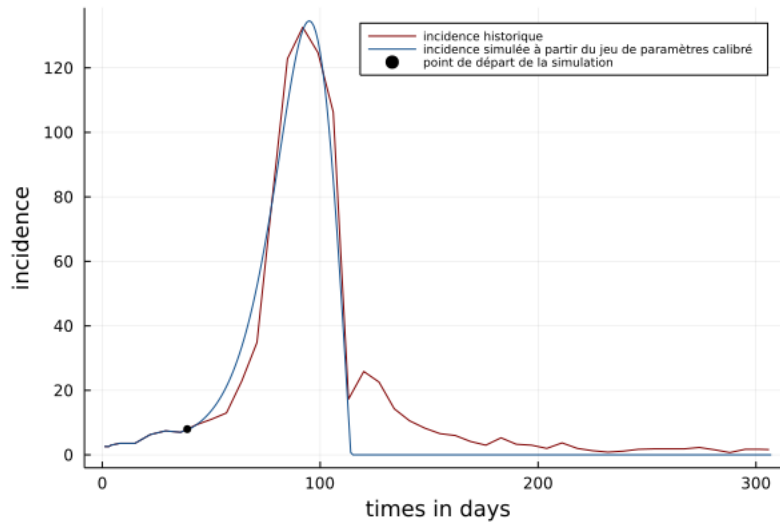


FIGURE 1 – Incidence historique et incidence du SIR^+ pour la saison 2003

Il en ressort que le SIR^+ fournit une modélisation adéquate sur les saisons considérées, après calibration, excepté un léger décalage au niveau de la fin de l'épidémie.

La transformation stochastique pour affiner le modèle déterministe

À ce stade du mémoire, la modélisation déterministe proposée par le SIR^+ représente une base solide pour une modélisation épidémiologique, mais demeure relativement imprécise et trop peu « souple » au regard de certaines caractéristiques fondamentales des épidémies. En effet, une épidémie est par nature aléatoire, et une modélisation déterministe ne permet pas de faire efficacement face à la volatilité très importante de ces phénomènes. Il est donc dès lors nécessaire d'opter pour une modélisation stochastique afin de capter au mieux les effets non pris en compte par le SIR^+ déterministe. Cette transition SIR^+_{det} , déterministe au SIR^+_{sto} , nom donné à la version stochastique du SIR^+ , a vu plusieurs approches explorées.

La première a été d'obtenir une distribution des paramètres du SIR^+ calibrés sur les données d'incidence de Sentinelles. Sur la distribution empirique des paramètres θ calibrés sur chaque saison épidémique *via ADAM* est ajustée une loi de probabilité. Les paramètres a et b , eux, sont ajustés par *moment-matching*. Il est alors possible de tirer aléatoirement trois paramètres pour alimenter le SIR^+ et obtenir une incidence. Cependant cette première approche de construction d'un modèle stochastique n'est pas satisfaisante, car aboutissant à des scénarios bien trop explosifs en comparaison des données

d'incidence réelles, ce qui est probablement lié au faible nombre d'observations à disposition rendant difficile l'estimation de la loi jointe.

La deuxième approche s'est portée sur une modification des dynamiques du modèle, en ajoutant à la dynamique d'incidence un processus stochastique, qui représente alors le « bruit » non pris en compte par le SIR_{det}^+ . Cette voie s'est avérée non concluante car la complexité des équations manipulées ne permettaient aucune adéquation entre le le modèle déterministe et le comportement en espérance des dynamiques stochastiques. Ne répondant pas aux attentes fixées au préalable, cette approche a donc été abandonnée, au profit d'une transformation stochastique plus aisée à mettre en œuvre et présentant les caractéristiques adéquates.

La méthode retenue consiste ainsi en une transformation du modèle déterministe grâce à un mouvement brownien géométrique d'espérance unitaire et de variance à calibrer. Le mouvement brownien géométrique possède deux paramètres : γ et σ , le premier étant nul et le deuxième restant à calibrer pour obtenir les propriétés mentionnées ci-dessus.

Le modèle SIR_{sto}^+ possède deux composantes : une déterministe, incarnée par le modèle SIR_{det}^+ , et une stochastique, représentée par le mouvement brownien géométrique. La composante déterministe, portée par les équations déterministes du SIR_{det}^+ , est obtenue en ne gardant qu'un seul jeu de paramètres parmi ceux obtenus lors de la phase de calibration *via* l'algorithme *ADAM*. Dans le cadre du cas pratique étudié ici, le jeu de paramètres unique retenu pour la suite des travaux est celui minimisant la *MSE* globale, c'est à dire permettant la meilleure performance du modèle toutes saisons confondues. Il s'agit en pratique du jeu de paramètres de la saison 1985, qui se présente comme suit :

Saison	θ^*	a^*	b^*
1985	0,0173	0,1658	28,7689

TABLE 2 – Jeu de paramètres optimaux du modèle SIR_{det}^+ minimisant la *MSE* globale

Le paramètre σ de la composante stochastique, elle, doit désormais être calibrée.

Calibration du SIR_{sto}^+ et cas pratique

La calibration du paramètre σ s'effectue par minimisation de la *MSE* selon une grille de recherche, mais ici en comparant le taux d'erreur théorique, qui est le rapport de l'incidence stochastique et de l'incidence déterministe, et le taux d'erreur empirique observé sur les données de Sentinelles. Le résultat de cette calibration donne une valeur $\sigma^* = 0,02776$. L'obtention de ce dernier paramètre du modèle SIR_{sto}^+ permet désormais de simuler des scénarios épidémiques.

La figure ci-après en présente 1000, accompagnés des enveloppes à 1% et 5% :

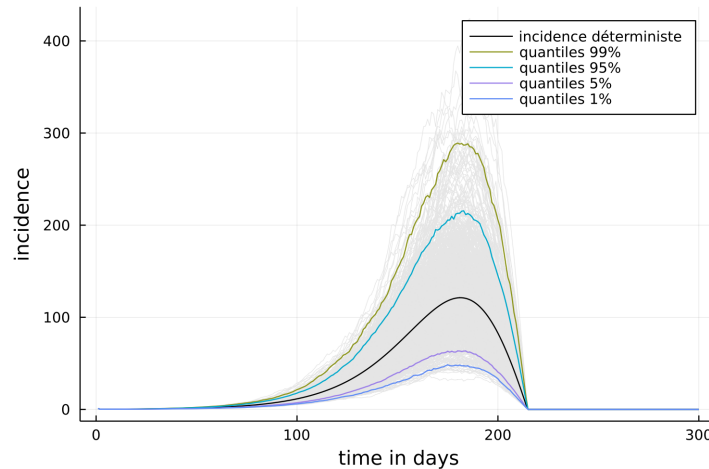


FIGURE 2 – 1 000 scénarios du SIR_{sto}^+ , enveloppes correspondantes et incidence déterministe - population totale de 100 000 individus

Les scénarios ainsi simulés peuvent désormais être utilisés dans le cadre d'un cas pratique qui, dans le cas de ce mémoire, est celui de la quantification des provisions associées à des arrêts de travail consécutifs à des infections par le virus de la grippe.

La garantie arrêt de travail proposée est une garantie à hauteur de 100% du salaire journalier de base (SJB), sans délai de carence. Les 3 premiers jours, l'employeur (ou la complémentaire) prend en charge 100% du SJB du salarié arrêté. À partir du 4^{ème} jour, la Sécurité sociale prend en charge 50% du SJB dans la limite de 1,8 SMIC, et l'employeur (ou la complémentaire) prend en charge le reste pour un maintien de salaire total de 100%.

Un portefeuille fictif d'assurés est créé : chaque individu du portefeuille est considéré comme identique. Aucune distinction n'est faite au niveau du sexe, de l'âge, des comportements etc. À chaque individu est associé un salaire, un numéro d'individu, une durée d'infection (initialisée à 0), une pondération de rétablissement (initialisée elle aussi à 0), ainsi qu'un état sanitaire de départ : S , I , ou R .

Un scénario épidémique est simulé grâce au modèle SIR_{sto}^+ construit précédemment. Les incidences ainsi que le nombre de nouveaux rétablis jours après jours sont arrondis, puis répartis au sein du portefeuille sur les populations correspondantes. L'incidence est distribuée uniformément au sein des individus susceptibles, et les nombre de nouveau rétablis est distribuée uniformément au sein des individus infectés, en tenant compte de la pondération de rétablissement de chacun d'entre eux (plus un individu a passé de temps dans l'état I et plus sa pondération sera élevée). À l'issue de cette

procédure, à chaque individu qui a été infecté lors de l'épidémie est associée une durée d'infection : mise en relation avec le salaire de cet individu, il est possible de déterminer la charge que représente l'arrêt de travail de ce salarié. En agrégeant les charges de l'ensemble des salariés du portefeuille, la charge totale, *i.e* la provision, est obtenue.

En répétant 1000 fois la procédure ci-dessus sur le même portefeuille de salariés, la distribution des provisions associées aux arrêts de travail consécutifs à des infections par le virus de la grippe est représentée par la figure ci-après :

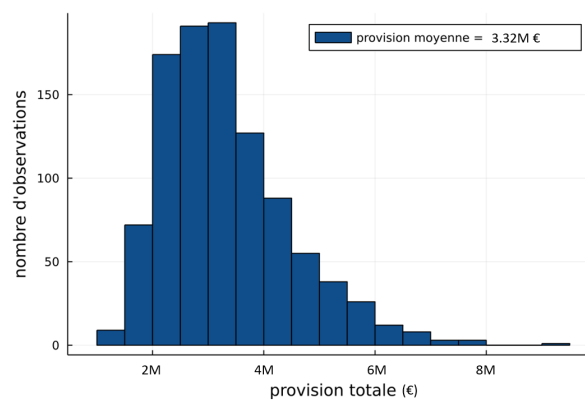


FIGURE 3 – Distribution des provisions maintien de salaire pour des arrêts de travail consécutifs à des infections par le virus de la grippe dans un portefeuille de 100 000 salariés, pour 1 000 scénarios

Summary

State of the art and insurance issues of epidemiological modeling

Over the years, epidemiological modeling has continued to attract the attention of the insurance industry, and this interest has increased tenfold in light of recent events. The consequences of pandemics or epidemics are numerous : loss of business, unemployment, work stoppages, major disruption of financial markets, etc.

From this observation, a certain necessity emerges : to develop robust and reliable models capable of helping insurers to better understand the risk that these phenomena convey, in order to better protect themselves against them.

Epidemiological models usually used fall into two categories : individual models and compartmental models. Compartmental models model the population by separating individuals into homogeneous health states. This is the case of the *SIR* model, which segments the population into « Susceptible », « Infected » and « Recovered » states. These three states are linked together by inter-state transition equations which represent the population movements or more precisely the changes of health state in the considered population. They are characterized by two parameters β corresponds to the average number of people that an infected individual infects during a unit of time and governs the passage from the *S* state to the *I* state, whereas α represents the proportion of infected people who recover per unit of time and is fundamental in the passage from the *I* state to the *R* state.

Recasting the dynamics of one of the best known deterministic models : construction of the *SIR*⁺ model

It appears from its equations that the *SIR* model as it stands is a model whose simplicity and the hypotheses on which it is based, such as not taking into account the time spent in the *I* state for what is the transition from the *I* state to the *R* state, make it significantly perfectible. The need to revise the dynamics of this model was then imposed, to succeed in better apprehending the dynamics of entry and exit of each state and to increase the precision of a model which seems until now too

coarse. By adopting a probabilistic approach, the equations of the new SIR^+ model were established with a structure similar to those of the SIR , and presented a notable and desirable complexification of the latter.

The SIR^+ model then has 3 parameters : the first one, noted θ , is the probability that a contact between a susceptible individual and an infected individual leads to the contamination of the susceptible individual knowing that such a contact has taken place. By noting m the average number of contacts that take place per unit of time, it appears that the parameter β of the SIR model is neither more nor less than $m \times \theta$ in the SIR^+ .

Finally, the second and third parameters of the SIR^+ model, noted respectively a and b , are the parameters of a *Gamma* law : this law represents the duration that an infected individual spends in the I state, and will influence the exits from the I state, and thus the entries into the R state.

Introduction of the illustrative case study, calibration of the SIR^+ and first result

The SIR^+ model thus constructed seems to overcome the shortcomings of the SIR mentioned above. It can therefore be calibrated according to the modeling needs. The use of the model as well as the calibration methodology are illustrated through a practical case : the estimation of the load distribution associated with a work stoppage guarantee due to an influenza infection. The data used are influenza incidence data in France from 1984 to nowadays, documented by Sentinelles network. These are open-data and they can be used for the calibration of the SIR^+ model. However, they need to be reprocessed : the Sentinelles observation grid is weekly, but influenza generally lasts less than a week : in order to model as finely as possible, it was necessary to switch from a weekly grid to a daily grid. To do this, a uniform distribution followed by a smoothing by moving average was performed.

Calibration, which aims at optimizing the parameters by minimizing the quadratic difference between the real data and the predictions of the SIR^+ , has here the particularity of exploiting an algorithm which is more usually used to answer problems of deep-learning : the *ADAM* (*Adaptive Moment Estimation*) algorithm. This algorithm has several interesting advantages in the work of this thesis. The first is its increased speed of convergence to the global minimum while avoiding local minima, and the second is its ability to calibrate a model with a lot of parameters. As there are only 3 parameters here, this aspect of the *ADAM* algorithm does not present a significant advantage, but its use opens the way to a potential added complexity to the model by adding states and/or parameters. Thus the use of the *ADAM* algorithm allows a certain generalization of this calibration method on this model. Classically, the minimized loss function *via* the *ADAM* algorithm is the *MSE* (*Mean Squared Error*), which represents the mean square error between the incidence data provided by

Sentinelles and the incidences modeled by the SIR^+ . At each iteration the *ADAM* algorithm uses the computation of the gradient of the loss function, here the gradient of the squared error, i.e. its partial derivatives according to each of its parameters. In order to be able to use this algorithm, it was necessary to demonstrate that this gradient exists, and thus, due to the lack of documentation on this precise subject, to propose an integral demonstration of the derivability of the loss function according to each of its parameters.

Once this demonstration was made, one obstacle remained : the impossibility of determining a closed formula for the partial derivatives of the loss function with respect to the three parameters. It was therefore necessary to numerically approximate the derivative by the differential quotient (or rate of increase) of the *MSE* according to these three parameters, a differential quotient which depends on a precision (also called the step) chosen beforehand. The choice of this step was made in accordance with a metric introduced to ensure the relevance of this numerical approximation. Once the numerical approximation methodology is in place, the calibration using the *ADAM* algorithm is possible and allows to obtain optimal parameters.

The results of the calibration in the illustrative case study give, for each epidemic season, a set of *theta*, *a* and *b* parameters that are optimal with respect to the *MSE*. They are as follows for some seasons :

Season	optimal θ	optimal a	optimal b
1996	0,0190	3,6678	0,8880
1998	0,0221	0,7466	4,3034
2008	0,0265	3,8223	0,6407

TABLE 3 – Optimal parameter sets of the SIR^+ model for some epidemic seasons

The results of the calibration can be visualized by comparing the historical incidence curves :

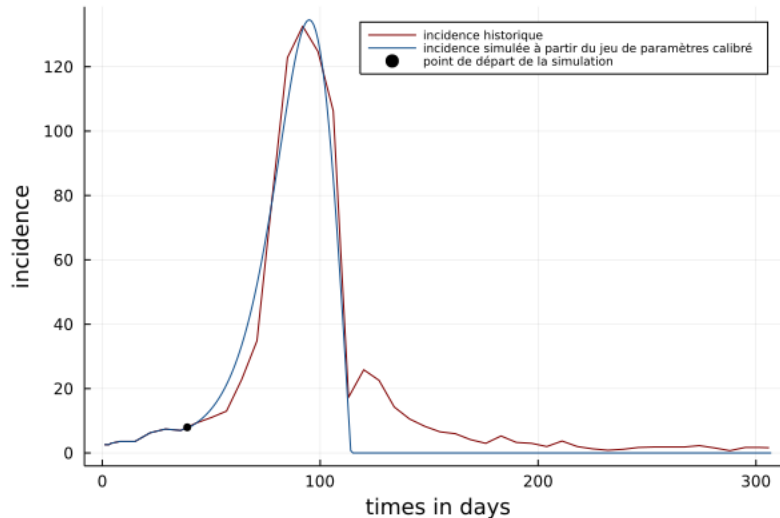


FIGURE 4 – Historical incidence and incidence of SIR^+ for the 2003 season

This shows that the SIR^+ provides an adequate model over the considered seasons, after calibration, except for a slight delay at the end of the epidemic.

The stochastic transformation to refine the deterministic model

At this stage of the dissertation, the deterministic modeling proposed by the SIR^+ represents a solid basis for epidemiological modeling, but remains relatively imprecise and too inflexible with respect to certain fundamental characteristics of epidemics. Indeed, an epidemic is by nature random, and a deterministic modeling does not allow to face efficiently the very important volatility of these phenomena. It is therefore necessary to opt for a stochastic model in order to capture the effects not taken into account by the deterministic SIR^+ . This transition from deterministic SIR^+_{det} to stochastic SIR^+_{sto} has been explored in several ways.

The first was to obtain a distribution of SIR^+ parameters calibrated on Sentinelles incidence data. A probability distribution is fitted to the empirical distribution of the θ parameters calibrated on each epidemic season *via ADAM*. The parameters a and b are adjusted by moment-matching. It is then possible to randomly draw three parameters to feed the SIR^+ and obtain an incidence. However, this first approach of building a stochastic model is not satisfactory, because it leads to scenarios that are too explosive compared to the real incidence data, which is probably linked to the too small number of observations available, making it difficult to estimate the joint distribution.

The second approach was to modify the dynamics of the model by adding a stochastic process to the incidence dynamics, which then represents the noise not taken into account by the SIR_{det}^+ . This approach proved to be inconclusive because the complexity of the equations manipulated did not allow any adequacy between the deterministic model and the behavior in expectation of the stochastic dynamics. As it did not meet the expectations set beforehand, this approach was therefore abandoned, in favor of a stochastic transformation that was easier to implement and had the appropriate characteristics.

The chosen method consists in a transformation of the deterministic model thanks to a geometrical Brownian motion of unit expectation and variance to be calibrated. The geometrical Brownian motion has two parameters : γ and σ , the first one being null and the second one remaining to be calibrated in order to obtain the properties mentioned above.

The SIR_{sto}^+ model has two components : a deterministic one, embodied by the SIR_{det}^+ model, and a stochastic one, represented by the geometric Brownian motion. The deterministic component, carried by the deterministic equations of the SIR_{det}^+ , is obtained by keeping only one set of parameters among those obtained during the calibration phase of the *ADAM* algorithm. In the context of the practical case studied here, the single set of parameters retained for the rest of the work is the one minimizing the global *MSE*, i.e. allowing the best performance of the model all seasons combined. In practice, this is the parameter set of the 1985 season, which is as follows :

Season	θ^*	a^*	b^*
1985	0,0173	0,1658	28,7689

TABLE 4 – Optimal parameter set of the SIR_{det}^+ model minimizing the global *MSE*.

The *sigma* parameter of the stochastic component must now be calibrated.

SIR_{sto}^+ calibration and practical case

The calibration of the *sigma* parameter is done by minimizing the *MSE* according to a search grid, but here by comparing the theoretical error rate, which is the ratio of the stochastic incidence and the deterministic incidence, and the empirical error rate observed on the Sentinelles data. The result of this calibration is a value of $sigma^* = 0.02776$. Obtaining this last parameter of the SIR_{sto}^+ model allows us to simulate epidemic scenarios.

The figure below shows 1000 simulations, with envelopes at 1% and 5% :

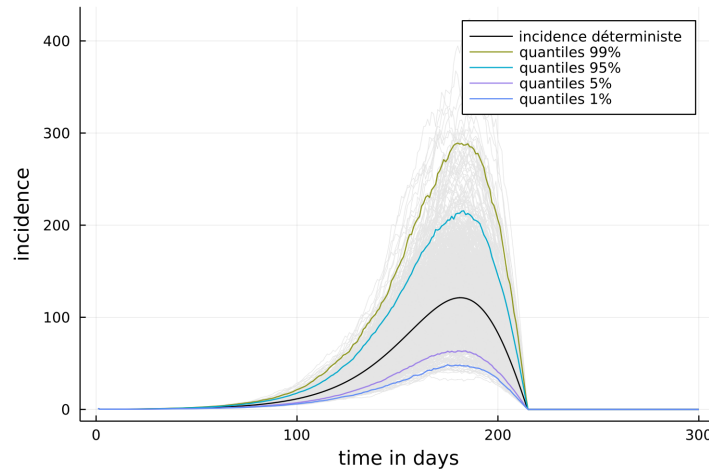


FIGURE 5 – 1000 SIR_{sto}^+ scenarios, corresponding envelopes and deterministic impact - total population of 100 000 individuals

The scenarios simulated in this way can now be used in the context of a practical case which, in the case of this dissertation, is that of quantifying the provisions associated with work stoppages due to infections by the influenza virus.

The proposed work stoppage coverage is a guarantee of 100% of the basic daily wage (BDW), with no waiting period. For the first 3 days, the employer (or the complementary health insurance company) covers 100% of the daily wage of the employee in work stoppage. As of the 4^{textth} day, the social Security covers 50% of the basic daily wage up to 1.8 SMIC, and the employer (or the complementary health insurance) covers the rest for a total salary maintenance rate of 100%.

A fictitious portfolio of insureds is created : each individual in the portfolio is considered identical. No distinction is made in terms of gender, age, behavior, etc. To each individual is associated a salary, an individual number, a duration of infection (initialized to 0), a recovery weight (also initialized to 0), as well as a starting health state : S , I , or R .

An epidemic scenario is simulated using the SIR_{sto}^+ model constructed earlier. Incidence and the number of new recoveries day by day are rounded and then distributed within the portfolio over the corresponding populations. The incidence is distributed uniformly among the susceptible individuals, and the number of new recoveries is distributed uniformly among the infected individuals, taking into account the recovery weight of each of them (the more time an individual has spent in the I state, the higher its weight will be). At the end of this procedure, each individual who was infected during

the epidemic has an associated duration of infection : when related to the salary of this individual, it is possible to determine the burden represented by the work stoppage of this employee. By aggregating the burdens of all the employees in the portfolio, the total expense, i.e. the provision, is obtained.

After repeating the above procedure 1000 times on the same portfolio of employees, the distribution of provisions associated with work stoppages due to influenza virus infections is represented by the figure below :

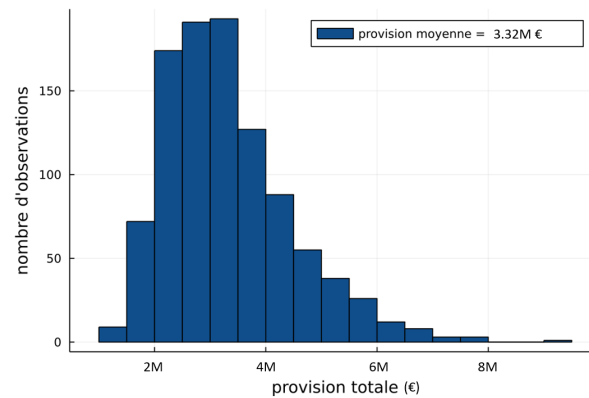


FIGURE 6 – Distribution of salary maintenance provisions for work stoppages due to influenza virus infections in a portfolio of 100 000 employees, for 1 000 scenarios

Première partie

Mise en contexte

Chapitre 1

Caractérisation du risque épidémiologique

1.1 Nomenclature et principaux concepts

1.1.1 Définition de l'épidémiologie

L'épidémiologie est la science qui étudie la fréquence et la répartition des problèmes de santé liés à la propagation d'un agent pathogène dans l'espace et dans le temps au sein d'une population, ainsi que les facteurs qui influent sur cette propagation. Cette notion est donc étroitement liée à la notion d'endémie, d'épidémie et de pandémie.

1.1.2 Notions d'endémie, d'épidémie et de pandémie

Le terme épidémie se rapporte au phénomène soudain d'apparition et de diffusion d'une maladie infectieuse, c'est à dire qui résulte de la multiplication d'une bactérie ou d'un virus dans un organisme vivant, et contagieuse, qui touche une large partie d'une population dans un intervalle de temps relativement court, et sur un espace restreint. Il est dès lors fait une distinction entre épidémie, endémie et pandémie.

Une endémie est la présence habituelle d'une maladie dans une région spécifique, sans expansion spatiale notable. Il n'est pas rare de rencontrer la formulation « maladies endémiques », comme le paludisme par exemple.

Une pandémie fait quant à elle référence à l'apparition et à la propagation d'une maladie infectieuse dans un délai très court, mais à l'échelle mondiale.

Il est à noter qu'aucun de ces trois termes ne véhicule la notion de gravité de la maladie à proprement parler : la sévérité de ces phénomènes est évaluée à travers l'impact qu'ils ont sur les interactions sociales, économiques et démographiques des zones touchées.

Une endémie peut devenir une épidémie si certaines conditions sont réunies (conditions environnementales, comportementales etc). Puis cette épidémie peut se transformer en une pandémie si elle se propage à une proportion plus élevée de la population, ou bien décliner jusqu'à redevenir endémique, ou même disparaître totalement si la maladie est éradiquée dans la zone concernée ou si elle est suffisamment maîtrisée pour devenir quasiment insignifiante.

1.2 Impacts historiques des pandémies et épidémies depuis le XX^{ème} siècle

Quelques pandémies qui ont sévi au cours du XX^{ème} siècle et au début du XXI^{ème} ont, par leur virulence, particulièrement marqué l'Histoire et significativement accentué l'attention portée par les autorités sur ces fléaux.

Grippe espagnole

Une des premières pandémies convenablement documentée de l'ère moderne, et probablement la plus virulente de l'époque contemporaine en termes de décès qui peuvent lui être imputés, est la grippe espagnole qui débuta en 1918 et prit fin en 1921. Cette pandémie tire son appellation du seul pays touché qui publia des bilans de décès et d'infections représentatifs de la réalité, l'Espagne, bien que les premiers cas furent vraisemblablement observés aux États-Unis. Les autres pays européens engagés dans la Première Guerre Mondiale, bien que très lourdement touchés par la grippe espagnole, préférèrent passer sous silence leurs bilans épidémiologiques.

Cette épidémie a décimé en l'espace de 3 ans entre 2,5% et 5% de la population mondiale. La souche, c'est à dire la version initiale du virus dont sont dérivés les variants, de grippe responsable de cette épidémie était la souche A (H1N1). « H1N1 » fait référence aux antigènes caractéristiques présents à la surface du virus. Ces notions propres à l'épidémiologie ne seront pas développées dans le cadre de ce mémoire. Des agents pathogènes de la sorte sont encore aujourd'hui retrouvés lors d'épidémies d'une moindre ampleur, le virus ayant évolué au cours de ces 100 dernières années. C'est notamment le cas lors de plusieurs épidémies de grippe, comme par exemple la grippe saisonnière, pendant lesquelles des

virus possédant cette structure H1N1 ont été détectés.

La particularité de la grippe espagnole ne réside pas tant dans sa contagiosité, qui n'est pas particulièrement remarquable, que dans la répartition de la gravité selon l'âge. En effet, les individus âgés de 20 à 40 ans représentaient 50% des décès, avec un pic de mortalité aux alentours de 30 ans, ce qui constitue une anomalie étant donné que cette tranche d'âge est généralement la moins sensible aux maladies.

Grippe asiatique

La grippe dite « asiatique » fait son apparition en Chine en 1956 et se répand rapidement à tous les continents, pour prendre fin en 1958. Une des raisons de la progression rapide de la pandémie fut vraisemblablement l'absence d'immunité des personnes de moins de 70 ans. La souche virale était cette fois A (H2N2), de type « grippe aviaire », souche déjà présente en Russie dès 1889, et potentiellement antérieurement. Le manque de données ne permet pas aux historiens de dater précisément l'origine de cette souche.

Grippe de Hong-Kong

L'épisode pandémique faisant suite à la grippe asiatique est la grippe de Hong-Kong, qui se propage en Asie durant l'année 1968, et s'étend au continent américain *via* les États-Unis à l'hiver 1968-1969. La grippe de Hong-Kong est considérée en 2003 par l'épidémiologiste *Antoine Flahault* comme « la première pandémie de l'ère moderne. Celle des transports aériens rapide [...] elle est la base de tous les travaux de modélisation visant à prédire le calendrier de la future pandémie. »

Pandémie de Covid-19

Enfin, peut être évoquée la pandémie de Covid-19, qui serait apparue à la fin de l'année 2019 à Wuhan en Chine, et qui s'est développée rapidement sur l'ensemble du globe. Cette pandémie se caractérise par des conséquences économiques particulièrement marquées, et une mise en place de mesures, tant de protection que de soins, difficile due à la mutation rapide de l'agent pathogène, entravant la recherche dans le domaine.

1.3 Focus sur les virus de la grippe

Le présent mémoire s'articulera autour de l'ajustement concret du modèle épidémiologique proposé par la suite à des données d'incidence grippale en France, comme il en sera fait mention ultérieurement.

1.3.1 Agent pathogène

La grippe saisonnière est une épidémie qui sévit chaque année en France, d'une ampleur variable d'une saison épidémique à une autre, et qui fait partie des maladies les plus surveillées dans l'Hexagone. Elle consiste en une infection virale causée par un virus grippal. 3 types de virus grippaux ont été identifiés : A, B et C.

Les souches A et B sont les plus répandues parmi les épidémies saisonnières, la souche C étant plus sporadique. La souche A est la seule des trois qui peut être à l'origine de pandémies.

1.3.2 Symptômes et transmission

La grippe saisonnière se caractérise par une apparition soudaine de fièvre, maux de tête, maux de gorge, écoulements nasaux et toux, cette dernière pouvant durer jusqu'à 2 semaines après l'infection, voire plus. La plupart des personnes contractant la grippe se rétablissent sans préoccupation particulière, mais chez les personnes à haut risque, elle peut provoquer une maladie grave et aller jusqu'au décès. Cette population de personnes à haut risque comprend notamment les personnes âgées de plus de 65 ans, les personnes souffrant d'asthme, d'une maladie chronique (statut ALD, Affectation de Longue Durée), du VIH (Virus d'Immunodéficience Humaine) etc. Elle se propage facilement par gouttelettes dans l'air, ou par le dépôt sur une surface (poignée de porte, mains etc.) après une toux ou un éternuement d'une personne infectée. La maladie dure en général de 3 à 7 jours, et peut atteindre jusqu'à deux semaines selon les cas. La moyenne se situe aux alentours de 4 jours. Cette durée d'infection fera l'objet d'une attention toute particulière lors de la construction du modèle épidémiologique.

1.3.3 Statistiques en France et dans le monde

Les périodes épidémiques varient selon les climats : dans les régions tropicales elles peuvent émerger tout au long de l'année, tandis que dans les climats tempérés elles sont généralement concentrées en hiver. Le virus de la grippe est responsable du développement de forme grave (notamment des surinfections) chez 5 millions de personnes et de la mort de 290 000 à 650 000 d'entre elles chaque année dans le monde. Concernant la France, le bilan annuel est de l'ordre de 2 à 7 millions d'infectés par an, avec un nombre de décès situé aux alentours de 10 000. Les personnes âgées représentent la majeure

partie des décès (environ 90% des décès concernent des individus âgés de plus de 65 ans).

1.3.4 Mesures de prévention mises en place

Des mesures de préventions sont entreprises par les pouvoirs publics afin d'éviter ou de retarder le début de l'épidémie, ou de limiter la propagation du virus et son impact si l'épidémie a déjà débuté.

À l'échelle individuelle, il est recommandé à un individu infecté de limiter ses contacts avec d'autres personnes et en particulier celles à risques (femmes enceintes, enfants en bas âge, personnes porteuses de certaines maladies chroniques, personnes âgées), d'utiliser un masque jetable lors de rencontre avec les types de personnes susmentionnés, de tousser et éternuer dans son coude, d'utiliser des mouchoirs à usage unique, et enfin de se laver les mains régulièrement au savon ou à défaut avec du gel hydroalcoolique.

À l'échelle collective, il existe depuis le XX^{ème} siècle un vaccin contre le virus de la grippe, composé de deux virus de souche A (généralement une souche H1N1 et une autre H3N2) et d'un virus de souche B, tous inactivés chimiquement. Ce vaccin est commercialisé en France par plusieurs laboratoires pharmaceutiques. Lors de la saison épidémique 2018/2019, 30% des personnes de moins de 65 ans et 51% des plus de 65 ans se sont vus administrer le vaccin, selon une étude de Santé publique France. Il prévient, chez les individus âgés de 18 à 64 ans n'étant pas à risque, environ 6 infections sur 10 en moyenne.

Lors de la récente pandémie de Covid-19, de nombreuses mesures et préconisations gouvernementales, respectées par la très grande majorité de la population française, ont permis de freiner la progression de cette épidémie, mais aussi en parallèle d'enregistrer en 2020 et 2021 des baisses significatives des cas de grippe saisonnière. Les confinements, couvre-feux, obligations de port du masque, attentions portées à l'hygiène et la désinfection régulière des mains et objets du quotidien, respects des distanciations sociales, restrictions de déplacement etc. ont permis en effet d'assister depuis le début de la pandémie à un net recul de la maladie. À titre d'exemple, en décembre 2020, les urgences et SOS médecin avaient enregistré une baisse du nombre d'interventions pour des cas de grippe de respectivement 79% et 83% par rapport à 2019.

Toutes ces mesures destinées à endiguer la pandémie de Covid-19 ont rendu la circulation du virus de la grippe dérisoire, si bien que les spécialistes ont d'abord pensé à une possible disparition, avant sa réapparition confirmée à l'hiver 2021. Cette soudaine chute dans les cas de grippe rendent ainsi toute analyse de cette épidémie biaisée si les années 2020 et 2021 font partie des observations.

1.4 Impacts possibles de ce type de catastrophes et enjeux assurantiels afférents

1.4.1 Surmortalité engendrée

Selon la maladie considérée, l'impact en termes de surmortalité induit par les symptômes de cette dernière peut devenir si élevé qu'il devient alors la préoccupation première lors de la surveillance de l'évolution de l'épidémie. Il est cependant à noter que les termes de mortalité et de surmortalité, bien qu'étroitement liées, ne véhiculent pas les mêmes notions. La mortalité désigne le nombre de personnes décédant pour une même raison (*i.e* de la même maladie dans le cas de la grippe), alors que la surmortalité correspond à un taux de mortalité anormalement élevé, comparé à un taux de référence. Cette dernière s'accompagne ainsi d'une diminution de l'espérance de vie. Dans le cadre de la grippe, la surmortalité engendrée est une augmentation du taux de létalité de la maladie sur la saison considérée comparé aux saisons précédentes, toutes choses égales par ailleurs. Il est donc primordial de capter correctement quels décès sont imputables à la maladie, en excluant ceux qui auraient eu lieu même en l'absence de l'agent pathogène dans la population. Cette tâche est complexifiée par le fait que la mortalité au sein d'une population est très variable selon les années, comme l'indique la figure présentée ci-après.

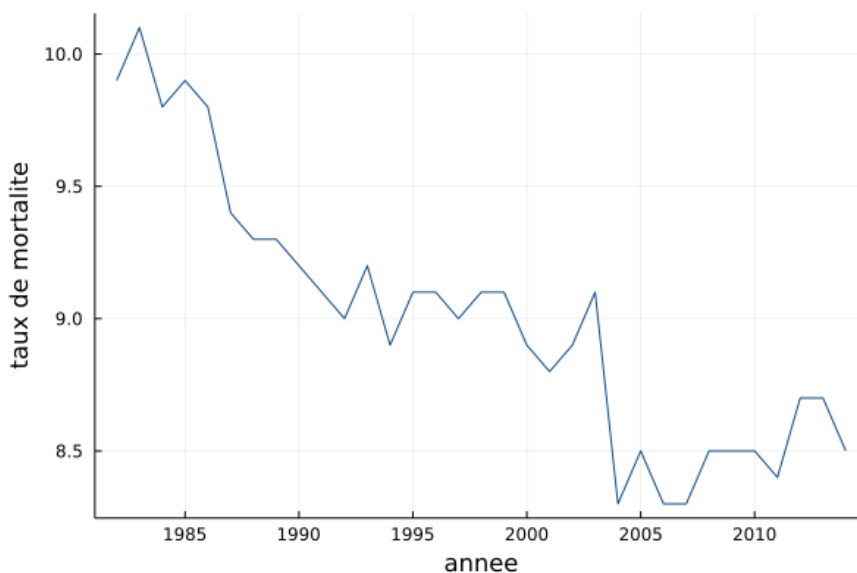


FIGURE 1.1 – Taux de mortalité toute causes confondues pour 1 000 habitants en France, de 1982 à 2014

En termes de mortalité, la grippe saisonnière est non significative si elle est comparée aux grandes pandémies du $XX^{\text{ème}}$ siècle précédemment détaillées :

Épidémie/pandémie	Période	Nombre d'infections	Nombre de décès
Grippe espagnole	1918-1921	500M	20M ~ 50M, jusqu'à 100M (réévaluation 2020)
Grippe asiatique	1956-1958	250M ~ 1Md	≈ 1,2M
Grippe de Hong-Kong	1968-1970	250M ~ 1Md	1M ~ 4M
Sida	1980-	56M ~ 110M	30M ~ 45M
Pandémie Covid-19	2019-	≈ 470M	≈ 6M (mars 2022)
Grippe saisonnière	Chaque année	2M ~ 5M*	290K ~ 650K

TABLE 1.1 – Bilans de quelques pandémies mondiales

*Uniquement les formes graves

Les chiffres avancés dans ce tableau sont à prendre avec précaution : leur marge d'erreur étant assez élevée (à l'image de la réévaluation de 2020 faisant état de 100M de décès causés par la grippe espagnole), les chiffres peuvent sensiblement différer d'une étude à l'autre. Le manque de documentation pour les pandémies les plus lointaines temporellement, ou le fait que certaines soient encore en cours (pandémie Covid-19) rendent ces estimations difficiles.

1.4.2 Impacts d'ordres économiques et sociaux

La surmortalité est très certainement un des aspects les plus médiatisés et marquants à l'issue d'une épidémie ou pandémie. Cependant il existe d'autres conséquences qui accompagnent très souvent ces catastrophes sanitaires, comme la récession économique souvent incarnée par une baisse du *PIB* (Produit Intérieur Brut) mondial :

Épidémie/pandémie	Période	Perte estimée (en % du PIB mondial)
Grippe espagnole	1918-1921	5%
Grippe asiatique	1956-1958	2%
Grippe de Hong-Kong	1968-1970	0,7%
Covid-19	2019 -	4,2% (2020)

Ces pertes financières, en partie caractérisées par la baisse du *PIB* viennent s'ajouter aux pertes humaines, et sont principalement le fruit des mesures prises afin d'endiguer l'épidémie ou la pandémie, comme le précisent *Andrew Burns, Dominique van der Mensbrughe et Hans Timmer* dans leur article « Evaluating the Economic Consequences of Avian Influenza ».

Une partie de l'étude fait le parallèle avec les conséquences possibles de la survenance d'une pandémie de l'ampleur de la grippe espagnole de 1918, dont les estimations sont présentées dans la figure ci-dessous :

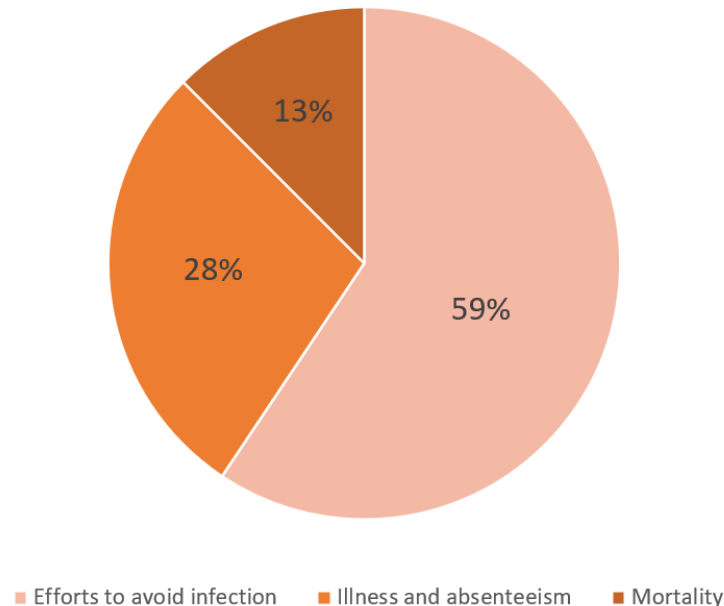


FIGURE 1.2 – Répartition estimée des causes de baisse du PIB consécutive à un épisode pandémique de type grippe espagnole

Il se dégage de cette figure que près des 2/3 de la baisse du PIB sont supposés être imputables aux efforts fournis pour limiter la propagation du virus, alors que la mortalité ne pourrait représenter « que » 13% de l'impact économique.

1.4.3 Le cas des épidémies de grippe saisonnière

La grippe saisonnière s'inscrit dans la lignée des épidémies ou pandémies ayant déjà frappé l'Humanité, bien que chaque virus soit différent et que les conséquences de ces épisodes réguliers n'atteignent pas les bilans des grandes pandémies les plus meurtrières. Non seulement les épidémies de grippe en France sont, comme cela a déjà été évoqué, responsables de milliers voire de dizaines de milliers de morts chaque année mais, en outre, elles représentent irrémédiablement un poids économique pour la société française. Mélanie Gomez, journaliste santé à Europe 1, estime que l'addition s'élève à environ 1 milliard d'euros pour l'épidémie de grippe saisonnière de 2015, soit 0,05% du PIB de la France. Cette somme comprend 300 millions d'euros de dépenses médicales directes, les 700 millions restant étant utilisés en tant que compensations de revenus, heures supplémentaires, remplacements des absents etc. En effet, l'Assurance Maladie aurait versé 300 millions au titre des maintiens de salaire, et les entreprises auraient vu le coût de l'épidémie se chiffrer à 400 millions au titre des conséquences des arrêts de travail consécutifs aux infections par le virus de la grippe. Cette épidémie qui revient année après

année représente une très importante perte financière, d'où la nécessité de mieux appréhender ce phénomène afin de mieux maîtriser les risques qu'il représente.

Chapitre 2

Introduction au cas pratique : enjeux assurantiels associés aux épidémies de grippe saisonnière sur une garantie « arrêt de travail »

2.1 Modalités de la garantie arrêt de travail

La garantie arrêt de travail est une garantie intervenant à la suite d'un évènement, dans la vie d'un individu, l'entravant dans l'accomplissement des tâches inhérentes à son emploi. Il peut s'agir de la survenance d'un accident, ou de la contraction d'une maladie. L'activité professionnelle ainsi mise en suspens, une perte de revenus vient généralement impacter la personne sinistrée. La garantie arrêt de travail vient alors compenser tout ou partie desdits revenus. Elle peut être mise en place *via* un contrat de prévoyance collective, ou bien être souscrite à titre individuel.

Il existe plusieurs organismes qui interviennent relativement au versement de ce remplacement de salaire : la sécurité sociale et l'employeur (ou un organisme complémentaire si l'employeur a souscrit un contrat auprès de celui-ci).

La Sécurité sociale applique un délai de carence, égal à trois jours, après lesquels le substitut de revenu est versé au sinistré. Cependant trois conditions permettent la suppression de ce délai de carence et en conséquence le versement de l'indemnité dès le début de l'incapacité. Ces trois conditions sont :

- reprise d'activité entre deux arrêts si la durée de la reprise est inférieure à 48 heures
 - arrêts de travail successifs dus à une ALD (Affectation Longue Durée)
 - prise en charge par l'employeur des 3 jours de carence stipulée explicitement dans l'accord d'entreprise ou de branche, si l'assuré est un salarié. Dans ce cas l'employeur (ou la complémentaire) prend en charge les 3 premiers jours d'arrêts, et à partir du 4^{ème} jour la Sécurité sociale et l'employeur se répartissent
-

les prises en charge

2.1.1 Modalités pour un salarié

Si l'assuré est salarié d'une entreprise, la Sécurité sociale lui verse une IJ (Indemnité Journalière) après le délai de carence, sauf cas de figure cité précédemment, à hauteur de 50% de son salaire journalier de base (SJB). Ce salaire est égal à la somme des trois derniers mois de salaire précédant l'arrêt et divisée par 91,25. Ce SJB est cependant plafonné à 1,8 fois le montant du SMIC (Salaire Minimum de Croissance, le salaire minimal en France) en vigueur au moment de l'arrêt. En considérant le SMIC en vigueur en 2022, l'IJ est plafonnée à $50\% \times \frac{1,8 \times 1\,603 \times 3}{91,25} = 50\% \times \frac{2\,885 \times 3}{91,25} = 47,43\text{€}$. Si l'assuré n'a pas exercé d'activité continue, le salaire brut de base est la somme des salaires des 12 derniers mois divisée par 365.

2.1.2 Loi de Mensualisation du 19 janvier 1978

Depuis le 19 janvier 1978, la Loi de Mensualisation vient davantage protéger financièrement les salariés en arrêt de travail. En effet, elle impose aux entreprises de prendre en charge une partie du salaire de l'employé touché par un accident ou une maladie. Cependant certaines catégories de salariés ne sont pas concernées par cette obligation : il s'agit des saisonniers, intermittents, travailleurs à domicile et intérimaires.

Si le salarié dispose d'au moins 1 an d'ancienneté dans l'entreprise, l'employeur a l'obligation de verser un complément de revenus à partir du 8^{ème} jour d'arrêt. Cette somme additionnée à l'IJ versée par la Sécurité sociale représente ainsi 90% du salaire brut de base, puis 66,6% du 38^{ème} jour au 68^{ème} jour. Chaque période est prolongée de 10 jours par tranche de 5 ans d'ancienneté, à partir de la 6^{ème} année d'ancienneté. À titre d'exemple, si un salarié a entre 11 et 15 ans d'ancienneté, il sera indemnisé à 90% puis aux deux tiers du salaire brut durant 100 jours au total, sauf convention collective plus favorable. Ces durées de maintien sont limitées à 90 jours chacune, soit 180 jours au total.

2.1.3 Modalités pour un indépendant

Si le travailleur est indépendant (TNS pour Travailleur Non Salarié), il perçoit également des IJ de la part de la Sécurité Sociale. Cependant à la différence du travailleur salarié, un TNS ne dispose pas d'une suppression du délai de carence dans le cas d'un accident du travail ou d'une maladie professionnelle.

2.1.4 Garanties arrêt de travail

La garantie arrêt de travail, il en a été fait mention précédemment, se décompose en une part prise en charge par la Sécurité sociale, une part prise en charge par l'entreprise et / ou une part prise en charge par la complémentaire. Le délai de carence de la Sécurité sociale est immuable ; il est fixé à 3 jours quelles que soient les garanties proposées par les autres intervenants quant au maintien de salaire.

Le schéma suivant fournit un exemple de la répartition des prises en charge par les différents acteurs de la garantie arrêt de travail :

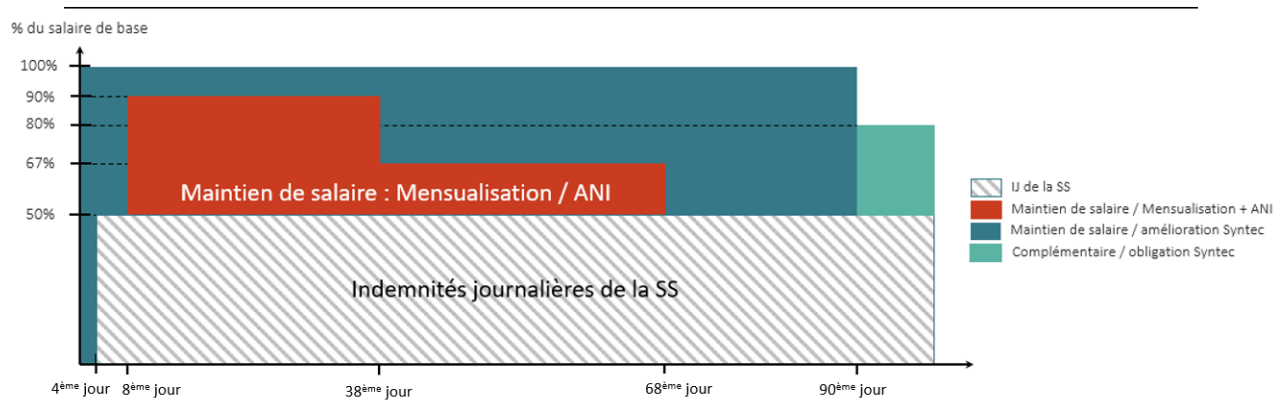


FIGURE 2.1 – Répartition de la prise en charge par la SS, l'entreprise, la CCN et la complémentaire

Cet exemple sera simplifié dans la suite, selon des hypothèses faites sur le portefeuille assuré. À noter que salarié, à titre individuel, ou employeur, à titre collectif, peuvent souscrire des contrats d'assurance qui permettent un maintien de salaire jusqu'à 100%, indemnités journalières de la Sécurité sociale incluses.

Deuxième partie

Vers une évolution du modèle épidémiologique SIR : valeur ajoutée du SIR^+ déterministe

La problématique de modélisation de la propagation tant temporelle que spatiale des épidémies a toujours été au cœur des recherches en épidémiologie. Il semblerait que le premier modèle en la matière, qui ne sera pas explicité dans le cadre de ce mémoire, ait été proposé par Daniel Bernoulli en 1760, en se basant sur une table de population mise au point par Edmond Halley en 1693. Ce modèle avait pour but de démontrer la pertinence de l'inoculation, à un âge précoce, d'une forme supposée bénigne de la variole (anciennement appelée « petite vérole »), afin de provoquer une immunité chez cet individu malade, à la manière du principe de la vaccination.

Il existe plusieurs types de modèles épidémiologiques, qui pourraient être regroupés en deux catégories :

- Les modèles compartimentaux
- Les modèles individus-centrés, ou TN-ILMs pour « Transmission Network for Individual Level Models »

Les modèles compartimentaux classiques sont construits autour de la théorie des chaînes de Markov, et constituent donc des modèles déterministes relativement simples dans leurs hypothèses : chaque individu du modèle possède les mêmes caractéristiques et réagit de la même manière face à la maladie (probabilité d'être infecté, durée de la maladie, nombre de contact avec les autres individus etc).

À l'inverse, les modèles à l'échelle de l'individu permettent de les distinguer et de les étudier au cas par cas, en leur attribuant des caractéristiques distinctes : âge, sexe, comportement face à l'épidémie etc. Ce deuxième type de modèle n'est donc plus déterministe, mais stochastique, car chaque individu aura une "trajectoire" au cours de l'épidémie qui lui sera propre. Cette façon de modéliser a l'avantage notamment de permettre l'étude spatiale de la maladie, et de rendre compte des déplacements de population et de l'impact de ces mouvements sur la diffusion de l'épidémie.

Le choix du type de modèle se porte ici sur l'approche déterministe compartimentale, davantage pertinente en vue d'une évaluation des risques assurables. Les modèles individuels font généralement de nombreuses hypothèses qui peuvent s'avérer complexes à prendre en compte quant au comportement de la population sous-jacente. Enfin, la multitude de paramètres qui entrent en jeu dans ces modèles rend leur calibration extrêmement difficile à mettre en œuvre. Toutes ces raisons dirigent naturellement cette étude actuarielle des épidémies vers une modélisation compartimentale.

Les modèles compartimentaux sont fondés sur les notions de compartiments et de règles : la population d'étude est séparée en « classes ». Chaque individu de la population totale ne peut appartenir qu'à un des compartiments du modèle, qui s'interprète comme un état sanitaire dans lequel se trouve l'individu. Enfin, le passage d'une classe à une autre est régi par des règles, *i.e* des relations de passage d'un état à un autre. Dans la suite de ce mémoire, seront utilisés sans distinction sémantique « classe » ou « état » comme synonymes du mot « compartiment » pour ce type de modèle.

Les présents travaux se réduiront uniquement aux modèles compartimentaux faisant l'hypothèse que la population est « fermée » : ni naissance, ni décès autres que provoqués par la maladie, ni porosité entre la population d'étude et d'éventuelles populations environnantes (tourisme, immigration etc.).

Ces hypothèses seront rappelées et développées ultérieurement.

Chapitre 3

État de l'art

3.1 Le modèle SI

Le modèle dit « SI », développé par Sir William Heaton Hammer en 1906, est composé de deux compartiments, S et I .

- S correspond à l'état « susceptible » : il regroupe tous les individus sains qui sont susceptibles de contracter la maladie étudiée
- I est lui l'état des infectés : dès lors qu'une personne devient malade, elle passe de l'état S à l'état I

Les états du modèle ainsi que leur dépendance, incarnée par le facteur β , peuvent être schématisés comme suit :

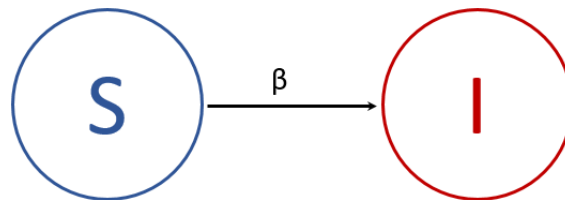


FIGURE 3.1 – Schéma des états et de la transition dans le modèle SI

Sont introduits $S(t)$ qui représente le nombre de personnes susceptibles d'être infectées à l'instant t , et $I(t)$ le nombre de personnes infectées à l'instant t .

Ces deux états S et I sont liés par une relation de passage, qui traduit la transition unidirectionnelle des personnes nouvellement infectées de S à I .

Cette relation de passage, également appelée dynamique de ces états est mathématiquement exprimée, dans le cadre d'une modélisation continue, sous la forme du système d'équations différentielles suivant :

$$\begin{cases} \frac{dS(t)}{dt} = -\beta \frac{S(t)I(t)}{N} \\ \frac{dI(t)}{dt} = \beta \frac{S(t)I(t)}{N} \end{cases}, \text{ avec :} \quad (3.1)$$

$$\begin{cases} S(0) = S_0 \\ I(0) = I_0 \\ S_0 + I_0 = N \end{cases} \quad (3.2)$$

Dans ces équations, $\frac{I(t)}{N}$ est la probabilité de rencontrer un individu infecté, donc $S(t) \times \frac{I(t)}{N}$ représente le nombre de contact entre l'ensemble des susceptibles et l'ensemble des infectés. Enfin, β représente le nombre moyen de personne qu'un individu infecté contamine pendant une durée dt .

Ce modèle fait l'hypothèse que la population est fermée (pas de naissance, décès, immigration etc.). Ceci implique :

$$\forall t \geq 0, \frac{dS(t)}{dt} + \frac{dI(t)}{dt} = 0 \quad , \text{ d'où :}$$

$$\forall t \geq 0, S(t) + I(t) = S_0 + I_0 = N \quad (3.3)$$

Les dynamiques de ce modèle, de par leur simplicité, permettent la résolution analytique (non détaillée ici) de ce système d'équations différentielles, et donc l'obtention d'une expression explicite de $S(t)$ et $I(t)$ qui se trouve être :

$$\forall t \geq 0, \begin{cases} S(t) = \frac{N}{1 + \frac{I(0)}{N-I(0)} e^{\beta N t}} \\ I(t) = \frac{N}{\frac{N-I(0)}{I(0)} e^{-\beta N t} + 1} \end{cases} \quad (3.4)$$

3.2 Le modèle SIR

3.2.1 Le modèle SIR dans la littérature

Le modèle *SIR* est un modèle compartimental à état, composé de trois états *S*, *I* et *R*.

- L'état *S* correspond à l'ensemble des individus qui n'ont jamais infecté, c'est à dire susceptibles d'être infecté
- L'état *I*, lui, regroupe les individus infectés
- L'état *R* est composé des individus qui ont été infectés par le passé et qui ont guéri de la maladie. Il est ici à noter que l'état *R* regroupe en réalité les individus guéris ou décédés, sans distinction. Ainsi par la suite il sera fait mention de « rétablissement » pour désigner à la fois les guérisons et les décès (*i.e* les sorties de l'état *I* au sens large)

Le schéma du modèle *SIR*, de la même forme que celui du modèle *SI*, est généralement présenté de la façon suivante :

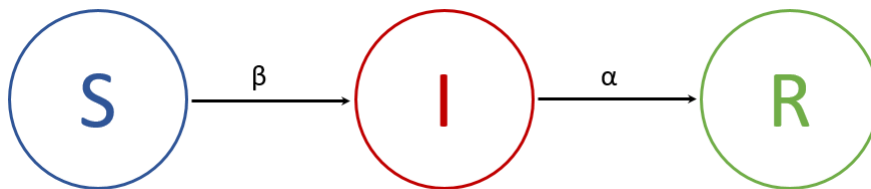


FIGURE 3.2 – Schéma des états et des transitions dans le modèle *SIR*

Dès lors qu'un individu susceptible est infecté, il passe de l'état *S* à l'état *I*, et y reste jusqu'à être guéri. Ainsi dès qu'il guérit, il quitte l'état *I* et rejoint l'état *R*.

Il est dès lors possible d'établir des relations de passage qui lient les 3 états de ce modèle, qui s'expriment mathématiquement par le système de trois équations différentielles suivant :

$$\begin{cases} \frac{dS(t)}{dt} = -\beta \times \frac{S(t)I(t)}{N} \\ \frac{dI(t)}{dt} = \beta \times \frac{S(t)I(t)}{N} - \alpha I(t) \\ \frac{dR(t)}{dt} = \alpha \times I(t) \end{cases} \quad (3.5)$$

Avec :

- *N* le nombre total d'individus considérés
- *S(t)* le nombre d'individus sains à l'instant *t*
- *I(t)* le nombre d'individus infectés à l'instant *t*
- *R(t)* le nombre d'individus rétablis à l'instant *t*
- β et α deux constantes, détaillées ci-après

Le paramètre de passage entre l'état S et l'état I , β , a la même interprétation que dans le cadre du modèle SI : il s'agit du nombre moyen de personnes qu'un individu infecté contamine pendant une unité de temps.

α , quant à lui, est plus facilement interprétable par son inverse : α^{-1} représente la durée moyenne d'infection d'un individu infecté. α est donc la proportion d'infectés se rétablissant par unité de temps.

Ainsi, des équations (3.5) et de l'interprétation des paramètres β et α peut être dégagé un indicateur important dans le suivi et la comparaison des épidémies entre elles : le nombre de reproduction de base, R_0 , qui s'écrit $R_0 = \frac{\beta}{\alpha}$. En multipliant la durée moyenne d'infections par le nombre moyen de personnes qu'un individu infecté contamine, R_0 peut alors s'interpréter comme le nombre moyen de personnes qu'un infecté va contaminer au cours de sa période de contagion au début de l'épidémie, c'est à dire si tous les individus sont susceptible sauf un qui est infecté. Cet individu est donc supposé ne pouvoir rencontrer que des susceptibles au début de l'épidémie. À un instant $t > 0$ de l'épidémie, la fraction de la population qui est infectée ou rétablie, c'est à dire qui n'est plus susceptible, est $1 - \frac{S(t)}{N}$: cette fraction de la population ne peut pas être infectée. Ainsi le nombre de reproduction effectif constitue un indicateur plus pertinent. Il est noté $R_t = R_0 \times \frac{S(t)}{N}$, et correspond, à l'instant t , au nombre moyen d'individus qu'un infecté peut contaminer pendant sa période d'infection.

Il l'a été évoqué les paramètres β et α sont, dans le modèle SIR , des constantes. Cette hypothèse très forte, remise en question par la suite, implique que les dynamiques restent somme toute assez rigides et donc que ce modèle ne permet pas, en l'état, de prendre en compte certains cas de figure pourtant fréquents qui apparaissent communément dans certaines épidémies. Il est en effet tout à fait envisageable, et ce cas de figure se vérifie sur certaines épidémies, que la proportion de personnes quittant l'état S pour l'état I varie au cours de l'épidémie, la même réflexion s'appliquant à la proportion de personnes quittant I pour R .

La structure du modèle SIR est un choix judicieux pour modéliser des épidémies saisonnières, parmi le panel de modèles compartimentaux classiques. En effet, elles confèrent en général une immunité de longue durée en comparaison de la durée de l'épidémie. Ainsi, si le choix est fait d'une modélisation de chaque saison épidémique, l'horizon de modélisation est suffisamment court pour valider l'hypothèse que les individus rétablis ne peuvent retourner dans l'état S , étant immunisés au moins jusqu'à la saison épidémique suivante.

3.2.2 Une discrétisation temporelle du modèle SIR plus cohérente avec la maille de surveillance de la grippe

Le modèle tel que présenté ci-dessus repose sur une hypothèse de continuité temporelle. Cependant, ce point d'attention sera détaillé par la suite, une telle modélisation avec des coefficients de passage constants - « constant-rate models » - ne permet pas de prendre en considération certaines informations souvent indispensables à une modélisation satisfaisante. C'est pourquoi il est parfois nécessaire de préférer une modélisation à pas de temps discret qui, contre intuitivement, permettra de complexifier le modèle dans le sens où il sera alors permis d'introduire la notion de durée passée dans l'état I .

Exprimer cette latence entre l'infection et le rétablissement semble primordial dans le cas du modèle *SIR*, qui sera le modèle de base retenu dans le cadre de ces travaux. Intuitivement, cela assurerait en quelques sortes qu'un individu nouvellement infecté a une propension plus faible de se rétablir qu'un individu qui est déjà infecté depuis plusieurs jours. Enfin, la maille de surveillance des épidémies est en règle générale de l'ordre de la semaine, voire du jour : une discrétisation du modèle fait alors sens, dans une optique de cohérence avec les données historiques et en l'absence de valeur ajoutée d'une modélisation continue dans le cadre présent.

Cette discrétisation des dynamiques du modèle appliquée au modèle *SIR* donne :

$$\forall t \in \mathbb{N}, \begin{cases} S(t+1) - S(t) = -\beta \times \frac{S(t)I(t)}{N} \\ I(t+1) - I(t) = \beta \frac{S(t)I(t)}{N} - \alpha I(t) \\ R(t+1) - R(t) = \alpha I(t) \end{cases} \quad (3.6)$$

Ce modèle *SIR* à pas de temps discret servira par la suite de base à partir de laquelle sera proposé un modèle plus cohérent au regard de l'interprétation de ses dynamiques inter-état.

Dorénavant et sauf mention contraire, le nombre d'individus dans chacun des états du modèle amélioré *SIR*⁺ à un instant donné t sera noté sans distinction d'interprétation $S_t = S(t)$, $I_t = I(t)$ et $R_t = R(t)$.

Chapitre 4

Le modèle SIR^+ , amélioration de l'existant SIR

Le modèle SIR en l'état, tel que décrit dans la littérature, ne semble pas présenter certaines propriétés primordiales dans le cadre de l'étude de la propagation temporelle discrète d'une épidémie de grippe en France, notamment quant au regard du manque de précision et d'interprétation de ses dynamiques. C'est pourquoi il est proposé un modèle alternatif, dénommé SIR^+ , se fondant sur les mêmes états que le modèle SIR , mais présentant des dynamiques de passage d'un état à un autre différentes, prenant en compte notamment les durées passées par les individus dans l'état transitionnel du modèle, l'état I .

4.1 Refonte des dynamiques déterministes de transition inter-état

4.1.1 Nécessité de correction des erreurs induites par les modèles « constant-rate »

Une des limites du modèle SIR réside dans son paramètre α . Il est rappelé que α , dans le cadre du SIR , représente la proportion d'individus infectés qui guérissent de la maladie par pas de temps. Son inverse, α^{-1} s'interprète comme la durée moyenne d'infection de la maladie. Ce paramètre, constant, pose certains problèmes de précision, notamment l'absence de prise en compte de la date d'infection dans la dynamique de la population rétablie entre t et $t + 1$.

En effet le modèle SIR ne distingue pas les proportions d'infectés rétablis en fonction de l'instant où ils ont été infectés : en d'autres termes, la dynamique des rétablis ne tient compte de la durée d'infection du virus sous-jacent qu'en espérance à l'aide d'un paramètre α constant, là où une définition plus précise pourrait venir améliorer la modélisation.

À titre d'exemple, si 10 individus sont infectés à un instant t et que 5 individus sont infectés à un instant $t + 1$, alors à l'instant $t + 2$, la proportion d'individus rétablis qui ont été infectés en t devrait être supérieure à la proportion d'individus rétablis qui ont été infectés en $t + 1$, car intuitivement plus on s'éloigne de la date d'infection et plus la probabilité de guérir est élevée.

Il ressort donc qu'une prise en compte de ces durées d'infections dans la dynamique de passage de l'état I à l'état R permet d'améliorer le modèle.

4.1.2 Modélisation plus fine du nombre de nouveaux infectés par pas de temps

Le modèle SIR fait état d'un premier paramètre, β , correspondant au nombre moyen de personnes qu'un individu infecté contamine par unité de temps. Il serait cependant intéressant de préciser quelles sont les composantes de ce paramètre : c'est ce qui est proposé par la suite.

Soient les notations suivantes :

- A l'évènement « un contact entre un individu susceptible et un individu infecté engendre une contamination de l'individu susceptible »
- X_t la variable aléatoire représentant le nombre de contacts entre tous les individus infectés et tous individus susceptibles entre les instants t et $t + 1$
- $Y = \mathbb{1}_A$

L'évènement A est supposé indépendant du temps, donc la variable aléatoire Y l'est aussi. La variable aléatoire X_t quant à elle dépend des effectifs des états S et I qui eux sont dépendants du temps. Les variables aléatoires X_t et Y sont supposées indépendantes : Y se rapportant à une caractéristique de la maladie sous-jacente, à savoir son pouvoir infectieux, elle est entièrement déterminée par celle-ci et est donc indépendante du nombre de contacts entre les individus.

Il est également supposé qu'un individu $i \in I$ peut rencontrer indépendamment tous les individus $s \in S$. Il y a donc $S(t) \times I(t)$ contacts possibles entre les individus susceptibles et les individus infectés.

Les variables aléatoires X_t et Y étant indépendantes, il est déduit que :

$$\mathbb{E}[X_t Y] = \mathbb{E}[X_t] \mathbb{E}[Y] = \mathbb{E}[X_t] \mathbb{E}[\mathbb{1}_A] = \mathbb{E}[X_t] \mathbb{P}(A)$$

Soient :

- $\theta = \mathbb{P}(A)$ la probabilité qu'un contact entre un individu susceptible et un individu infecté entraîne une contamination de l'individu susceptible
- m le nombre de contact moyen qu'un individu a avec d'autres personnes au cours d'une semaine unité de temps

Tous les individus, qu'ils soient dans l'état S , I ou R , sont supposés avoir le même nombre de contacts avec d'autres individus lors d'une semaine. Ainsi :

$$\begin{aligned} \mathbb{E}[X_t Y] &= \mathbb{E}[X_t] \times \mathbb{P}(A) \\ &= \theta \sum_{i=1}^{S(t)} [(\text{nombre de contacts de l'individu } i \text{ sur une unité de temps}) \\ &\quad \times \mathbb{P}(\text{un de ces contacts est contaminé})] \\ &= \theta \sum_{i=1}^{S(t)} m \times \frac{I(t)}{N-1} \\ &= m\theta \times \frac{S(t)I(t)}{N-1} \\ &\boxed{\underset{N \rightarrow +\infty}{\sim} m\theta \times \frac{S(t)I(t)}{N}} \end{aligned}$$

N est dorénavant supposé suffisamment grand pour approximer $N - 1$ par N dans les dynamiques des états du modèle.

Le paramètre β du modèle SIR a donc été affiné par approche probabiliste, avec $\beta = m\theta$. Ici θ peut être interprété comme le pouvoir infectieux de la maladie étudiée, et m , le nombre de contacts moyen d'un individu, est lui dépendant d'autres facteurs. Ces facteurs sont très variés : l'âge de la personne qui entre en contact, l'âge des personnes avec qui elle entre en contact, la période temporelle (semaine, week-ends, vacances scolaires etc.) ou encore la durée des contacts. De tels paramètres permettraient de segmenter la population pour modéliser de manière plus fine la propagation de la maladie. Ces nombres de contacts sont documentés dans l'étude « Social contact data for France » publiée en 2015.

Dans le cadre de ce mémoire cependant, choix a été fait de considérer m constant, correspondant au nombre moyen de contacts au cours d'une unité de temps, qui fait donc l'approximation d'une population homogène en termes d'âge, de sexe, de comportement de contacts etc.

4.1.3 Modélisation plus fine du nombre de nouveaux rétablis par pas de temps

Comme évoqué en amont, un modèle *constant-rate* induit une interprétation erronée des dynamiques du modèle. Ne pas tenir compte des dates de contamination aboutit en effet à une faible précision du modèle. Après avoir apporté les modifications nécessaires à la dynamique de passage entre l'état S et l'état I , il est nécessaire de faire de même pour la dynamique de passage de l'état I à l'état R .

Sont introduites en amont du développement de la dynamique de rétablissement les notations suivantes :

$$\left\{ \begin{array}{l} -A_i^t \text{ l'évènement « l'individu } i \text{ est dans l'état infecté à l'instant } t \text{ »} \\ -B_i^t \text{ l'évènement « l'individu } i \text{ se rétablit entre } t \text{ et } t + 1 \text{ »} \\ -C_i^k \text{ l'évènement « l'individu } i \text{ est infecté à l'instant } k \text{ »} \\ -D_i^t \text{ l'évènement « l'individu } i \text{ ne n'est pas rétabli avant } t \text{ »} \\ -T_i \text{ la variable aléatoire représentant le temps passé par l'individu } i \text{ dans l'état infectieux.} \\ \quad \text{Ces variables aléatoires sont supposées } i.i.d \text{ donc } \forall i \neq j, T_i \stackrel{iid}{=} T_j \stackrel{iid}{=} T \\ -F_T(t) = \mathbb{P}(T \leq t) \text{ la fonction de répartition de la variable aléatoire } T \end{array} \right.$$

Pour $t = 1$, l'expression de la dynamique des nouveaux rétablis s'exprime comme suit :

$$\mathbb{E}[R(1) - R(0)] = (F_T(1) - F_T(0)) \times I(0) = F_T(1) \times I(0)$$

qui correspond au nombre d'individu dans l'état I à l'instant 0 multiplié par la probabilité pour ces individus infectés en $t = 0$ d'être rétablis en $t = 1$.

En se plaçant dans le cas $t > 1$:

$$\begin{aligned}
\mathbb{E}[R(t+1) - R(t)] &= \sum_{i=1}^N \mathbb{P}[A_i^t \cap B_i^t] \\
&= \sum_{i=1}^N \mathbb{P}\left[\left(\bigcup_{k=0}^t (C_i^k \cap D_i^t)\right) \cap B_i^t\right] \\
&= \sum_{i=1}^N \mathbb{P}\left[\bigcup_{k=0}^t (C_i^k \cap D_i^t \cap B_i^t)\right] \\
&= \sum_{i=1}^N \mathbb{P}\left[\bigcup_{k=0}^t (C_i^k \cap B_i^t)\right] \quad (\text{car } B_i^t \subset D_i^t) \\
&= \sum_{i=1}^N \sum_{k=0}^t \mathbb{P}[C_i^k \cap B_i^t] \quad (\text{car les évènements } (C_i^k \cap B_i^t) \text{ sont indépendants}) \\
&= \sum_{i=1}^N \sum_{k=0}^t \mathbb{P}[B_i^t | C_i^k] \mathbb{P}[C_i^k] \\
&= \sum_{i=1}^N \left[\mathbb{P}[B_i^t | C_i^0] \mathbb{P}[C_i^0] + \sum_{k=1}^t (\mathbb{P}[B_i^t | C_i^k] \mathbb{P}[C_i^k]) \right] \\
&= \sum_{i=1}^N \left[\mathbb{P}[B_i^t | C_i^0] \times \frac{I(0)}{N} + \sum_{k=1}^t \left(\mathbb{P}[B_i^t | C_i^k] \times \frac{m\theta}{N} I(k-1) \right) \right] \\
&\stackrel{iid}{=} \sum_{i=1}^N \left[\mathbb{P}(T \in [t; t+1]) \times \frac{I(0)}{N} + \sum_{k=1}^t \left(\mathbb{P}(T \in [t-k; t+1-k]) \times \frac{m\theta}{N} I(k-1) \right) \right] \\
&= \sum_{i=1}^N \left[[F_T(t+1) - F_T(t)] \times \frac{I(0)}{N} + \sum_{k=1}^t \left((F_T(t+1-k) - F_T(t-k)) \times \frac{m\theta}{N} I(k-1) \right) \right] \\
&= [F_T(t+1) - F_T(t)] \times I(0) + \sum_{k=1}^t [(F_T(t+1-k) - F_T(t-k)) \times m\theta I(k-1)] \\
&= (F_T(t+1) - F_T(t)) \times I(0) + m\theta \sum_{k=1}^t [F_T(t+1-k) - F_T(t-k)] I(k-1)
\end{aligned}$$

Cette nouvelle définition de la dynamique de passage de l'état I à l'état R propose de partitionner les populations infectées selon leur date d'infection, pour ensuite appliquer une pondération à leur rétablissement, pondération qui dépend de la durée passée dans l'état I .

Finalement, les équations suivantes, représentations mathématiques des dynamiques de passage entre chaque

état du modèle, sont obtenues :

$$\begin{cases} S(t+1) - S(t) = -m\theta \times \frac{S(t)I(t)}{N} \\ I(t+1) - I(t) = m\theta \times \frac{S(t)I(t)}{N} - \left[[F_T(t+1) - F_T(t)] \times I(0) + m\theta \sum_{k=1}^t [F_T(t+1-k) - F_T(t-k)] I(k-1) \right] \\ R(t+1) - R(t) = [F_T(t+1) - F_T(t)] \times I(0) + m\theta \sum_{k=1}^t [F_T(t+1-k) - F_T(t-k)] I(k-1) \end{cases} \quad (4.1)$$

Avec des conditions initiales, qui sont les compositions des états à l'instant 0 et à l'instant 1 :

$$\begin{cases} S(0) = S_0 \\ I(0) = I_0 \\ R(0) = R_0 \\ S(1) = S(0) - m\theta \times \frac{S(0)I(0)}{N} \\ I(1) = I(0) + m\theta \times \frac{S(0)I(0)}{N} - F_T(1) \times I(0) \\ R(1) = R(0) + F_T(1) \times I(0) \end{cases} \quad (4.2)$$

Dans la suite le temps de contagion d'un individu, modélisé par la variable aléatoire notée T précédemment, sera supposé suivre une loi Gamma de paramètres a et b . Ce choix de loi de probabilité pour modéliser cette durée est motivé par l'article « Modélisation du risque de pandémie dans Solvabilité 2 » publié par Frédéric Planchet en 2012.

Quelques rappels sur la loi Gamma :

Si $T \hookrightarrow \Gamma(a, b)$, avec $(a, b) \in (\mathbb{R}^{+*})^2$, alors :

$\mathbb{E}(T)$	$\mathbb{V}(T)$	densité $f_T(x, a, b)$
$a \times b$	$a \times b^2$	$\frac{x^{a-1} e^{-x/b}}{\Gamma(a) b^a}$

où $\forall a > 0, \Gamma : a \mapsto \int_0^{+\infty} t^{a-1} e^{-t} dt$ est la fonction Gamma d'Euler.

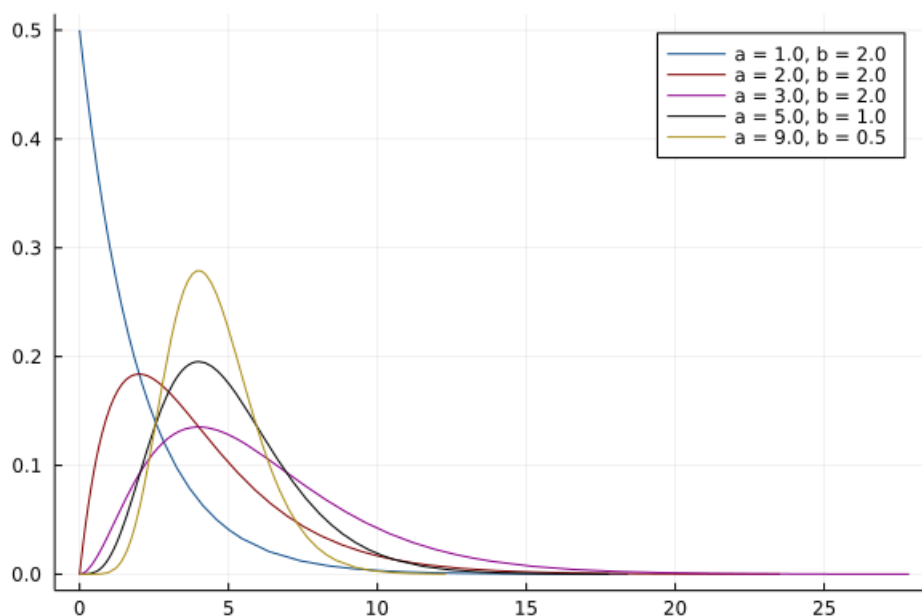


FIGURE 4.1 – Densités de la loi Gamma pour différents jeux de paramètres

Le modèle SIR^+ déterministe est désormais construit. Des équations 4.1, il est possible de déduire que ce modèle comporte 3 paramètres :

- θ , la probabilité qu'un contact entre un individu susceptible et un individu infecté engendre l'infection de l'individu susceptible
- a , le paramètre de forme de la loi de rétablissement T
- b , le paramètre d'échelle de cette loi

Ainsi le modèle SIR^+ construit dans le cadre de ce mémoire est entièrement caractérisé par ses conditions initiales et par le jeu de paramètres (θ, a, b) qui lui est associé.

L'enjeu désormais réside dans la construction d'une procédure de calibration cohérente au regard de la structure du modèle SIR^+ .

4.2 Données utilisées pour l'ajustement du modèle : historique de l'incidence grippale en France

Le modèle déterministe SIR^+ étant désormais entièrement construit, il est pertinent de proposer un exemple d'utilisation au regard d'une réponse à une problématique actuarielle concrète. Une proposition de calibration est donc faite, et le choix des données utilisées pour la calibration du modèle SIR^+ proposé précédemment se

porte sur des données d'incidence de la grippe en France.

L'incidence, en épidémiologie, représente le nombre de nouveaux cas recensés par unité de temps. À titre d'exemple, une incidence de 10 entre t et $t + 1$ signifie que 10 individus ont été infectés entre t et $t + 1$. L'incidence correspond ainsi à l'opposé de la dynamique de l'état S dans le modèle SIR^+ . En effet, tous les individus quittant l'état S entre t et $t + 1$ migrent vers l'état I . Ainsi l'incidence entre t et $t + 1$ vaut :

$$S_t - S_{t+1} = m\theta \times \frac{S(t)I(t)}{N}$$

4.2.1 Le réseau Sentinelles

Le réseau Sentinelles, créé en 1984 conjointement par la Sorbonne Université et l'INSERM (Institut national de la santé et de la recherche médicale), a pour but le suivi et la surveillance de la diffusion de nombreuses maladies infectieuses en France (exception faite des actes de suicides qui sont eux aussi documentés par le réseau) et est alimenté par des informations fournies par des centaines de médecins généralistes.

Ce réseau fournit ainsi des données hebdomadaires concernant la grippe, à la maille région ou à la maille France métropolitaine, qui comprennent les variables suivantes :

- *week*, le numéro de la semaine concernée dans l'année, au format AAAASS (201926 correspond par exemple à la semaine 26 de 2019, *i.e* du 24 au 30 juin 2019)
- *inc*, correspondant à l'estimation du nombre de nouveaux infectés sur la population considérée (*inc* signifiant incidence) par semaine
- *inc_low*, la borne inférieure de l'incidence
- *inc_up*, la borne supérieure de cette même incidence
- *inc100*, *inc100_low* et *inc100_up*, qui sont les mêmes variables que ci-dessus, mais ramenés à 100 000 individus
- *geo_insee*, l'indicateur Insee de la zone géographique étudiée
- *geo_name*, est le nom de la zone géographique étudiée

Les données se présentent comme suit, à la maille France métropolitaine :

week	inc	inc_low	inc_up	inc100	inc100_low	inc100_up	geo_insee	geo_name
202105	29198	23610	34786	44	36	52	FR	France
202104	26159	21751	30567	40	33	47	FR	France
202103	21810	17894	25726	33	27	39	FR	France
202102	17320	13906	20734	26	21	31	FR	France
202101	21799	17778	25820	33	27	39	FR	France
202053	21220	16498	25942	32	25	39	FR	France
202052	16428	12285	20571	25	19	31	FR	France
202051	21619	17370	25868	33	27	39	FR	France

FIGURE 4.2 – Extrait de la base de données de Sentinelles, au format Excel

Les deux dernières variables des bases de données de Sentinelles ne présentent pas d'intérêt particulier dans le cadre de l'étude de l'épidémie sur le territoire français, car il n'est pas fait d'étude de l'évolution de l'épidémie

d'un point de vue spatial.

Le périmètre des saisons épidémiques répertoriées par Sentinelles s'étend de 1984 à 2020 : la population française est passée en 36 ans de 56 à 67 millions d'habitants. Or, il en a été fait mention précédemment, le modèle SIR^+ fait, comme le modèle SIR , l'hypothèse que la population est fermée, c'est à dire à N constant : il est donc dorénavant et déjà possible de retenir que l'incidence pour 100 000 habitants au détriment de l'incidence sur l'ensemble de la population dans les données fournies par Sentinelles. En effet la première est une variable normalisée par 100 000, là où la seconde est calculée sur l'ensemble de la population réelle de la zone considérée (région ou France métropolitaine dans son ensemble), qui varie au cours des semaines. Il est donc plus aisé de manipuler la variable d'incidence pour 100 000 habitants, pour ne pas avoir à prendre en compte les variations de la population totale.

Par la suite, la variable $inc100$, appelée également « taux d'incidence », sera celle retenue. Les termes « incidence » et « taux d'incidence » seront utilisés sans différence sémantique, par souci de concision, les deux faisant référence au nombre de nouveaux infectés à la maille temporelle considérée et ramenés à 100 000 personnes.

La base de données de Sentinelles est segmentée par saison épidémique, ce qui est indispensable pour la cohérence des hypothèses du modèle utilisé. En effet, l'hypothèse structurelle du modèle selon laquelle il n'y a pas de retour possible dans l'état S à partir de l'état R ne peut se faire sur plusieurs saisons épidémiques : si certaines personnes infectées redeviennent susceptibles, comme c'est le cas d'une année à une autre pour la grippe, les hypothèses du modèle ne sont plus respectées et la modélisation n'est alors plus fiable. Pour une saison épidémique de l'année n , la plage d'étude est ainsi réduite à [semaine 36 n ; semaine 29 $n + 1$], qui correspond à peu près à [septembre n ; juillet $n + 1$]. En effet, Santé Publique France avance que les épidémies de grippe saisonnière en France s'étalent de novembre à avril, avec un début autour de fin décembre voire début janvier. La courbe d'incidence extraite du réseau Sentinelles, pour la saison 2018, se présente comme suit :

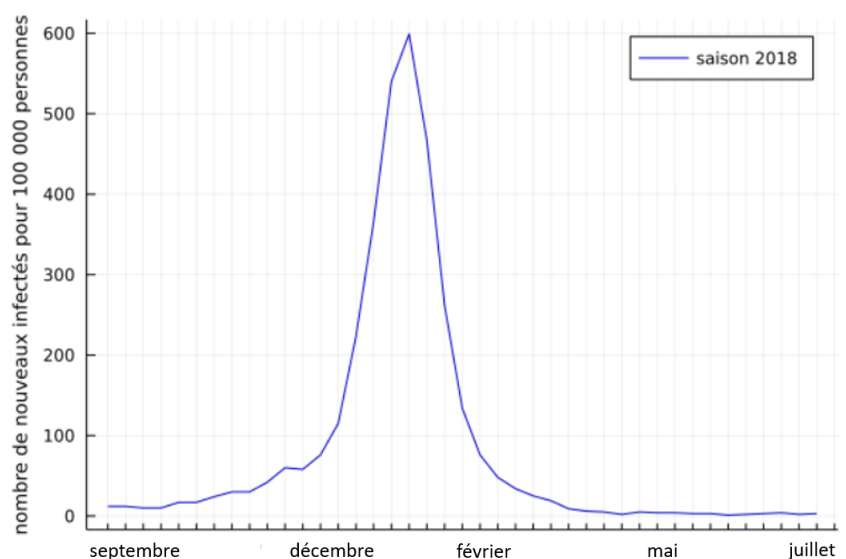


FIGURE 4.3 – Courbe d'incidence de la grippe en France, en fonction du temps

En pratique, toutes les saisons épidémiques fournies par le réseau Sentinelles débutent ultérieurement au mois de septembre, et s'estompent avant juillet : c'est pourquoi l'intervalle d'observation est [septembre n ; juillet $n + 1$], qui permet de tenir compte d'éventuelles épidémies plus précoces, ou qui tarderaient à s'éteindre. L'hypothèse inductive selon laquelle aucune épidémie de grippe en France ne sort de ce cadre temporel est faite ici. De plus désormais et sauf mention contraire, les abscisses de toutes les figures présentant des courbes d'incidence sont des entiers, qui correspondent au nombre d'unité de temps écoulées depuis la semaine 36 de l'année considérée.

4.2.2 Retraitement des données

Affinage de la granularité temporelle des données d'incidence

Les données issues du réseau Sentinelles sont hebdomadaires, ce qui permet d'avoir l'information du nombre d'individus nouvellement infectés chaque semaine, et ce pour chaque saison. Cependant, la répartition de cette

incidence à la maille jour est inconnue. Or parfois la connaissance de cette incidence quotidienne peut être nécessaire pour certaines problématiques actuarielles. C'est d'ailleurs le cas dans le cadre du cas pratique proposé au sein de ce mémoire, la quantification temporelle des arrêts de travail se faisant au jour près. Il est donc nécessaire de retraiter les données pour qu'elles soient quotidiennes.

La saison 2018 est ici prise à titre d'exemple illustratif de la méthode de retraitement ; celle-ci sera ensuite appliquée à l'ensemble des saisons retenues.

A partir des données hebdomadaires, l'incidence totale est répartie uniformément sur tous les jours de la semaine. En prenant la semaine 2 de 2019 par exemple, 115 nouveaux cas de grippe pour 100 000 habitants sont dénombrés sur l'ensemble du territoire français métropolitain. En répartissant uniformément ces 115 infectés il ressort que chaque jour de la semaine n° 2, du lundi au dimanche, $\frac{115}{7} \approx 16,42$ personnes ont été nouvellement infectées.

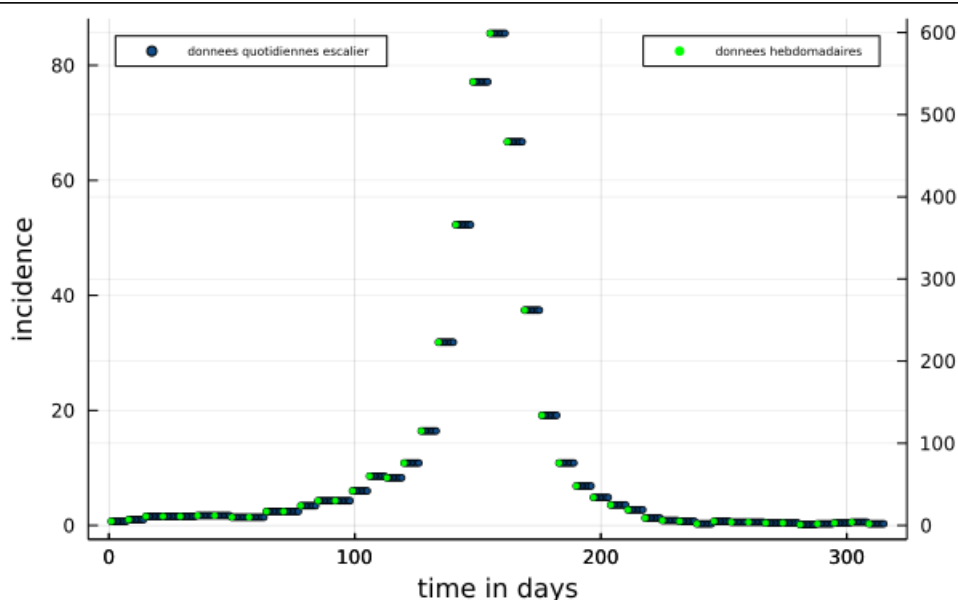


FIGURE 4.4 – Répartition uniforme maille jour des données d'incidence maille semaine de la saison 2018

La figure (4.4) ci-dessus fait état en abscisse du temps exprimé en jours à partir de la semaine 36 de 2018. L'axe des ordonnées de gauche correspond au nombre de nouveaux cas quotidiens « en escalier », tandis que l'axe de droite correspond lui au nombre de nouveaux cas hebdomadaires. Les abscisses des marqueurs verts, rattachés à l'axe des ordonnées de droite, correspondent au premier jour de la semaine considérée. Le numéro de la semaine est ainsi l'abscisse divisée par 7 pour les données hebdomadaires.

Un lissage par moyenne mobile uniforme d'ordre 7 est ensuite appliquée sur les données uniformément réparties.

En d'autres termes, en notant I_n (resp. \hat{I}_n) l'incidence quotidienne obtenue (resp. l'incidence quotidienne lissée), et N_{jours} la fenêtre d'étude de la saison épidémique, la formule suivante est appliquée :

$$\forall n \in \llbracket 3, N_{jours} - 3 \rrbracket, \hat{I}_n = \frac{1}{7} \sum_{i=-3}^3 I_{n+i} \quad (4.3)$$

Les nuages de points suivants sont obtenus :

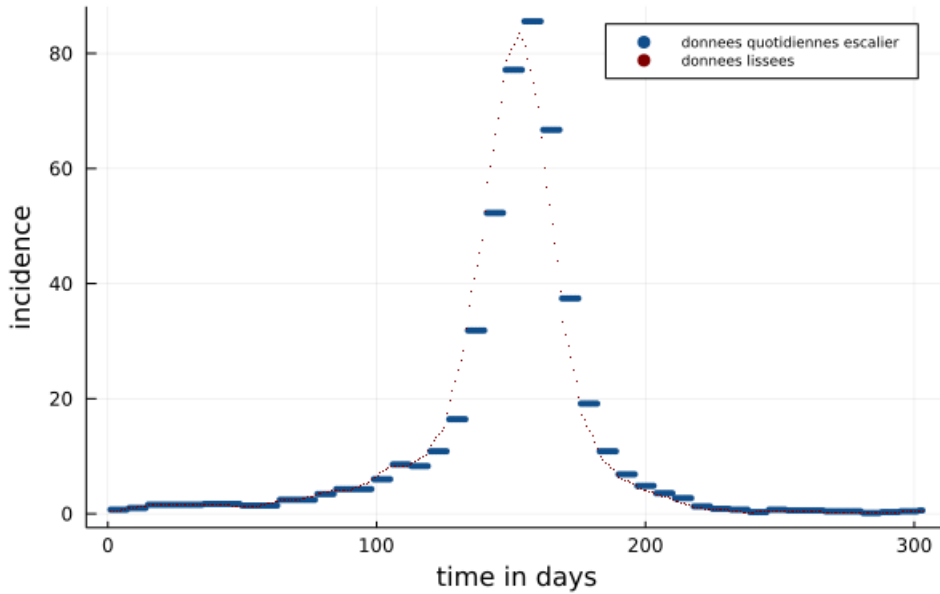


FIGURE 4.5 – Lissage par moyennes mobiles d'ordre 7

La répartition telle qu'effectuée ici n'est néanmoins pas satisfaisante car si la courbe est croissante, toutes les données lissées seront supérieures à leur pendant en escalier (et donc l'incidence cumulée sur la semaine sera plus conséquente avec les données lissées qu'avec les données en escalier). À l'inverse, si la courbe est décroissante, le phénomène inverse sera observé : les données lissées seront toutes inférieures aux données en escalier, et le cumulé également. Pour pallier ce problème, un simple décalage du graphe des données lissées, c'est-à-dire l'introduction d'un "retard" dans le lissage, est effectué. Afin de déterminer de combien de jours le lissage doit être décalé, le calcul de la MSE (pour Mean Squared Error, ou Erreur quadratique moyenne en français) est effectué. Cette métrique servira de critère de décision quant à cet indice de décalage. En reprenant les notations de (4.3) et en notant $j \in \llbracket 1, 6 \rrbracket$ le décalage opéré, la MSE s'écrit :

$$MSE(j) = \frac{1}{N_{jours} - j} \sum_{n=1}^{N_{jours}-j} (\widehat{I}_{n+j} - I_n)^2 \quad (4.4)$$

Le nuage de points représentant la MSE en fonction du décalage réalisé est tracé :

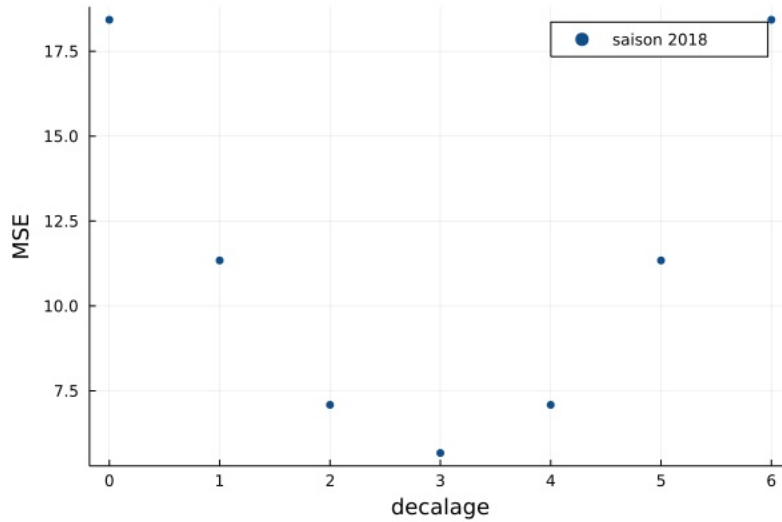


FIGURE 4.6 – Critère de choix du décalage de lissage

Quelle que soit la saison étudiée, le décalage renvoyant la MSE minimale est 3. La valeur de la fonction de perte pour la saison 2018 est d'environ 5,67. Ainsi, le lissage retardé de 3 jours pour cette saison est :

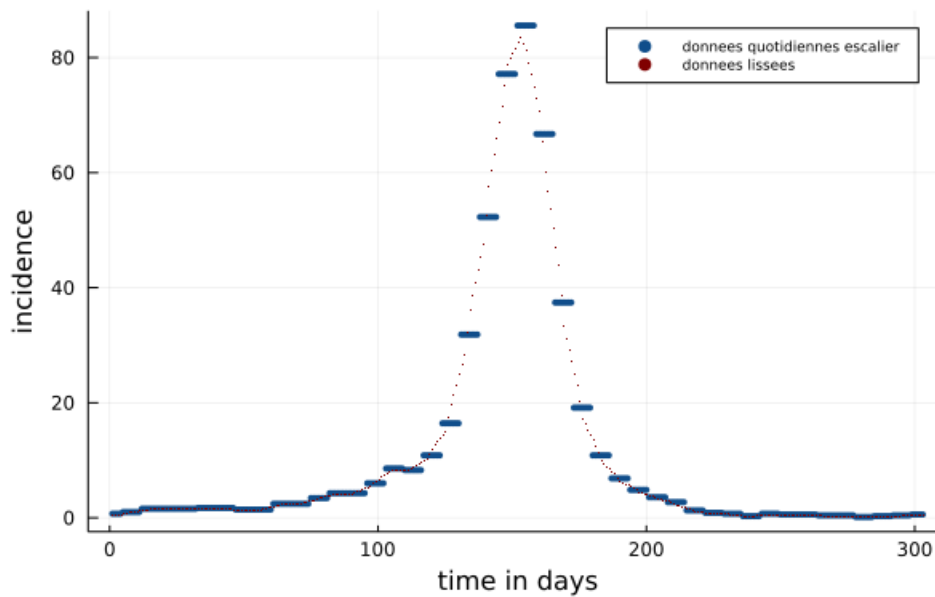


FIGURE 4.7 – Lissage retardé de 3 jours

Étant donné que la courbe a été décalée de 3 jours vers la droite, cela revient à considérer désormais que l'axe des abscisses représente le nombre de jours écoulés depuis le 4^{ème} jour de la semaine 36 de la saison considérée.

Exclusion des saisons 2020 et 2021

Parmi les saisons épidémiques documentées par le réseau Sentinelles, les saisons 2020 et 2021 sont volontairement exclues, car impactées et donc biaisées par l'apparition du virus SARS-CoV-2. En effet comme il en a été fait mention dans la première partie de ce mémoire, des mesures gouvernementales visant à endiguer cette pandémie combinées à des changements comportementaux de la population ont également fait drastiquement chuter les cas de grippe.

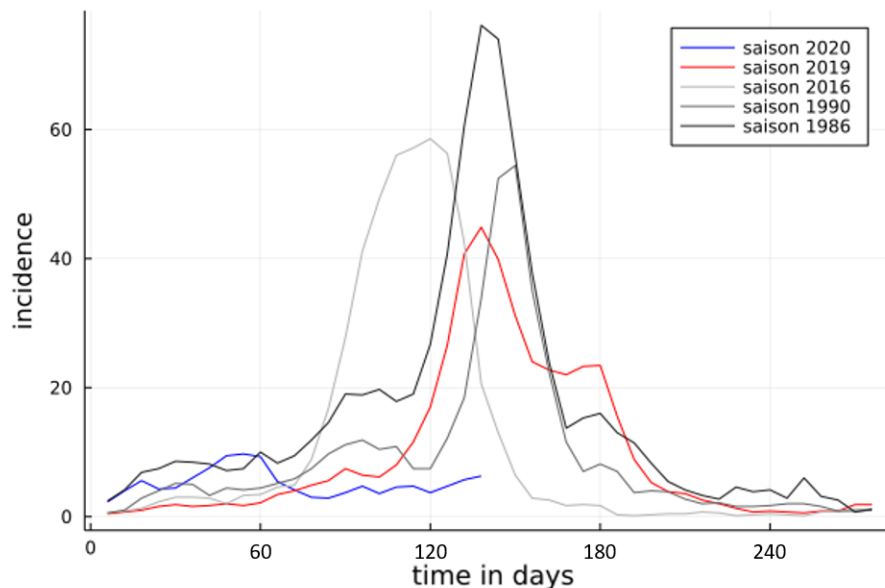


FIGURE 4.8 – Comparaison des incidences de la saison 2020 et la saison 2021 avec quelques autres saisons

Exclusion de certaines saisons épidémiques jugées « irrégulières »

Seront également exclues de la suite des travaux les saisons 1989, 1991, 1997, 2000, 2015, 2017 et 2019, car le manque de « régularité » de leur courbe d'incidence en fait des saisons trop atypiques.

La courbe d'incidence de la saison 2015, qui est retenue ici à titre d'exemple et qui est considérée représentative des autres saisons exclues, possède 2 « pics », comme le montre la figure ci-après :

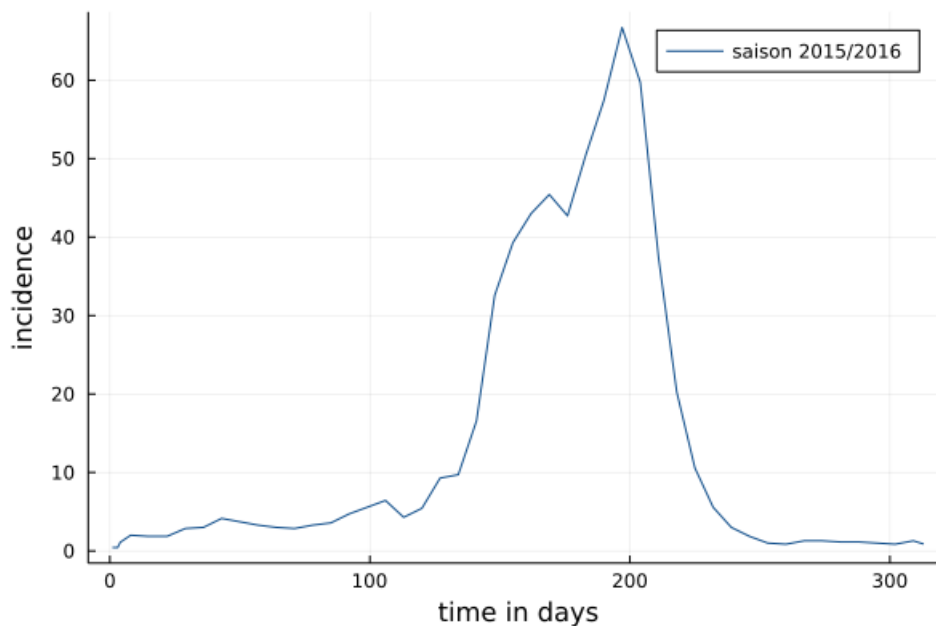


FIGURE 4.9 – Incidence de la saison 2015 en fonction du temps

Une raison possible à cette irrégularité serait la présence de 2 souches différentes du virus de la grippe : la souche *A* et la souche *B*, une des deux se propageant avec un léger retard par rapport à l'autre. Ces deux souches pourraient de plus ne pas avoir les mêmes caractéristiques intrinsèques (contagiosités différentes, temps d'infection également etc.), d'où l'exclusion des saisons épidémiques de cette sorte de notre modèle, qui ne permet pas la prise en compte de tels cas de figure.

Chapitre 5

Adaptation d'une technique d'optimisation de l'ajustement issue du *Machine Learning : Adaptive Moment Estimation*

Comme il l'a été évoqué précédemment, le modèle SIR^+ possède 3 paramètres, que sont θ , a et b . Dorénavant le vecteur de ces paramètres sera noté $\Lambda = (\theta, a, b)$.

Les équations (4.1) régissant les états de ce modèle permettent, à partir des conditions initiales $S(0) = S_0$, $I(0) = I_0$ et $R(0) = R_0$, et les paramètres θ , a et b , de fournir la composition des 3 états S , I et R à chaque instant t , moyennant la connaissance des compositions de ces états à tout instant $t' \in \llbracket 0, t - 1 \rrbracket$.

Dans le cadre du cas pratique qui a été choisi pour illustrer l'utilisation du modèle SIR^+ , à savoir une application concrète à la grippe saisonnière, le modèle nécessite une phase de calibration. La méthode consiste à utiliser les trois degrés de liberté que représentent les trois paramètres θ , a et b pour faire varier la sortie du modèle (l'incidence quotidienne), et faire en sorte que cette sortie du modèle se rapproche au mieux des données historiques observées extraites du réseau Sentinelles.

L'ajustement des sorties du modèle SIR^+ aux données historiques issues de Sentinelles repose sur le calcul de la MSE , métrique déjà introduite précédemment. Son expression, avec les paramètres du SIR^+ , s'exprime comme suit :

$$MSE : (\theta, a, b) \mapsto \frac{1}{n} \sum_{t=1}^n (incidence_t^{modelisee}(\theta, a, b) - incidence_t^{observee})^2$$

Cette MSE est la moyenne des écarts entre l'incidence modélisée par le SIR^+ et l'incidence réellement observée, *i.e* l'incidence historique de Sentinelles.

L'objectif est maintenant de déterminer le vecteur $\Lambda^* = (\theta^*, a^*, b^*)$ tel que :

$$\Lambda^* = \underset{\Lambda}{\operatorname{argmin}} MSE(\Lambda)$$

Le vecteur Λ^* ne peut cependant pas être calculé analytiquement car il n'existe pas de formule fermée de cet *argmin*. La mise en place d'une méthode de résolution numérique pour approcher au mieux ce minimum théorique est donc indispensable.

5.1 Spécificités de l'algorithme de descente de gradient *ADAM*

Algorithme classique de la descente de gradient

Afin de mener à bien cette résolution numérique, et obtenir trois paramètres satisfaisants et suffisamment « proches » de θ^* , a^* et b^* , la méthode utilisée sera une méthode dérivée de celle appelée « descente de gradient », le gradient étant le vecteur des dérivées partielles de la fonction à minimiser défini en (5.1). Il s'agit d'un algorithme d'optimisation différentiable dont le principe est de minimiser une fonction (fonction de perte, ici la *MSE*) en effectuant à chaque itération un déplacement dans la direction du gradient (la direction de chacune des composantes du gradient, pour chaque dimension), et dans le sens opposé à celui-ci.

Il est rappelé que le gradient d'une fonction à plusieurs variables est le vecteur colonne composé des dérivées partielles de la fonction par rapport à chacune de ses variables. En d'autres termes, le gradient en (x, y, z) d'une fonction f à trois variables s'écrit :

$$\nabla f(x, y, z) = \begin{pmatrix} \frac{\partial f(x, y, z)}{\partial x} \\ \frac{\partial f(x, y, z)}{\partial y} \\ \frac{\partial f(x, y, z)}{\partial z} \end{pmatrix} = \begin{pmatrix} \lim_{h_1 \rightarrow 0} \frac{f(x+h_1, y, z) - f(x, y, z)}{h_1} \\ \lim_{h_2 \rightarrow 0} \frac{f(x, y+h_2, z) - f(x, y, z)}{h_2} \\ \lim_{h_3 \rightarrow 0} \frac{f(x, y, z+h_3) - f(x, y, z)}{h_3} \end{pmatrix} \quad (5.1)$$

Le jeu de paramètres associé à l'itération k de la descente de gradient est noté $\Lambda_k = (\theta_k, a_k, b_k)$.

L'algorithme de la descente de gradient est le suivant : un premier itéré est créé, $\Lambda_0 \in]0; 1[\times (\mathbb{R}^{++})^2$, et un critère d'arrêt est choisi. Celui-ci est libre, et peut porter sur le nombre d'itération, sur la valeur de la *MSE* etc. Ainsi à chaque itération, Λ_{k+1} est calculé à partir de Λ_k de la manière suivante :

- $[\nabla MSE](\Lambda_k)$ est calculé
- L'itéré $\Lambda_{k+1} = \Lambda_k - \lambda [\nabla MSE](\Lambda_k)$ est créé
- Si le critère d'arrêt est vérifié, alors $\Lambda_{final} \approx \Lambda_{k+1}^*$

À chaque itération, chacun des « bonds » sur l'hypersurface de la fonction à minimiser sera de norme égale au produit de la norme du gradient et d'un hyperparamètre nommé « learning-rate », et noté λ . Cet hyperparamètre est constant dans l'algorithme de descente de gradient classique, mais fait l'objet d'adaptations pour optimiser la convergence de l'algorithme. Dans le cas d'un *learning-rate* constant, il en existe une valeur optimale. Le choisir trop faible ralentit fortement la vitesse de convergence, alors qu'en utiliser un trop élevé risque d'engendrer une oscillation de l'algorithme autour du minimum de la fonction, et dégraderait ainsi la qualité de la convergence.

L'enjeu d'un choix optimal de ce *learning-rate* est illustré par les trois figures suivantes :

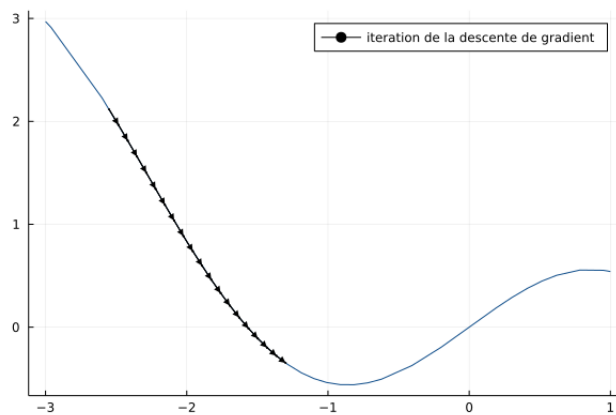


FIGURE 5.1 – Cas d'un learning-rate faible (convergence trop lente)

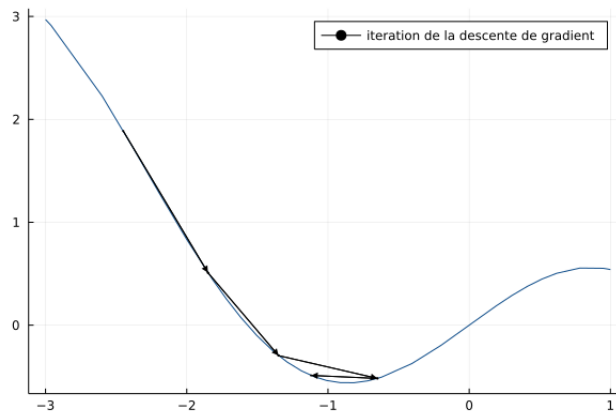


FIGURE 5.2 – Cas d'un learning-rate élevé (oscillations autour du minimum)

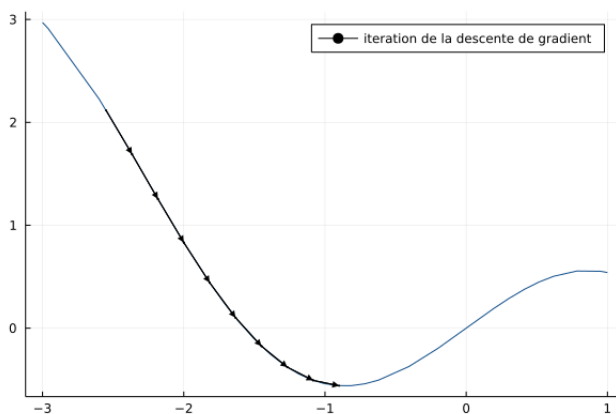


FIGURE 5.3 – Cas d'un learning-rate adapté

ADAM, amélioration de la descente de gradient classique

Diederik P. Kingma et Jimmy Ba introduisent en 2015 une amélioration de cette méthode de descente de gradient : l'algorithme ADAM (pour *Adaptive Moment Estimation*).

L'algorithme *ADAM* ressemble dans sa structure à l'algorithme classique de descente de gradient, à ça près que la partie $\lambda \times [\nabla MSE](\Lambda)$ est ici adaptative (elle est modifiée à chaque itération pour tenir compte de l'avancée de la descente), ce qui lui permet, à chaque itération k , de prendre en compte la vitesse et l'accélération cumulées lors des itérations $j \in \llbracket 1, k-1 \rrbracket$, d'où l'appellation « Adaptive Moment ». L'algorithme *ADAM* présente ainsi l'avantage d'une convergence plus rapide vers les paramètres optimaux dans le cadre de l'ajustement de réseaux de neurones, et il est supposé que cette propriété bienvenue est transposable à l'ajustement du modèle *SIR*⁺. Un des critères qui explique son utilisation concernant des calibrations de réseaux de neurones s'explique également par sa robustesse face à un grand nombre de paramètres à ajuster. Ceci permet la généralisation de la méthodologie de calibration dans la mesure où si une modification du modèle était opérée, entraînant l'ajout d'un paramètre, alors *ADAM* pourrait vraisemblablement toujours être utilisé pour ajuster le modèle.

Les étapes de l'algorithme se présentent sous la forme :

$\forall k \in \mathbb{N}^*$,

$$\begin{aligned}
 & - g_k = [\nabla MSE](\Lambda_{k-1}) \\
 & - m_k = \beta_1 \times m_{k-1} + (1 - \beta_1) \times g_k \\
 & - v_k = \beta_2 \times v_{k-1} + (1 - \beta_2) \times g_k^2 \\
 & - \hat{m}_k = \frac{m_k}{1 - \beta_1^k} \\
 & - \hat{v}_k = \frac{v_k}{1 - \beta_2^k} \\
 & - \Lambda_k = \Lambda_{k-1} - \lambda_0 \times \frac{\hat{m}_k}{\sqrt{\hat{v}_k + \epsilon}}
 \end{aligned} \tag{5.3}$$

Avec :

- $(\beta_1, \beta_2) \in]0; 1[^2$
- λ_0 le *learning-rate* initial, à déterminer selon les mêmes critères et avec les mêmes enjeux que pour la descente de gradient classique
- ϵ une constante, fixée dans le cas présent à 10^{-24} , qui évite une division par 0 dans le cas où $\hat{v}_k = 0$
- $m_0 = v_0 = 0$ les valeurs initiales de $(m_k)_{k>0}$ et $(v_k)_{k>0}$

Les hyperparamètres β_1 et β_2 sont fixés à respectivement 0,9 et 0,99 ; ces valeurs sont celles largement plébiscitées dans les utilisations d'ADAM en *machine learning*.

5.2 Démonstration de l'existence du gradient de l'écart quadratique

Pour pouvoir appliquer une méthode de descente de gradient, qu'il s'agisse de la méthode classique ou d'ADAM, il est nécessaire que le gradient soit bien défini : en d'autres termes que la fonction de perte à minimiser soit dérivable. Dans le cadre de ce mémoire, il est nécessaire d'avoir l'information que la *MSE* est bien dérivable par rapport à ses trois paramètres θ , a et b .

La démonstration qui suit, dont la mise en œuvre a été entièrement réalisée dans le cadre de ce mémoire, a été motivée par l'absence de démonstration de la sorte dans la littérature, et par sa nécessité pour une utilisation rigoureuse de l'algorithme ADAM.

La démonstration est menée à bien par récurrence simple, et comportera les étapes suivantes :

1. Une phase d'initialisation
2. Une phase démontrant la dérivabilité de la fonction de répartition de la loi *Gamma* par rapport à son paramètre de forme a et à son paramètre d'échelle b . La principale complexité d'une telle démonstration réside dans cette preuve de dérivabilité, qui est nécessaire car la fonction de répartition est présente dans l'expression de l'incidence découlant des dynamiques du modèle
3. Une phase d'hérédité
4. Une phase de conclusion quant à la dérivabilité de la *MSE*

Sont introduites, pour commencer, les notations suivantes :

$$- g(\theta, a, b, t) = \begin{cases} m\theta \sum_{k=1}^t [F_T(t+1-k) - F_T(t-k)]I_{k-1} + [F_T(t+1) - F_T(t)]I_0 & \text{si } t > 0 \\ F_T(1)I_0 & \text{si } t = 0 \end{cases}$$

– m le nombre moyen de contacts qu'a un individu pendant un jour

– N la taille de la population

$$- \mu = \frac{m\theta}{N}$$

– $F_T(x)$ la fonction de répartition de la variable aléatoire T qui représente le temps passé dans l'état I , avec $T \hookrightarrow \Gamma(a, b)$

– Une fonction sera dite $der(\mathbf{V})^D$ si elle est dérivable sur D par rapport à chacun des éléments du vecteur V . À titre d'exemple si $V = (x, y, z)$ et $D = \mathbb{R}^2 \times \mathbb{R}^{+*}$, alors « f est $der(V)^D$ » \Leftrightarrow « f est dérivable par rapport à x sur \mathbb{R} , par rapport à y sur \mathbb{R} et par rapport à z sur \mathbb{R}^{+*} ». Cette notation est créée ici par souci de clarté et de lisibilité de la preuve.

$$- \mathcal{F} = [0; 1] \times (\mathbb{R}^{+*})^2$$

$$- \Lambda = (\theta, a, b) \in \mathcal{F}$$

5.2.1 Initialisation de la récurrence

La fonction de perte qui doit être minimisée *via* l'algorithme *ADAM* est ici rappelée :

$$MSE(\Lambda) = \frac{1}{n} \sum_{t=1}^n (incidence_t^{simulee}(\theta, a, b) - incidence_t^{observee})^2 \quad (5.4)$$

Ainsi, pour prouver que MSE est dérivable par rapport à chacun des paramètres θ , a et b (notation « $der(\Lambda)^{\mathcal{F}}$ » introduite précédemment), il suffit de prouver la dérivabilité de $incidence_t^{simulee}(\theta, a, b)$, qui est la seule partie de cette expression dépendant des paramètres. Or :

$$incidence_t^{simulee} = \mu S_{t-1} I_{t-1} \quad (5.5)$$

donc il suffit de prouver que μ , S_{t-1} et I_{t-1} sont $der(\Lambda)^{\mathcal{F}}$; la MSE sera alors dérivable par somme et produit de fonctions dérivables. Enfin, de manière immédiate, $\mu = \frac{m\theta}{N}$ est $der(\Lambda)^{\mathcal{F}}$, ce qui réduit la démonstration à prouver que S_t et I_t sont $der(\Lambda)^{\mathcal{F}}$.

La relation de récurrence est $\forall t \in \mathbb{N}$, $H(t) : S_t$ et I_t sont $der(\Lambda)^{\mathcal{F}}$.

$\forall t \in \mathbb{N}^*$:

$$\begin{cases} S_{t+1} = S_t - \mu I_t S_t = S_t(1 - \mu I_t) \\ I_{t+1} = I_t + \mu I_t S_t - g(\theta, a, b, t) \end{cases}$$

Initialisation :

$H(0)$

$$S_0 \text{ et } I_0 \text{ sont } \text{der}(\Lambda)^{\mathcal{F}} \text{ (de manière immédiate car ce sont des constantes)} \quad (5.6)$$

$H(1)$

$$S_1 = S_0 - \mu S_0 I_0 \text{ est } \text{der}(\Lambda)^{\mathcal{F}} \text{ (de manière immédiate)} \quad (5.7)$$

$$I_1 = I_0 + \mu S_0 I_0 - g(\theta, a, b, 0) = I_0(1 + \mu S_0 - F_T(1)) \quad (5.8)$$

5.2.2 Dérivabilité de la fonction de répartition mesurant la durée d'infection

Il est désormais nécessaire de montrer que F_T est $\text{der}((a, b))^{(\mathbb{R}^{++})^2}$ (F_T ne dépend pas de θ). Cette propriété n'ayant pas été démontrée dans la littérature, ce qui suit en est une proposition de démonstration faisant partie intégrante des travaux réalisés au cours de ce mémoire.

Soient $t \in \mathbb{N}, k \in \llbracket 1, t \rrbracket$. Est notée J l'expression suivante :

$$\begin{aligned} J &= F_T(t+1-k) - F_T(t-k) \\ &= \int_0^{t+1-k} f_T(x) dx - \int_0^{t-k} f_T(x) dx \\ &= \int_{t-k}^{t+1-k} \frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a} dx \end{aligned}$$

La démonstration de la dérivabilité de l'intégrale ci-dessus est entreprise *via* le théorème de dérivation sous le signe intégrale.

$$\begin{aligned} \text{Soit } h \text{ la fonction } h : (\mathbb{R}^{++})^3 &\longrightarrow \mathbb{R}^{++} \\ (x, a, b) &\longmapsto \frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a} \end{aligned}$$

Existence de J

La loi Gamma est une loi dite à densité, donc sa densité est *de facto* intégrable sur \mathbb{R}^{+*} . Ainsi, il est immédiat que $\forall (a, b) \in (\mathbb{R}^{+*})^2$

$$x \mapsto h(x, a, b) \text{ est intégrable sur } \mathbb{R}^{+*}, \text{ donc en particulier sur } [t - k; t + 1 - k] \subset \mathbb{R}^{+*}$$

Dérivabilité de h par rapport à a

$$\text{Soit } (x, b) \in (\mathbb{R}^{+*})^2, h(x, a, b) = \frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a} = e^{-\frac{x}{b}} \times x^{a-1} \frac{1}{\Gamma(a) b^a}$$

Il est ici admis que la fonction $a \mapsto \Gamma(a)$ est dérivable pour $a > 0$.

Il est nécessaire de montrer que $a \mapsto \frac{1}{\Gamma(a)}$ est dérivable pour $a > 0$ (qui est bien définie car la fonction Γ ne s'annule jamais sur \mathbb{R}^{+*}).

Soit $(a, \epsilon) \in (\mathbb{R}^{+*})^2$:

$\frac{\frac{1}{\Gamma(a+\epsilon)} - \frac{1}{\Gamma(a)}}{\epsilon} = -\frac{\Gamma(a+\epsilon) - \Gamma(a)}{\epsilon} \times \frac{1}{\Gamma(a)\Gamma(a+\epsilon)}$ or Γ est dérivable en $a > 0$, et $x \mapsto \frac{1}{\Gamma(x)}$ est continue en $a > 0$, donc $\epsilon \mapsto \frac{\frac{1}{\Gamma(a+\epsilon)} - \frac{1}{\Gamma(a)}}{\epsilon}$ admet une limite finie en 0, qui vaut :

$$\lim_{\epsilon \rightarrow 0} \frac{\frac{1}{\Gamma(a+\epsilon)} - \frac{1}{\Gamma(a)}}{\epsilon} = -\frac{\Gamma'(a)}{\Gamma(a)^2}$$

$\forall (x, b) \in (\mathbb{R}^{+*})^2$, les fonctions $a \mapsto x^{a-1}$ et $a \mapsto \frac{1}{b^a}$ sont dérivables sur \mathbb{R}^{+*} donc finalement, en tant que produit de fonctions dérivables par rapport à a , la fonction $a \mapsto h(x, a, b)$ est bien dérivable relativement à a sur \mathbb{R}^{+*} .

Dérivabilité de h par rapport à b

$$\text{Soit } (x, a) \in (\mathbb{R}^{+*})^2, h(x, a, b) = \frac{x^{a-1}}{\Gamma(a)} \times e^{-\frac{x}{b}} \times \frac{1}{b^a}.$$

$\forall (x, a) \in (\mathbb{R}^{+*})^2$, les fonctions $b \mapsto e^{-\frac{x}{b}}$ et $b \mapsto \frac{1}{b^a}$ sont dérivables sur \mathbb{R}^{+*} donc finalement, en tant que produit de fonctions dérivables par rapport à b , la fonction $b \mapsto h(x, a, b)$ est bien dérivable relativement à b sur \mathbb{R}^{+*} .

Domination des dérivées partielles de h selon a et b

Il est désormais nécessaire de montrer qu'il existe deux applications ϕ et ψ de $(\mathbb{R}^{+*})^2$ dans \mathbb{R}^{+*} intégrables sur leur ensemble de définition, et telles que :

$$\left\{ \begin{array}{l} \forall a > 0, \forall (x, b) \in (\mathbb{R}^{+*})^2, \left| \frac{\partial h(x, a, b)}{\partial a} \right| \leq \phi(x, b) \quad \text{(i)} \\ \forall b > 0, \forall (x, a) \in (\mathbb{R}^{+*})^2, \left| \frac{\partial h(x, a, b)}{\partial b} \right| \leq \psi(x, a) \quad \text{(ii)} \end{array} \right.$$

Preuve de (i) :

Il est rappelé que $\forall b \in \mathbb{R}^{+*}$, la fonction $a \mapsto b^a$ est dérivable selon a sur \mathbb{R}^{+*} et admet pour dérivée $a \mapsto b^a \ln(b)$.

$$\begin{aligned}
 \left| \frac{\partial h(x, a, b)}{\partial a} \right| &= \left| \frac{\partial}{\partial a} \left(\frac{x^{a-1} e^{-\frac{x}{b}}}{\Gamma(a) b^a} \right) \right| \\
 &= \left| e^{-\frac{x}{b}} \left(\frac{\ln(x) x^{a-1}}{\Gamma(a) b^a} - x^{a-1} \frac{\Gamma'(a) b^a + \ln(b) b^a \Gamma(a)}{(\Gamma(a) b^a)^2} \right) \right| \\
 &= \left| e^{-\frac{x}{b}} \frac{x^{a-1}}{\Gamma(a) b^a} \left(\ln(x) - \frac{\Gamma'(a)}{\Gamma(a)} - \ln(b) \right) \right| \\
 &= \left| e^{-\frac{x}{b}} \frac{x^a}{x \Gamma(a) b^a} \left(\ln(x) - \frac{\Gamma'(a)}{\Gamma(a)} - \ln(b) \right) \right| \\
 &= e^{-\frac{x}{b}} \frac{1}{x \Gamma(a)} \left(\frac{x}{b} \right)^a \left| \ln(x) - \frac{\Gamma'(a)}{\Gamma(a)} - \ln(b) \right| \\
 &\leq e^{-\frac{x}{b}} \frac{1}{x \Gamma(a)} \left(\frac{x}{b} \right)^a \left(\left| \ln \left(\frac{x}{b} \right) \right| + \left| \frac{\Gamma'(a)}{\Gamma(a)} \right| \right) \quad \text{d'après l'inégalité triangulaire} \\
 &\leq e^{-\frac{x}{b}} \frac{1}{x} \left(\frac{x}{b} \right)^a \left(\frac{\left| \ln \left(\frac{x}{b} \right) \right|}{\Gamma(a)} + \left| \frac{\Gamma'(a)}{\Gamma(a)^2} \right| \right)
 \end{aligned}$$

La fonction Γ est minorée sur \mathbb{R}^{+*} donc $a \mapsto \frac{1}{\Gamma(a)}$ est majorée sur \mathbb{R}^{+*} , i.e :

$$\exists K_1 \geq 0 : \forall a > 0, \left| \frac{1}{\Gamma(a)} \right| \leq K_1$$

Une majoration par une constante doit être déterminée pour l'expression $\left| \frac{\Gamma'(a)}{\Gamma(a)^2} \right|$ sur \mathbb{R}^{+*} , i.e

$$\exists K_2 \geq 0 : \forall a > 0, \left| \frac{\Gamma'(a)}{\Gamma(a)^2} \right| \leq K_2$$

Il est supposé connu que $\Gamma'(a) = \int_0^\infty \ln(t) t^{a-1} e^{-t} dt$.

Soit $a > 0$ et $0 < t < e$.

$$\ln(t) t^{a-1} e^{-t} < t^{a-1} e^{-t} \quad (5.9)$$

Le terme de droite est intégrable sur $]0; e[$ de manière immédiate, et celui de gauche également car :

$$\begin{cases} \ln(t) t^{a-1} \xrightarrow[t \rightarrow 0]{} 0 & \text{pour } a \neq 1 \\ \forall x \in]0; e[, \int_x^e \ln(t) t^{a-1} dt = \int_x^e \ln(t) dt = [t \ln(t) - t]_x^e = e - e - x \ln(x) + x \xrightarrow[x \rightarrow 0]{} 0 & \text{pour } a = 1 \end{cases}$$

L'intégration de l'inégalité (5.9) sur $]0; e[$ donne ainsi :

$$\int_0^e \ln(t)t^{a-1}e^{-t} dt < \int_0^e t^{a-1}e^{-t} dt \quad (5.10)$$

En remarquant que :

$$\int_0^e \ln(t)t^{a-1}e^{-t} dt = \Gamma'(a) - \int_e^\infty \ln(t)t^{a-1}e^{-t} dt$$

et

$$\int_0^e t^{a-1}e^{-t} dt = \Gamma(a) - \int_e^\infty t^{a-1}e^{-t} dt$$

l'inégalité (5.10) devient :

$$\begin{aligned} \Gamma'(a) - \int_e^\infty \ln(t)t^{a-1}e^{-t} dt &< \Gamma(a) - \int_e^\infty t^{a-1}e^{-t} dt \\ \Gamma'(a) &< \Gamma(a) + \left[\int_e^\infty \ln(t)t^{a-1}e^{-t} dt - \int_e^\infty t^{a-1}e^{-t} dt \right] \end{aligned} \quad (5.11)$$

$$\frac{\Gamma'(a)}{\Gamma(a)^2} < \frac{1}{\Gamma(a)} + \frac{\left[\int_e^\infty \ln(t)t^{a-1}e^{-t} dt - \int_e^\infty t^{a-1}e^{-t} dt \right]}{\Gamma(a)^2} = \frac{1}{\Gamma(a)} + \frac{\Omega_1}{\Gamma(a)^2} \quad (5.12)$$

en notant $\Omega_1 = \int_e^\infty \ln(t)t^{a-1}e^{-t} dt - \int_e^\infty t^{a-1}e^{-t} dt$.

En regroupant ces intégrales, Ω_1 devient :

$$\Omega_1 = \int_e^\infty [\ln(t) - 1]t^{a-1}e^{-t} dt \quad (t \mapsto [\ln(t) - 1]t^{a-1}e^{-t} \text{ est bien intégrable sur } [e; +\infty[)$$

$\forall t > e$, $\ln(t) - 1 < \ln(t) < t$ car la fonction \ln est concave sur \mathbb{R}^{+*} .

Donc, en passant à l'intégrale $(\ln(t) - 1$ et t sont strictement positifs sur $]e; +\infty[)$ Ω_1 peut se réécrire :

$$\Omega_1 = \int_e^\infty [\ln(t) - 1]t^{a-1}e^{-t} dt < \int_e^\infty t \times t^{a-1}e^{-t} dt = \int_e^\infty t^a e^{-t} dt < \int_0^\infty t^a e^{-t} dt = \Omega_2 \quad (\text{car } t^a e^{-t} > 0 \text{ sur }]0; e])$$

Sont introduites les fonctions suivantes : $u : t \mapsto t^a$ et $v : t \mapsto -e^{-t}$. u et v sont dérivables par rapport à t , et u' et v' sont continues sur \mathbb{R}^{+*} avec pour expressions $u' : t \mapsto at^{a-1}$ et $v' : t \mapsto e^{-t}$. Ω_2 peut être intégré par partie :

$$\Omega_2 = [-t^a e^{-t}]_0^\infty - \int_0^\infty -at^{a-1}e^{-t} dt = \int_0^\infty at^{a-1}e^{-t} dt = a\Gamma(a) \quad (5.13)$$

L'intégration par parties est effectuée directement sur $]0; +\infty[$, en supposant qu'elle le soit sur $]x; y[$ et que x tende vers 0 et y vers $+\infty$, opération mathématiquement réalisable car les intégrales convergent.

Donc, d'après (5.12) et (5.13) :

$$\frac{\Gamma'(a)}{\Gamma(a)^2} < \frac{1}{\Gamma(a)} + \frac{\Omega_1}{\Gamma(a)^2} < \frac{1}{\Gamma(a)} + \frac{\Omega_2}{\Gamma(a)^2} = \frac{1}{\Gamma(a)} + \frac{a\Gamma(a)}{\Gamma(a)^2} = \frac{1+a}{\Gamma(a)} \quad (5.14)$$

La fonction Gamma possède le comportement asymptotique suivant :

$$\forall \alpha > 0, a^\alpha = o_{+\infty}(\Gamma(a)) \Rightarrow \lim_{a \rightarrow +\infty} \frac{a}{\Gamma(a)} = 0 \text{ (d'après la formule de Stirling)}$$

De plus :

$$\lim_{a \rightarrow 0} \Gamma(a) = +\infty \Rightarrow \lim_{a \rightarrow 0} \frac{1+a}{\Gamma(a)} = 0$$

La fonction Γ est strictement positive sur \mathbb{R}^{++} : $a \mapsto \frac{1+a}{\Gamma(a)}$ est donc continue et strictement positive sur \mathbb{R}^{++} .

Il en est conclu que le quotient $\left| \frac{\Gamma'(a)}{\Gamma(a)^2} \right|$ est bien majoré par une constante positive K_2 .

Enfin, si $x \leq b$, alors $\left(\frac{x}{b}\right)^a$ est majoré par 1 donc finalement, en prenant l'hypothèse que $x \leq b$, alors :

$$\left| \frac{\partial h(x, a, b)}{\partial a} \right| \leq e^{-\frac{x}{b}} \frac{1}{x} (K_1 \left| \ln\left(\frac{x}{b}\right) \right| + K_2) = \phi(x, b) \text{ avec } \phi \text{ qui est bien intégrable sur } (\mathbb{R}^{++})^2.$$

La preuve telle que rédigée ci-dessus n'est valable que pour $x \leq b$. Cependant, dans le cadre de ce mémoire, le domaine d'étude de a et b est plus restreint que \mathbb{R}^{++} . Ce domaine peut donc être réduit à $[\xi; \Xi]$ avec $(\xi, \Xi) \in (\mathbb{R}^{++})^2$ et $\xi < \Xi$, ce qui rend la preuve valable indépendamment de la condition $x \leq b$.

En effet, si $b \in [\xi; \Xi]$, alors :

$$\begin{cases} \left(\frac{x}{b}\right)^a \leq 1 & \text{si } 0 < \frac{x}{b} \leq 1 \\ \left(\frac{x}{b}\right)^a \leq \left(\frac{x}{b}\right)^\Xi = K_{3,1}(x, b) & \text{si } \frac{x}{b} > 1 \end{cases}$$

et en posant $K_3(x, b) = \max(1; K_{3,1}(x, b))$, $\left(\frac{x}{b}\right)^a$ est bien majoré par $K_3(x, b) \geq 0$.

Preuve de (ii) :

$$\begin{aligned} \left| \frac{\partial h(x, a, b)}{\partial b} \right| &= \left| \frac{x^{a-1}}{\Gamma(a)} \times \frac{\frac{x}{b^2} \times \exp\left(-\frac{x}{b}\right) \times b^a - \exp\left(-\frac{x}{b}\right) \times a \times b^{a-1}}{(b^a)^2} \right| \\ &= \left| \frac{x^{a-1}}{\Gamma(a)} \times \frac{\exp\left(-\frac{x}{b}\right) \left(\frac{x}{b} - a\right)}{b^{a+1}} \right| \\ &\leq \frac{x^{a-1}}{\Gamma(a)} \times \frac{\exp\left(-\frac{x}{b}\right)}{b^{a+1}} \times \left(\frac{x}{b} + a\right) \quad \text{d'après l'inégalité triangulaire} \end{aligned}$$

De la même manière que précédemment, le domaine de dérivabilité est arbitrairement restreint à $[\xi; \Xi]$, avec $(\xi, \Xi) \in (\mathbb{R}^{++})^2$ et $\xi < \Xi$.

Alors, $\forall b \in [\xi; \Xi]$:

$$\begin{aligned} \left| \frac{\partial h(x, a, b)}{\partial b} \right| &\leq \frac{x^{a-1}}{\Gamma(a)} \times \frac{\exp\left(-\frac{x}{b}\right)}{b^{a+1}} \times \left(\frac{x}{b} + a\right) \\ &\leq \frac{x^{a-1}}{\Gamma(a)} \times \frac{\exp\left(-\frac{x}{\Xi}\right)}{\xi^{a+1}} \times \left(\frac{x}{\xi} + a\right) \\ &= \psi(x, a) \text{ intégrable sur } (\mathbb{R}^{+*})^2 \end{aligned}$$

Conclusion sur la dérivabilité de l'intégrale J

D'après le théorème de dérivation sous le signe intégrale, la fonction $(a, b) \mapsto \int_{t-k}^{t+1-k} \frac{x^{a-1} \exp\left(-\frac{x}{b}\right)}{\Gamma(a) b^a} dx$ est $der((a, b))^{[\xi; \Xi]^2}$, avec ξ et Ξ choisis arbitrairement strictement positifs et tels que $\xi < \Xi$, et ses dérivées partielles sont :

$$\begin{aligned} - \frac{\partial J}{\partial a} &= \int_{t-k}^{t+1-k} \exp\left(-\frac{x}{b}\right) \frac{1}{x\Gamma(a)} \left(\frac{x}{b}\right)^a \left(\ln\left(\frac{x}{b}\right) - \frac{\Gamma'(a)}{\Gamma(a)}\right) dx \\ - \frac{\partial J}{\partial b} &= \int_{t-k}^{t+1-k} \exp\left(-\frac{x}{b}\right) \frac{1}{x\Gamma(a)} \left(\frac{x}{b}\right)^a \frac{1}{b} \left(\frac{x}{b} - a\right) dx \end{aligned}$$

5.2.3 Finalisation de la récurrence

D'après les éléments de preuve qui précèdent :

$$\begin{aligned} \forall t > 0, g_t : \mathcal{F} &\longrightarrow \mathbb{R}^{+*} \\ \Lambda &\longmapsto m\theta \sum_{k=1}^t [F_T(t+1-k) - F_T(t-k)] I_{k-1} + [F_T(t+1) - F_T(t)] I_0 \end{aligned} \quad (5.15)$$

est $der(\Lambda)^{\mathcal{F}}$ comme somme et produit de fonctions dérivables par rapport à ces paramètres.

En reprenant la récurrence (5.8), il ressort que S_1 et I_1 sont bien $der(\Lambda)^{\mathcal{F}}$.

Hérédité :

Soit $t \geq 1$. En supposant $H(t)$ vrai, alors :

$$\begin{aligned} S_{t+1} &= S_t(1 - \mu I_t) \text{ est } der(\Lambda)^{\mathcal{F}} \text{ car } S_t \text{ et } I_t \text{ le sont (d'après } H(t)) \\ I_{t+1} &= I_t + \mu I_t S_t - g(\theta, a, b, t) \text{ est } der(\Lambda)^{\mathcal{F}} \text{ car } S_t, I_t \text{ et } g(\theta, a, b, t) \text{ le sont (d'après } H(t) \text{ et (5.15))} \end{aligned}$$

D'où $H(t+1)$

Etant donné que $MSE(\theta, a, b) = \frac{1}{n} \sum_{t=1}^n \left(incidence_t^{simulee} - incidence_t^{observee} \right)^2$, la propriété suivante est obtenue

par somme et produit de fonctions $der(\Lambda)^{\mathcal{F}}$:

La MSE est $der(\Lambda)^{\mathcal{F}}$, et son gradient est donc bien défini

5.3 Approximation du gradient de la MSE

Calculer les dérivées partielles exactes, par rapport à θ , a , et b de la fonction MSE nécessite d'obtenir au préalable une expression explicite du nombre d'individu dans chacun des 3 états, et ce à chaque instant t .

Cependant, la forme des dynamiques temporelles liant les états S , I et R , dont les équations associées sont (4.1), rendent très difficile l'obtention d'une formule fermée pour ces dérivées partielles. En effet la dynamique des nouveaux infectés à l'instant t est donnée par :

$$\begin{aligned} incidence_t^{modelisee} &= \mu \times S_{t-1} \times I_{t-1} \\ &= \mu \times (S_{t-2} - \mu S_{t-2} I_{t-2}) \times (I_{t-2} + \mu I_{t-2} S_{t-2}) \end{aligned}$$

et ainsi de suite.

La complexification très rapide des expressions manipulées et la difficulté à obtenir *in fine* une expression de $incidence_t^{modelisee}$ en fonction des seules valeurs θ , a , b , S_0 , I_0 , R_0 et N rendent ainsi le calcul théorique de la dérivée partielle très complexe, et ne fait donc pas partie des travaux de ce mémoire.

Ainsi, il apparaît judicieux d'approximer numériquement les dérivées partielles de la MSE par rapport à ses trois paramètres.

5.3.1 Estimation des dérivées partielles par quotient différentiel

L'expression formelle du gradient est rappelée :

$$[\nabla MSE](\Lambda) = \begin{pmatrix} \frac{\partial MSE(\Lambda)}{\partial \theta} \\ \frac{\partial MSE(\Lambda)}{\partial a} \\ \frac{\partial MSE(\Lambda)}{\partial b} \end{pmatrix} = \begin{pmatrix} \lim_{h_1 \rightarrow 0} \frac{MSE(\theta+h_1, a, b) - MSE(\theta, a, b)}{h_1} \\ \lim_{h_2 \rightarrow 0} \frac{MSE(\theta, a+h_2, b) - MSE(\theta, a, b)}{h_2} \\ \lim_{h_3 \rightarrow 0} \frac{MSE(\theta, a, b+h_3) - MSE(\theta, a, b)}{h_3} \end{pmatrix}$$

Une nouvelle notation est introduite pour les taux d'accroissements, ou quotients différentiels, de la MSE relativement à chacun de ses trois paramètres.

Seront donc notés, pour tout vecteur $\Lambda = (\theta, a, b)$ dans $\mathcal{F} =]0; 1[\times (\mathbb{R}^{+*})^2$ et tout vecteur (h_1, h_2, h_3) dans $(\mathbb{R}^{+*})^3$:

$$\begin{cases} q_{h_1}^{MSE}(\theta, a, b) = \frac{MSE(\theta+h_1, a, b) - MSE(\theta, a, b)}{h_1} \\ q_{h_2}^{MSE}(\theta, a, b) = \frac{MSE(\theta, a+h_2, b) - MSE(\theta, a, b)}{h_2} \\ q_{h_3}^{MSE}(\theta, a, b) = \frac{MSE(\theta, a, b+h_3) - MSE(\theta, a, b)}{h_3} \end{cases}$$

Une propriété sur h_1 doit être vérifiée, à savoir $h_1 < 1 - \theta$ car θ est une probabilité.

L'objectif de l'approximation est ici de déterminer un triplet (h_1, h_2, h_3) suffisamment « petit » pour que les $q_{h_i}^{MSE}$ ne soient pas trop éloigné des dérivées partielles de la MSE , et qui est satisfaisant au regard d'un critère qui sera introduit par la suite.

Les valeurs $h_i \in]10^{-2}; +\infty[$, pour $i \in (1, 2, 3)$ seront considérées comme non satisfaisantes, car elles ne permettent pas une précision suffisante ; c'est pourquoi l'espace de recherche de ces valeurs est restreint à l'intervalle $]0; 10^{-2}]$.

Le critère utilisé pour comparer la pertinence des valeurs de h_i est l'écart relatif entre deux valeurs successives de q_{h_i} . En d'autres termes, si les h_1 candidats sont $\{10^{-7}; 10^{-6}; 10^{-5}; 10^{-4}; 10^{-3}; 10^{-2}; 10^{-1}\}$, alors comparer la pertinence de l'utilisation de $h_1^2 = 10^{-6}$ et $h_1^3 = 10^{-4}$ revient à comparer les valeurs de :

$$\frac{q_{10^{-6}} - q_{10^{-5}}}{q_{10^{-5}}}$$

et

$$\frac{q_{10^{-4}} - q_{10^{-3}}}{q_{10^{-3}}}$$

Il ressort de ces comparaisons que les variations des écarts relatifs entre deux valeurs successives de h_i sont relativement constantes, indépendamment des saisons ou du vecteur des paramètres Λ .

Pour la saison 2008, et le vecteur $\Lambda = (0.0265, 3.8223, 0.6407)$, les courbes suivantes sont obtenues :

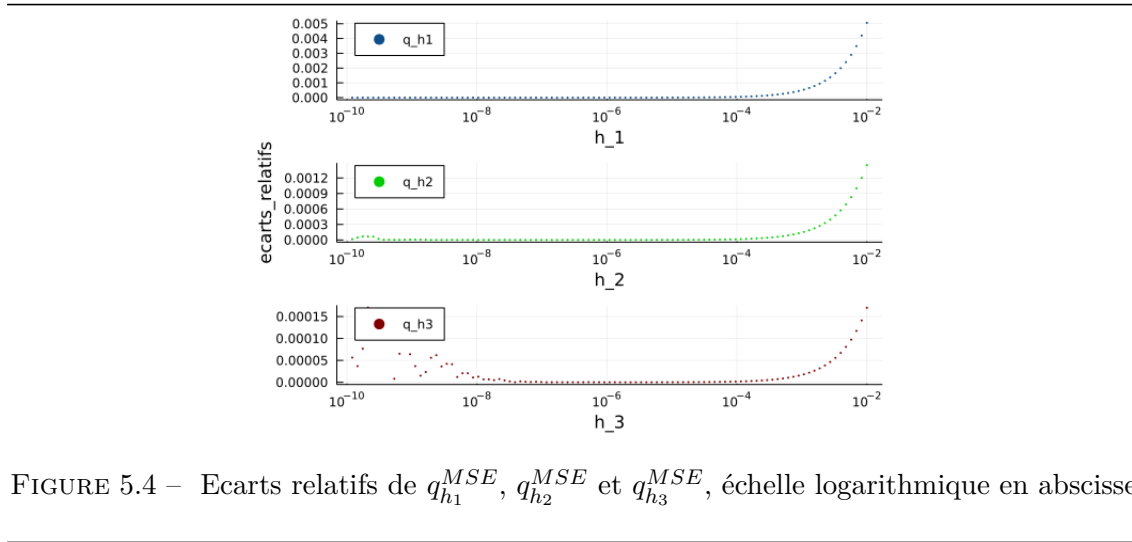


FIGURE 5.4 – Ecart relatifs de $q_{h_1}^{MSE}$, $q_{h_2}^{MSE}$ et $q_{h_3}^{MSE}$, échelle logarithmique en abscisse

Il ressort de la figure 5.4 que l'écart relatif n'est pas croissant sur $[10^{-10}; 10^{-8}]$, puis il semble l'être à partir de 10^{-7} . Ces oscillations pour des faibles valeurs de h peuvent s'expliquer par la représentation faite par le langage de programmation utilisé. Ainsi les calculs de sont plus convenablement effectués pour des valeurs trop faibles. Il n'est donc pas suffisant de choisir h_i les plus petits possibles ; il faut également que l'approximation numérique aie du sens pour pouvoir conclure quant à une valeur de h_i à retenir. L'introduction d'une métrique permettant une telle conclusion est donc nécessaire pour concilier précision de l'approximation et maintien dans le domaine de validité.

La notation suivante est introduite, et correspond à un intervalle entre x_0 et x_l , d'une longueur l et avec un pas constant de $\frac{x_l - x_0}{l - 1}$:

$$K_l(x_0, x_l) = \left\{ x_0 ; x_0 + \frac{x_l - x_0}{l - 1} ; x_0 + 2 \times \frac{x_l - x_0}{l - 1} ; \dots ; x_0 + (l - 2) \times \frac{x_l - x_0}{l - 1} ; x_l \right\}$$

L'espace de recherche des h_i est noté $(h_{i,n}) = (h_{i,1}, h_{i,2}, \dots, h_{i,n-1}, h_{i,n})$. En pratique, cet espace de recherche est commun aux trois dérivées partielles, :

$$(h_{1,n}) = (h_{2,n}) = (h_{3,n}) = (h_n) = (10^{-x})_{x \in K_n(2,10)} \tag{5.16}$$

À titre d'exemple :

$$h_{1,100} = \{10^{-2} ; 10^{-2.08} ; 10^{-2.16} ; \dots ; 10^{-9.92} ; 10^{-10}\}$$

L'espace de recherche des h_i optimaux est le même pour les dérivées selon chacun des trois paramètres du modèle, mais trois recherches distinctes sont bien effectuées : en effet, comme cela est visible sur la figure (5.4), la convergence n'est pas la même pour les trois. Ainsi un h_i pourra être satisfaisant pour une des dérivées partielles mais pas pour les deux autres. Il est donc nécessaire de cloisonner la démarche, et il n'est pas possible

à ce stade de conclure sur le choix d'un seul et unique h , commun aux trois dérivées.

5.3.2 Introduction d'une métrique sur les quotients différentiels, critère de précision

Afin de déterminer les h_i satisfaisants pour approcher la dérivée partielle de la MSE , une métrique est introduite. Elle permettra de conclure, pour un h candidat noté $h_c \in (h_n)$, s'il s'avère être satisfaisant ou bien insuffisamment proche de 0. Il est ici nécessaire de se placer au voisinage de h_c : le taux $dh_c = x\% \times h_c$ est défini, avec x le rayon du voisinage normalisé par la valeur de h_c . Les éléments de $J = (h_n) \cap [h_c - dh_c; h_c + dh_c]$ sont ici considérés.

La métrique retenue est, pour $(\Lambda, h_c) \in \mathcal{F} \times (h_n)$:

$$\Upsilon_i(\Lambda, h_c) = \frac{\sqrt{\frac{1}{\text{card}(J)} \sum_{h \in J} (q_{h_i}^{MSE}(\Lambda) - q_{h_c}^{MSE}(\Lambda))^2}}{|q_{h_c}^{MSE}(\Lambda)|} \quad (5.17)$$

En notant \mathcal{Q} la dimension du quotient différentiel,

$$\dim(\Upsilon) = \frac{\sqrt{\frac{1}{\text{card}(J)} \sum_{h \in J} \mathcal{Q}^2}}{\mathcal{Q}} = \frac{\sqrt{\frac{1}{\text{card}(J)} \times \text{card}(J) \times \mathcal{Q}^2}}{\mathcal{Q}} = \frac{\sqrt{\mathcal{Q}^2}}{\mathcal{Q}} = \frac{\mathcal{Q}}{\mathcal{Q}} = 1$$

La métrique Υ ainsi construite est sans unité, et représente la variance partielle et normalisée par le quotient différentiel $q_{h_c}^{MSE}(\Lambda)$ de l'estimateur des dérivées partielles de la MSE centré en $q_{h_c}^{MSE}(\Lambda)$.

Il est à noter que la valeur de x retenue, 5%, est arbitraire. Cette valeur de x ne peut être choisie trop élevée, surtout si la valeur de h_c considérée est proche de 10^{-2} . En effet, si $dh_c = x\% \times h_c > 10^{-2} - h_c$, alors les valeurs utilisées pour le calcul de la métrique, à savoir la partie $J \setminus \{h \in h_n | h > 10^{-2}\}$ qui n'est pas vide dans ce cas particulier, ne seraient pas complètes.

Soient 100 triplets de paramètres $(\Lambda_1, \Lambda_2, \dots, \Lambda_{100})$ avec chaque triplet $\Lambda_i = (\theta_i, a_i, b_i)$ simulé comme suit :

$$\begin{aligned} \theta_i &\hookrightarrow \mathcal{U}([0; 1]) \\ a_i &\hookrightarrow \mathcal{U}([0; 10]) \\ b_i &\hookrightarrow \mathcal{U}([0; 100]) \end{aligned} \quad (5.18)$$

Le choix de l'espace de simulation de θ est trivial. C'est une probabilité, et l'intervalle n'est pas davantage restreint : $[0; 1]$ est retenu. Concernant a (resp. b), le domaine est restreint à $[0; 10]$ (resp. $[0; 100]$) après avoir effectué des calibrations sur plusieurs saisons. De plus, supposition est faite que ces paramètres ne varient pas significativement d'une saison à l'autre.

Ainsi pour $h_c = 10^{-6}$, $x = 5\%$ la relation $h_c + dh_c = 1,05 \times h_c = 1,05 \times 10^{-6} < 10^{-2}$ est vérifiée.

La métrique pour chaque quotient différentiel est affichée : l'estimateur de la dérivée partielle par rapport à θ , celui par rapport à a et celui par rapport à b .

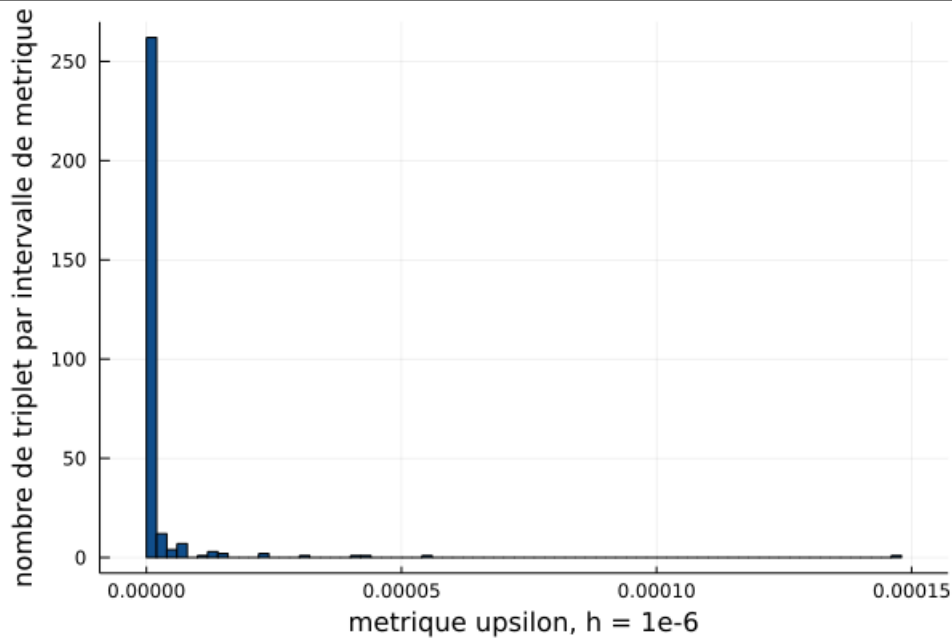


FIGURE 5.5 – Histogramme des métriques epsilon des trois quotients différentiels confondus pour 100 triplets avec $h = 10^{-6}$, pour la saison 2008

Pour 100 triplets Λ_i uniformément simulés et sur les données de la saison grippale 2008, la métrique Υ avec $h_c = 10^{-6}$ parmi $h \in K_{10^3}(10, 2)$ prend des valeurs entre 0 et $1,5 \times 10^{-3}$ avec une forte concentration des valeurs en dessous de 5×10^{-5} ce qui peut être considéré comme satisfaisant au regard de la construction de cette métrique. Ainsi, il est conclu que la valeur de h peut être fixée à 10^{-6} dans les trois estimateurs des dérivées partielles pour toute la suite des travaux.

5.4 Implémentation et mise en application de l'algorithme d'ADAM

En amont de la mise en application de l'algorithme *ADAM*, plusieurs travaux ont été réalisés. Dans un premier temps, il était primordial de vérifier que le gradient est bien défini. Cette vérification a été réalisée au travers de la démonstration de la dérivabilité de la fonction de répartition d'une loi *Gamma* par rapport à ses paramètres, ce qui implique la dérivabilité de la *MSE* dans le cadre de ce mémoire. Puis les dérivées partielles de la *MSE*, dont une formule fermée s'avère trop complexe à obtenir, ont été approchées numériquement par des estimateurs. Enfin, cette approximation numérique a nécessité de choisir avec quelle « précision » acceptable les dérivées sont approchées. Une fois ces étapes réalisées, il est alors possible d'implémenter l'algorithme *ADAM* afin de

calibrer le modèle SIR^+ sur les données de grippe saisonnière de Sentinelles.

5.4.1 Détermination des paramètres initiaux d'ADAM

Un jeu de paramètres initiaux est nécessaire à l'initialisation de l'algorithme $ADAM$. Celui-ci est déterminé par recherche itérative dans une grille de recherche, en retenant le jeu de paramètres initiaux renvoyant la plus faible MSE .

Les grilles de recherche des combinaisons exhaustives de θ , a et b sont représentés par la grille de recherche

$$\begin{cases} \theta \in K_{100}(10^{-4}, 10^{-1}) \\ a \in K_{100}(10^{-2}, 10) \\ b \in K_{100}(1, 100) \end{cases} \quad (5.19)$$

Les ensembles de recherches de ces trois paramètres sont déterminés suite à des recherches « naïves », et reprennent la notation introduite précédemment.

La non optimalité de la méthode de détermination des paramètres initiaux d' $ADAM$ constitue une perte d'efficacité de l'algorithme relativement peu significative du fait du grand nombre d'itérations qui seront effectuées, et de la rapidité de convergence de l'algorithme.

5.4.2 Détermination de l'hyperparamètre λ , le *learning-rate* initial

Comme évoqué plus tôt dans ce mémoire, la mise en pratique de l'algorithme $ADAM$ nécessite un hyperparamètre dénommé le *learning-rate* initial et noté λ . Celui-ci est déterminé par recherche dichotomique. Deux valeurs extrémales que peut prendre λ sont d'abord choisies arbitrairement. Dans le cas présent ces deux valeurs sont 10^{-6} et 10^{-3} . Puis pour chacune de ces deux valeurs de *learning-rate*, $ADAM$ est lancé sur 100 itérations à partir des paramètres initiaux introduits précédemment et la MSE obtenue à la fin de ces 100 itérations est gardée en mémoire. Cette MSE permet de comparer l'efficacité des *learning-rate* quant à la convergence de l'algorithme. Un *learning-rate* λ_i sera considéré comme meilleur qu'un *learning-rate* λ_j s'il induit une MSE plus faible au bout de 100 itérations d' $ADAM$.

Les étapes d'une itération de la recherche dichotomique se décomposent de la manière détaillée ci-après.

Au départ de l'algorithme, deux moyennes géométriques pondérées sont calculées à partir des deux valeurs initiales possibles de λ notées $borne_1 = 10^{-6}$ et $borne_4 = 10^{-3}$ selon la formule suivante :

$$borne_2 = borne_1^{2/3} \times borne_4^{1/3}$$

$$borne_3 = borne_1^{1/3} \times borne_4^{2/3}$$

La $borne_2$ est, par le biais de la pondération $2/3$ plus « proche » de $borne_1$ que de $borne_4$ et inversement pour $borne_3$. Les MSE après 100 itérations d' $ADAM$ avec $borne_2$ et $borne_3$ sont également calculées. La notation

$MSE(borne_i)$ sera utilisée par la suite pour désigner la MSE renvoyée par ADAM après 100 itérations avec le $learning-rate$ $borne_i$. Les 4 MSE sont comparées, et les critères suivants sont utilisés :

- si $\min_{i \in \{1,4\}} MSE(borne_i) = MSE(borne_1)$: $borne_1$ reste inchangée, $borne_4$ prend la valeur de $borne_2$
- si $\min_{i \in \{1,4\}} MSE(borne_i) = MSE(borne_2)$: $borne_1$ reste inchangée, $borne_4$ prend la valeur de $borne_3$
- si $\min_{i \in \{1,4\}} MSE(borne_i) = MSE(borne_3)$: $borne_4$ reste inchangée, $borne_1$ prend la valeur de $borne_2$
- si $\min_{i \in \{1,4\}} MSE(borne_i) = MSE(borne_4)$: $borne_4$ reste inchangée, $borne_1$ prend la valeur de $borne_3$

A l'issue de la redéfinition de $borne_1$ et $borne_4$, l'itération est répétée tant que le critère d'arrêt choisi n'est pas vérifié. L'espace de recherche $[borne_1 ; borne_4]$ se réduit ainsi autour du minimum recherché.

Le schéma suivant fournit un exemple de quelques itérations de la recherche dichotomique, et permet de mieux appréhender son principe :

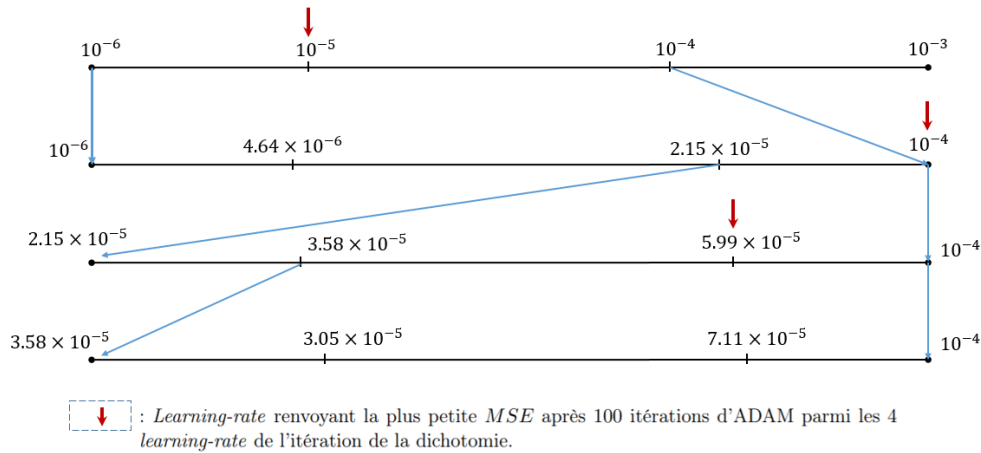


FIGURE 5.6 – Exemple de trois itérations de la recherche dichotomique

5.4.3 Introduction d'un retard dans le début de simulation

Les données d'incidence à disposition présentent, pour la plupart, la particularité d'osciller au départ avec une faible amplitude, puis d'augmenter très vite (peu de temps avant le pic épidémique) pour ensuite retomber tout aussi brutalement et marquer ainsi la fin de l'épidémie. Cette période d'oscillation au départ est globalement mal modélisée par le SIR^+ , ce qui induit une mauvaise calibration dans les faits. Il apparaît donc important d'en tenir compte.

Il est donc proposé de considérer comme point de départ de l'épidémie non plus l'instant $t = 0$, mais un instant défini comme celui qui induit la MSE la plus faible après 100 itérations d'ADAM. Il est toutefois plafonné car si ce n'était pas le cas, alors la MSE minimale serait celle obtenue en débutant l'épidémie à $t = t_{max}$, ce qui serait un non-sens car aucune modélisation par le SIR^+ n'interviendrait.

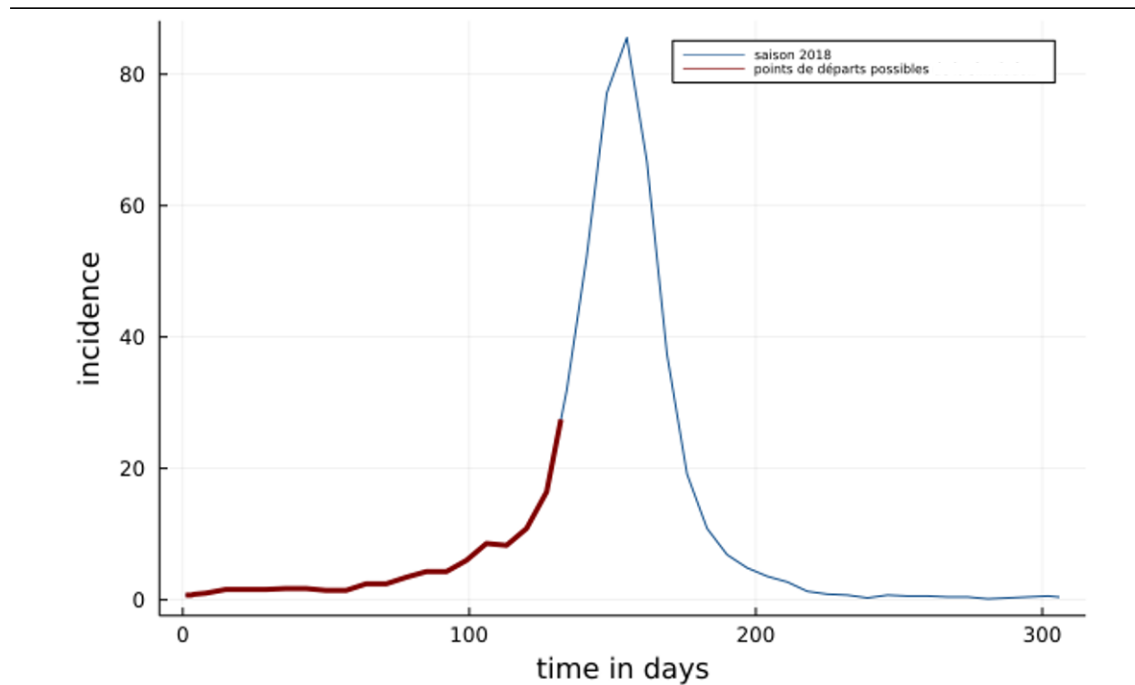


FIGURE 5.7 – Ensemble des points de départ possibles pour la saison 2018

L'ensemble des éléments préalablement nécessaires au lancement de la calibration des paramètres du modèle SIR^+ sont définis. L'algorithme peut donc désormais être appliqué au modèle, et fournira en sortie un jeu de paramètres calibrés pour chaque saison épidémique.

5.5 Résultats de la calibration du modèle SIR^+

Pour chaque saison épidémique, le nombre d'itérations qu'opère *ADAM* est fixé à 1500 afin de déterminer les paramètres optimaux, c'est à dire minimisant la fonction de perte retenue, la *MSE*. En sortie de l'algorithme, la *MSE* selon l'itération présente une convergence nette, puisqu'un palier est atteint pour chaque saison. Ces paliers ne correspondent pas à la même valeur de *MSE* d'une saison à l'autre et n'interviennent pas à la même itération, mais il est déduit des graphes que pour chaque saison un palier est présent avant la fin des 1500 itérations.

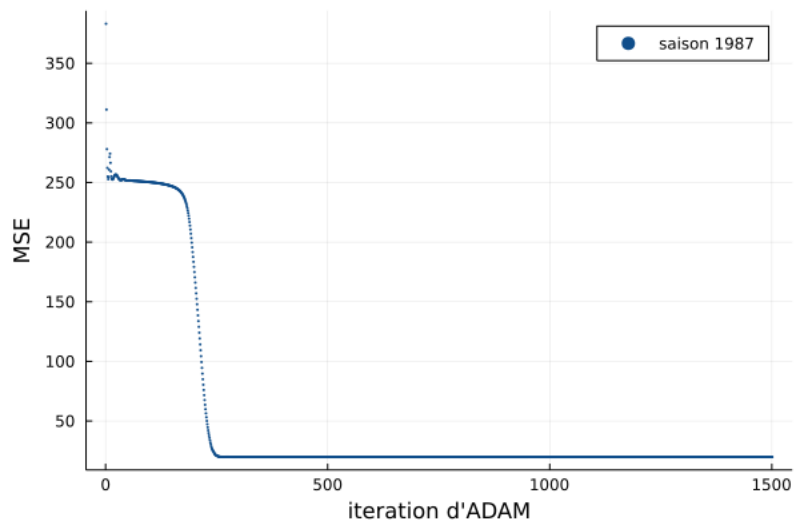


FIGURE 5.8 – Convergence d'ADAM pour la saison 1987

La figure 5.8 illustre la validation de la convergence de l'algorithme *ADAM*. En se restreignant aux itérations d'*ADAM* entre 500 et 1500, la figure devient :

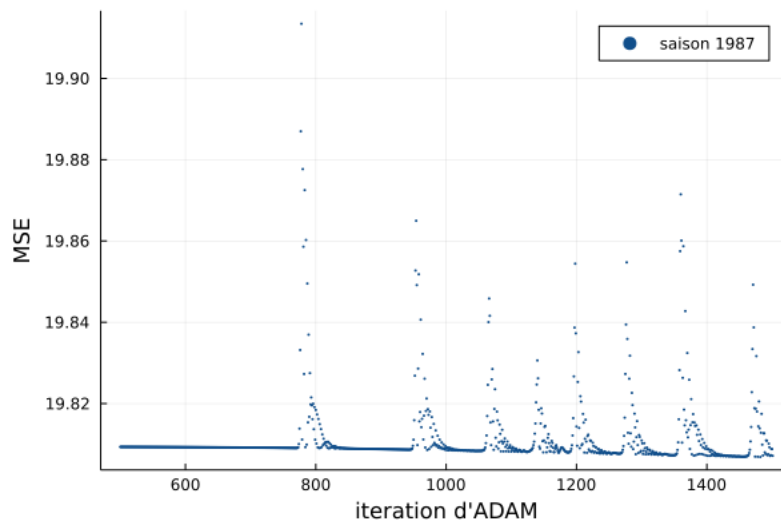


FIGURE 5.9 – Convergence d'ADAM au-delà de l'itération 500 pour la saison 1987

A la vue de ce graphe tronqué, il apparaît qu'*ADAM* converge davantage par « bonds » que linéairement, ce qui ne représente néanmoins pas une précision dégradée.

Ces « bonds » récurrents dans la calibration sont issus de la capacité d'ADAM à « sortir des minima locaux ».

Cette observation pourrait être due à une détermination non optimale du *learning-rate* initial dans un cas de figure similaire à celui présenté en figure 5.2 avec un *learning-rate* trop élevé.

La mise en application de l'algorithme ADAM permet d'obtenir une liste de jeux de paramètres $\Lambda = (\theta, a, b)$, chacun optimal pour la saison épidémique correspondante.

Saison	θ optimal	a optimal	b optimal
1985	0,0173	0,1658	28,7689
1996	0,0190	3,6678	0,8880
1998	0,0221	0,7466	4,3034
2008	0,0265	3,8223	0,6407

TABLE 5.1 – Jeux de paramètres optimaux pour quelques saisons

Pour la saison 2003, qui est ici prise à titre d'exemple illustratif des résultats de la calibration, la courbe d'incidence obtenue en sortie du modèle SIR^+ déterministe, obtenue grâce au jeu de paramètres calibré par ADAM, est comparée dans la figure suivante à la courbe d'incidence réelle observée lors de cette saison :

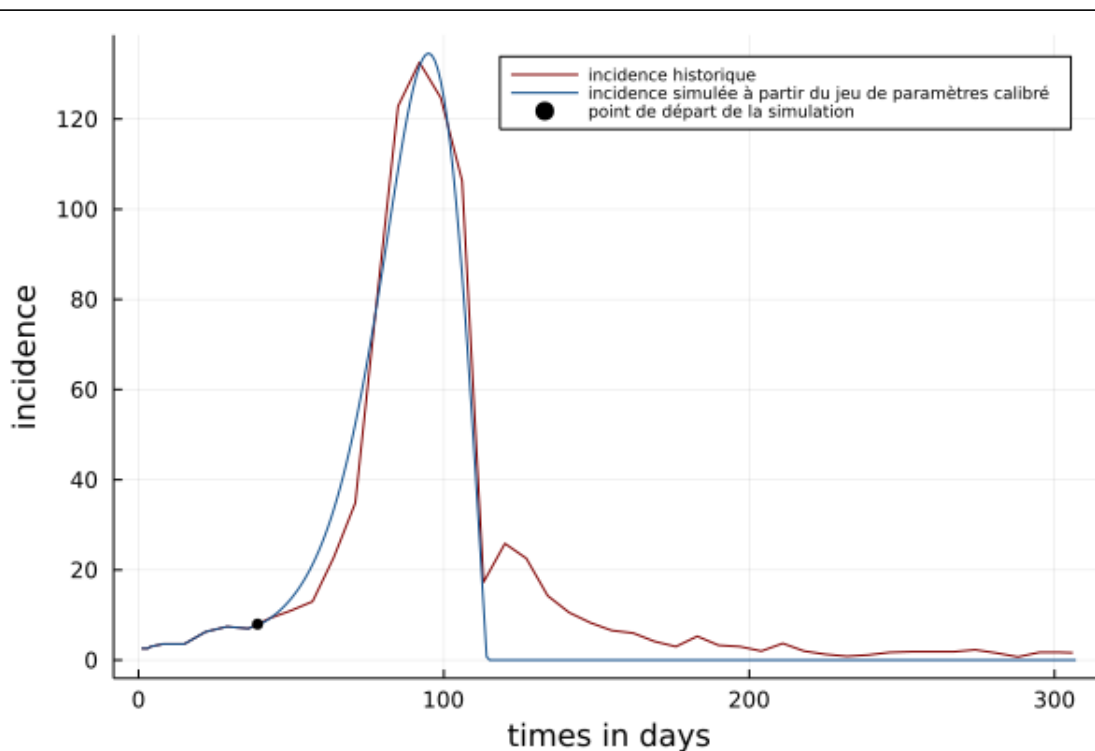


FIGURE 5.10 – Incidence historique et incidence du SIR^+ pour la saison 2003

Le point de départ, ici à $t = 39$, est celui retenu dans la calibration : c'est à partir de ce point de départ que le modèle est entré en jeu, après avoir pris en compte les données historiques pour $t < 39$.

Il est désormais possible de simuler les épidémies grippales que fournit le modèle, et ce pour chaque jeu de paramamètres.

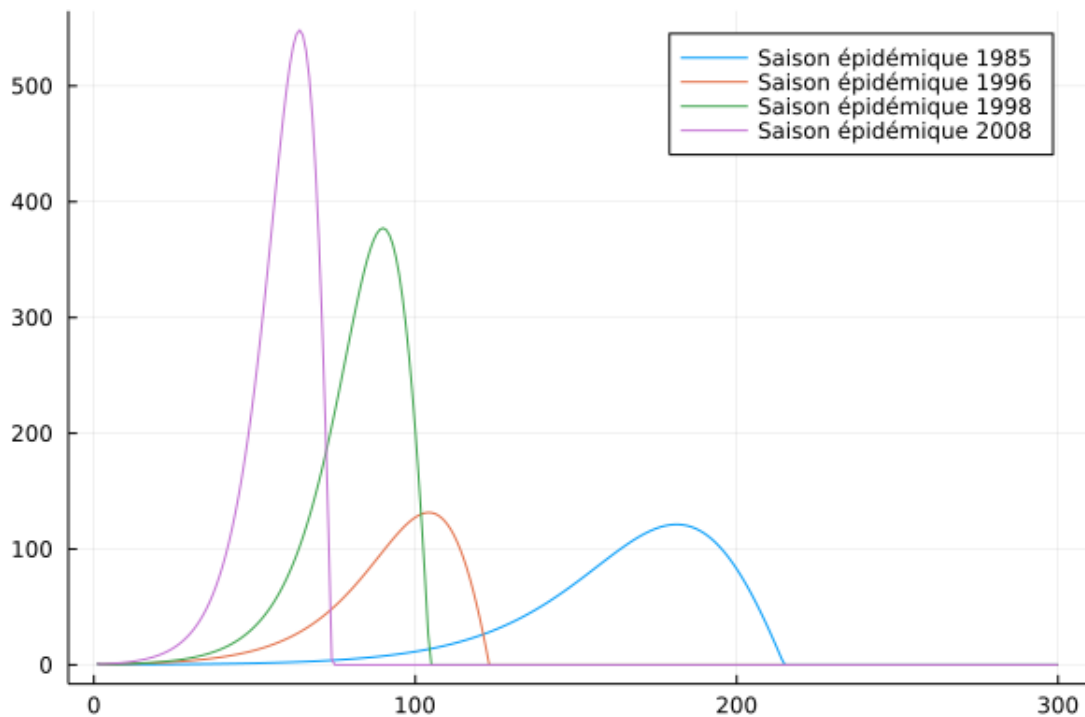


FIGURE 5.11 – Exemples de courbes déterministes obtenues selon quatre jeux de paramètres calibrés

Les simulations ci-dessus sont produites en faisant l'hypothèse qu'à « l'instant initial » de l'épidémie, $t_0 = 0$, la population totale, ici 100 000 personnes, se répartit parmi les états du modèle de la façon suivante :

- S contient 999 999 individus
- I contient 1 individu
- R contient 0 individu

Ainsi l'incidence initiale vaut 1 pour chacune de ces simulations.

Plusieurs informations peuvent être déduites de la figure 5.11 : tout d'abord le modèle déterministe est très sensible au jeu de paramètres en entrée. Chaque courbe de cette figure est la courbe obtenue avec un jeu de paramètre calibré sur l'incidence d'une saison épidémique. Cependant les courbes présentent une grande variété, à la fois dans la durée de l'épidémie (certaines s'éteignent au bout de seulement 80 jours, d'autres durent au-delà de 200 jours) mais aussi dans leur « pic ». Certains jeux de paramètre produisent une incidence qui peut s'élever à plus de 500 nouveaux infectés par jour, alors que d'autres ne dépassent jamais les 120 nouveaux cas par jour. De plus, le modèle déterministe présente une trajectoire similaire quel que soit le jeu de paramètre utilisé.

Le nombre total d'infectés est également obtenu en sortie de modèle pour chacune des saisons considérées. En voici quelques exemples :

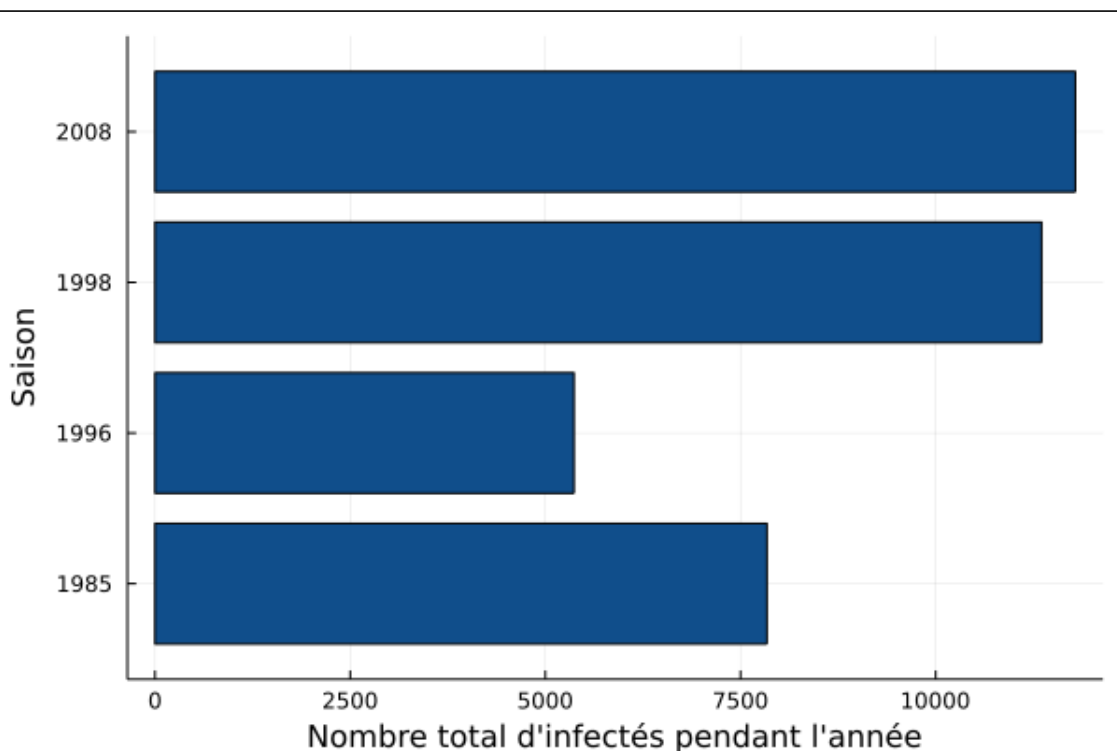


FIGURE 5.12 – Nombre d'infectés total par saison

Le modèle déterministe est désormais complètement construit, et ses paramètres optimaux au regard des données d'incidence de la grippe en France ont été déterminés grâce à l'adaptation d'un algorithme d'optimisation initialement utilisé pour des problématiques de *machine-learning*, *ADAM*.

Il est proposé dès maintenant, ce point sera motivé par la partie suivante de ce mémoire, de déterminer un jeu de paramètres $\Lambda = (\theta, a, b)$ « central » qui sera celui retenu si besoin est de ne retenir qu'un seul jeu de paramètres. Chaque jeu de paramètres va donc être testé sur l'ensemble des saisons épidémiques, et le jeu de paramètres qui minimise la MSE globale sera retenu comme le jeu de paramètres de référence dans la suite des travaux réalisés. Ainsi pour un jeu de paramètres Λ_i donné, la MSE_{1984} est calculée sur la saison 1984, et cette MSE_{1984} est gardée en mémoire. Puis le même jeu de paramètres est utilisé pour calculer la MSE_{1985} sur la saison suivante, et cette MSE_{1985} est ajoutée à MSE_{1984} pour obtenir la MSE globale sur les deux saisons. Cette procédure est répétée pour toutes les saisons épidémiques, ce qui permet de connaître l'erreur globale, incarnée par cette $MSE_{globale}$ qu'engendre ce jeu de paramètres Λ_i .

Enfin, la $MSE_{globale}$ est calculée pour chaque Λ_i , et le jeu de paramètre retenu est celui qui induit la $MSE_{globale}$

la plus faible, à savoir le jeu de paramètre correspondant à la saison 1985, et qui vaut

$$\Lambda_* = \begin{pmatrix} \theta_* \\ a_* \\ b_* \end{pmatrix} = \begin{pmatrix} 0,0173 \\ 0,1658 \\ 28,7689 \end{pmatrix} \quad (5.20)$$

La simulation de l'épidémie à partir de ce jeu de paramètres est la suivante :

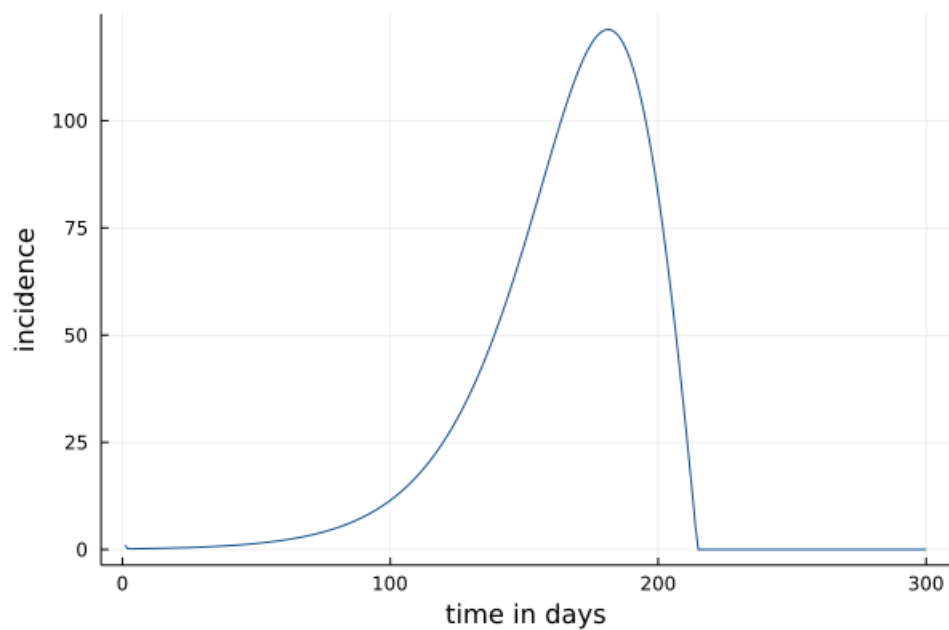


FIGURE 5.13 – Incidence du modèle SIR^+ , avec le jeu de paramètres (5.20)

La figure qui suit présente donc chacune des épidémies réelles observées par le réseau Sentinelles et conservées pour les besoins de la calibration, ainsi que la courbe 5.13 :

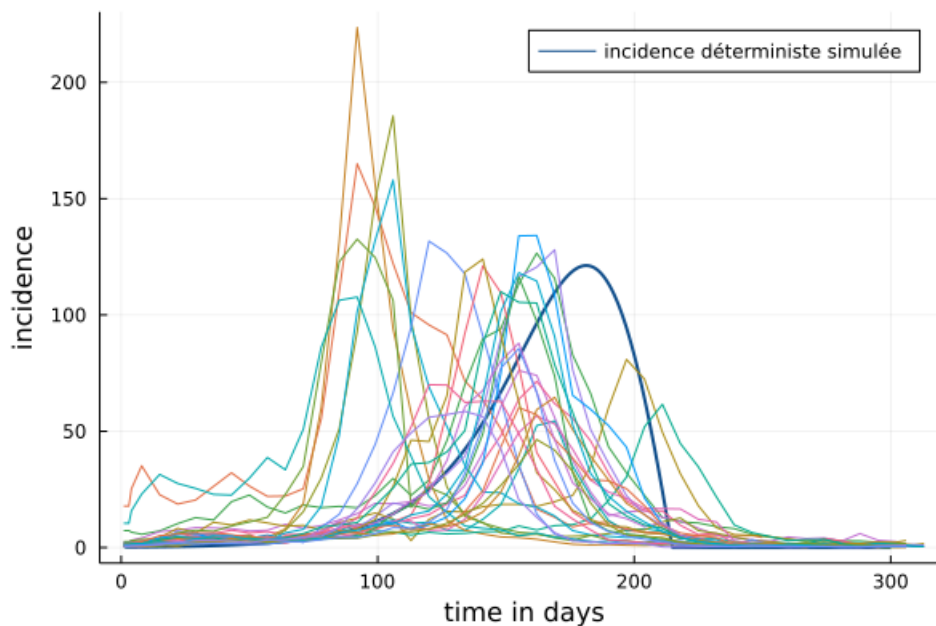


FIGURE 5.14 – Incidence du modèle SIR^+ et incidences historiques

Il ressort de cette figure que les épidémies sont décalées les unes par rapport aux autres, et donc que leurs pics d'incidence ne se superposent pas : ce point d'attention sera traité par la suite.

La courbe d'incidence du SIR^+ semble plus « écrasée » que sur la figure 5.13, mais cela provient uniquement du fait que la figure 9.1 présente un intervalle de temps plus restreint.

Le modèle SIR^+ développé jusqu'à présent tout au long de ce mémoire représente une modélisation convenable de l'incidence moyenne attendue des épidémies de grippe saisonnière. Cependant une épidémie présente par définition un caractère aléatoire, et ce même sans considérer par exemple que plusieurs souches du virus circulent. Il semble donc pertinent de chercher à complexifier le modèle SIR^+ déterministe proposé jusqu'ici, en y intégrant une composante stochastique qui représenterait le caractère aléatoire du phénomène étudié.

Troisième partie

Construction du modèle SIR^+ stochastique

Chapitre 6

Tentative d'obtention d'un modèle stochastique par ajustement d'une loi sur la distribution des paramètres

La première tentative d'introduction d'un phénomène aléatoire s'est d'abord portée sur l'obtention d'une distribution des paramètres du modèle SIR^+ . En effet, l'algorithme *ADAM* fournit en sortie de calibration 28 jeux de paramètres, détaillés dans la partie précédente. 28 valeurs de θ , 28 valeurs de a ainsi que 28 valeurs de b sont ainsi disponibles. Il est donc possible de tracer les histogrammes des 28 valeurs de θ de la manière suivante :

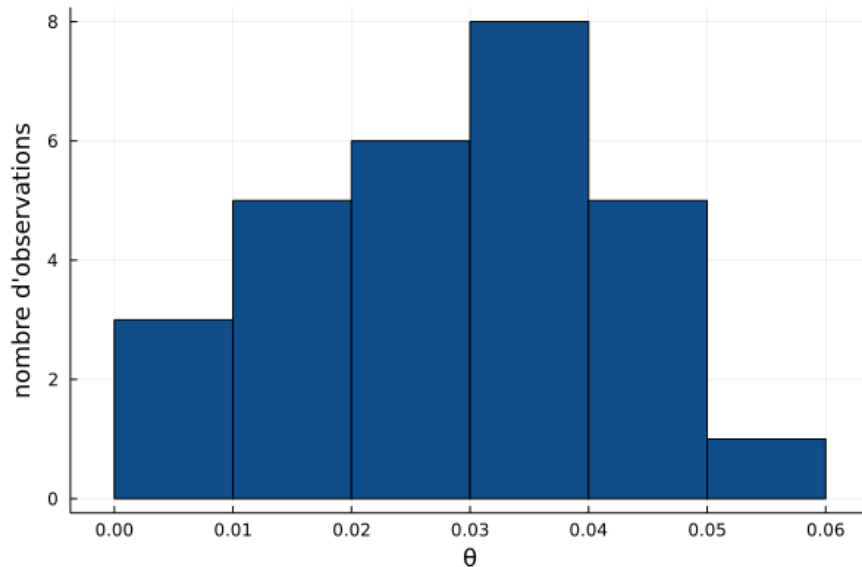


FIGURE 6.1 – Distribution empirique des valeurs de θ

L'idée est alors d'ajuster une loi sur ces données, pour avoir une distribution de ce paramètre à partir de la distribution empirique.

Si un ajustement en loi peut être effectué sur le paramètre θ , ce n'est pas le cas pour les paramètres a et b pris indépendamment. Le couple (a, b) est utilisé dans la calibration au niveau de la détermination des dynamiques de sorties de l'état I et d'entrée dans l'état R , et ces deux paramètres sont dépendants l'un de l'autre.

Pour réussir à obtenir un tirage de ces paramètres a et b , une procédure de *moment-matching* peut être mise en place. Les espérances et variances de la variable aléatoire T représentant la durée passée dans l'état I sont calculées, pour chaque couple de paramètres (a, b) relatif à une saison épidémique, selon les formules usuelles de la loi *Gamma* :

$$\begin{cases} \mathbb{E}(T) &= a \times b \\ \mathbb{V}(T) &= a \times b^2 \end{cases} \quad (6.1)$$

Les distributions empiriques des espérances et variances des lois *Gamma* calibrées peuvent ainsi être obtenues :

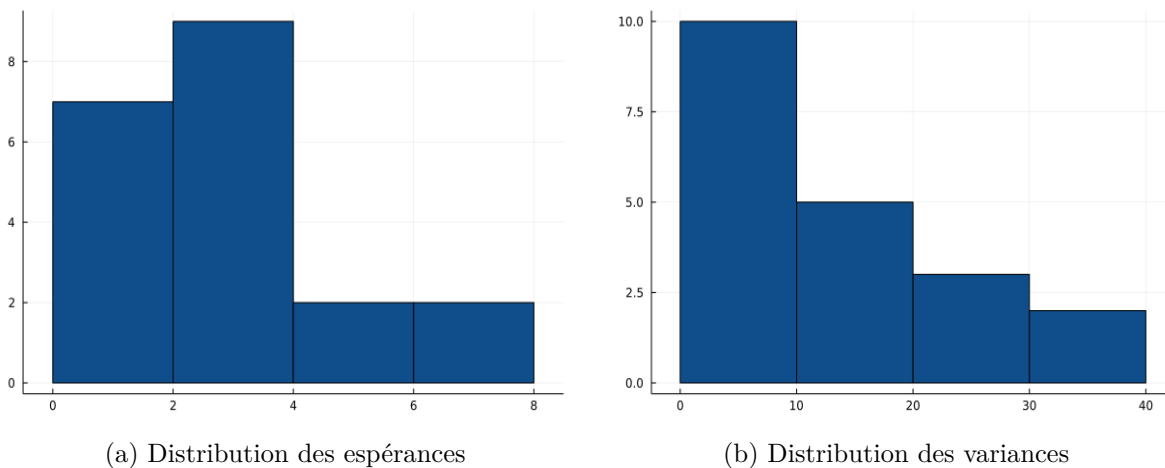


FIGURE 6.2 – Distributions des moments empiriques des lois *Gamma* calibrées

Après ajustement d'une loi sur la distribution des espérances, et ajustement d'une autre loi sur la distribution des variances, il est possible d'obtenir une valeur de l'espérance et une valeur de la variance par tirage aléatoire selon ces deux lois. En notant \mathcal{E} et \mathcal{V} ces dernières, les valeurs de a et de b correspondantes sont obtenues grâce aux relations suivantes, implications directes des propriétés (6.2) :

$$\begin{cases} a &= \frac{\mathcal{E}^2}{\mathcal{V}} \\ b &= \frac{\mathcal{E}}{\mathcal{V}} \end{cases} \quad (6.2)$$

Un scénario stochastique correspond donc ici à un modèle SIR^+ alimenté des paramètres θ , a et b obtenus par tirage aléatoire selon les lois obtenues par les procédures citées ci-dessus.

Cependant la mise en pratique de ces procédures aboutit à des scénarios stochastiques bien trop explosifs au re-

gard des incidences historiques. En d'autres termes, les courbes d'incidences simulées de cette manière présentent un écart-type bien trop élevé en comparaison de l'écart-type observé sur l'échantillon des données Sentinelles. Ce phénomène peut vraisemblablement s'expliquer par la faible taille de l'échantillon des paramètres : 28 observations peuvent difficilement aboutir à un ajustement de qualité sur la loi jointe des paramètres, d'autant plus que le modèle est très sensible aux variations de ses paramètres.

Il serait donc préférable, dans la mesure du possible, d'introduire une composante stochastique directement dans les équations régissant les dynamiques inter-états du modèle, tout en conservant comme scénario central le scénario déterministe.

Chapitre 7

Tentative d'obtention d'un modèle stochastique par refonte des dynamiques de transition déterministes

7.1 Proposition de refonte de la dynamique d'incidence

Suite à l'impossibilité de prendre en compte le caractère aléatoire intrinsèque au phénomène épidémique à posteriori, il est tout naturel d'essayer d'introduire cette propriété au sein-même des dynamiques du modèle.

7.1.1 Rappel de la dynamique déterministe

Les équations modélisant les dynamiques du modèle déterministe SIR^+ sont rappelées ici :

$$\left\{ \begin{array}{l} \mathbb{E}[S_{t+1} - S_t] = -m\theta \times \frac{S_t I_t}{N} \\ \mathbb{E}[I_{t+1} - I_t] = m\theta \times \frac{S_t I_t}{N} - \left[m\theta \sum_{k=1}^t [(F_T(t+1-k) - F_T(t-k))I_{k-1}] + (F_T(t+1) - F_T(t)) \times I_0 \right] \\ \mathbb{E}[R_{t+1} - R_t] = m\theta \sum_{k=1}^t [F_T(t+1-k) - F_T(t-k)]I_{k-1} + [F_T(t+1) - F_T(t)] \times I_0 \end{array} \right. \quad (7.1)$$

Avec :

$$\left\{ \begin{array}{l} -S_t \text{ le nombre de personnes susceptibles d'être infectées à l'instant } t \\ -I_t \text{ le nombre de personnes infectées à l'instant } t \\ -R_t \text{ le nombre de personnes déjà rétablies à l'instant } t \\ -N \text{ le nombre total d'individus} \\ -m \text{ le nombre moyen de contacts qu'a une personne pendant } \Delta_t = 1 \text{ jour} \\ -\theta \text{ la probabilité qu'un contact entre un individu susceptible et un individu infecté} \\ \quad \text{engendre la contamination de l'individu susceptible} \\ -T \sim \Gamma(a, b) \text{ la variable aléatoire représentant la durée passé dans l'état infecté} \\ -F_T(t) \text{ la fonction de répartition de cette variable aléatoire au temps } t \end{array} \right. \quad (7.2)$$

Et les conditions initiales :

$$\left\{ \begin{array}{l} S(0) = S_0 \\ I(0) = I_0 \\ R(0) = R_0 \\ S(1) = S(0) - m\theta \times \frac{S(0)I(0)}{N} \\ I(1) = I(0) + m\theta \times \frac{S(0)I(0)}{N} - F_T(1) \times I(0) \\ R(1) = R(0) + F_T(1) \times I(0) \end{array} \right. \quad (7.3)$$

Dorénavant, le modèle déterministe sera noté SIR_{det}^+ , et le modèle stochastique proposé sera noté SIR_{sto}^+ .

Les nombres d'individus dans chaque état du modèle déterministe à un instant t seront notés S_t^{det} , I_t^{det} , R_t^{det} (respectivement S_t^{sto} , I_t^{sto} , R_t^{sto} pour le modèle stochastique).

Les incidences déterministes (resp. stochastiques), correspondant au nombre de nouveaux infectés, à un instant t seront elles notées $X_t^{det} = S_t^{det} - S_{t+1}^{det}$ (resp. $X_t^{sto} = S_t^{sto} - S_{t+1}^{sto}$).

7.1.2 Introduction d'un processus stochastique

La proposition de construction du modèle stochastique SIR_{sto}^+ dans le cadre de ce mémoire prend la forme de l'ajout, à la dynamique d'incidence, d'un processus stochastique centré, qui pourrait être par exemple un bruit blanc. En notant $(\epsilon_t)_{t>0}$ ce bruit blanc, alors $\forall (t_i, t_j) \in (\mathbb{N}^{+*})^2, t_i \neq t_j$:

$$\left\{ \begin{array}{l} \mathbb{E}[\epsilon_{t_i}] = 0 \\ \mathbb{V}[\epsilon_{t_i}] = \sigma^2 \\ \mathbb{E}[\epsilon_{t_i}\epsilon_{t_j}] = 0 \end{array} \right. \quad (7.4)$$

La propriété de centralité évoquée précédemment permet que les dynamiques du modèle SIR_{sto}^+ soient égales en espérance aux dynamiques déterministes. Il est donc désormais nécessaire de vérifier que le modèle stochastique se comporte en moyenne de la même manière que le modèle déterministe, ou du moins qu'il vérifie certaines propriétés suffisantes à la validation de sa cohérence avec le modèle déterministe.

L'expression de l'incidence stochastique rattachée au modèle SIR_{sto}^+ s'exprime alors comme :

$$X_t^{sto} = S_{t+1}^{sto} - S_t^{sto} = -m\theta \frac{S_t^{sto} I_t^{sto}}{N} + \epsilon_t \quad (7.5)$$

Pour simplifier la lecture dans la suite des travaux, la notation $\mu = \frac{m\theta}{N}$ est introduite.

7.2 Adéquation entre la dynamique déterministe et l'espérance de la dynamique stochastique : incompatibilité avec la structure du modèle

Comme énoncé précédemment, il est important d'avoir une égalité entre l'espérance du processus stochastique et le déterministe. Afin de déterminer si l'adéquation entre la dynamique déterministe et l'espérance de la dynamique stochastique est vérifiée, les premières expressions théoriques des équations (7.1) sont établies, puis leur espérance est calculée. Les 3 premières expressions de S_t^{sto} et I_t^{sto} sont :

$$\left\{ \begin{array}{l} S_1 = S_0 - \mu S_0 I_0 + \epsilon_1 \\ I_1 = I_0 (1 + \mu S_0 - F_T(1)) \\ \\ S_2 = f_1 + f_2 \epsilon_1 + \epsilon_2 \\ I_2 = f_3 + f_4 \epsilon_1 \\ \\ S_3 = f_5 + f_6 \epsilon_1 + f_7 \epsilon_1^2 + f_8 \epsilon_2 + f_9 \epsilon_1 \epsilon_2 + \epsilon_3 \\ I_3 = f_{10} + f_{11} \epsilon_1 - f_7 \epsilon_1^2 + (1 - f_8) \epsilon_2 - f_7 \epsilon_1 \epsilon_2 \end{array} \right.$$

Avec les fonctions déterministes suivantes :

$$\left\{ \begin{array}{l} f_5 = f_1 - \mu f_1 f_3 \\ f_6 = f_2 - \mu f_2 f_3 - \mu f_1 f_4 \\ f_7 = -\mu f_2 f_4 \\ f_8 = 1 - \mu f_3 \\ f_9 = -\mu f_4 \\ f_{10} = f_3 + \mu f_1 f_3 - [m\theta(F_T(2) - F_T(1)) \\ \quad + m\theta F_T(1)(1 + \mu S_0 - F_T(1)) - (F_T(3) - F_T(2))] I_0 \\ f_{11} = f_4 + \mu f_1 f_4 + \mu f_2 f_3 \end{array} \right.$$

Peuvent en être déduites les expressions des trois premières valeurs de $\mathbb{E}[X_t^{sto}]$, qui sont ensuite comparées avec leur pendant déterministe :

$$\mathbb{E}[X_0^{sto}] = \mathbb{E}[X_0^{det}]$$

$$\mathbb{E}[X_1^{sto}] = \mathbb{E}[X_1^{det}]$$

$$\mathbb{E}[X_2^{sto}] = f_5 + f_7\sigma^2 - f_1$$

Or $\mathbb{E}[X_2^{det}] = f_5 - f_1$ donc :

$$\begin{aligned} \mathbb{E}[X_2^{sto}] = \mathbb{E}[X_2^{det}] &\Leftrightarrow f_7 = 0 \\ &\Leftrightarrow [1 + \mu S_0 - F_T(1) = 0] \cup [1 - \mu I_0 - \mu^2 S_0 I_0 + \mu I_0 F_T(1) = 0] \end{aligned}$$

$\forall x \in \mathbb{R}^{+*}, F_T(x) < 1$ donc $F_T(1) < 1$ et donc $F_T(1) < 1 + \mu S_0$ car $\mu > 0$ et $S_0 > 0$. Ainsi :

$$\begin{aligned} f_7 = 0 &\Leftrightarrow 1 - \mu I_0 - \mu^2 S_0 I_0 + \mu I_0 F_T(1) = 0 \\ &\Leftrightarrow -\frac{m^2 S_0 I_0}{N^2} \theta^2 - \frac{m I_0}{N} \theta + [1 + \mu I_0 F_T(1)] = 0 \end{aligned}$$

qui est une équation du second degré en θ .

Le déterminant vaut $\Delta = [-\frac{m I_0}{N}]^2 - 4 \times [-\frac{m^2 S_0 I_0}{N^2}] \times (1 + \mu I_0 F_T(1)) = \frac{m}{N} \sqrt{I_0^2 + 4 S_0 I_0 (1 + \mu I_0 F_T(1))}$

$r = \sqrt{\Delta} = \frac{m}{N} \sqrt{I_0^2 + 4 S_0 I_0 (1 + \mu I_0 F_T(1))}$

Et donc : $\theta_{1,2} = \frac{\frac{m I_0}{N} \pm r}{-2 \frac{m^2 S_0 I_0}{N^2}} = \frac{m N I_0 \pm r N^2}{-2 m^2 S_0 I_0}$. Or $\theta > 0$ donc finalement la seule et unique solution est donnée par :

$\mathbb{E}[X_2^{sto}] = \mathbb{E}[X_2^{det}]$ est $\theta_s = \frac{m N I_0 - r N^2}{-2 m^2 S_0 I_0}$

$$\mathbb{E}[X_2^{sto}] = \mathbb{E}[X_2^{det}] \Leftrightarrow \theta_s = \frac{m N I_0 - r N^2}{-2 m^2 S_0 I_0} \quad (7.6)$$

Dans l'expression de cette relation, seul r dépend des deux paramètres a et b .

Il ressort de cette étude de l'adéquation en espérance entre le modèle stochastique et le modèle déterministe que le paramètre θ doit impérativement être fixé en fonction des paramètres initiaux pour espérer obtenir cette propriété, ne serait-ce que jusqu'à $t = 2$. Or le paramètre θ est un paramètre fondamental du modèle SIR_{det}^+ , et il a été calibré plus en amont dans ce mémoire. Le fixer selon (7.6) modifierait grandement l'interprétation qui a été faite de ce paramètre, et reviendrait à considérer que toutes les épidémies de grippe possèdent la même

capacité d'infection. En d'autres termes l'épidémie serait entièrement caractérisée par la durée d'infection par le virus sous-jacent.

Enfin, cette relation déjà contraignante sur θ n'est valable que pour $t = 2$. Si la réalisation de la démarche ci-dessus pour les $t > 2$ aboutit à une accumulation des conditions sur θ , alors la limite de cette méthode n'en serait que renforcée.

Si cette piste ne s'avère pas concluante, il pourrait cependant potentiellement être possible de calculer les espérances du modèle stochastique qui remplaceraient alors le déterministe, et auraient fait l'objet d'une nouvelle calibration.

7.3 Difficultés à obtenir une relation de récurrence entre les espérances stochastiques successives

A défaut d'une égalité parfaite en espérance entre les deux modèles SIR_{sto}^+ et SIR_{det}^+ , il serait convenable d'obtenir une relation de récurrence entre l'espérance du modèle stochastique à l'instant $t + 1$ et son espérance à l'instant t afin de globalement être capable de se ramener au déterministe.

Cette expression prendrait la forme $\forall t \geq 0, \boxed{f(\mathbb{E}[S_t^{sto}]) = \mathbb{E}[S_{t+1}^{sto}]}$ où f est une transformation quelconque à déterminer, indépendante des états du modèle SIR_{sto}^+ .

Si une telle relation peut être déterminée, la récurrence donnerait $\mathbb{E}[S_t^{sto}] = f^t(\mathbb{E}[S_0^{sto}]) = f^t(S_0)$, et l'espérance du processus stochastique à tout instant t serait alors connue, sans avoir besoin de connaître toutes valeurs prise par cette espérance entre 0 et t . La relation précédente pourrait alors être transposée à X_t^{sto} par le biais de la relation $\mathbb{E}[X_t^{sto}] = \mathbb{E}[S_t^{sto} - S_{t+1}^{sto}] = \mathbb{E}[S_t^{sto}] - \mathbb{E}[S_{t+1}^{sto}] = f^t(S_0) - f^{t+1}(S_0)$

L'hypothèse $\forall t > 0, \epsilon_{(t-1,t)} \sim \mathcal{N}(0, \sigma^2)$ est ici faite. Ceci induit les propriétés suivante, M_k désignant le moment d'ordre k de ϵ_t :

$$\begin{cases} M_{2k} = \mathbb{E}[\epsilon_{(t-1,t)}^{2k}] = \frac{(2k)!}{2^k k!} \sigma^{2k} \\ M_{2k+1} = \mathbb{E}[\epsilon_{(t-1,t)}^{2k+1}] = 0 \end{cases} \quad (7.7)$$

Les premières expressions d'une relation liant $\mathbb{E}[S_{t+1}^{sto}]$ et $\mathbb{E}[S_t^{sto}]$ sont détaillées ci-après. Par souci de clarté des expressions, $\mathbb{E}[S_t^{sto}]$ sera désormais notée Φ_t .

$$\begin{cases} \Phi_1 = \Phi_0 - \mu I_0 \Phi_0 \\ \Phi_2 = \Phi_1 - \mu I_0 (1 + \mu I_0 - F_T(1)) \Phi_1 \\ \Phi_3 = \Phi_2 - \mu f_3 \Phi_2 - \mu f_4 (1 - f_4) \sigma^2 \\ \Phi_4 = \Phi_3 - \mu (f_5 f_{10} - f_7 \sigma^2 (\Phi_3 + f_{10} + 2f_7 \sigma^2 + f_9 \sigma^2) \\ \quad + f_6 f_{11} \sigma^2 + f_8 (1 - f_8) \sigma^2) \end{cases} \quad (7.8)$$

Bien qu'il soit possible de réécrire, pour ces 4 premières expressions, de manière globale $\Phi_{t+1} = \Phi_t - \mu F(\Phi_t)$, il s'avère très vite difficile de déterminer une expression générique de la transformation F , et la complexification très rapide des expressions, pour seulement les 4 premières valeurs de t , ne permettent pas de conclure à l'existence d'une formule de récurrence liant Φ_{t+1} et Φ_t .

A la suite de cette nouvelle tentative infructueuse, la méthode de construction d'un modèle SIR_{sto}^+ par le biais de l'ajout d'un processus stochastique au sein des équations des dynamiques de ce modèle est abandonnée, au profit d'une méthode présentée par la suite qui elle permet l'adéquation souhaitée entre l'espérance de l'incidence du SIR_{sto}^+ et l'incidence du SIR_{det}^+ .

Chapitre 8

Transformation du déterministe SIR_{det}^+ par le biais d'un mouvement brownien géométrique

La cohérence entre le modèle SIR_{sto}^+ et le modèle SIR_{det}^+ en espérance ne semble pas obtainable en passant à une dynamique de transition stochastique entre les états du modèle pour les raisons qui viennent d'être détaillées. C'est pourquoi le choix est fait d'appliquer une transformation stochastique au modèle SIR_{det}^+ via un mouvement brownien géométrique.

Le modèle SIR_{sto}^+ devient alors un *proxy* stochastique du modèle déterministe : tout l'intérêt de ce *proxy*, qui est lié au modèle déterministe car dérivé de ce dernier, est de réussir à capter le phénomène aléatoire présent au sein d'une épidémie et dont le modèle SIR_{det}^+ ne parvient pas à rendre compte.

Dans un premier temps, quelques rappels théoriques sur les mouvements browniens géométriques sont exposés, puis la mise en application du *proxy* est détaillée.

8.1 Rappel des propriétés des mouvements browniens géométriques

Un processus stochastique $(Z_t)_{t>0}$ est appelé mouvement brownien géométrique s'il vérifie l'équation différentielle stochastique :

$$dZ_t = \gamma Z_t dt + \sigma Z_t dW_t \tag{8.1}$$

Ici γ et σ sont des constantes, respectivement appelées « dérive en pourcentage » et « pourcentage de volatilité »,

et W_t est un processus de Wiener.

La résolution de cette équation différentielle stochastique, qui ne sera pas détaillée dans le cadre de ce mémoire, est donnée par la formule :

$$\boxed{Z_t = Z_0 e^{((\gamma - \frac{\sigma^2}{2})t + \sigma W_t)}} \quad (8.2)$$

Où Z_0 est la valeur initiale du mouvement brownien géométrique $(Z_t)_{t>0}$.
Les moments du processus $(Z_t)_{t>0}$ sont :

$$\begin{cases} \mathbb{E}[Z_t] &= Z_0 e^{\gamma t} \\ \mathbb{V}[Z_t] &= Z_0^2 e^{2\gamma t} (e^{\sigma^2 t} - 1) \end{cases} \quad (8.3)$$

Une fois ce processus stochastique géométrique construit, il est utilisé pour transformer la dynamique déterministe des incidences et ainsi obtenir une version stochastique de celle-ci.

8.2 Transformation appliquée à l'incidence déterministe

La construction du modèle stochastique SIR_{sto}^+ est étroitement liée au modèle déterministe SIR_{det}^+ proposé en amont et au jeu de paramètres retenus (5.20). Il contiendra ainsi deux composantes : une déterministe, portée par le modèle SIR_{det}^+ , et une stochastique qui est le mouvement brownien géométrique $(Z_t)_{t>0}$ présenté précédemment. Le mouvement brownien géométrique correspond en fait à l'écart modélisé entre le déterministe et la réalité et sera par la suite nommé « taux d'erreur ».

Ainsi, la dynamique des nouveaux infectés du modèle SIR_{sto}^+ est donnée par la relation :

$$X_t^{sto} = X_t^{det} \times Z_t \quad (8.4)$$

La valeur initiale Z_0 du processus $(Z_t)_{t>0}$ est fixée à 1 et la dérive en pourcentage, γ , est elle fixée à 0. Cela a pour conséquence de faire coïncider X_t^{sto} et X_t^{det} en espérance.

Enfin σ n'est pas fixé a priori, et fera l'objet d'une calibration sur les données d'incidence de Sentinelles.

Ainsi :

$$\boxed{X_t^{sto} = X_t^{det} \times e^{(-\frac{\sigma^2}{2}t + \sigma W_t)}} \quad (8.5)$$

Le choix de cette modélisation qui s'apparente à un taux d'erreur, par multiplication du déterministe par un processus stochastique géométrique, est motivé par l'interprétation qu'il est possible de faire d'une telle transformation. Bien qu'étant un *proxy* et non un processus issu d'une révision stochastique des dynamiques, cette transformation permet de répliquer les caractéristiques importantes attendues d'une telle révision.

L'erreur de modélisation en épidémiologie a intuitivement plus d'impact sur le reste de l'épidémie si elle est commise au début de l'épidémie qu'à la fin : si le nombre d'infectés est surestimé au début, l'épidémie peut s'accélérer brutalement, alors que si ce nombre est surestimé peu de temps avant la fin de l'épidémie, cela n'aura pas d'impact aussi significatif sur l'incidence totale. Les erreurs d'une modélisation par un mouvement brownien géométrique se composent. Ainsi une modélisation de la forme $X_t^{sto} = X_t^{det} + \epsilon_t$ avec ϵ_t un mouvement brownien ne permettrait pas de rendre compte de cette répercussion multiplicative d'une éventuelle erreur commise au début de l'épidémie.

De plus, d'un point de vue opérationnel, le mouvement brownien géométrique étant toujours positif, une telle modélisation assure que l'incidence simulée aura la même propriété, là où un mouvement brownien classique pourrait prendre des valeurs négatives.

Les expressions des moments et de l'écart-type de $(X_t^{sto})_{t>0}$ peuvent s'écrire $\forall t > 0$,

$$\begin{cases} \mathbb{E}[X_t^{sto}] &= X_t^{det} \\ \mathbb{V}[X_t^{sto}] &= (X_t^{det})^2 \times (e^{\sigma^2 t} - 1) \\ \nu_t &= \sqrt{\mathbb{V}[X_t^{sto}]} = X_t^{det} \sqrt{e^{\sigma^2 t} - 1} \end{cases} \quad (8.6)$$

Le modèle SIR_{sto}^+ est désormais construit, et apparaît cohérent avec le modèle SIR_{det}^+ dont il est une transformation. Comme attendu, l'espérance de la modélisation stochastique est bien égale au déterministe.

La figure suivante illustre des scénarios stochastiques de l'incidence en sortie du modèle SIR_{sto}^+ , pour un paramètre σ choisi arbitrairement pour le moment.

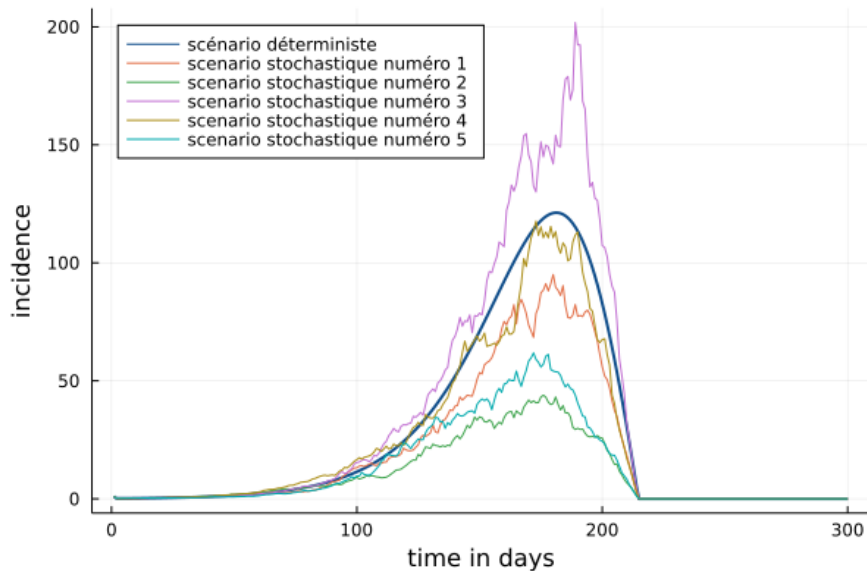


FIGURE 8.1 – 5 scénarios stochastiques issus de la transformation du scénario central, pour $\sigma = 0,05$

Le paramètre σ du processus stochastique sur la figure 8.1 n'est pas choisi ici de manière optimale. Il convient donc de le calibrer à partir des données d'incidence de Sentinelles, afin que l'écart à la moyenne du modèle stochastique se rapproche au mieux de l'empirique.

Chapitre 9

Calibration du taux d'erreur du modèle SIR_{sto}^+

9.1 Minimisation de l'écart quadratique entre les taux d'erreur théoriques et les taux d'erreur empiriques

La métrique choisie pour la calibration de la volatilité σ du modèle stochastique SIR_{sto}^+ est la MSE , qui a déjà été utilisée précédemment pour calibrer le modèle déterministe. Ici un seul paramètre doit être ajusté, et les MSE pour des valeurs de σ prises dans une grille de recherche sont comparées.

Le processus stochastique géométrique $(Z_t)_{t \geq 0}$ modélise le taux d'erreur entre les données réelles et le modèle déterministe X_t^{det} . Il est possible de calculer d'une part l'écart-type empirique entre les données réelles et le modèle déterministe, et d'autre part l'écart type théorique du processus stochastique géométrique Z_t . En faisant varier le paramètre σ de ce processus stochastique, il est alors possible de comparer l'écart-type empirique à chaque pas de temps t avec l'écart-type théorique.

Cet écart-type, ou taux de volatilité noté τ_t , permet de rendre compte du pourcentage de volatilité entre le modèle stochastique et le modèle déterministe. Son expression prend la forme :

$$\begin{cases} Z_t &= e^{(-\frac{\sigma^2}{2}t + \sigma W_t)} \\ \Rightarrow \tau_t &= \sqrt{\mathbb{V}[Z_t]} \\ &= \sqrt{e^{\sigma^2 t} - 1} \end{cases} \quad (9.1)$$

Pour effectuer cette calibration, les incidences historiques et l'incidence déterministe sont translatées pour que leur pic d'incidence coïncide. En effet, comme cela a été mentionné dans la partie II, les épidémies recensées par Sentinelles ne débutent pas au même moment, ne prennent pas fin au même moment, et l'incidence simulée par le modèle déterministe présente également un décalage par rapport aux données historiques.

Après les avoir toutes centrées en leur pic épidémique, la représentation suivante des données utilisées pour cette calibration est obtenue :

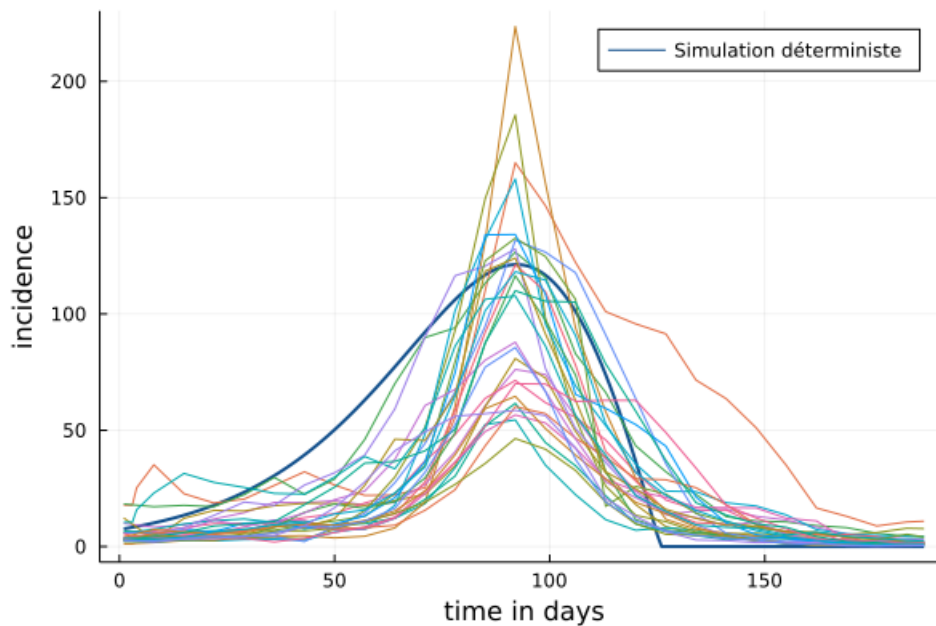


FIGURE 9.1 – Incidence issue du SIR^+ et incidences historiques avec coïncidence de leur maximum

Il est important de noter ici que l'axe des abscisses ne correspond plus, comme c'était le cas jusqu'à présent, à la durée en jour écoulée depuis la semaine 36 de l'année considérée. Il s'agit davantage d'un repère temporel permettant de comparer les durées entre les épidémies, tous leurs pics étant centrés à l'unité 92 de l'axe des abscisses.

Les incidences historiques normalisées par les incidences déterministes sont ensuite obtenues, jusqu'à $t = 125$, l'incidence déterministe étant nulle au-delà.

Une grille de valeurs possibles de σ est créée : elle sera fixée à $K_{1000}(10^{-3}, 10^{-1})$, notation introduite en partie 4.3.1. Pour rappel, cette notation représente 1 000 valeurs entre 10^{-3} et 10^{-1} avec un pas constant entre ces deux valeurs. Pour chaque valeur de σ dans cette grille de recherche, le taux d'erreur théorique est calculé. Ce taux d'erreur est ensuite « translaté » pour coïncider avec les pics épidémiques présentés en figure 9.1. Puis la MSE entre ce taux d'erreur et le taux d'erreur empirique est calculée. La valeur de sigma pour laquelle cette MSE est la plus faible est enfin retenue.

9.2 Résultats de la calibration et trajectoires stochastiques des épidémies simulées

Après mise en place de la procédure développée dans la partie précédente, la valeur de σ retenue est :

$$\sigma^* = 0,02776 \quad (9.2)$$

Le modèle stochastique tel que présenté en (8.5) est désormais construit en intégralité, d'expression finale

$$X_t^{sto} = X_t^{det} \times e^{\left(-\frac{(\sigma^*)^2}{2}t + \sigma^* W_t\right)} \quad (9.3)$$

Les résultats de la calibration du modèle SIR_{det}^+ et celle du taux d'erreur associé au modèle SIR_{sto}^+ peuvent être synthétisés par le tableau ci-dessous.

θ^*	a^*	b^*	σ^*
0,0173	0,1658	28,7689	0,02776

La figure suivante illustre les trajectoires d'incidence obtenues pour 1 000 scénarios du SIR_{sto}^+ , en gris, accompagnés de la courbe déterministe ainsi que de l'enveloppe à 1% et l'enveloppe à 5%, sur une population totale de 100 000 personnes :

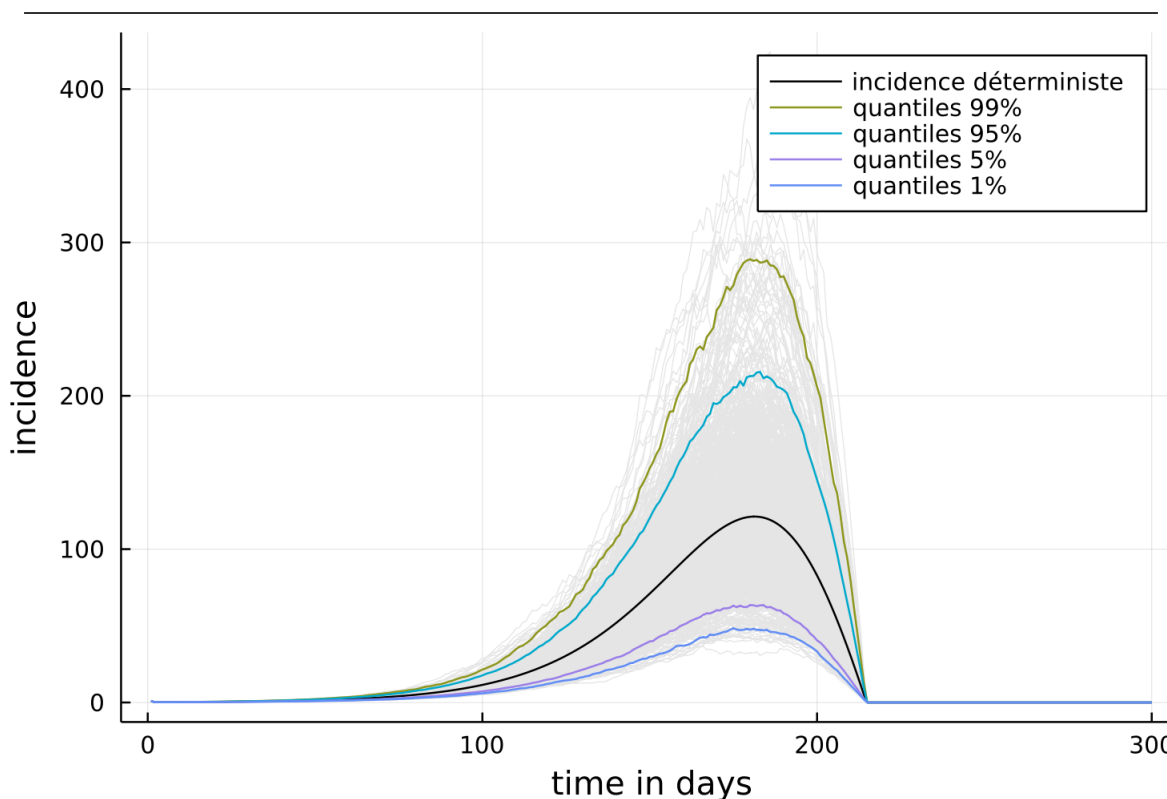


FIGURE 9.2 – 1 000 scénarios du SIR_{sto}^+ , enveloppes correspondantes et incidence déterministe - population totale de 100 000 individus

La visualisation des enveloppes du processus stochastique SIR_{sto}^+ permettent de valider la non explosivité des scénarios, qui est cohérente avec l'observation des données historiques sur les saisons retenues dans les travaux de ce mémoire.

Le modèle en l'état permet donc de simuler un scénario épidémique, et d'obtenir le nombre de nouveaux infectés jour après jour, sur une population donnée, qui est dans la figure 9.2 de 100 000 personnes.

Le SIR_{sto}^+ permet également d'obtenir le nombre de personnes présents dans chacun des états sanitaires S , I et R à chaque instant t de l'épidémie.

Enfin, il permet d'obtenir la durée d'infection *i.e* la durée passée dans l'état I , qui est représentée par la variable aléatoire $T \hookrightarrow \Gamma(a^*, b^*)$.

Le modèle SIR_{sto}^+ fournit donc des informations macroscopiques, *i.e* sans distinction entre les individus, quant au nombre de personnes dans chaque état du modèle et à tout instant de l'épidémie t . Une illustration d'une application du modèle SIR_{sto}^+ est proposée dans la partie suivante, portant sur l'obtention d'une distribution des arrêts de travail dus à des infections par le virus de la grippe.

Quatrième partie

Cas pratique : distribution de provisions en arrêts de travail engendrés par des infections par virus grippal

Chapitre 10

Processus de simulation d'épidémies et distribution des provisions

Il est rappelé qu'une garantie « arrêt de travail » prévoit le maintien de tout ou partie du Salaire Journalier de Base (SJB) de la personne arrêtée. Les spécificités de cette garantie ont été introduites dans le chapitre 2 de ce mémoire, et ces modalités ne sont pas rappelées ici.

Le cadre de ce cas pratique est celui de la grippe saisonnière : seuls les arrêts pour cause d'infection par le virus de la grippe seront ici pris en compte.

Un portefeuille fictif d'assuré est dans un premier temps simulé.

Le modèle SIR_{sto}^+ précédemment construit est appliqué à ce portefeuille, pour obtenir parmi les assurés un nombre de nouveaux infectés par le virus de la grippe, ainsi qu'un nombre de nouveaux rétablis, à chaque pas de temps de la simulation.

Une répartition du nombre de nouveaux infectés simulé parmi la population susceptible, ainsi qu'une répartition du nombre de nouveaux rétablis simulé parmi la population infectée est ensuite mise en place, à l'issue de laquelle il est possible de déterminer quels sont les individus qui ont été infectés, et pendant combien de temps ils l'ont été.

L'information sur la durée d'infection de chaque infecté est ensuite mise en relation avec son salaire, pour calculer la charge que représente le maintien de ce salaire durant l'arrêt de travail. Ces charges sont ensuite agrégées pour obtenir une provision au titre des arrêts de travail consécutifs aux infections par le virus de la grippe.

Plusieurs simulations de la sorte sont réalisées, pour dégager *in fine* une distribution de ces provisions.

10.1 Hypothèses retenues

Plusieurs hypothèses sont introduites préalablement à la simulation effective du portefeuille, de l'épidémie et au calcul des provisions associées à cette épidémie grippale.

Ces hypothèses ont pour principal intérêt de simplifier la compréhension des calculs et la démarche qui s'apprête à être présentée.

10.1.1 Portefeuille

Le portefeuille simulé est composé de 100 000 salariés supposés appartenir à une même entreprise, qui a souscrit un contrat d'assurance collective pour se prémunir du risque de survenance des arrêts de travail parmi ses employés. Ce nombre d'individus permet une cohérence avec le développement du modèle, qui a été calibré à partir de données d'incidence pour 100 000 personnes. Les individus du portefeuille sont indifférenciables d'un point de vue épidémiologique donc leurs variables intrinsèques suivantes ne sont pas prises en compte :

- le sexe
- l'âge
- la résistance au virus
- la capacité à se rétablir de la maladie en cas d'infection

Les variables exogènes suivantes seront elles aussi écartées :

- la période de l'année (vacances scolaires ou non)
- la période de la semaine (du lundi au vendredi ou samedi et dimanche)
- la durée des contacts sociaux entre les salariés
- le cadre du contact (sphère personnelle ou professionnelle)
- l'intensité du contact (physique ou non)

Ces hypothèses éludent pour la plupart des facteurs qui impactent la propagation d'une épidémie : par exemple l'hypothèse selon laquelle le nombre de contacts qu'une personne a au cours d'une journée est constant est une hypothèse très forte, dans la mesure où une personne infectée qui est en arrêt de travail rencontre inévitablement moins de personnes que si elle continuait son activité professionnelle.

L'arrêt de travail est lui supposé consécutif à l'infection, c'est à dire que dès lors qu'un individu est infecté, il cesse son activité professionnelle.

De telles hypothèses quant à la population considérée et quant aux variables retenues permettent de positionner l'étude de ce portefeuille de salariés dans l'espace de validité du modèle SIR_{sto}^+ développé dans la partie précédente. Ce modèle est en effet adapté à la modélisation de la propagation d'une épidémie au sein d'une population homogène, au sens des hypothèses détaillées précédemment, et fermée.

Pour chacun des employés du portefeuille, un salaire est affecté par extrapolation de la répartition salariale en France en 2019 fournie par le site Statistita.

Les résultats de Statistita sont considérés constants entre 2019, année de l'étude, et 2022.

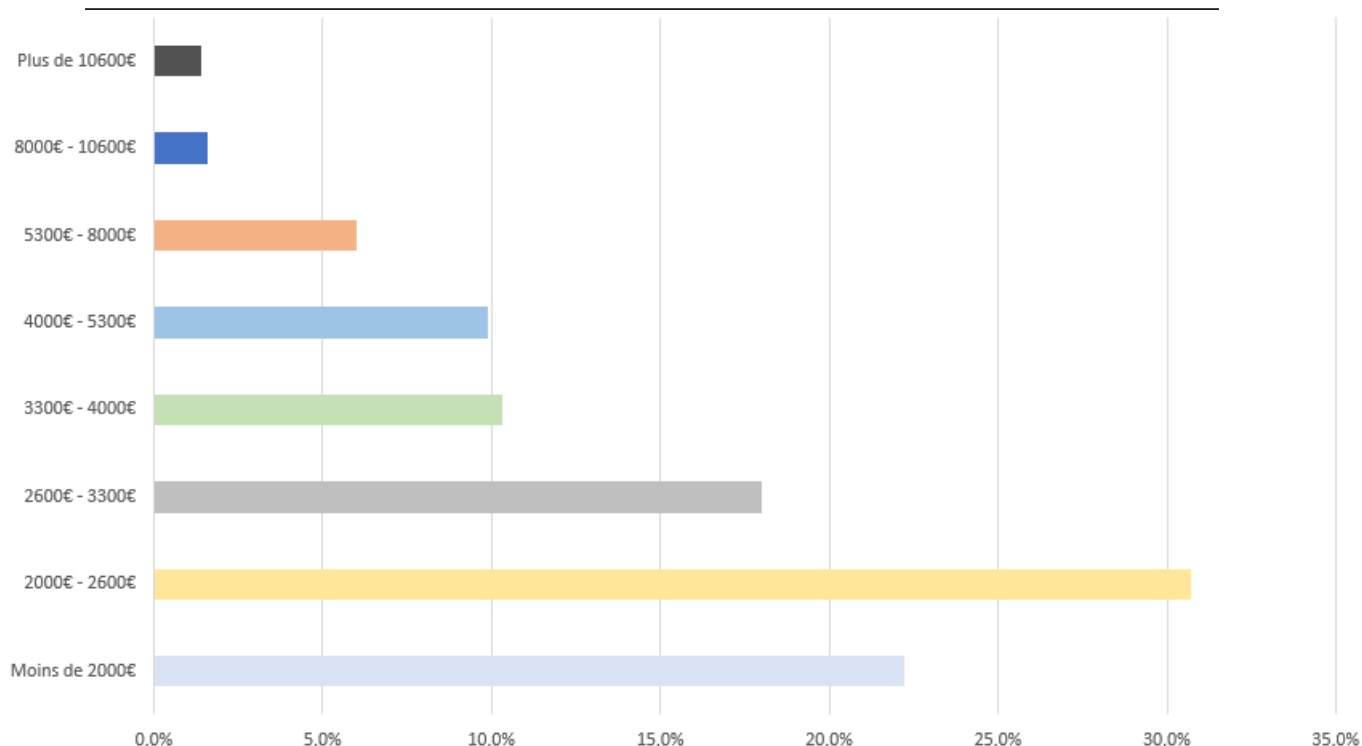


FIGURE 10.1 – Répartition des salaires en France selon la tranche salariale

Ces données – arrondies dans la figure par souci de clarté – concernent la France hors Mayotte et des salariés en équivalent temps plein du secteur privé et des entreprises publiques. Il est également à préciser que les deux extrêmes, c'est à dire les tranches « Plus de 10 600 € » et « Moins de 2000 € » ne peuvent opérationnellement pas rester telles quelles. Sont introduites une borne inférieure correspondant à la valeur du SMIC 2022 pour la première tranche, et une borne supérieure à hauteur de 20000 € (choisie arbitrairement) pour la deuxième tranche. Ainsi la première tranche devient [1603€; 2000€] et la deuxième [10 600€; 20000€].

Une simulation macroscopique est d'abord effectuée : elle fournit un nombre de nouveaux infectés ainsi qu'un nombre de nouveaux rétablis jour après jour, à la maille portefeuille sans distinction entre les individus. Puis pour chaque jour j , l'incidence est « répartie » au sein de la population saine en j . Enfin, toujours au jour j , le nombre de nouveaux rétablis est « réparti » au sein de la population infectée, grâce à une pondération qui dépendra de la durée passée dans l'état I par chaque infecté.

10.1.2 Garantie arrêt de travail

La garantie arrêt de travail qui est retenue est une garantie à hauteur de 100% du salaire, et ce sans délai de carence (type « CCN Syntec »).

Le schéma de cette garantie est une version tronquée de la figure 2.1 :



FIGURE 10.2 – Répartition tronquée à 38 jours du maintien de salaire en arrêt de travail

Cette troncature est cohérente au regard des données de la grippe saisonnière, dans la mesure où la grippe ne dure en général pas plus d'une semaine. Les rares cas de grippe de longue durée ne sont cependant pas totalement écartés : la seule hypothèse faite ici est que pour ces cas de figure, la durée d'arrêt ne dépasse jamais 38 jours.

La garantie retenue est donc un maintien de salaire à hauteur de 100% du salaire de base les 3 premiers jours d'infection, puis 50% les 35 jours suivant, les autres 50% étant pris en charge par la Sécurité sociale. Il est à préciser que le remboursement de la Sécurité sociale ne peut dépasser 50% du salaire journalier de base (SJB) maximum. Ainsi, la Sécurité sociale prend en charge au maximum 47,43€ par jour, après le délai de carence de 3 jours évoqué précédemment. Il est important de garder à l'esprit que cette garantie est la plus haute possible, et qu'elle représente en conséquence un coût très élevé pour l'organisme qui versera ces compensations journalières.

10.2 Création du portefeuille fictif

La première phase de la simulation macroscopique est la création d'un portefeuille, fictif dans le cadre de ce cas pratique, qui contiendra les informations suivantes sur les individus qui le composent :

- Le numéro de l'individu, qui servira de clé
- Le salaire, simulé uniformément parmi les tranches de salaires introduites précédemment. Cette simulation uniforme est une estimation perfectible dans la mesure où elle ne tient pas compte des potentielles disparités de répartition des salaires au sein d'une tranche. Il est par exemple intuitif que le nombre de personnes dont le salaire est le *SMIC* représente une forte proportion des salariés de la première tranche.

- Le statut : S pour susceptible, I pour infecté ou R pour rétabli
- La durée passée dans l'état I : cette durée vaut -1 tant que l'individu est susceptible, puis est incrémentée de 1 chaque jour supplémentaire qu'un individu passe dans l'état I , et reste fixe une fois que l'individu passe dans R . La construction de cette durée est développée dans la simulation individuelle
- La pondération de rétablissement : elle équivaut à la fonction de répartition, évaluée en la valeur de la durée d'infection, de la loi *Gamma* suivant les paramètres a et b utilisés pour la simulation. Cette pondération sera également construite dans la simulation individuelle, et servira de pondération à la répartition des nouveaux rétablis parmi les infectés

Le portefeuille prend la forme suivante, une ligne correspondant aux caractéristiques d'un individu :

numeros	salaires	statut	duree_infection	ponderation_retablissement
1	1393,59	S	-1,0	0,0
2	1232,82	S	-1,0	0,0
3	1247,12	S	-1,0	0,0
4	1468,61	S	-1,0	0,0
5	1447,82	S	-1,0	0,0

10.3 Simulation macroscopique d'une épidémie de grippe

Les dynamiques des équations (8.6) permettent de simuler une épidémie grippale, et ce à partir des conditions initiales fournies par (4.1).

Dans tout ce qui va suivre, les conditions initiales seront les mêmes que celle que soit la simulation faite, à savoir :

Population totale N	Population susceptible initiale S_0	Population infectée initiale I_0	Population rétablie initiale R_0
100 000	999 999	1	0

TABLE 10.1 – Conditions initiales des simulations épidémiques

Le processus débute tout d'abord, comme il l'a été évoqué, par une simulation macroscopique de l'épidémie de grippe saisonnière. La simulation se base sur les équations (8.6), après obtention de la trajectoire déterministe X_t^{det} via les paramètres optimaux déterminés antérieurement.

À chaque pas de temps t , c'est à dire chaque jour de l'épidémie, une simulation du nombre de personnes susceptibles (S_t), du nombre de personnes infectées (I_t), du nombre de personnes rétablies (R_t), de l'incidence ainsi que du nombre de nouveaux rétablis est obtenue. L'incidence à l'instant t sera notée i_t , et le nombre de

nouveaux rétablis en t sera notée r_t .

Ces valeurs prennent la forme de réels ; or pour l'application à un portefeuille composé d'individus distincts il est nécessaire que les valeurs utilisées soient des entiers : 10,86 nouveaux infectés ne ferait pas sens ici. Les valeurs S_t , I_t , R_t , i_t et r_t sont ainsi arrondies a posteriori de la simulation, en s'assurant que $S_t + I_t + R_t = N$, $i_t \leq S_t$ et $r_t \leq I_t$ à chaque instant t , étant donné qu'il ne peut y avoir plus de nouveaux infectés que de susceptibles, ou plus de nouveaux rétablis que d'infectés. Il est important d'effectuer ces arrondis à posteriori car s'ils étaient faits au fur et à mesure la simulation en serait grandement impactée.

La simulation s'étale sur 300 jours, qui est la durée de découpage des saisons épidémiques utilisée lors de la phase de calibration du modèle déterministe.

10.4 Simulation individuelle : processus de répartition des incidences et des nouveaux rétablis

À partir des résultats de la simulation macroscopique, il est désormais nécessaire de déterminer quels individus du portefeuille sont dans chaque état du modèle. Par exemple si la simulation macroscopique indique :

- $S_t = 999\ 500$
- $I_t = 150$
- $R_t = 350$
- $i_t = 50$
- $r_t = 20$

Il est impossible en l'état de déterminer quels sont les 50 individus du portefeuille, en S à l'instant $t - 1$, qui viennent de passer dans l'état I , ou les 20, en I à l'instant $t - 1$, qui viennent de passer dans R . C'est pourquoi une procédure de répartition de i_t et de r_t à chaque instant t est mise en place.

Le numéro du premier infecté est choisi uniformément parmi les 100 000 numéros du portefeuille. Son statut devient I et non plus S , sa durée d'infection passe à 0 et sa pondération reste à 0.

À chaque instant t , r_t numéros d'individus sont tirés uniformément sans remise parmi les I_t numéros d'individus correspondant à des individus infectés. Ce tirage uniforme est pondéré par la durée passée par l'individu dans l'état I , via la fonction de répartition de la loi *Gamma* évaluée en cette durée. En d'autres termes, si l'individu numéro 100 est infecté depuis 3 jours, sa pondération de rétablissement vaudra $F_T(3) = 0,7299$ avec $T \hookrightarrow \Gamma(a^*, b^*)$. Cette pondération se comprend de la manière suivante : un infecté qui est dans I depuis 3 jours aura une probabilité d'en sortir plus élevée qu'un individu qui vient d'être infecté. Si un individu est tiré, son état passe de I à R , sa pondération ainsi que sa durée passée dans I sont gardées en mémoire, et il n'intervient plus dans le processus.

A la suite de ce tirage des nouveaux rétablis, i_t numéros d'individus sont tirés uniformément sans remise parmi les S_t numéros d'individus correspondant aux individus susceptibles. Ici pas de pondération : chaque individu susceptible possède une probabilité $\frac{i_t}{S_t}$ d'être choisi, puis $\frac{i_t - 1}{S_t - 1}$ (un des susceptibles étant passé dans I) etc.

jusqu'à $\frac{1}{S_t - i_t + 1}$ pour le dernier tirage sans remise. Si un individu est tiré, son statut passe de S à I , sa durée d'infection passe à 0 et sa pondération reste égale à 0.

Avant de passer à la répartition à l'instant $t + 1$, la durée d'infection des individus présents dans l'état I est incrémentée de 1, et sa pondération est la fonction de répartition évaluée en cette durée incrémentée.

À la fin de ce processus de répartition, chaque ligne du portefeuille renseigne le numéro de l'individu qui n'a pas changé, son état (S , I ou R), son salaire qui n'a pas changé lui non plus, sa durée d'infection qui est positive si l'individu est dans I ou R et négative sinon, et sa pondération de rétablissement qui est positive si la durée d'infection l'est et négative sinon.

10.5 Calcul de la charge totale des infectés dans le portefeuille

À l'issue de cette procédure de répartition de l'incidence ainsi que des nouveaux rétablis tout au long de l'épidémie, la durée d'infection ainsi que le salaire de chaque individu est obtenue.

Soient G et H les ensembles composés respectivement des numéros des individus dans I et des numéros des individus dans R à l'instant $t = 300$. En pratique aucun individu n'est présent dans l'état I à la fin de la simulation car l'incidence du modèle déterministe (et donc du stochastique) est nulle à partir du jour 215 et ce jusqu'à la fin, *i.e* jusqu'au jour 300. Or historiquement aucun individu nouvellement infecté en $t = 215$ reste infecté 85 jours, sauf cas extrêmement rare. Ainsi il sera considéré que $G \cup H = H$.

Est considéré ici un salarié numéro $x \in H$, qui a donc été infecté. Le salaire mensuel brut de cet individu est noté W_x , sa durée d'infection $\delta_x \in \mathbb{N}$ et la charge de son arrêt pour l'employeur est notée C_x . Son SJB (évoqué plus en amont dans ce mémoire) est calculé de la manière suivante :

$$SJB_x = \frac{3 \times K_x}{91,25}$$

Premier cas de figure :

Si $\delta_x \leq 3$, alors l'employeur prend en charge l'intégralité de ce SJB, d'après la garantie présentée dans la figure 10.2. La charge de cet arrêt de travail est alors

$$C_x = \delta_x \times SJB_x$$

Second cas de figure :

Si $\delta_x > 3$, alors l'employeur prend en charge l'intégralité du SJB pendant les 3 premiers jours. Puis pour chacun des $\delta_x - 3$ jours restant, la Sécurité sociale prend en charge une IJ_x égale à 50% du SJB dans la limite de 1,8 fois

le SMIC tel que décrit précédemment, et l'employeur prend en charge $SJB_x - IJ_x = SJB_x - \min\left(47, 43, \frac{SJB_x}{2}\right)$.
La charge totale de cet arrêt pour l'employeur est ainsi :

$$C_x = 3 \times SJB_x + (\delta_x - 3) \times \left[SJB_x - \min\left(47, 43, \frac{SJB_x}{2}\right) \right]$$

En conclusion :

La charge totale de l'employeur est définie selon la formule suivante générique qui suit.

$$C_x = \mathbb{1}_{\delta_x \leq 3} \times \delta_x \times SJB_x + \mathbb{1}_{\delta_x > 3} \times \left(3 \times SJB_x + (\delta_x - 3) \times \left[SJB_x - \min\left(47, 43, \frac{SJB_x}{2}\right) \right] \right)$$

La charge totale de l'employeur au regard de l'ensemble des individus ayant été infectés au cours de l'épidémie est

$$C = \sum_{x \in H} C_x$$

10.6 Distributions des durées d'infection et distributions des provisions

La simulation stochastique d'un scénario, suivie de la répartition des incidences et des nouveaux rétablis au sein du portefeuille d'un point de vue individuel permet d'obtenir pour chaque individu qui a été infecté la durée pendant laquelle il est resté infecté. Il est donc possible de tracer la distribution de ces durées d'infection, toutes simulations confondues :

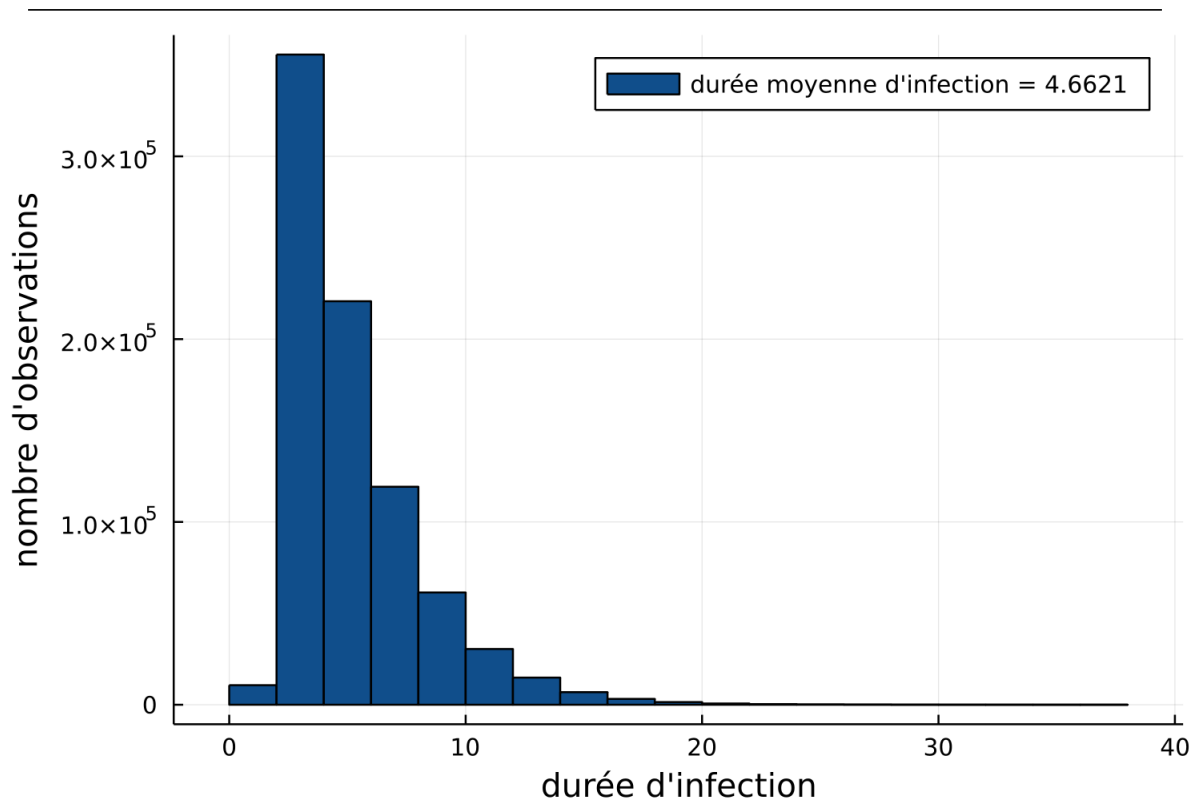


FIGURE 10.3 – Distribution des durées d'infection de la grippe sur 100 000 individus, pour 100 scénarios

Cette distribution apparaît cohérente avec les paramètres du modèle calibré. En effet, il est rappelé que la durée moyenne d'infection théorique est donnée par le moment d'ordre 1 de la loi représentant la durée passée dans l'état I , à savoir $\mathbb{E}[T] = a^* \times b^* = 0,1658 \times 28,7689 = 4,7698$, ce qui est proche de la valeur 4,6621 affichée dans la figure 10.3 ci-dessus, et qui correspond à peu près, à dire d'expert, à la durée moyenne d'infection de la grippe.

Les quantiles de ces durées sont ici intéressants à tracer, puisqu'ils fournissent une information quant au nombre d'individus infectés dont les durées d'infections sont très éloignées de la moyenne.

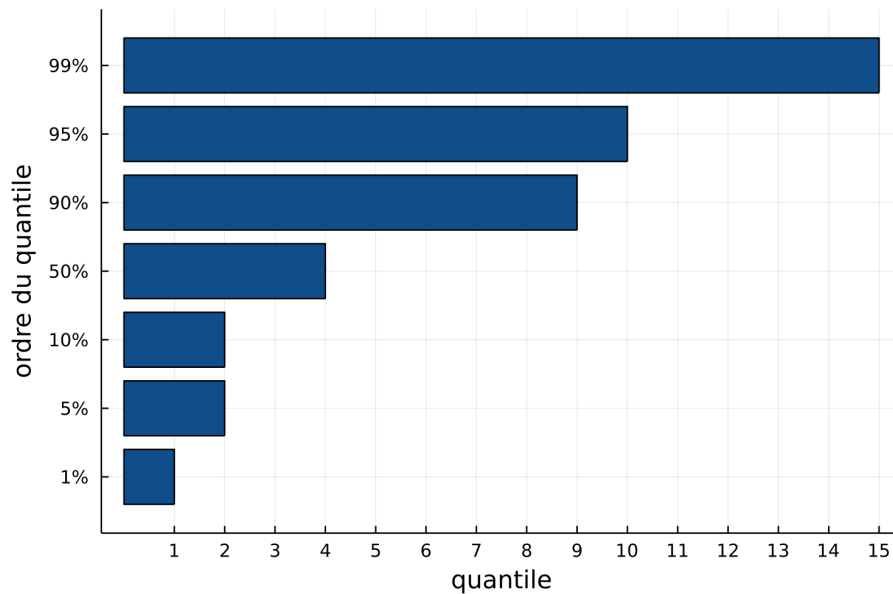


FIGURE 10.4 – Quantiles des durées d'infection pour 1 000 scénarios

Il ressort de ces quantiles que 99% des infectés se rétablissent au bout d'une durée inférieure ou égale à 15 jours. La durée maximale observée sur ces simulations est de 37 jours d'infection : cela valide à la fois l'hypothèse faite dans la figure 10.2 selon laquelle les durées prises en compte sont limitées à 38 jours, mais aussi l'hypothèse évoquée plus selon laquelle $G \cup H = H$. Ainsi, les modalités retenues de répartition de la prise en charge du salaire entre la Sécurité sociale et l'employeur sont représentatives des scénarios obtenus.

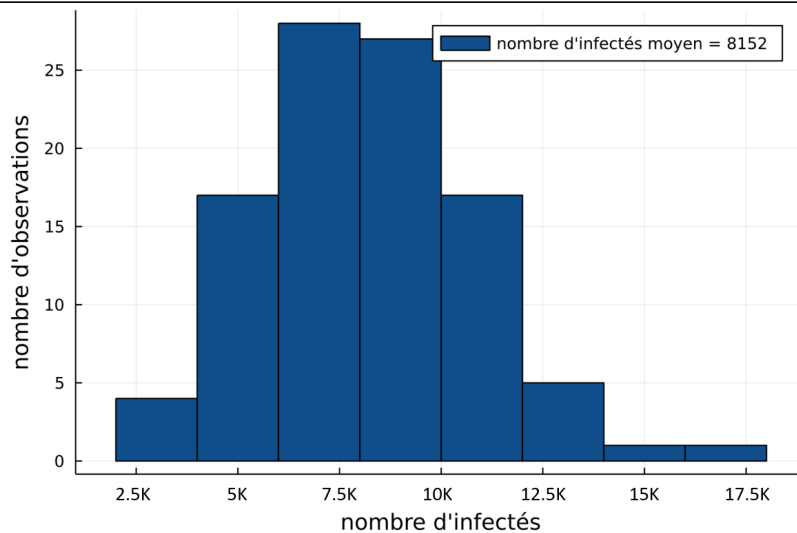


FIGURE 10.5 – Distribution du nombre d’infectés par le virus de la grippe pour 100 scénarios

Cette distribution représente le nombre total d’infectés pas le virus de la grippe au cours d’une saison, et ce pour 100 simulations. La moyenne s’établit ici à 8 152 personnes. Ramenés à une population de 67,39M (population française en 2022), cela représente environ 5,5M d’infectés, qui est cohérent avec la fourchette 2M ~ 8M souvent annoncée. Enfin, il est possible de tracer la distribution des charges pour l’employeur, assimilables à des provisions en arrêt de travail consécutifs à des infections par le virus de la grippe :

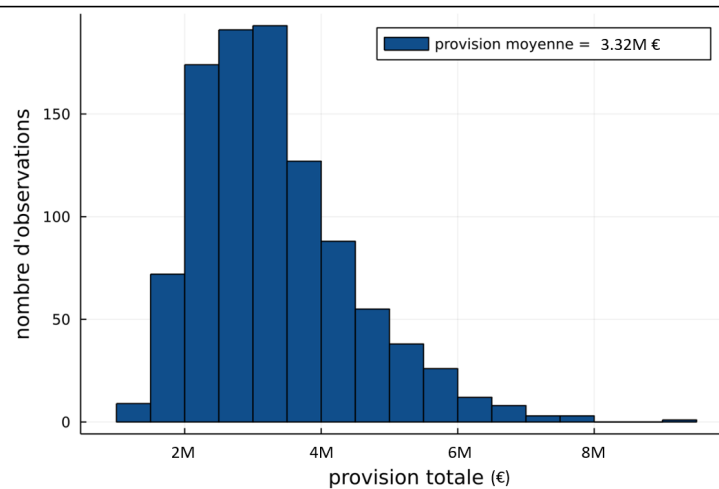


FIGURE 10.6 – Distribution des provisions maintien de salaire pour des arrêts de travail consécutifs à des infections par le virus de la grippe dans un portefeuille de 100 000 salariés, pour 1 000 scénarios

Ces provisions représentent plusieurs millions d’euros à la charge des employeurs et/ou de la complémentaire le

cas échéant.

L'application qui a été faite dans le cadre de ce mémoire de l'exploitation des simulations du modèle stochastique construit est un exemple relativement simplifié qui utilise certaines hypothèses fortes concernant la population d'étude. Le modèle pourrait être de bien des manières amélioré et/ou adapté en fonction des besoins.

À titre illustratif, la valeur de m , représentant le nombre de contacts qu'un individu a au cours d'une unité de temps est fixé ici à $m = 16,42$. Or l'étude dont cette valeur est tirée précise que ce nombre de contact dépend de plusieurs facteurs (âge de la personne considérée, âge des personnes avec qui elle entre en contact, période de l'année - scolaire ou vacances- etc.). Un modèle segmenté par exemple en classes d'âge pourrait permettre de modéliser la propagation de l'épidémie de manière davantage précise.

Une infection par le virus de la grippe est en général accompagné d'une période durant laquelle l'individu infecté ne présente pas les symptômes de la maladie et n'est pas contagieux (période d'incubation), qui n'est pas prise en compte par le modèle proposé. Une introduction d'un tel état, intermédiaire entre S et I , présent dans les modèles $SEIR$ (Susceptibles, Exposed, Infectés, Rétablis) constituerait indéniablement une amélioration du modèle.

L'effet d'une potentielle vaccination, qui est en pratique non négligeable dans le cadre du virus de la grippe, est également possible. Il pourrait venir impacter le paramètre θ par exemple, qui représente la probabilité qu'un contact entre un individu infecté et un individu susceptible engendre la contamination de l'individu susceptible : le paramètre θ pourrait ici être plus faible si l'individu susceptible était vacciné.

De plus l'application qui est proposée porte sur les arrêts de travail, et donc utilise directement l'incidence simulée. Une donnée qui n'est pas prise en compte par le modèle construit est le nombre de personnes décédées du fait de la maladie. Il en a été fait mention, ces personnes sont confondues ici avec les rétablis. Cependant une scission de l'état R en deux états que sont R' (rétablis et vivants) et D (décédés) pourrait permettre, moyennant une revoiture des dynamiques du modèle, d'étudier la surmortalité liée à la maladie considérée. Un tel modèle est appelée $SIRD$. Les modèles SIR^+ et SIR_{sto}^+ construits dans ce mémoire sont donc limités à une étude épidémiologique portant sur le nombre d'infectés, ou la durée d'infection de la maladie sous-jacente. Une prise en compte des décédés par exemple pourrait être étudiée par ramification du modèle et adaptation des dynamiques de passage entre les états en conséquence.

Enfin, le virus étudié tout au long de ce mémoire est le virus de la grippe : il est tout à fait envisageable de transposer les travaux de ce mémoire à une autre maladie transmissible, en faisant des hypothèses similaires à celles qui ont été faites ici.

Conclusion

Les travaux de ce mémoire découlent d'un constat sans équivoque : la place qu'occupe les épidémies dans les préoccupations des acteurs de l'assurance n'a vocation qu'à devenir, au fil du temps, encore plus importante qu'elle ne l'est déjà à l'heure actuelle. Cette affirmation a trouvé un écho récemment avec la survenance de la pandémie de Covid-19, d'une ampleur inégalée depuis au moins une centaine d'année. Cette dernière constitue un rappel précieux des potentielles conséquences associées aux phénomènes épidémiques, qui sont à même d'impacter durablement la société, à l'image de la grippe saisonnière ou encore du SIDA. Ce rappel est d'autant plus important qu'une recrudescence des phénomènes épidémiques est anticipée pour les décennies à suivre.

La quasi totalité des secteurs d'activité sont amenés à subir les conséquences d'une épidémie mondiale, ou même d'une épidémie, et les institutionnels du secteur de l'assurance y sont inéluctablement exposés. Au-delà de l'impact direct de la pandémie sur les sociétés d'assurance en tant qu'entreprise, le fondement même de leur activité nécessite une prise en compte de ces phénomènes. Ceux-ci doivent donc en conséquence se munir d'outils adéquats pour appréhender au mieux ces phénomènes. Ce mémoire a donc permis la construction de modèles, permettant de modéliser la propagation temporelle d'une épidémie, plus efficaces que les modèles documentés dans la littérature. En effet, les modèles existants présentent certaines imprécisions qu'il s'est avéré nécessaire de corriger.

Le modèle épidémiologique le plus connu est le modèle *SIR*, qui sépare la population en trois états sanitaire : Susceptibles d'être infectés, Infectés par la maladie, Rétablis de la maladie. Le *SIR* modélise l'évolution du nombre de personnes se trouvant dans chacun de ces trois états au cours du temps, en s'appuyant sur des relations de passage d'une partie de la population de chaque état à un autre état, à chaque pas de temps. Il ressort néanmoins d'une interprétation de ce modèle que ses relations de passages ne sont pas adéquates, et qu'elles ne permettent pas de prendre en compte des informations pourtant indispensables à une modélisation acceptable du phénomène. Il a donc été fait le choix tout au long de ce mémoire de proposer des améliorations de ce modèle qui permettraient une meilleure appréciations des risques assurantiels liés aux épidémies.

Dans un premier temps donc, la nécessité de révision du modèle *SIR* a aboutit, en adoptant une approche probabiliste, à l'obtention d'un modèle dénommé *SIR*⁺, lui aussi déterministe, qui vient pallier certains défauts du *SIR*. Le *SIR*⁺ devient dès lors un modèle fiable lorsqu'il s'agit de modéliser temporellement une épidémie dont les états sanitaires peuvent se résumer aux trois états *S*, *I* et *R*. Il permet ainsi de mesurer le nombre de personnes infectés, la durée d'infection ou encore la propension à guérir de la maladie. Il est cependant à noter que ce modèle n'est pas adapté à toutes les épidémies, ni à tous les impacts que celles-ci pourraient avoir. Un exemple réside dans l'absence de prise en compte du nombre de décès imputables à l'épidémie : ainsi le modèle

SIR^+ est inadapté à l'étude de l'impact de la mortalité d'une maladie infectieuse, ou du moins en l'état. Il nécessiterait l'ajout d'un état "décédé", et donc une ramification accompagnée d'une nouvelle définition de ses dynamiques.

Une caractéristique intrinsèque aux épidémies n'est cependant pas prise en compte par le SIR^+ : leur caractère incertain et imprévisible. Ils se vérifient en pratique, les saisons épidémiques saisonnières comme la grippe en France se ressemblent généralement peu, en termes de durée, de nombre total de personnes infectées etc. La modélisation déterministe présente donc ici une limite : elle ne permet pas de capter la nature intrinsèquement aléatoire des épidémies. Des tentatives de modification des dynamiques du déterministe pour y intégrer une composante stochastiques ont été étudiées, puis rejetées faute de possibilité de les mettre en pratique et d'adéquation suffisante avec le modèle déterministe. Le choix s'est donc porté sur une transformation de la dynamique de l'incidence déterministe *via* un mouvement brownien géométrique, qui représente ainsi le taux d'erreur entre le déterministe et les données réelles. Le modèle ainsi construit, noté SIR_{sto}^+ permet l'obtention de scénarios stochastiques représentés par le nombre de personnes dans chaque état S , I et R à chaque instant de la modélisation.

La méthode de calibration utilisée pour calibrer le SIR^+ est une méthode issue du domaine du *deep-learning* : il s'agit de l'algorithme *ADAM* qui assure une convergence rapide vers les paramètres optimaux. Le choix de cet algorithme est motivé par sa capacité à calibrer des modèles composés d'une multitude de paramètres : il permet à la méthodologie de ce mémoire d'envisager des modèles avec un nombre d'états plus important que le SIR^+ qui posséderaient de nombreux paramètres qui pourraient être alors plus difficiles à calibrer si l'algorithme retenu était un autre qu'*ADAM*.

Afin de valider que le SIR^+ permet de modéliser convenablement une épidémie, un cas pratique est proposé : le SIR^+ a été dans un premier temps calibré sur des données d'incidence, c'est à dire le nombre de nouveaux infectés par unité de temps, de la grippe saisonnière en France. Ces données ont été extraites des bilans épidémiologiques du réseau Sentinelles, qui recense les cas de grippe sur le territoire national semaine après semaine depuis 1984. Ces données ont nécessité un retraitement afin d'obtenir des données quotidiennes. Les résultats de cette calibration démontrent la capacité du SIR^+ à modéliser convenablement des épidémies grippales. Cependant la forme même de la dynamique d'incidence semble présenter une potentielle limite : il semble que l'épidémie s'arrête de manière relativement brutale après avoir atteint son pic, comportement qui n'est pas toujours observé en pratique. Ainsi le modèle SIR^+ pourrait potentiellement être amélioré en complexifiant davantage les dynamiques de transition entre les états.

Le modèle stochastique a été défini comme une transformation du modèle déterministe : le processus de taux d'erreur, à partir de la calibration du SIR^+ , est calibré à son tour sur les mêmes données d'incidence, par minimisation de l'erreur quadratique entre son taux d'erreur théorique et le taux d'erreur observé sur les données de Sentinelles. Les scénarios issus de 1 000 simulations du SIR^+ calibré permettent de conclure quant à la capacité du modèle à prendre en compte des scénarios avec une forte déviation à la moyenne, sans pour autant présenter un caractère trop « explosif ».

Le modèle stochastique SIR_{sto}^+ ainsi calibré sur les données d'incidence de la grippe, une application de celui-ci à un portefeuille d'assurés couverts par une garantie arrêt de travail est proposée. De la simulation macroscopique est extrapolée une simulation individuelle au jour le jour, c'est à dire une affectation de l'incidence ainsi que du nombre de personnes nouvellement rétablies au sein du portefeuille. Les durées d'infections sont mises à jour

et gardées en mémoire tout au long de l'épidémie. À la fin de l'épidémie simulée, les salaires des personnes qui ont été infectées ainsi que leur durée d'infection permettent de déterminer la charge totale que représentent les maintiens de salaires que prévoient la garantie arrêt de travail. Les hypothèses qui ont été retenues afin de réaliser ces simulations sont des hypothèses qui s'avèrent être très fortes.

Les modèles, déterministe et stochastique, qui ont été construits et développés tout au long de ce mémoire s'avèrent fiables et d'une précision acceptable, malgré la nécessité d'hypothèses qui demeurent fortes. Ils sont donc susceptibles d'être amendés et/ou améliorés selon les utilisations qui pourraient en être faites, et constituent ainsi une première contribution au perfectionnement d'une modélisation épidémiologique pertinente au regard de la quantification actuarielle des risques.

Bibliographie

- [BVT08] Andrew BURNS, Dominique VAN DER MENSBRUGGHE et Hans TIMMER. « Evaluating the Economic Consequences of Avian Influenza(1) ». In : (jan. 2008).
- [JM09] Teri JOHNSON et Barry MCQUARRIE. « Mathematical modeling of diseases : Susceptible-infected-recovered (sir) model ». In : *University of Minnesota, Morris, Math 4901 Senior Seminar*. 2009.
- [BCC12] Fred BRAUER, Carlos CASTILLO-CHAVEZ et Carlos CASTILLO-CHAVEZ. *Mathematical models in population biology and epidemiology*. T. 2. Springer, 2012.
- [YH12] Kang-Hung YANG et Jin-Yuan HSU. « A new SIR-based model for influenza epidemic ». In : *International Journal of Health and Medical Engineering* 6.7 (2012), p. 701-706.
- [Pla13] Frédéric PLANCHET. « Modélisation du risque de pandémie dans Solvabilité 2. » In : *Assurances et Gestion des Risques* 81.3/4 (2013), p. 225-234. ISSN : 17057299. URL : <https://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=93455877&site=ehost-live>.
- [Lóp16] Maria Jesús LÓPEZ HERRERO. « Epidemic transmission on SEIR stochastic models with nonlinear incidence rate ». In : *Mathematical Methods in the Applied Sciences* 40.7 (2016), p. 2532-2541.
- [Bér+20] Guillaume BÉRAUD et al. *Social contact data for France*. Version 3. Zenodo, juin 2020. DOI : 10.5281/zenodo.3886590. URL : <https://doi.org/10.5281/zenodo.3886590>.
- [Cha+20] CHARPENTIER, ARTHUR et al. « COVID-19 pandemic control : balancing detection policy and lockdown intervention under ICU sustainability* ». In : *Math. Model. Nat. Phenom.* 15 (2020), p. 57. DOI : 10.1051/mmnp/2020045. URL : <https://doi.org/10.1051/mmnp/2020045>.