

**Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuares**

Par : Madame / Monsieur André Stephen BALANDA

**Titre du mémoire : Elaboration et impact de nouveaux chocs biométriques dans la
formule standard pour le risque pandémie**

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

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

Membres présents du jury de la Signature : Entreprise : Crédit Agricole Assurances
filiale :

Nom :

Signature :

Membres présents du jury de
l'Institut des Actuares :

Signature :

Directeur de mémoire en
entreprise

Nom :

Signature :

*ROURON-LEVI
Ribecca*



Invité :

Nom :

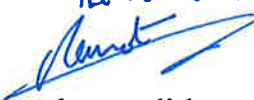
Signature :

**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 :

Ribecca Rand-Ling



Signature du candidat :



Résumé

Une pandémie est une épidémie qui touche plusieurs populations à l'échelle mondiale. Sachant qu'une épidémie est une croissance continue des cas de contamination d'une maladie infectieuse dans une région donnée, il est clair que ce phénomène représente un risque important pour les compagnies d'assurance. En effet, les agents pathogènes à l'origine de ces phénomènes provoquent des maladies chez les individus qui vont demander des soins, des arrêts de travail, des hospitalisations et même dans certains cas malheureux, cela peut aboutir au décès. Or tous ces risques sont des garanties couvertes par les assureurs, et la pandémie ne fait pas partie des exclusions de couverture.

Bien que l'histoire démontre que ce phénomène a une fréquence faible, les grands changements d'ordre climatiques notamment opérés dans le monde au fil du temps l'ont rendu beaucoup plus fréquent qu'à l'accoutumée par son lien avec la biodiversité et cette tendance pourrait bien devenir la norme au cours des décennies à venir donc il est important pour les compagnies d'assurance de savoir évaluer ce risque du mieux possible.

A l'heure actuelle, les compagnies d'assurance et de réassurance peuvent tenter une première approche de l'évaluation de ce risque en suivant les recommandations de la réglementation prudentielle Solvabilité 2. Cette réglementation consiste notamment à uniformiser l'évaluation du besoin en capital nommé Capital de Solvabilité Requis (SCR pour « Solvency Capital Requirement » en anglais) qui serait nécessaire pour absorber un choc bicentenaire c'est-à-dire se réalisant une fois tous les 200 ans. Ce SCR se calcule en faisant une approche par module de risque qui eux-mêmes sont constitués de sous-modules. Les modules concernés pour la pandémie sont d'une part le module santé et son sous-module CAT et d'autre part le module vie et son sous-module CAT. Des chocs fournis par l'EIOPA sont appliqués au sein de chaque sous-module puis les sous-modules sont agrégés par une matrice de corrélation aussi fournie par l'EIOPA.

Cependant, on peut se demander si les chocs appliqués sont toujours cohérents ou s'il peut être judicieux d'utiliser des chocs beaucoup plus nuancés. Par exemple, le choc du SCR Vie CAT consiste à considérer une surmortalité uniforme de 0.15% à toutes les classes d'âge. Ainsi, dans ce mémoire, nous allons présenter des approches basées sur des outils mathématiques de nature différente pour discuter des chocs qui en résultent par rapport aux chocs fixés dans la réglementation prudentielle Solvabilité 2.

Mots-clés : Pandémie, Solvabilité 2, formule standard, SCR

Abstract

A pandemic is an epidemic that affects several populations worldwide. Knowing that an epidemic is a continuous growth of cases of contamination of an infectious disease in a given region, it is clear that this phenomenon represents a significant risk for insurance companies. Indeed, the pathogens at the origin of these phenomena cause illnesses in individuals that will require care, work stoppages, hospitalizations and even, in some unfortunate cases, death. All these risks are covered by many insurance policies. However, all these risks are covered by insurers, and the pandemic is not one of the exclusions of coverage.

Although history shows that this phenomenon has a low frequency, the major climatic changes in the world over time have made it much more frequent than usual because of its link with biodiversity, and this trend could well become the norm in the decades to come, so it is important for insurance companies to know how to assess this risk as best they can.

At present, insurance and reinsurance companies can attempt a first approach to the evaluation of this risk by following the recommendations of the Solvency 2 prudential regulations. This regulation consists in particular in standardizing the evaluation of the capital requirement called Solvency Capital Requirement (SCR) which would be necessary to absorb a bicentennial shock, i.e. one that occurs once every 200 years. This SCR is calculated using a risk module approach, which themselves are made up of sub-modules. The modules concerned for the pandemic are on the one hand the health module and its CAT sub-module and on the other hand the life module and its CAT sub-module. Shocks provided by EIOPA are applied within each sub-module and then the sub-modules are aggregated by a correlation matrix also provided by EIOPA.

However, one may ask whether the shocks applied are always consistent or whether it may be appropriate to use much more nuanced shocks based on a partial internal model. Thus, in this paper, we will present two internal models based on different mathematical tools and discuss the resulting shocks in relation to the shocks set in the Solvency 2 prudential regulation.

Keywords : Pandemic, Solvency 2, standard formula, SCR , partial internal model

Note de Synthèse

Une situation jusqu'ici inimaginable comme la covid19 a pris le monde par surprise et peut se reproduire à tout moment dans le futur. La réglementation donne aux organisations financières une chance de réagir de manière flexible et sûre à un événement pandémique. En vertu de la directive Solvabilité 2, les compagnies d'assurance établies au sein de l'Union Européenne doivent détenir en permanence des fonds propres éligibles suffisants pour couvrir leur besoin en capital. Cela permet aux entreprises d'assurance d'absorber des pertes importantes et de donner confiance aux assurés et aux bénéficiaires que les paiements seront effectués à l'échéance. Il existe des risques de souscription dans toutes les assurances santé, vie et non-vie, à savoir les pertes potentielles qu'un assureur peut subir en raison d'une souscription défectueuse ou encore d'une dérive de la sinistralité. L'impact d'un éventuel événement catastrophique est pris en compte dans le calcul du SCR dans les trois modules de risque de souscription : le risque de souscription non-vie, le risque de souscription vie et le risque de souscription santé. Tous ces modules sont calculés dans un sous-module catastrophe (CAT) mais nous allons regarder les événements catastrophes liés aux pandémies.

En partant du focus sur le risque pandémie, l'objectif du travail va être de proposer des chocs au sein de la formule standard qui sont différents de ceux fixés uniformément par la directive solvabilité 2 et d'en tirer les impacts sur le SCR du Crédit Agricole Assurances dont nous rappellerons le profil de risque. Les compartiments concernés dans la formule standard sont le module Santé et le module Vie. Au sein du module santé, on retrouve 3 sous modules qui sont le sous module santé pour les risques assimilables à la vie, le sous-module santé pour les risques assimilables à la non-vie et le sous-module CAT. C'est ce dernier sous-module qui nous intéresse et il est lui-même formé de 3 compartiments qui sont le risque d'accident de masse, le risque de concentration d'accident et le risque pandémie, c'est donc le choc du risque Pandémie que l'on va tenter de redéfinir. Du côté du module Vie, c'est une proposition différente du choc appliqué au sous-module CAT que nous allons faire. Nous rappelons que le SCR Vie CAT est pris en compte dans le SCR en considérant une hausse de 0,15 point en pourcentage des taux de mortalité et le SCR Santé CAT Pandémie est pris en compte en appliquant un coefficient de 0,075 pour mille à l'exposition au risque de protection des revenus en cas de pandémie et un coefficient de 40% à l'exposition aux frais d'hospitalisation en cas de pandémie.

Pour tester des chocs différents de ceux de la formule standard, nous allons utiliser deux outils mathématiques. D'une part nous utiliserons le modèle épidémiologique pour le calcul du besoin en capital du SCR Santé CAT Pandémie. Cela va nous permettre de déterminer le choc du montant à charge des hospitalisations dans un contexte de pandémie. D'autre part, nous utiliserons la théorie des valeurs extrêmes (TVE) pour le SCR Vie CAT dans le but de modéliser les quantiles extrêmes de surmortalité dû à une pandémie.

Le fonctionnement d'un modèle épidémiologique basique est qu'il repose sur le passage d'un état à un autre : un individu commence dans l'état sain car il n'a pas encore été contaminé puis au moment de la contamination, il devient exposé et commence alors sa période d'incubation et à la fin de cette période, il devient à son tour un infecté capable de contaminer les autres individus sains. A partir de ce moment, il lui reste deux issues possibles, soit il se rétablit au bout d'un moment, soit il décède.

Dans ce mémoire, nous allons utiliser le modèle SEIR sous forme stochastique qui simule les contacts de manière probabiliste entre les individus. Il évite de rendre tous les contacts systématiquement contagieux et ne fixe pas de temps d'incubation et d'infection pour tous les individus. Pour simuler les contacts, le modèle utilise un graphe dit d'Erdős-Rényi.

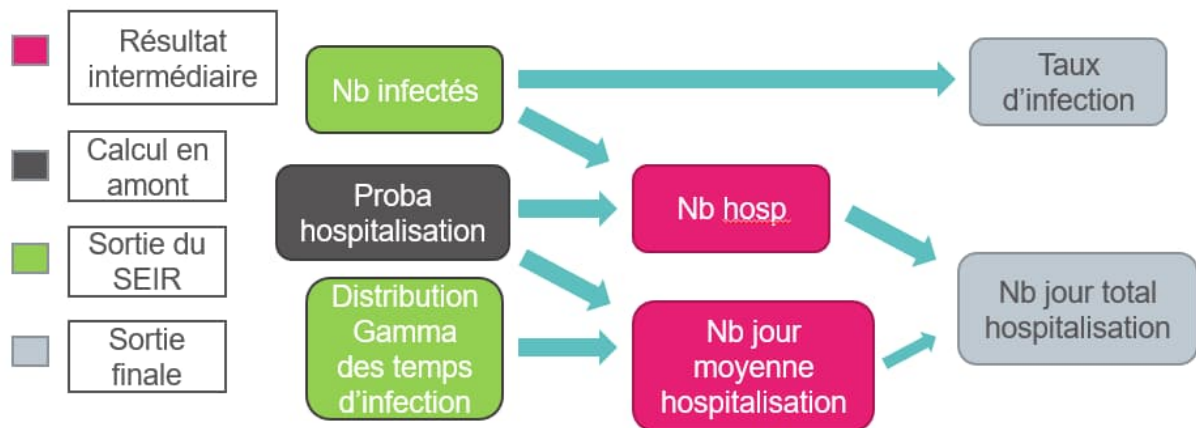
Ensuite, à partir de ces contacts, l'épidémie évolue avec les temps d'incubation T_E et d'infection T_I qui sont ici aléatoires de loi gamma. Pour éviter que tous les contacts donnent lieu à une infection, une variable de temps de contamination X_c est utilisée.

Le choix de la loi gamma est motivé par le fait qu'elle est adaptée à des phénomènes temporels et qu'elle a des réalisations plus crédibles que la loi exponentielle qui est aussi adaptée pour ce type de phénomènes.

Le modèle de base a été pensé pour ne prendre en compte que les épidémies sans létalité, de plus il ne donne pas d'informations sur les hospitalisations éventuelles or nous voulons utiliser le modèle épidémiologique pour un risque comme l'hospitalisation. Pour le prendre en compte, nous nous basons sur une étude de l'Institut Nationale de Veille Sanitaire (INVS) qui a évalué le niveau de risque en cas de pandémie, cette étude nous donne des informations sur la répartition des cas par âge, le pourcentage de la population à risque et le taux d'hospitalisation.

A partir de ces données, tout l'enjeu est de déterminer des probabilités d'hospitalisation à appliquer sur le nombre d'infectés en sortie du modèle SEIR stochastique puis d'obtenir le nombre de jour d'hospitalisations. La probabilité d'hospitalisation nous permet de déterminer le nombre de jours d'hospitalisations associé sur la distribution des temps d'infection : on prend le quantile correspondant.

Notre modèle se présente comme suit :



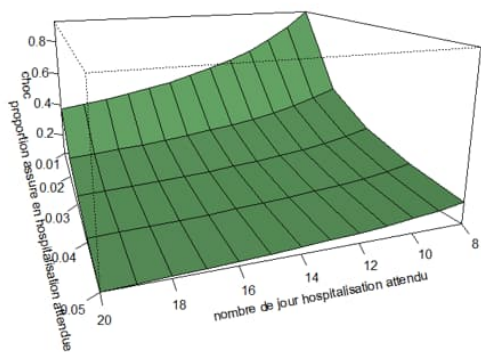
Toutes les pandémies du modèle SEIR stochastique ne donnent pas en sortie un nombre d'infectés considérable car beaucoup de pandémies avortent assez tôt. De plus, l'INVS, le CDC et l'European Actuarial Consultive Group estiment que la taille finale d'une épidémie est de l'ordre de 20 à 35% donc le modèle est calibré pour avoir des taux d'infection globaux proches de ces estimations.

Pour déterminer le quantile d'une pandémie correspondant à un scénario bicentenaire, on se base sur une fréquence annuelle de pandémie avancée par Swiss Re (2007) qui est de 3,33% soit une pandémie toutes les 30 années. Après calcul, on obtient un quantile bicentenaire de 85% pour une pandémie basée sur cette fréquence.

Nous évaluons par la suite le montant en sortie du modèle avec l'exposition donnée dans la directive S2 afin d'avoir le choc à appliquer en cas de pandémie sur l'exposition en hospitalisation. Cette exposition étant dépendante de plusieurs variables attendues par l'assureur, nous regardons comment le choc évolue en fonction de celles-ci :

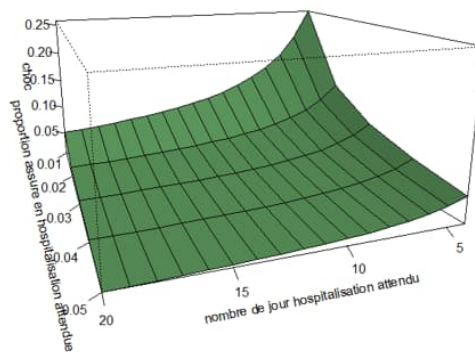
Scénario moyen

Niveau de choc en fonction des variables d'exposition attendues



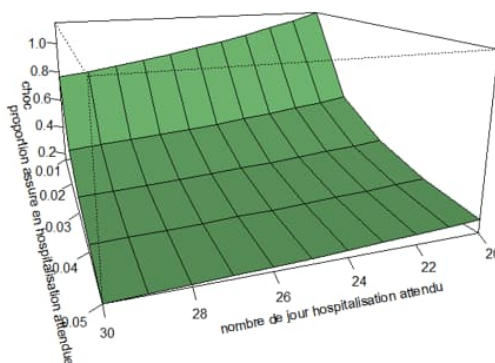
Scénario faible

Niveau de choc en fonction des variables d'exposition attendues



Scénario dur

Niveau de choc en fonction des variables d'exposition attendues



En TVE, on s'intéresse à la queue de distribution c'est-à-dire les observations qui ont les statistiques d'ordre les plus élevées. L'intérêt étant d'avoir une idée sur l'épaisseur de la queue de distribution pour modéliser les quantiles extrêmes.

Deux méthodes phares sont régulièrement utilisées : le Bloc Maxima et le Peak Over Threshold (POT). La méthode Bloc Maxima en théorie des valeurs extrêmes consiste comme son nom l'indique à diviser les observations par blocs temporels de même taille et considérer uniquement les maximums de chacun des blocs. Dans la méthode POT, l'objectif est de déterminer un seuil u optimal, et par la suite estimer les paramètres du modèle dont le paramètre de queue.

La détermination du seuil u optimal est crucial afin d'avoir un compromis biais/variance raisonnable. Notons également que le seuil choisi ne doit pas être trop petit afin de ne pas considérer des observations qui ne seraient pas extrêmes, tout en n'étant pas trop élevé afin d'avoir suffisamment d'observations. La détermination du seuil repose sur des méthodes graphiques qui ne sont pas toujours évidentes surtout en présence d'une quantité de données assez faible.

Afin d'améliorer la qualité de la modélisation, on renforce les modèles en considérant le comportement moyen existant dans les données qui a pour effet de faire une pondération satisfaisante entre les observations de la queue et le reste. Ce modèle amélioré est le modèle G-E-GPD (Gaussien, Exponentiel, Pareto Généralisée).

Par manque de données françaises sur les épidémies passées qui sont notamment constituées d'épidémies de grippe annuelle, nous avons utilisé les données fournies par le CDC qui remontent à 1900 et vont jusqu'à 2015 mais concernent la population américaine. L'année 2020 sert d'année 2016 pour capter les effets du covid19.

Démarche :

- Pour obtenir les résultats, nous prenons les variations de mortalité annuelles d'épidémie que nous modélisons par une G-E-GPD
- La G-E-GPD modélisée sert ensuite à établir des réalisations via des simulations de Monte-Carlo. Rappelons que les simulations de Monte-Carlo sont utiles pour obtenir un certain nombre de réalisations probables qui ne se sont pas forcément réalisées.
- Sur la base de ces réalisations, nous prenons ensuite le quantile correspondant à une épidémie bicentenaire

Pour l'ensemble des classes d'âge avec et sans prise en compte de la grippe 1918 (grippe espagnole), on obtient donc nos différents chocs de mortalité auxquels on applique un coefficient de 60% comme l'a fait l'European Actuarial Consultative Group sur les taux européens pour obtenir un choc spécifique aux assurés. On voit que globalement on serait en-dessous d'un choc à 0,15 point.

Classe d'âge	Choc de mortalité	
	Avec grippe 1918	Sans grippe 1918
15-24	0.12	0.1
25-34	0.15	0.16
35-44	0.17	0.16
45-54	0.19	0.2
55-64	0.21	0.20
65-74	0.27	0.28

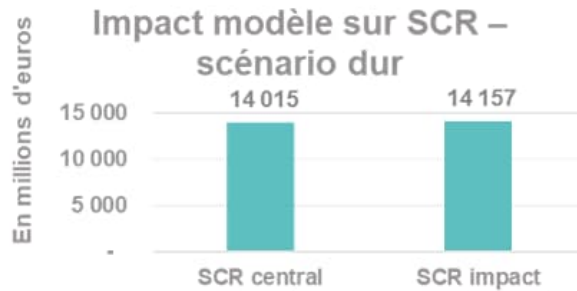
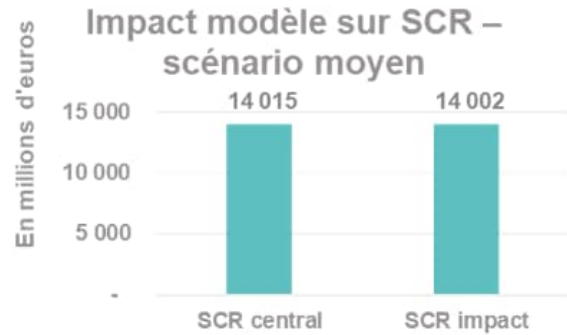
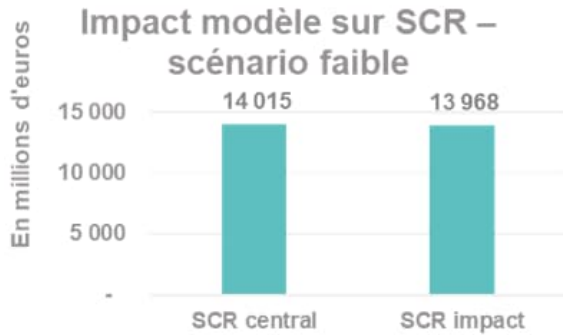
Choc de mortalité par classe d'âge

Classe d'âge	Choc de mortalité	
	Avec grippe 1918	Sans grippe 1918
15-24	0.072	0.06
25-34	0.09	0.096
35-44	0.1	0.096
45-54	0.11	0.12
55-64	0.13	0.12
65-74	0.16	0.17

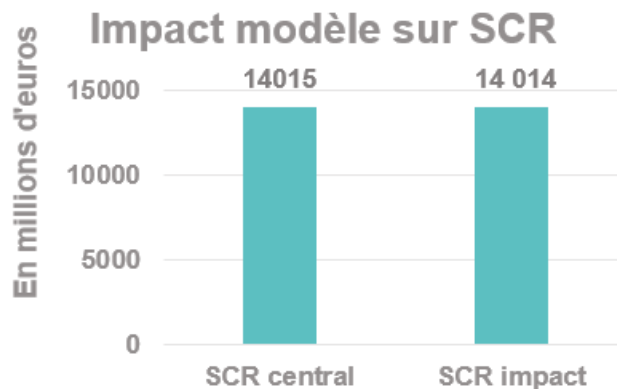
Choc de mortalité par classe d'âge avec coefficient de 60%

A partir des chocs obtenus par modélisation épidémiologique pour l'hospitalisation et par théorie des valeurs extrêmes pour la surmortalité, nous sommes en mesure de déterminer l'impact de ces nouveaux chocs sur le SCR du Crédit Agricole Assurances au 31/12/2020 puis de conclure.

Pour le SCR Santé CAT Pandémie, en faisant un traitement nous permettant de passer d'un choc de population générale à la population assurée, nous obtenons les résultats suivants sur le SCR de la compagnie :



Pour le SCR Vie CAT, nous appliquons les différents taux sur mortalité par classe d'âge en sortie du modèle et nous obtenons ce résultat :



Dans les deux études, les nouveaux chocs ont très peu d'impact sur le capital de solvabilité requis de la compagnie car ce sont avant tout des chocs construits pour deux risques biométriques alors que le Crédit Agricole Assurances a un profil de risque très exposé au risque de marché. Nous concluons donc qu'au vu de la faible exposition de la compagnie aux risques biométriques, une étude spécifique pour mieux appréhender ces risques dans un contexte de pandémie n'est pas nécessaire et le mémoire a permis de le démontrer de manière expérimentale. Cela évite de passer à un modèle interne qui est un processus contraignant et très coûteux.

Synthesis Note

An until now unimaginable situation like covid19 took the world by surprise and can happen again anytime in the future. Regulation give financial organisations a chance to flexibly and safely react to a pandemic event. Under Solvency II, EU-based insurance companies must hold sufficient eligible own funds on an ongoing basis to cover their Solvency Capital Requirement. It enables insurance businesses to absorb significant losses and give confidence to policyholders and beneficiaries that payments will be made as they fall due. There are underwriting risks in all health, life and non-life insurance, namely, potential loss an insurer may have resulting from faulty underwriting. The impact of a possible catastrophic event is captured in the SCR calculation in all three underwriting risk modules: the non-life underwriting risk, the life underwriting risk and the health underwriting risk. All of them are calculated in a catastrophic (CAT) submodule but we are going to look at disaster events related to pandemics.

Starting from the focus on the pandemic risk, the objective of the work will be to propose shocks within the standard formula that are different from those set uniformly by the Solvency 2 Directive and to draw the impacts on the SCR of Crédit Agricole Assurances, whose risk profile we will recall. The compartments concerned in the standard formula are the Health module and the Life module. Within the health module, there are 3 sub-modules: the health sub-module for life risks, the health sub-module for non-life risks and the CAT sub-module. It is this last sub-module that interests us and it is itself made up of 3 compartments which are the risk of mass accident, the risk of accident concentration and the pandemic risk, it is therefore the shock of the pandemic risk that we will try to redefine. On the Life module, it is a different proposal from the shock applied to the CAT sub-module that we will make. We recall that the Life CAT SCR is taken into account in the standard formula by considering a 0.15 percentage point increase in mortality rates and the Health CAT Pandemic SCR is taken into account by applying a coefficient of 0.075 per thousand to the exposure to income protection risk in the event of a pandemic and a coefficient of 40% to the exposure to hospitalization costs in the event of a pandemic.

To propose different shocks from those of the standard formula, we will use two mathematical tools. On the one hand, we will use the epidemiological model to calculate the capital requirement of the SCR Health CAT Pandemic. This will allow us to determine the shock of the cost of hospitalization in a pandemic context. On the other hand, we will use the extreme value theory (EVT) for the Pandemic Life SCR in order to model the extreme quantiles of excess mortality due to a pandemic.

The functioning of a basic epidemiological model is that it is based on the passage from one state to another: an individual starts in the healthy state because he has not yet been contaminated, then at the moment of contamination, he becomes exposed and begins his incubation period and at the end of this period, he becomes in turn an infected person able of contaminating other healthy individuals. From this moment on, he has two possible outcomes, either he recovers after a while or he dies.

In this dissertation, we will use the SEIR model in a stochastic form that simulates contacts in a probabilistic way between individuals. It avoids making all contacts systematically contagious and does not set incubation and infection times for all individuals. To simulate contacts, the model uses an Erdős-Rényi graph.

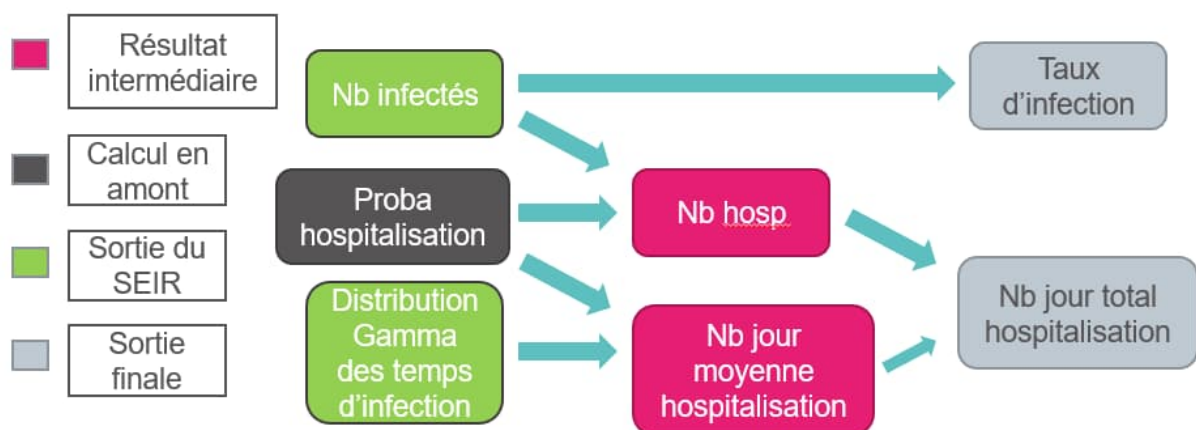
Then, from these contacts, the epidemic evolves with the incubation time T_E and the infection time T_I which are here random with gamma distribution. To avoid that all contacts result in an infection, a contamination time variable X_C is used.

The choice of the gamma distribution is motivated by the fact that it is adapted to temporal phenomena and that it has more credible realizations than the exponential distribution which is also adapted for this type of phenomena.

The basic model has been designed to take into account only epidemics without lethality, and it does not provide information on possible hospitalizations, but we want to use the epidemiological model for a risk like hospitalization. To take it into account, we base ourselves on a study of the Institut Nationale de Veille Sanitaire (INVS) which evaluated the level of risk in case of a pandemic, this study gives us information on the distribution of cases by age, the percentage of the population at risk and the hospitalization rate.

From these data, the challenge is to determine the probability of hospitalization to be applied to the number of infected persons at the output of the stochastic SEIR model and then to obtain the number of days of hospitalization. The probability of hospitalization allows us to determine the number of days of hospitalization associated with the distribution of infection times: we take the corresponding quantile.

Our model is presented as follows :



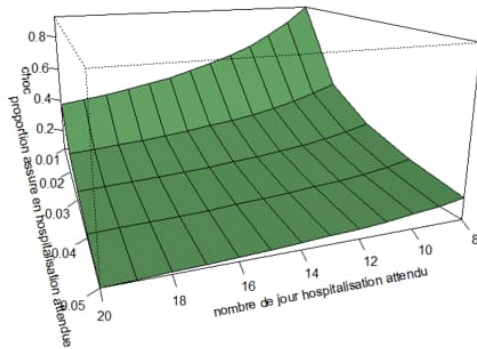
Not all pandemics in the stochastic SEIR model result in a significant number of infected people as many pandemics abort early. Moreover, the INVS, the CDC and the European Actuarial Consultative Group estimate that the final size of an epidemic is in the order of 20 to 35%, so the model is calibrated to have global infection rates close to these estimates.

To determine the quantile of a pandemic corresponding to a bicentennial scenario, we base ourselves on an annual frequency of pandemic put forward by Swiss Re (2007) which is 3.33%, i.e. a pandemic every 30 years. After calculation, we obtain a bicentennial quantile of 85% for a pandemic based on this frequency.

We then evaluate the output of the model with the exposure given in the S2 guideline in order to have the shock to be applied in case of a pandemic on the hospitalization exposure. As this exposure is dependent on several variables expected by the insurer, we look at how the shock evolves according to them:

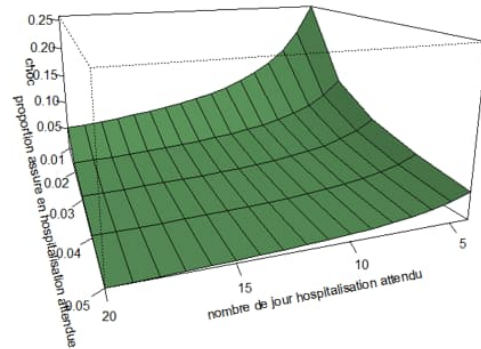
Scénario moyen

Niveau de choc en fonction des variables d'exposition attendues



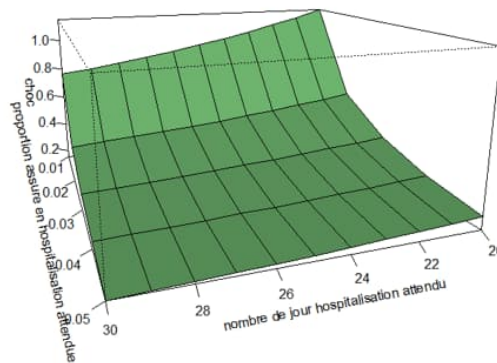
Scénario faible

Niveau de choc en fonction des variables d'exposition attendues



Scénario dur

Niveau de choc en fonction des variables d'exposition attendues



In EVT, we are interested in the tail of the distribution, i.e. the observations with the highest order statistics. The interest is to have an idea on the thickness of the tail of distribution to model the extreme quantiles.

Two main methods are regularly used: the Block Maxima and the Peak Over Threshold (POT). The Block Maxima method in extreme value theory consists, as its name indicates, in dividing the observations by temporal Blocks of the same size and considering only the maximums of each block. In the POT method, the objective is to determine an optimal threshold u , and then to estimate the parameters of the model including the tail parameter.

The determination of the optimal threshold u is crucial in order to have a reasonable bias/variance compromise. Note also that the chosen threshold should not be too small in order not to consider observations that would not be extreme, while not being too high in order to have enough observations. The determination of the threshold relies on graphical methods that are not always obvious, especially in the presence of a small amount of data.

In order to improve the quality of the modeling, the models are strengthened by considering the average behavior existing in the data which has the effect of making a satisfactory weighting between the observations of the tail and the rest. This improved model is the G-E-GPD (Gaussian, Exponential, Generalized Pareto) model.

Because of the lack of French data on past epidemics, which include annual influenza epidemics, we used data provided by the CDC that go back to 1900 and go to 2015 but are for the US population. The year 2020 is used as the year 2016 to capture the effects of covid19. The results can therefore only be interpreted as an upper bound of what could be obtained with French data since these are not globally larger from the years of availability.

Approach :

- To obtain the results, we take annual epidemic mortality variations that we model by a G-E-GPD
- The modeled G-E-GPD is then used to establish realizations via Monte-Carlo simulations. Recall that Monte-Carlo simulations are useful to obtain a number of probable realizations that may not have occurred.
- On the basis of these realizations, we then take the quantile corresponding to a bicentennial epidemic

For all the age groups with and without taking into account the 1918 flu (Spanish flu), we obtain our different mortality shocks to which we apply a coefficient of 60% as the European Actuarial Consultative Group did on the European rates to obtain a shock specific to the insured population. We see that globally we would be below a shock of 0.15 points.

Classe d'âge	Choc de mortalité	
	Avec grippe 1918	Sans grippe 1918
15-24	0.12	0.1
25-34	0.15	0.16
35-44	0.17	0.16
45-54	0.19	0.2
55-64	0.21	0.20
65-74	0.27	0.28

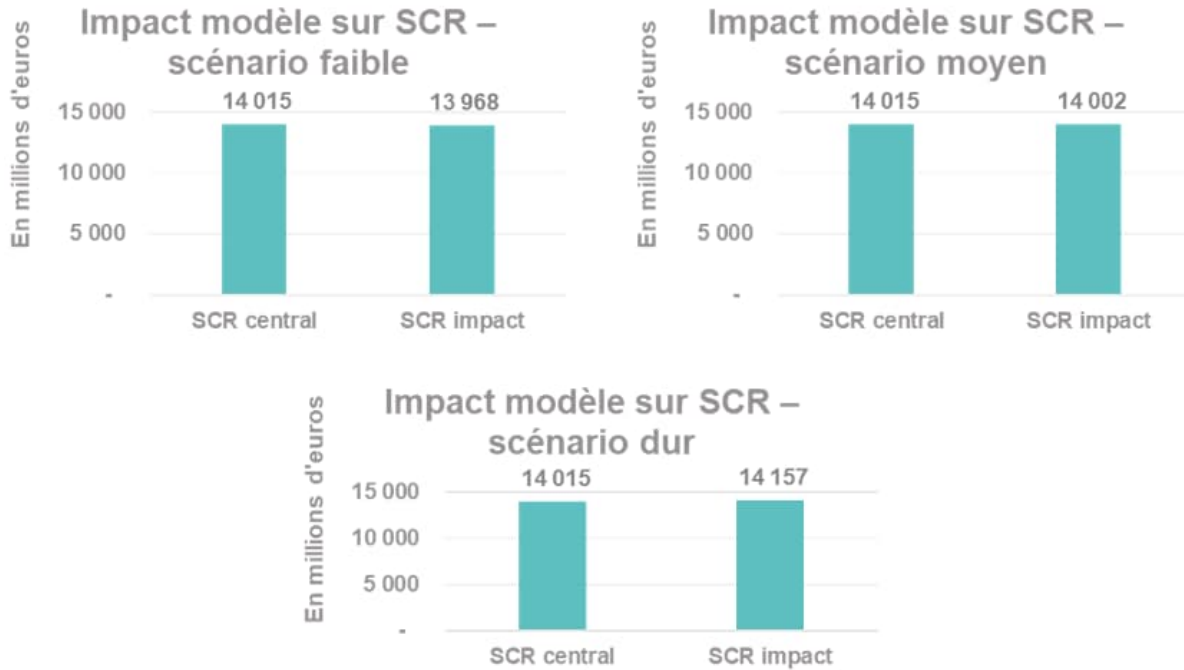
Choc de mortalité par classe d'âge

Classe d'âge	Choc de mortalité	
	Avec grippe 1918	Sans grippe 1918
15-24	0.072	0.06
25-34	0.09	0.096
35-44	0.1	0.096
45-54	0.11	0.12
55-64	0.13	0.12
65-74	0.16	0.17

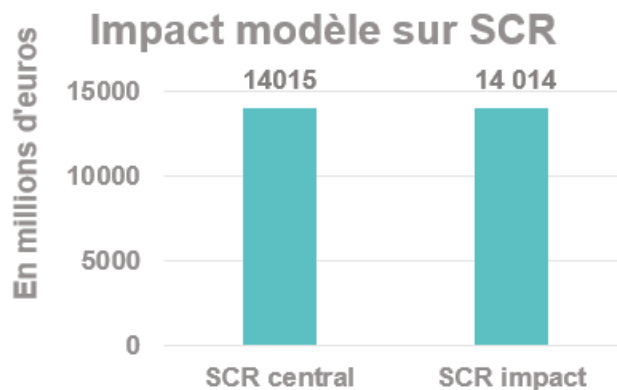
Choc de mortalité par classe d'âge avec coefficient de 60%

From the shocks obtained by epidemiological modeling for hospitalization and by extreme value theory for excess mortality, we are able to determine the impact of these new shocks on the SCR of Crédit Agricole Assurances at 31/12/2020 and to conclude.

For the Pandemic CAT Health SCR, by doing a treatment that allows us to go from a general population shock to the insured population, we get the following results on the company's SCR:



For the SCR Life CAT, we apply the different rates on mortality by age class at the output of the model and we obtain this result:



In both studies, the new shocks have very little impact on the company's solvency capital requirement because they are primarily shocks constructed for two biometric risks, whereas Crédit Agricole Assurances has a risk profile that is highly exposed to market risk. We therefore conclude that given the company's low exposure to biometric risks, a specific study to better understand these risks in a pandemic context is not necessary and the thesis has allowed us to demonstrate this experimentally. This avoids the need to switch to an internal model, which is a cumbersome and very expensive process.

Table des matières

Résumé.....	1
Abstract.....	2
Note de Synthèse.....	3
Synthesis Note.....	8
Table des matières.....	13
INTRODUCTION.....	16
PARTIE I : Pandémie.....	18
1. Notion de pandémie.....	18
1.1 Définition.....	18
1.1 L'ampleur géographique d'une pandémie.....	18
1.3 La vitesse de propagation d'une pandémie.....	19
1.4 Traitement de la pandémie par l'OMS.....	19
2. Les exemples historiques de pandémie.....	20
2.1 La peste noire (1347-1352).....	21
2.2 La grippe espagnole A H1N1 (1918-1919).....	21
2.3 Grippe asiatique A H2N2 (1957-1958).....	23
2.4 Grippe de Hong-Kong A H3N2 (1968-1969).....	23
2.5 Grippe A H1N1 (2009).....	24
3. Lien entre la biodiversité, les pandémies et le secteur de l'assurance.....	24
4. De la perte de biodiversité aux pandémies.....	25
5. Prise en compte du risque pandémie au sein de solvabilité II.....	29
5.1 Solvabilité II.....	29
5.2 Les 3 piliers de Solvabilité II.....	30
5.2.1 La formule standard du SCR.....	31
5.2.2 Le risque de pandémie dans la formule standard.....	31
6. Profil CAA.....	32
6.1 Les risques.....	32
6.2 Méthodes utilisées pour la réduction du risque.....	34
PARTIE II : Modélisation épidémiologique pour le SCR Santé CAT Pandémie.....	35
1. Modèle déterministe.....	35
1.1 Intérêt d'un modèle épidémiologique.....	35
1.2 Hypothèses et Notations.....	36
1.3 Le modèle SIR.....	37

1.4 Le modèle SEIRD.....	39
2. Les paramètres des modèles	40
2.1 Taux de reproduction et contact.....	40
2.2 Le taux de létalité	42
3. Modèle Stochastique.....	43
3.1 Réseau de contact	43
3.2 Les temps d'incubation et d'infection	45
3.3 La transmission entre deux individus	48
3.4 Calcul du taux de reproduction de base \mathcal{R}_0	49
4. Données et construction de scénario.....	50
4.1 Les cas par âge	50
4.2 Les cas à risque par âge	51
4.3 Taux d'hospitalisation par âge	51
4.4 Modélisation des hospitalisations dans le modèle SEIR stochastique	52
4.4.1 Les probabilités.....	53
4.4.2 Probabilité d'hospitalisation	53
4.5 Nombre de jours d'hospitalisation.....	55
5. Choix du paramètre \mathcal{R}_0	56
5.1 Calibrage de β	56
5.2 Calibrage de p	57
6. Modélisation des pandémies.....	58
7. Approche pour l'évaluation du besoin en capital	61
8. Application de la formule de Guilbert	62
9. Le choc du SCR	63
PARTIE III : Modélisation du choc SCR Vie CAT.....	65
1. Théorie des valeurs extrêmes.....	65
1.1 Loi GEV.....	65
1.2 Blocs Maxima	68
1.3 Loi GPD et POT	69
1.3.1 GPD	69
1.3.2 POT	70
2. Méthodes statistiques pour l'estimation de l'indice de queue	71
2.1 Rappel sur le maximum de vraisemblance	71
2.2 Maximum de vraisemblance pour la loi GPD.....	73
3. Estimateurs spécifiques de la queue dans la GPD	74
4. Méthodes pour le choix du seuil.....	75

4.1 Mean excess plot.....	75
4.2 Hill plot.....	77
5. G-E-GPD.....	78
5.1 Intérêt et fonctionnement d'un modèle hybride.....	78
5.2 Formalisation du modèle hybride à trois composantes.....	80
5.3 Les hypothèses du modèle.....	81
5.4 Algorithme itératif pour le modèle hybride.....	82
5.5 Performance du G-E-GPD avec des données simulées.....	84
5.6 Application à des taux bruts de mortalité dû à une épidémie.....	86
5.6.1 Démarche.....	87
5.6.2 Résultats.....	88
PARTIE IV : Impact des chocs estimés.....	90
1. Indicateurs et traitements.....	90
1.1 Les indicateurs de base.....	90
1.2 De la population générale à la population assurée.....	91
2. Analyse des résultats.....	93
2.1 Résultats impact Santé CAT.....	93
2.2 Résultats impact Vie CAT.....	95
CONCLUSION.....	98
BIBLIOGRAPHIE.....	100
ANNEXE.....	102

INTRODUCTION

En avril 2020, quand les effets du covid19 ne faisaient que commencer, certains acteurs du marché de l'assurance s'attendaient à ce que ce soit l'évènement le plus coûteux de l'histoire du marché des assurances. Il est donc important de voir comment les compagnies d'assurance se préparent pour les évènements de type pandémie qui apportent un vent de panique et si la réglementation prévoit suffisamment de capital.

D'autant plus que le risque pandémie peut voir sa fréquence augmentée dans les années à venir à cause de la biodiversité et les migrations d'agents pathogènes.

Les grandes pandémies sont par nature rares mais peuvent être sévères. Elles représentent un des risques les plus menaçants pour les sociétés et par extension pour les entités qui sont garantes du risque sous-jacent c'est-à-dire les compagnies d'assurance. Les grandes pandémies se caractérisent par leur imprévisibilité aussi bien pour le moment de leur apparition que pour leur durée et par leur échelle géographique non localisée mais de grande ampleur. Leur spécificité réside aussi dans la nature exogène et endogène des facteurs de propagation du risque. La propagation du virus dépend du comportement des individus et des choix collectifs.

Un tel évènement affecte directement le business des compagnies d'assurance notamment sur les lignes de business en lien avec la santé et la vie. Par sa nature, une pandémie va premièrement engendrer une dégradation de la santé puis peut translater vers une déviation significative de la mortalité. L'impact d'une pandémie chez un assureur va largement dépendre de son portefeuille notamment de sa constitution par classe d'âge.

Concernant l'activité vie, les assureurs sont exposés au risque de mortalité et de longévité par leurs polices d'assurance. Dépendant de la balance de chaque produit dans le portefeuille de l'assureur, le risque de sinistre plus élevé dû à un taux de mortalité plus élevé peut être compensé par une diminution des paiements futurs des prestations de longévité.

Concernant la santé, l'impact va beaucoup dépendre des conséquences de la maladie elle-même. Toutefois, en cas de très forte mortalité, les effets sur la ligne de business de la santé seront moindres par rapport à celle du business sur la vie. Les assureurs font face à l'augmentation de la sinistralité sur les soins et les hospitalisations.

Le risque de pandémie n'est pas nouveau et a déjà été évalué et modélisé par des actuaires. Cette évaluation a permis d'intégrer le risque de pandémie dans les stratégies préventives des assureurs qui sont en particulier en conformité avec le capital de solvabilité requis recommandé par la directive Solvabilité II.

Par ailleurs, la formule standard de solvabilité 2 permet de déterminer le capital de solvabilité requis en cas de pandémie en appliquant des chocs par rapport au scénario central. Ces chocs sont à utiliser de manière uniforme pour tous les assureurs bien qu'ayant été déterminés par une étude unique remontant à plusieurs années maintenant. Il peut donc être intéressant d'être à même de faire une autre étude pour calibrer des chocs différents et voir à quel point le scénario choqué dévie de celui issu de la formule standard.

Dans notre mémoire, nous tenterons donc de modéliser ces nouveaux chocs de pandémie en santé et en vie après avoir rappelé le lien existant entre biodiversité et pandémie qui peut être une réelle

source de sinistralité à l'avenir. Cela permettra en partie de répondre au questionnement sur la nécessité ou non d'utiliser un modèle interne pour ces risques.

PARTIE I : Pandémie

1. Notion de pandémie

1.1 Définition

Avant de commencer par directement définir le terme « pandémie », il est nécessaire de bien comprendre les différences existantes dans le langage épidémiologique quand il est question d'endémie, épidémie et pandémie car pour le lecteur non averti la distinction n'est pas évidente bien qu'il s'agisse de transmission de maladie.

Tout d'abord, l'endémie se définit par la présence habituelle d'une maladie, en générale infectieuse, dans une population déterminée ou une région précise, avec une incidence stable. Typiquement, le paludisme est endémique dans de nombreux pays africains et l'hépatite A est endémique en Thaïlande. Le caractère endémique fait qu'il est bien souvent recommandé de se faire vacciner avant d'aller dans les pays africains concernés et en Thaïlande. En France, on parlerait de la grippe saisonnière.

Ensuite, nous avons l'épidémie qui se définit par la croissance rapide de nouveaux cas de contaminations d'une maladie dans une région donnée et pendant une période donnée. Au départ, seules les maladies infectieuses comme la variole ou encore la fièvre hémorragique due au virus Ebola étaient associées à des épidémies. Depuis le début de 21^{ème} siècle, on assiste à l'émergence de nouveaux virus comme le coronavirus.

Enfin, une pandémie (du grec ancien *pân* « tout » et *demós* « peuple ») est une épidémie présente sur une large zone géographique internationale. Ainsi, quand on parle de pandémie, on parle du développement et de la propagation rapide d'une maladie à une échelle suffisamment grande pouvant dépasser une région du globe où se concentrent plusieurs pays ou encore tout un continent entier.

1.1 L'ampleur géographique d'une pandémie

Comme nous l'avons vu la pandémie est à distinguer d'une simple épidémie ou encore d'une endémie. Dans sa définition, nous avons précisé qu'elle correspond à une épidémie ayant un caractère géographique international très prononcé là où une simple épidémie et une endémie sont très statiques dans leur configuration géographique.

L'ampleur du phénomène de transmission sous-jacent à une épidémie est donc importante à cartographier pour savoir s'il s'agit en réalité d'une pandémie. A cet effet, la définition donnée à ce terme par l'OMS a varié à plusieurs reprises. Aujourd'hui, il est admis que, pour parler de pandémie, la maladie doit toucher au moins deux continents. Pour exemple, nous avons la pandémie due au virus Sars-Cov2 plus connue sous le nom de pandémie de la Covid19 mais aussi la pandémie de la grippe espagnole.

La présence géographique étendue ne justifie pas à elle seule le caractère pandémique. Si on prend l'exemple du Chikungunya qui est une maladie virale se manifestant en Afrique, en Asie, dans le sous-continent indien et même récemment en Europe on pourrait croire qu'il s'agit d'une pandémie si on s'en tient à la présence sur au moins deux continents mais il s'agit d'une endémie car bien que présent sur différents continents, le Chikungunya est très prévisible par son caractère habituel et sa faible propagation au niveau régional.

1.3 La vitesse de propagation d'une pandémie

Comme nous l'avons vu, la géographie joue un rôle dans la définition d'une pandémie par l'étendue de celle-ci. Mais bien que la surface totale sur laquelle s'étend le phénomène soit conséquente, la rapidité à laquelle cela se produit est assez variable. Quand on regarde un virus comme le Syndrome de l'Immunodéficience Acquise (SIDA) et une maladie chronique comme l'obésité qui sont considérés comme à l'origine d'un phénomène de pandémie par l'OMS, on s'aperçoit que ces derniers sont considérablement présents à la surface du globe mais qu'en terme de propagation ils sont lents et très prévisibles. Leur propagation n'est pas vraiment sujette à une transmission hasardeuse mais est surtout liée aux comportements des individus. C'est d'ailleurs pourquoi il est possible de les ralentir par de la prévention afin que les individus soient au fait des bonnes pratiques.

Une fois qu'on a mis en évidence ces éléments, on se rend bien compte que pour parler de la rapidité de propagation, il faut connaître l'origine de la pandémie et son mode de transmission.

Généralement, les pandémies ont une origine bactérienne ou sont le fruit d'une mutation sur des maladies déjà existantes faisant l'objet de zoonoses c'est-à-dire des maladies issues d'un virus ayant muté au sein d'un organisme animal puis transmis à l'homme. Ainsi, une pandémie sera rapide dans sa propagation à partir du moment où les organismes atteints par sa souche virale sont non immunisés. Il est donc question d'avoir des individus tous susceptibles de se faire infectés au sein d'une population.

1.4 Traitement de la pandémie par l'OMS

Par sa portée internationale, l'émergence d'une pandémie se doit d'être surveillée avec un système d'information entre les différents pays, l'OMS se porte garante de cette surveillance. À tout moment, l'OMS suit l'évolution des maladies infectieuses, donne l'alerte le cas échéant, apporte ses compétences et prend les mesures qui s'imposent pour protéger les populations des conséquences des épidémies, quelle qu'en soit l'origine.

Le Règlement Sanitaire Internationale ou RSI (2005), est entré en vigueur le 15 juin 2007. Son objet et sa portée consistent à prévenir la propagation internationale des maladies, à s'en protéger, à la maîtriser et à y réagir par une action de santé publique proportionnée et limitée aux risques qu'elle présente pour la santé publique, en évitant de créer des entraves inutiles au trafic et au commerce internationaux. Le RSI (2005) constitue un cadre dans lequel s'inscrivent les opérations d'alerte et d'actions rapides en cas d'épidémie que mène déjà l'OMS en collaboration avec les pays pour endiguer les flambées d'ampleur internationale et renforcer la sécurité sanitaire internationale.

Les données épidémiologiques et les informations opérationnelles sur les flambées changent constamment et rapidement. L'OMS a élaboré un système de gestion global permettant de traiter les

informations cruciales sur les flambées et d'assurer leur communication rapide et fiable entre les principaux intervenants dans le domaine de la santé publique internationale (bureaux régionaux, bureaux de pays et centres collaborateurs de l'OMS, partenaires du réseau mondial d'alerte et d'action en cas d'épidémie).

Le système de gestion des événements de l'OMS permet de suivre en permanence l'évolution des opérations d'alerte et d'action et fournit des informations pour agir de manière systématique, ce qui permet à l'OMS et au réseau mondial d'alerte et d'action en cas d'épidémie de mieux se préparer, de réagir plus vite et de gérer les ressources plus efficacement. Ce système est actuellement renforcé pour intégrer les aspects opérationnels de l'alerte et de l'action du Règlement sanitaire international révisé.

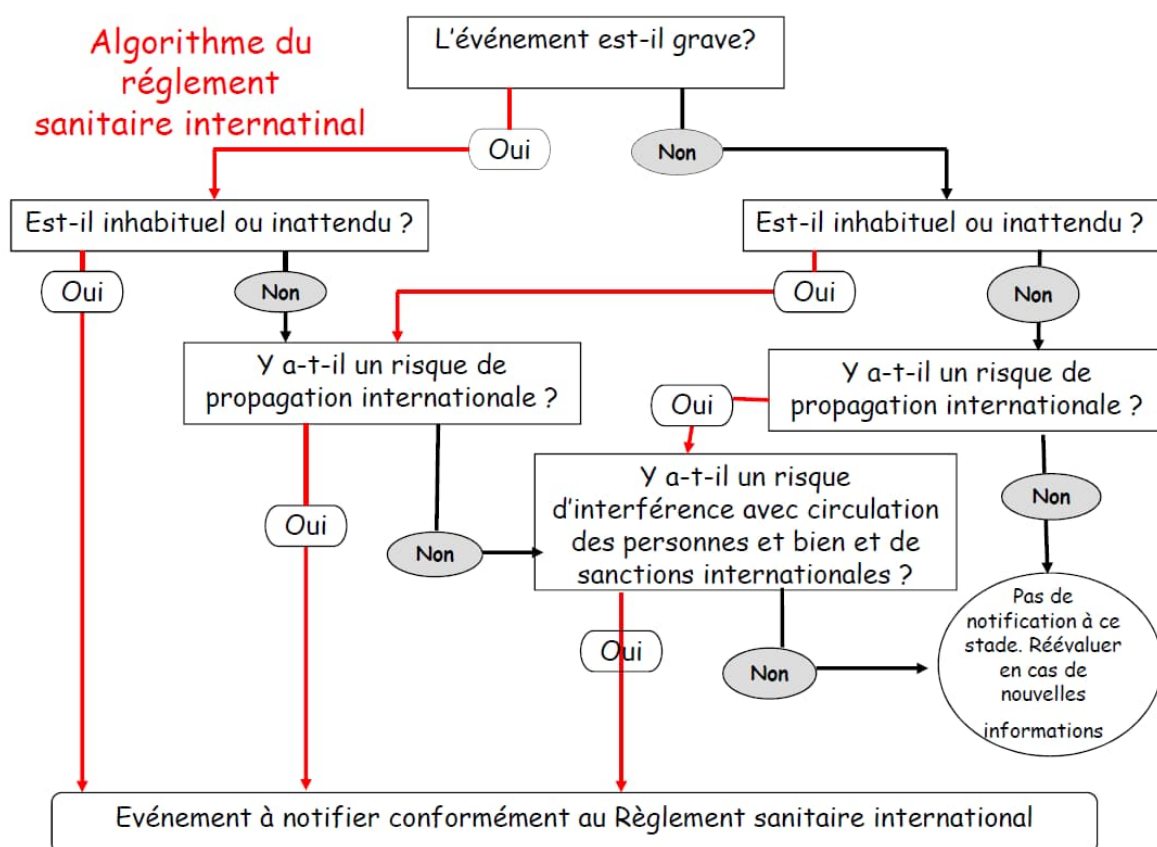


Schéma du traitement de l'information pour alerter la pandémie au sein de l'OMS

2. Les exemples historiques de pandémie

Dans l'histoire de l'humanité, les épidémies et pandémies sont devenues possibles à partir du moment où les groupes humains ont atteint une taille suffisante avec des contacts plus fréquents entre différents groupes. Les pandémies étaient peu nombreuses au début et même inexistantes

laissant la place à de petites épidémies mais par la suite leur occurrence est devenue une réalité avec les liaisons commerciales impliquant des villes de plus en plus peuplées et connectées par des routes. Dans la suite, nous allons parler de certaines d'entre-elles qui ont laissé des traces dans l'histoire.

2.1 La peste noire (1347-1352)

La peste noire est une pandémie de peste (principalement bubonique) qui a sévi au milieu du 14^{ème} siècle (pendant le Moyen Âge). Cette pandémie touche l'Eurasie, l'Afrique du Nord et peut-être l'Afrique subsaharienne. Elle n'est ni la première ni la dernière pandémie de peste, mais c'est la première pandémie à avoir été bien décrite par les chroniqueurs de l'époque. Elle est parfois également appelée grande peste.

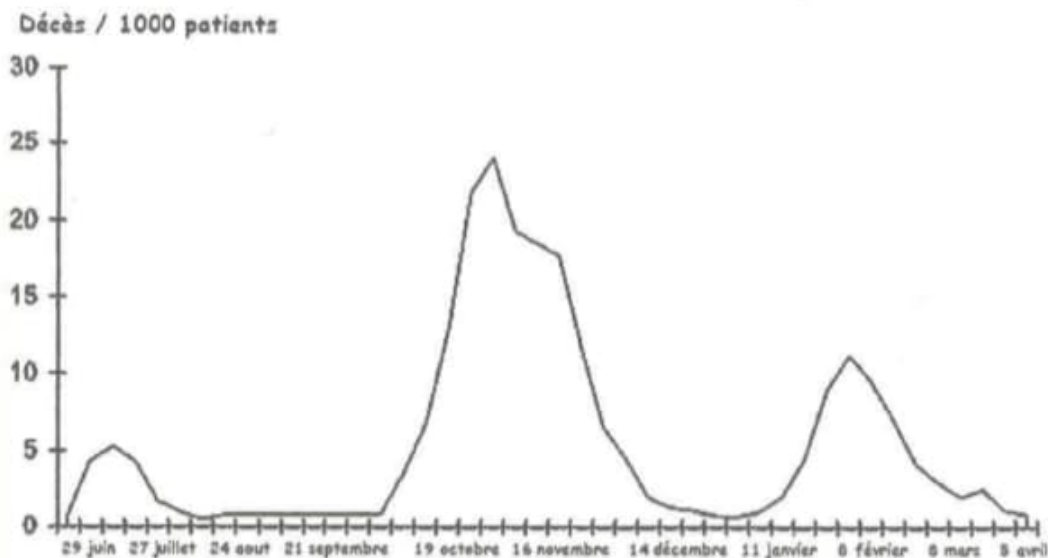
Après avoir sévi en Chine, la pandémie de peste noire arrive en 1346 en Asie centrale, parmi les troupes mongoles assiégeant le port de Caffa, sur la mer Noire, tenu par des marchands génois. La maladie, se manifestant par d'horribles bubons, se propage ensuite à l'Afrique du Nord puis à l'Italie et à la France, où elle arrive par le port de Marseille via des navires génois.

Elle tue entre 30 et 50 % des Européens en cinq ou six ans (1347-1352), faisant environ 25 millions de victimes. Ses conséquences sur la civilisation européenne sont sévères et longues.

2.2 La grippe espagnole A H1N1 (1918-1919)

Il s'agit de la pandémie la plus meurtrière qu'ait connue l'humanité et même largement devant la récente pandémie de covid19 qui a moins de victimes alors que la population mondiale est 4 fois plus grande qu'à la période de la grippe espagnole. Cette pandémie a touché près d'un tiers de la population mondiale de l'époque. Il est difficile d'estimer le nombre de décès liés à la grippe car les décès étaient surtout liés aux complications pulmonaires qui la suivirent. En effet, avec les anciennes gripes, seuls 1% des grippés présentaient des complications pulmonaires plus ou moins graves et parmi ceux-ci seuls 1% des cas étaient mortels. Mais avec la grippe espagnole, ce fut près de 15 à 30% des grippés qui représentaient des complications pulmonaires et environ 10% de ces cas décédèrent. Quoi qu'il en soit, la plupart des estimations dont celle de l'Institut Pasteur estimait un nombre de décès de 20 à 50 millions et de récentes réévaluations estiment jusqu'à 100 millions de morts.

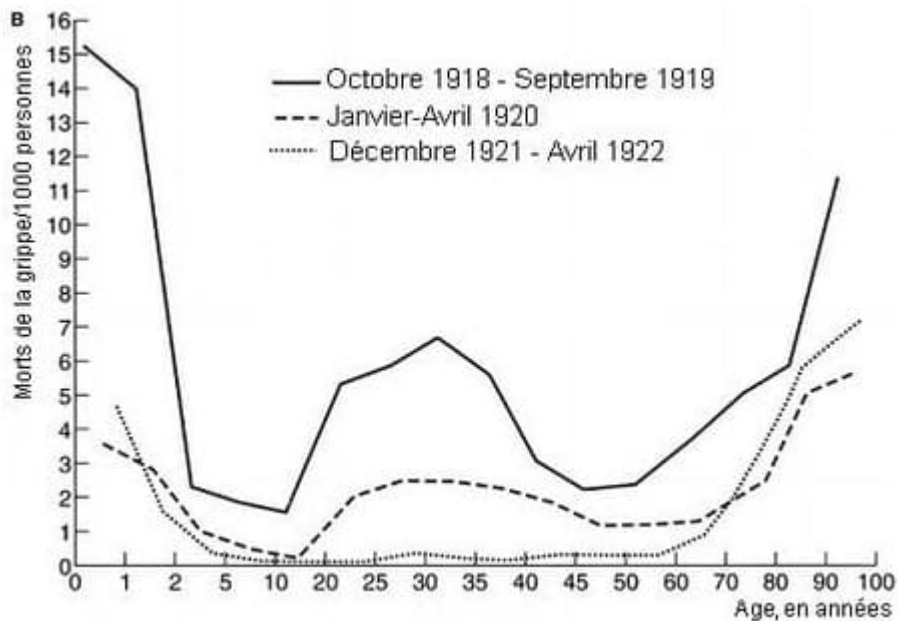
Cette pandémie a eu trois vagues principales ayant duré au total un an sur les différents continents où elle a sévi. Les trois vagues sont illustrées par ce graphique :



Nous voyons que la première vague a eu lieu au printemps 2018 et est la moins meurtrière des trois avec un taux de mortalité inférieur à 1%. Elle concernait principalement l'Europe de l'Ouest et les Etats-Unis. A cette période, les problèmes liés à cette grippe étaient largement minimisés surtout dans un contexte de première guerre mondiale. La deuxième vague va de septembre à novembre 1918 et est la plus meurtrière avec un virus désormais propagé dans le monde entier. Le virus a eu le temps de muter pour mieux s'adapter à l'humain. Le taux de mortalité qui s'était stabilisé à moins d'1% passe à 3% voire 5%. Enfin, la troisième vague commence début 1919 et est moins meurtrière que la seconde mais plus meurtrière que la première.

Concernant l'origine de virus, il est impossible de connaître la source du virus originel car il n'existe pas d'études sur les souches originelles faute de conservation de celles-ci. Il existe cependant deux hypothèses pour expliquer l'origine du virus responsable de cette pandémie. La première indique qu'il s'agirait d'une mutation d'un virus humain déjà existant à l'époque mais cette hypothèse est assez limitée car alors une bonne partie des populations auraient été immunisées ce qui n'était pas le cas. L'autre hypothèse qui revient dans la plupart des études stipule que le virus proviendrait d'une souche nouvelle, provenant elle-même d'une autre espèce, notamment les espèces aviaires qui sont des réservoirs naturels de beaucoup de virus.

Cette pandémie a aussi eu une particularité au niveau des différentes classes d'âge touchées par celle-ci. Quand il s'agit de grippe, généralement on observe ce qu'on appelle un schéma en « U » traduisant le fait que les personnes en très jeune âge et les personnes les plus âgées ont une proportion de décès significativement plus grande que le reste de la population. Or avec cette pandémie, c'est un schéma en « W » qui fut observé comme illustré ci-dessous. Autrement dit en plus des catégories habituelles, il s'ajoutait aussi les jeunes adultes.



Une explication à la mortalité inhabituelle chez les jeunes adultes serait à la fois leur période de naissance et le moment auquel prend place leur vie de jeune adulte. Etant nés vers la fin du 19^{ème} siècle, ils n'ont pas connu le virus de la grippe A H3N8 et donc n'avaient pas un système immunitaire préparé pour le virus H1N1. De plus, on ne peut pas mettre de côté le conflit mondial qui a réduit l'accès aux soins pour cette tranche d'âge après la guerre.

2.3 Grippe asiatique A H2N2 (1957-1958)

Le virus responsable a été identifié pour la première fois dans la province du Guizhou en Chine. Il s'agit du virus H2N2 issu d'une mutation d'un virus humain chez le canard.

Bien qu'elle soit moins virulente que la grippe espagnole, elle reste encore aujourd'hui une des pandémies les plus meurtrières de l'histoire avec 1 à 4 millions de morts répartis dans le monde selon l'OMS.

2.4 Grippe de Hong-Kong A H3N2 (1968-1969)

Elle tient son origine de la grippe asiatique car le virus H2N2 va muter quelques années plus tard en H3N2 pour provoquer une nouvelle grippe surnommée « grippe de Hong-Kong ». Cette dernière marque les débuts des premiers vaccins antigrippaux efficaces.

Cette pandémie sera classée au niveau 2 de l'indice de gravité de la pandémie, qui en compte 5, de 1 (bénigne) à 5 (très grave). Dans l'immédiat, elle mobilise les systèmes de surveillance et renforce les réseaux internationaux de laboratoire (centres mondiaux et nationaux de référence sur les virus grippaux). Après un ralentissement des découvertes, la recherche sur la structure des virus et le mécanisme de leurs variations est relancée.

A l'échelle mondiale, le bilan humain est d'environ un million de morts selon les estimations de l'OMS. Sur le plan temporel, il s'agit de la dernière pandémie du 20^{ème} siècle.

2.5 Grippe A H1N1 (2009)

L'épidémie de grippe H1N1 a été provoquée par une souche H1N1 du virus influenza de type A, qui est génétiquement une association de virus grippal, porcin, aviaire et humain. Il se transmet facilement d'une personne à l'autre. L'infection n'est pas contractée par l'ingestion de viande de porc et n'est que très rarement contractée par contact avec des porcs infectés.

Après les premières flambées en Amérique du Nord en avril 2009, le nouveau virus grippal s'est propagé rapidement dans le monde. Au moment où l'OMS a déclaré officiellement la situation de pandémie, en juin 2009, 74 pays et territoires avaient notifié des infections confirmées en laboratoire.

Contrairement au tableau classique de la grippe saisonnière, le nouveau virus a provoqué un grand nombre d'infections estivales dans l'hémisphère Nord, puis les niveaux d'activité se sont encore accrus pendant les mois plus froids dans cette région du monde.

Le profil de la mortalité et de la morbidité a également différé de celui observé normalement pour les infections grippales. La plupart des décès dus à la grippe pandémique sont survenus chez des personnes jeunes, y compris chez des sujets normalement en bonne santé. Les femmes enceintes, les jeunes enfants et les personnes de n'importe quel âge souffrant de certaines affections pulmonaires ou médicales chroniques semblaient également exposés à un risque plus élevé de complications ou de formes graves de la maladie. De nombreux cas graves ont été dus à une pneumonie virale, plus difficile à traiter que les pneumonies bactériennes communément associées à la grippe saisonnière. De nombreux patients ont dû être placés en soins intensifs.

Des études ont cependant montré qu'une proportion importante des personnes âgées de 65 ans et plus avaient une certaine immunité. Cela semble indiquer qu'un certain nombre de personnes, parmi les plus âgées, pourraient bénéficier d'une protection croisée, résultant de l'exposition à des virus qui ont circulé dans un passé relativement lointain.

3. Lien entre la biodiversité, les pandémies et le secteur de l'assurance

L'épidémie de COVID-19 a indéniablement remis en question nos habitudes de vie et ébranlé nos croyances. La faune sauvage, en particulier les chauves-souris et peut-être les pangolins ou d'autres carnivores, ont fait la une des journaux pour avoir été potentiellement à l'origine d'une crise d'ampleur mondiale (LeMonde, 2021). Cela a mis en lumière la vulnérabilité de nos sociétés face aux dangers du monde naturel dans lequel nous vivons. Cette prise de conscience a été une surprise dans les pays développés, car une fausse impression de protection s'était répandue avec le sentiment qu'on avait développé des organisations et des technologies suffisamment avancées pour se mettre à l'abri de telles menaces naturelles.

Après la surprise sont venues les questions pratiques. Comment faire face à cet épisode sanitaire sans précédent qui s'est transformé en une crise économique et sociale mondiale ? Quel rôle et quelles responsabilités chacun doit-il assumer ? En raison de sa mission très théorique d'atténuer et

de transférer les risques au sein des sociétés, le secteur de l'assurance a été pris dans la tourmente et blâmé pour son inaction face à son mandat présumé.

Dans la suite, on tentera, dans un premier temps, de faire la lumière sur les connaissances scientifiques qui soutiennent l'existence d'une corrélation concrète entre la perte de biodiversité et l'apparition de pandémies, puis d'expliquer dans quelle mesure le secteur de l'assurance peut être affecté par un épisode pandémique (en s'appuyant sur l'expérience du COVID-19 et les enseignements empiriques qui en ont été tirés).

Il est à noter qu'au moment de la rédaction de ce mémoire, les informations sur l'origine de l'épidémie de COVID-19 sont encore insuffisantes pour que la communauté scientifique puisse la qualifier de zoonose avec certitude. L'OMS poursuit actuellement l'enquête.

4. De la perte de biodiversité aux pandémies

Dans son dernier rapport d'atelier, l'IPBES a averti que nous entrons dans une ère de pandémies plus fréquentes et plus virulentes, corrélées à la perte de biodiversité.

Au cours des 50 dernières années, il a été prouvé qu'il y avait une augmentation des épidémies dans le monde entier (Fondation pour la Recherche sur la Biodiversité, 2020). Selon des estimations scientifiques, il pourrait y avoir 1,7 million de virus non découverts, dont plus de 45 % seraient potentiellement dangereux pour l'homme (Fondation pour la recherche sur la Biodiversité, 2020).

Environ 70 % des épidémies qui sont apparues au cours du dernier demi-siècle avaient des origines zoonotiques (IPBES, 2020c).

D'après l'OMS, une zoonose est une maladie infectieuse qui est passée d'un animal non humain à l'homme. Les agents pathogènes zoonotiques peuvent être bactériens, viraux ou parasitaires, ou peuvent impliquer des agents non conventionnels et peuvent se propager à l'homme par contact direct ou par la nourriture, l'eau ou l'environnement.

Les exemples les plus connus de ces maladies sont les virus Ebola, Zika, Nipah, le VIH/SIDA, les gripes aviaire et porcine ainsi que les SRAS et le MERS également causés par des coronavirus. Le COVID-19 pourrait éventuellement s'ajouter à cette liste. Les zoonoses nous obligent à comprendre la santé non seulement d'un point de vue anthropique mais aussi de celle d'un écosystème d'êtres vivants. C'est la base de l'approche One Health.

One Health est un programme de collaboration multisectorielle et transdisciplinaire travaillant aux niveaux local, régional, national et mondial pour obtenir des résultats optimaux en matière de santé et de bien-être en reconnaissant les interconnexions entre les personnes, les animaux, les plantes et leur environnement partagé.

On peut alors se demander comment la perte continue de la biodiversité est-elle liée à l'apparition de zoonoses et de pandémies potentielles.

La transmission d'une maladie infectieuse d'origine zoonotique est déterminée par la confluence de trois facteurs principaux (Fondation pour la Recherche sur la Biodiversité, 2020) :

- Le danger : la présence de micro-organismes, d'espèces hôtes et de vecteurs ;

- L'exposition : le comportement ou les activités qui favorisent les interactions entre les communautés locales, le bétail et la faune sauvage ;

- Vulnérabilité : aussi bien pour la résilience du système immunitaire (patrimoine génétique, facteurs physiologiques et microbiote) que pour les facteurs socio-économiques (pauvreté, accès aux services médicaux, alimentation, services médicaux, régimes alimentaires, etc.)

Les interactions écologiques au sein des écosystèmes permettent la régulation des agents pathogènes. En écologie des maladies, ce concept est appelé "l'effet de dilution" (Schmidt et al., 2001).

Par conséquent, un niveau élevé de biodiversité spécifique et génétique est considéré comme une garantie contre l'émergence et la propagation des agents pathogènes (Fondation pour la Recherche sur la Biodiversité, 2020).

Cependant, l'effet de dilution est encore contesté par certains scientifiques. Et le mécanisme d'émergence des zoonoses peut également s'expliquer simplement par la destruction de l'environnement qui augmente les interfaces entre la faune sauvage et les humains dans les endroits pleins de pathogènes.

La dégradation d'un écosystème sous tous ses aspects (par exemple pour la perte d'espèces ou de diminution de la diversité génétique) est un moteur de l'émergence de maladies zoonotiques par le biais de plusieurs dynamiques (Fondation pour la Recherche sur la Biodiversité, 2020) :

- Facteurs écologiques : destruction et fragmentation des habitats naturels, perturbation de la chaîne alimentaire, pollution, stress, etc.
- Facteurs épidémiologiques : comment la fragmentation et la réorganisation des populations influencent la dilution ou la propagation des agents pathogènes ;
- Facteurs adaptatifs : comment les changements de comportement augmentent les interactions entre la faune sauvage et les populations humaines. Lorsque les habitats sont détruits ou que les ressources alimentaires se raréfient, la faune migre vers les villes. Par exemple, la déforestation et les incendies de forêt provoqués par l'homme conduisent les chauves-souris à migrer vers les structures humaines, en particulier les élevages de porcs, ce qui a entraîné l'apparition du virus Nipah en 1998-1999. Le même processus est impliqué dans le virus de la dengue voyageant via les structures humaines.
- Facteurs évolutifs : pour l'homme et la plupart des autres êtres vivants, les agents pathogènes sont un moteur de sélection et d'évolution de notre système immunitaire. La réponse et le façonnement de notre système dans l'interaction avec les agents pathogènes sont appelés processus de coévolution. Il existe des preuves d'une corrélation entre la virulence d'une épidémie d'agents pathogènes et la distance phylogénétique entre les humains et l'hôte pathogène. Par conséquent, le passage d'un hôte à un autre peut perturber les relations de coévolution entre l'hôte et l'agent pathogène, ce qui entraîne une grave maladie.

Ces quatre dynamiques sont induites par les activités humaines (Fondation pour la Recherche sur la Biodiversité, 2020) :

La déforestation réduit les habitats naturels, accroît la présence humaine dans les forêts et les interfaces entre les humains et la faune sauvage, augmentant ainsi les interactions potentielles. La déforestation perturbe également les chaînes alimentaires, notamment en ayant un impact sur les prédateurs, ce qui permet à d'autres espèces, c'est-à-dire à des hôtes potentiels, de proliférer. La corrélation entre la déforestation et l'apparition de zoonoses a été prouvée.

La chasse, la consommation et le commerce de la viande de brousse augmentent l'interaction avec la faune sauvage et a été identifiée dans plusieurs cas comme étant à l'origine de l'émergence de pathogènes tels que le VIH, Ebola et le SRAS. Le commerce de la viande de brousse sur les marchés urbains et l'expansion vers les marchés nationaux et internationaux augmentent le risque de transmission.

L'urbanisation favorise également l'émergence et la propagation de pathogènes compte tenu des fortes densités de population et est favorable au développement des populations de rongeurs et d'insectes dans les villes. L'étalement urbain réduit également l'interface entre les populations humaines et la faune sauvage, et par conséquent le risque d'interactions.

Il a également été prouvé que l'élevage intensif est une source de pathogènes, comme la grippe aviaire et le virus Nipah. La proximité entre le bétail et la faune sauvage, la forte densité d'animaux en contact avec les populations humaines, la perte de diversité génétique et le stress généré par les conditions de vie sont des facteurs favorisant l'émergence et la transmission d'agents pathogènes.

Le changement climatique, en modifiant l'environnement naturel, influence le comportement, l'activité et la distribution des espèces, et donc l'émergence et la propagation des agents pathogènes. Le changement climatique entraîne également le dégel du permafrost, une couche de glace vieille de plusieurs milliers ou millions d'années et contenant des bactéries et des virus, c'est-à-dire des agents pathogènes potentiels pour la faune sauvage, le bétail et les humains.

Pour mieux visualiser à quels pathogènes s'exposent les individus et les compagnies d'assurance en termes de potentielle pandémie, nous avons ci-dessous deux tableaux produits par l'Institut de Veille Sanitaire (INSV). L'INSV a pour mission comme son nom l'indique de faire une veille épidémiologique sur les potentielles maladies pouvant affecter les individus sur le territoire français. Il y a donc une grande partie de leur travail visant à cibler les zoonoses en les classifiant selon qu'elles sont prioritaires ou à risque de propagation en métropole très faible. Nous avons donc :

Classement des maladies par groupe de priorité – Rapport priorisation des zoonoses 2009

Prioritaires	Importantes	Moyennement importantes
Borréliose de Lyme	Encéphalite à tiques	Anaplasmose humaine
Chikungunya	Fièvre jaune	Babésiose
Dengue	Fièvre Q	Bartonelloses
Fièvre Crimée-Congo	Fièvre de la Vallée du Rift	Ébola
	Psittacose	Fièvre boutonneuse méditerranéenne
	Tularémie	Fièvre hémorragique de Marburg
		Fièvre de Lassa
		Hépatite E
		Rickettsioses (autres)
		Typhus exanthématique

Nous observons donc que, par exemple, les agents pathogènes porteurs de Dengue et de Chikungunya sont classés comme prioritaires tandis que celui de l'Anaplasmose humaine est moyennement importante.

Nous avons plus de détails de ces choix de classification dans le tableau qui suit :

Maladies	Homme	Priorités				Animal-vecteur	Priorités	
	Arguments	De surveillance	De développement biologique	De contrôle et prévention	De recherche et connaissances	Arguments	Animales	Vectérielles
Anaplasmose humaine (A. phagocytophilum)	Rare, sévère et évolution possible liée au vecteur/climat	1	2	1	1	Émergence, impact élevage, peu de connaissances	2	2
Babésiose (B. divergens)	Rare, très sévère et évolution possible liée au vecteur/climat + augmentation liée à drépanocytaires	1	1	1	1	Impact possible sur élevage bovin	2	2
Bartonelloses <i>B. henselae</i>	Fréquence moyenne, pfs sévère, évolution possible liée au vecteur/climat	BH 1	BH 1	BH 1	BH 1	Possibilité de transmission vectorielle en Europe	1	1
<i>B. quintana</i>		BQ 1	BQ 1	BQ 2	BQ 1-2	Infestation et réinfestation de poux chez des SDF	50	2
Borreliose de Lyme	Fréquence très importante, sévérité pour 1/3 des cas, évolution possible liée au vecteur/climat	4	4	4	4	Rôle de la faune sauvage dans le cycle des tiques et de la transmission de B mal connu et indispensable pour évaluation des risques % homme	3	3
Chikungunya	Risques*** émergence, parfois sévère	4	3	4	4	Présence du vecteur en Métropole et possibilité d'implantation du virus	1	4
Dengue	Risques*** émergence, parfois très sévère	4	3	4	4	Présence du vecteur en Métropole et possibilité d'implantation du virus	1	4
Ébola	Rare, potentiel émergence en Métropole faible	4	1	4	1	Pas de virus circulant en France	1	50
Fièvre boutonneuse méditerranéenne	Rare mais augmentation ? régional et augmentation de la sévérité. Action ponctuelle? (surveillance faisable : DO ? ou outil de surveillance pour action)	3	1	3	2	Possibilité d'extension géographique du vecteur (importation, implantation déjà décrite) Mesure de contrôle par prévention tiques chez chien	1	2
Fièvre de la Vallée du Rift	Risques d'émergence possible lié à épidémie animale/vecteur/climat	4	1	2	3	Manque connaissance vecteur potentiel en France	3	3

Priorités : Prioritaire 4 ; Importante 3 ; Importance moyenne 2 ; Peu importante 1.

A la lecture de ce tableau, nous voyons que la veille et le suivi régulier de l'INSV ont mis en évidence le fait que le Chikungunya et la Dengue représentent bien une priorité car ils ont un risque d'émergence non négligeable, qu'il y a déjà une présence des virus en Métropole et que ces derniers pourraient s'implanter de manière durable tandis que l'Anaplasmose humaine est jugée moins importante car elle est rare bien que sévère et de plus il y a assez peu de connaissances sur elle actuellement.

5. Prise en compte du risque pandémie au sein de solvabilité II

Dans cette section, nous allons voir comment le risque de pandémie est traité par la directive Solvabilité II. Nous allons d'abord rappeler ce que sont Solvabilité II et ses trois piliers puis nous ciblerons les compartiments qui traitent de la pandémie au sein de la formule standard.

5.1 Solvabilité II

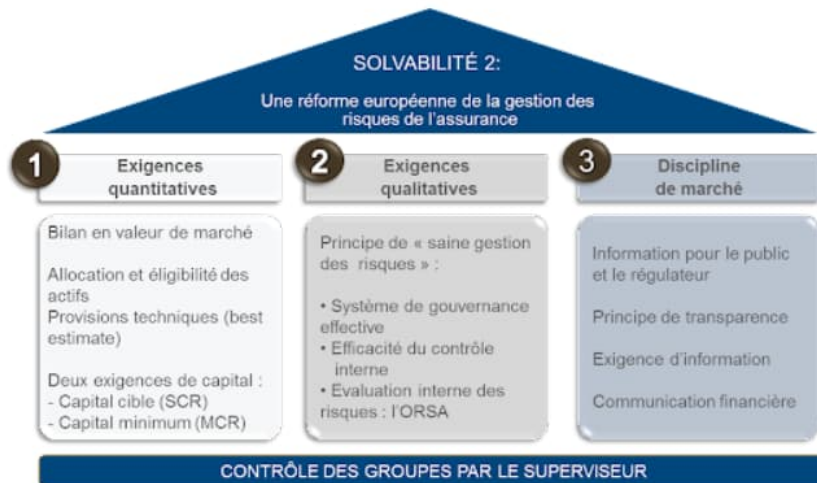
La directive Solvabilité II, adoptée par le conseil de l'Union Européenne et le Parlement Européen en novembre 2009 et détaillée dans le règlement délégué (UE) 2015/35 de la commission du 10 octobre 2014 est entrée en vigueur le 1^{er} janvier 2016. La directive n'établit pas seulement un ensemble de règles d'adéquation des fonds propres pour les compagnies d'assurance et de réassurance mais aussi un ensemble d'exigences pour la gouvernance et la divulgation au public. Les principaux objectifs de Solvabilité II sont d'apporter une meilleure protection aux assurés contre la défaillance des compagnies d'assurance et de réassurance et d'assurer une plus grande cohérence des exigences prudentielles à travers l'espace économique européen.

L'adéquation des fonds propres fondée sur les risques sous solvabilité II est plus solide et plus complète que les règles de solvabilité I. Avec les nouvelles règles, le niveau minimum de fonds propres qu'une compagnie d'assurance et de réassurance doit posséder est déterminé sur la base de son profil de risque et la manière à laquelle chacun de ses risques est piloté.

La directive Solvabilité II reconnaît et souligne le fait que la réduction du risque de défaillance d'un assureur demande plus que détenir un niveau minimum de capital. Les crises financières ne sont généralement pas seulement le résultat d'un niveau de fonds propres inadéquat. Ils découlent également de stratégies et d'activités inefficaces ou mal pilotées dans l'entreprise, par exemple, en ce qui concerne la gestion des risques, l'investissement, la tarification, les réserves ou la croissance de l'entreprise. La détention d'un niveau de fonds propres est juste une ultime protection contre les pertes dû à un mauvais pilotage de l'activité. Ainsi, en addition d'un capital requis, Solvabilité II donne des recommandations pour un système de gouvernance efficace. Le règlement exige aux compagnies d'assurance et de réassurance de consacrer des ressources significatives pour l'identification, la mesure et le pilotage actif des risques auxquels elles sont exposées. Cette mise en évidence des exigences sur une bonne gouvernance et un bon pilotage du risque, en accord avec l'exigence de capital minimum vise à réduire la probabilité de défaillance de l'entreprise renforçant ainsi la stabilité du marché européen des assurances.

5.2 Les 3 piliers de Solvabilité II

La directive Solvabilité II a un système de gestion autour de 3 piliers illustrés comme suit :



Le premier pilier s'adresse aux exigences quantitatives des compagnies d'assurance et de réassurance. Il établit les règles pour déterminer le montant minimum de capital à détenir et pour évaluer les fonds propres éligibles d'une compagnie d'assurance afin de couvrir l'exigence de capital minimum. Cela permet une escalade d'échelle d'intervention. Le premier niveau d'intervention est le capital de solvabilité requis (SCR). Si une compagnie d'assurance n'est pas en mesure de démontrer qu'elle détient suffisamment de fonds propres éligibles pour couvrir son SCR, l'autorité de supervision (l'ACPR en France) demandera à l'entreprise de prendre les mesures nécessaires pour se conformer au SCR dans un délai de six mois. La seconde exigence de capital est qu'une entreprise d'assurance détienne suffisamment de fonds propres de base éligibles pour couvrir son exigence de capital minimum (MCR), qui est plus basse que le SCR. Une défaillance pour cette seconde exigence résulte d'une sérieuse intervention de l'autorité de supervision incluant un possible retrait de l'autorisation de poursuivre des activités d'assurance.

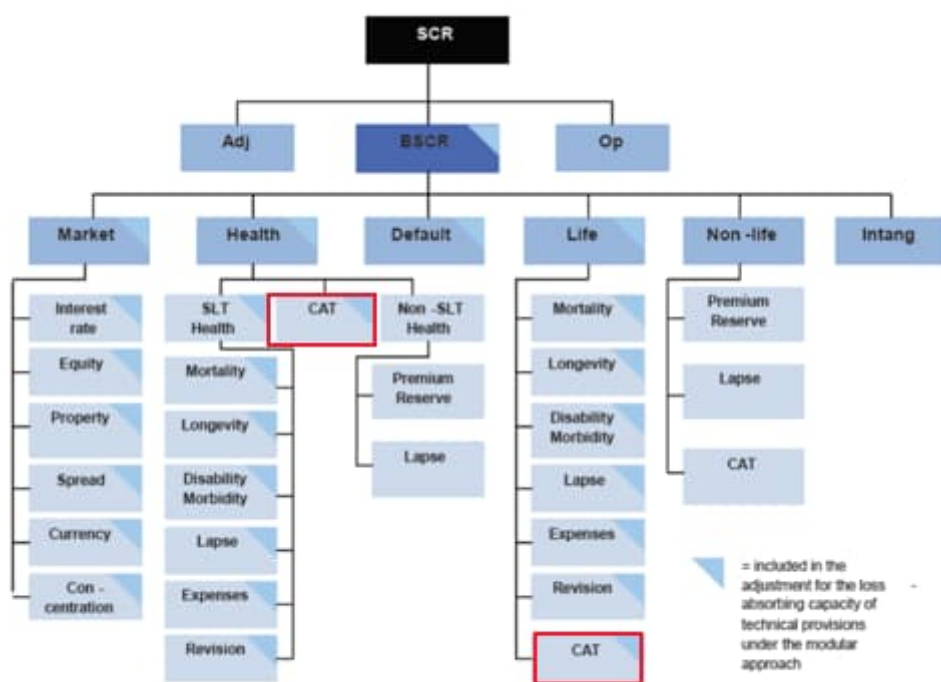
Le second pilier s'adresse aux exigences qualitatives pour l'efficacité du pilotage de l'activité d'assurance et le contrôle de celle-ci. D'une part, on a la gouvernance interne qui doit être mise en place par chaque assureur. Ils doivent aussi être garant des quatre fonctions clés (gouvernance, audit interne, conformité et fonction actuarielle). La directive impose en outre aux entreprises de définir des règles pour l'externalisation de toute activité ou fonction et de garantir que leur management est efficace. D'autre part, les régulateurs sont tenus d'examiner et d'évaluer la conformité des entreprises aux exigences quantitatives et qualitatives de Solvabilité II. La directive décrit le pouvoir et les devoirs des régulateurs à l'égard de ces fonctions.

Le troisième pilier s'adresse à la divulgation de l'information aux régulateurs et au public. La transparence en matière de solvabilité est assurée par la remise régulière par les entreprises d'assurance de deux séries de rapports aux autorités de contrôle. Le rapport sur la solvabilité et la situation financière (SFCR) divulgue au public et aux autorités de contrôle l'information requise pour analyser la solvabilité et la situation financière de la compagnie d'assurance. Les standards techniques réglementaires (RTS) sont uniquement à destination des contrôleurs, les assistent dans

l'exécution du processus de supervision prudentielle (PSR) et constituent la base du dialogue entre l'entreprise et le contrôleur.

5.2.1 La formule standard du SCR

La directive Solvabilité II propose une méthode de calcul de SCR qui s'appuie sur l'agrégation de certains risques auxquels les entreprises du secteur assurantiel sont exposées. C'est la formule standard. Elle se compose des risques de marché, santé, contrepartie, vie, non-vie, intangible et opérationnel dont certains sont eux-mêmes composés de sous-modules comme-ci-dessous :



Les modules qui vont nous intéresser dans le cadre d'une pandémie sont le module Santé CAT et le module Vie CAT encadrés en rouge dans la figure ci-dessus.

5.2.2 Le risque de pandémie dans la formule standard

Dans la formule standard, il n'y a pas de module ayant directement pour nomination le risque pandémie. Ce risque est en réalité pris en compte dans le module de Santé CAT et dans le module Vie CAT.

Le SCR Vie CAT se traduit par une hausse de 0,15 point des taux mortalité servant au calcul des provisions techniques et reflétant la mortalité estimée des assurés sur les 12 prochains mois. De manière simplifiée, on a :

$$SCR \text{ Vie CAT} = 0,15 \% \cdot \text{Capital sous risque}$$

Où le *Capital sous risque* est la somme des capitaux dépendant de la mortalité.

Le choc 0,15 % est issu de la publication de Swiss Re (2007) dans laquelle Swiss Re utilise un modèle épidémiologique pour arriver à une surmortalité dû à une pandémie allant de 0,1 à 0,15% dans les pays développés avec une population générale répartie par âge comme celle d'un portefeuille d'assurance. Nous tenterons de modéliser ce choc en prenant en compte l'historique des épidémies disponible.

Le SCR Santé CAT est composé de trois modules qui sont le module accident de masse, le module accident de concentration et le module pandémie. Nous nous intéressons uniquement au module pandémie dont la formule est :

$$SCR\ Santé\ CAT\ Pandémie = 0,000075 \cdot E + 0,4 \cdot F$$

Où E est l'exposition au risque de protection des revenus en cas de pandémie et F l'exposition aux frais d'hospitalisation en cas de pandémie. Etant donné que la plupart des contrats de PREDICA ont des franchises de 90 jours, E est considéré comme nul et dans notre mémoire, nous travaillerons notamment sur l'exposition hospitalière et chercherons à remettre en question le choc de 40% qui semble aléatoire.

6. Profil CAA

Dans cette section, nous allons décrire le profil de risque du Crédit Agricole Assurances.

6.1 Les risques

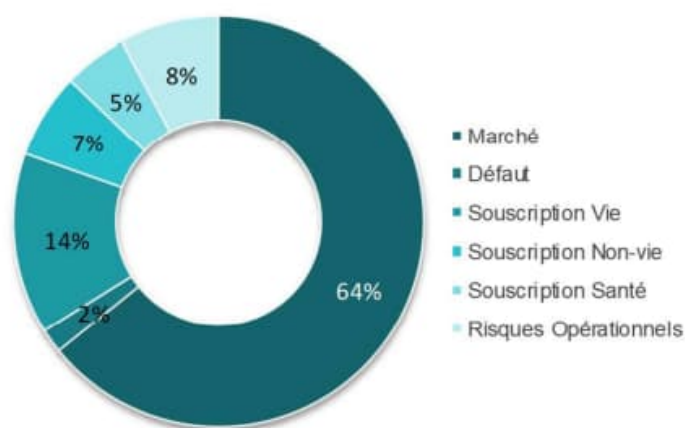
Ce profil s'établit à travers un processus permettant d'identifier les risques clés auxquels la compagnie est exposée et d'ensuite les évaluer. Ce profil sert de base pour l'évaluation du besoin en capital du groupe.

Au vu du poids des activités d'épargne et de retraite, le Crédit Agricole Assurances est principalement exposé aux risques de marché et de souscription vie puis dans une moindre mesure aux autres risques qui sont les risques de souscription non-vie, la santé, la contrepartie, les risques intangibles et les risques opérationnels. Tous ces risques rentrent dans le cadre de la formule standard pour déterminer le capital de solvabilité requis.

En plus de l'approche quantitative, une approche qualitative est effectuée pour identifier les risques clés. Cela a pour objectif d'avoir la main sur certains risques qui ne sont pas explicités au sein de la formule standard (risque de liquidité par exemple) afin que la gouvernance soit alertée en cas de déviation significative du cadre fixé.

A fin 2020, le SCR et le BSCR du groupe Crédit Agricole Assurances sont respectivement de 14 Milliards d'euros et de 28 milliards d'euros. L'exposition au risque qui définit le profil de risque de la compagnie dans le BSCR après absorption des pertes par les provisions techniques et intégrant le risque opérationnel est composée majoritairement des exigences de capital au risque de marché (64%) et suivi des risques de souscriptions (26%) comme on peut le voir ci-dessous :

31/12/2020



Contribution des modules de la formule du SCR

Notons que le risque de souscription santé ne représente que 5% cette exposition. Ainsi le sous-module CAT du module Santé a un poids encore moins important. Cela fait que la composante pandémie de ce sous-module est relativement négligeable du point de vue du groupe Crédit Agricole Assurances.

Les activités d'épargne et de retraite expose le groupe à l'insuffisance des chargements prélevés pour couvrir les charges d'exploitation et les commissions, au risque de mortalité, au risque de longévité et bien entendu aussi de rachats. Il faut aussi prendre en compte les risques de gestion actif/passif adossés à une collecte significative sur les fonds euros qui pourrait mettre en difficulté l'assureur face à ses engagements dans un contexte de taux bas.

Comme ça a été dit dans la section précédente, nous nous intéressons dans ce mémoire à la pandémie et donc naturellement ce sont les risques de santé et de souscription que nous allons regarder de plus près.

En prévoyance et assurance des emprunteurs, l'exposition du groupe Crédit Agricole est le fait des risques biométriques et des risques liés à la santé. En théorie, le risque catastrophe dû à un choc de mortalité pourrait impacter les résultats de la prévoyance individuelles et collectives. Cependant, le covid19 qui a touché la France en 2020 a eu impact très limité sur les risques biométriques.

Le risque de souscription santé est principalement le fait de l'assurance des emprunteurs pour les risques assimilables à la vie et par les contrats santé et Garanties des Accidents de la Vie pour les risques non assimilables à la vie. Ces risques sont secondaires pour le groupe. Nous tenterons néanmoins d'évaluer l'impact des hospitalisations comprises dans les contrats santé dans un contexte de pandémie.

6.2 Méthodes utilisées pour la réduction du risque

En épargne, le principal risque identifié par le groupe est le risque de rachat. Le groupe maîtrise ce risque via une politique de participation aux bénéfices très intéressante sur le marché et via une politique financière prudente notamment dans la gestion des réserves.

Le suivi des taux de rachat s'effectue à plusieurs niveaux :

- Un suivi mensuel pour cibler les dérives conjecturelles
- Un suivi annuel pour s'assurer que les taux servis concurrencent bien le marché

En assurance dommages, assurances des emprunteurs et en prévoyance, les risques d'anti-sélection et d'insuffisance de tarification sont gérés par :

- Une politique tarifaire appropriée
- La politique de souscription principalement le fruit des réseaux bancaires et des partenaires financiers
- La politique de gestion de sinistres principalement le fruit des unités de gestion dédiées ou des prestataires locaux.

Les risques de catastrophes et de dérive de sinistralité sont maîtrisés par la réassurance qui sert à protéger les fonds propres et réduire la volatilité des résultats.

Le suivi des risques extrêmes et des risques de provisionnement est géré par le suivi d'un budget des risques climatiques et des sinistres graves.

PARTIE II : Modélisation épidémiologique pour le SCR

Santé CAT Pandémie

Dans cette partie, nous verrons ce qu'est un modèle épidémiologique aussi bien déterministe que stochastique. Ensuite nous verrons les différentes étapes nous permettant d'utiliser le modèle stochastique pour obtenir des résultats.

1. Modèle déterministe

Il existe trois types de modèles déterministes pour modéliser la propagation d'une épidémie par les différents contacts entre les personnes au sein d'une population. Comme nous allons le voir, ces modèles se formalisent mathématiquement par des systèmes d'équations différentielles traduisant l'évolution de la dynamique de l'épidémie. Bien que ces modèles soient simples, ils prennent en compte des concepts clés pour l'analyse d'une épidémie tels que le niveau de transmission, la létalité ou encore dans certains cas la période d'incubation. Il est donc important de comprendre le comportement de ces modèles.

1.1 Intérêt d'un modèle épidémiologique

Le mécanisme de transmission d'un individu infecté à un individu qui n'a pas encore été malade est généralement connu pour les maladies infectieuses cependant les interactions de transmission sont très complexes ce qui fait qu'il est très difficile d'appréhender la dynamique des maladies à une large échelle sans la structure d'un modèle mathématique. Le principe d'un modèle épidémiologique est qu'il va utiliser une petite information, comme par exemple, le niveau d'infection individuelle pour prédire un comportement de la maladie plus grand au sein de la population.

Les expérimentations d'une propagation de maladie dans une population sont éthiquement impossibles à réaliser c'est pourquoi les données disponibles sont toujours issues de pandémie ou épidémies historiques mais ces dernières sont bien souvent incomplètes dû à une sous déclaration du nombre réel de cas contaminés. Ce manque d'information fait que la précision des paramètres du modèle lié à une maladie n'est pas toujours aisée donc un intervalle de valeurs est estimé pour certains paramètres. Puisque les expérimentations sont difficiles et les données souvent peu précises, les modèles mathématiques couplés à la puissance de calcul des ordinateurs sont utilisés pour faire des expérimentations théoriques.

Les comparaisons peuvent mener à une meilleure compréhension de la propagation de la maladie. La modélisation peut être utilisée pour comparer différentes maladies au sein de la même population, au sein de populations différentes ou la même maladie à des époques différentes.

Les modèles sont aussi utiles pour évaluer l'impact de la prévention et des contrôles. Hethcote et Yoke (1984) ont utilisé des modèles pour comparer les différentes mesures de contrôle pour la

gonorrhée telles que le dépistage, le traçage des infectieux, la vaccination post-traitement et la vaccination générale.

Un élément fondamental des modèles épidémiologiques est qu'ils permettent des hypothèses sur les mécanismes biologiques et sociaux qui influencent la propagation de l'épidémie. C'est ainsi que les paramètres utilisés dans le modèle doivent avoir une claire interprétation comme le taux de contact ou la durée de l'infection. Un modèle peut aussi servir à rendre compte de la sensibilité des prédictions dû au changement de la valeur des paramètres. Nous verrons que le taux de transmission et les durées d'incubation et d'infection sont cruciales.

1.2 Hypothèses et Notations

La population considérée dans un modèle se retrouve distribuée dans plusieurs états disjoints qui changent au temps t . L'état des susceptibles ou des personnes saines noté S_t est composé des individus qui ne sont pas encore tombés malades. L'état des exposés noté E_t correspond aux personnes saines qui sont rentrées en contact avec une personne infectée et commencent leur phase d'incubation. L'état infecté noté I_t correspond aux personnes dont la période d'incubation est finie et sont désormais à même de contaminer les personnes saines. Enfin nous avons l'état après l'infection qui est soit le rétablissement soit le décès. Dans la littérature épidémiologique, généralement seul l'état rétabli noté R_t est considéré comme seule issue possible mais nous faisons la distinction entre les rétablis et les décédés dans le mémoire. Ainsi les rétablis sont notés R_t et les décès D_t .

Dans un modèle épidémiologique, plusieurs hypothèses simplificatrices sont faites :

- La population est supposée avoir une taille constante N et suffisamment grande au cours de l'épidémie autrement dit, les naissances ne sont pas comptées. Si le modèle est dynamique avec des naissances et des décès à chaque classe, il faudrait considérer les nouveaux nés comme des susceptibles et évaluer des décès à chaque état de la maladie et pas seulement après la période d'infection. D'une part, les nouveaux nés sont par nature limités par les déplacements et ont très peu d'influence sur la contamination, et d'autre part nous verrons par la suite que la durée totale de la maladie (en prenant en compte temps d'incubation et temps d'infection) pour un individu est généralement très courte à tel point que son décès est lié à la maladie et non une autre cause. Au vu de ces éléments, nous conserverons l'hypothèse d'une taille de population constante.
- La population est mélangée de manière homogène. Le taux de contact noté β est le nombre de contact moyen journalier. Un contact avec un infecté résulte en une contamination de l'autre individu si ce dernier est dans l'état susceptible. Ainsi le nombre moyen de personnes susceptibles qui se retrouvent contaminées par un infecté chaque jour est de βS_t et le nombre moyen dû à tous les infectés est de $\beta S_t I_t$. β est fixe et ne dépend pas de la saisonnalité.
- Les individus passent d'état d'exposition (incubation) à état d'infection chaque jour proportionnellement au nombre de personnes exposés avec une constante ε . La durée

d'exposition est donc $\frac{1}{\varepsilon}$. Dans le cas où la maladie n'a pas de période d'exposition, les individus sains sont directement infectés après contamination sans étape d'incubation.

- Les individus se rétablissent ou décèdent chaque jour proportionnellement au nombre de personnes infectés avec une constante γ pour les rétablissements et une constante μ pour les décès. La durée de la période d'infection est alors $\frac{1}{\mu+\gamma}$.

Dans le cas où il n'y a pas de période d'incubation et que le rétablissement ne donne pas lieu à une immunité, on parle de modèle SIS pour Susceptible-Infecté-Susceptible puisque les individus passent de l'état susceptible à l'état infectés puis bien que guéris sont à nouveau susceptibles. Si le rétablissement donne lieu à une immunité permanente, on parle de modèle SIR pour Susceptible-Infecté-Rétabli. Enfin, s'il y a une période d'incubation, on parle de modèle SEIR mais comme on fait la distinction entre les décès et les rétablissements, nous allons l'appeler SEIRD pour Susceptible-Exposé-Infecté-Rétabli-Décédé.

1.3 Le modèle SIR

Il s'agit d'un modèle épidémiologique très classique. Dans ce modèle, les populations susceptibles et infectées sont supposées en contact, chaque individu infecté peut contaminer les individus sains avec un taux de contact β . Les individus infectés se rétablissent suivant un taux γ souvent appelé taux de rétablissement et donc il fait décroître le nombre d'infectés journaliers de γI_t . Dans ce modèle, on considère que les individus ne décèdent pas et que donc seul le rétablissement est l'état possible après l'état infectieux. Ce modèle capture ainsi la dynamique d'infections graves qui confèrent une immunité pour le reste de la vie. Il est naturellement applicable aux maladies non létales comme la variole. Schématiquement, le modèle SIR peut être illustré comme suit :



Modèle SIR

A chaque pas de temps, on a :

$$S_{t+1} = S_t - \beta S_t I_t$$

$$I_{t+1} = I_t + \beta S_t I_t - \gamma I_t$$

$$R_{t+1} = R_t + \gamma I_t$$

Le système d'équations différentielles traduisant la dynamique de la maladie qui en découle est le suivant :

$$\begin{cases} \frac{dS_t}{dt} = -\beta S_t I_t \\ \frac{dI_t}{dt} = \beta S_t I_t - \gamma I_t \\ \frac{dR_t}{dt} = \gamma I_t \end{cases}$$

Comme nous avons émis l'hypothèse que la population est constante alors nous avons :

$$N = S_t + I_t + R_t, \quad \forall t$$

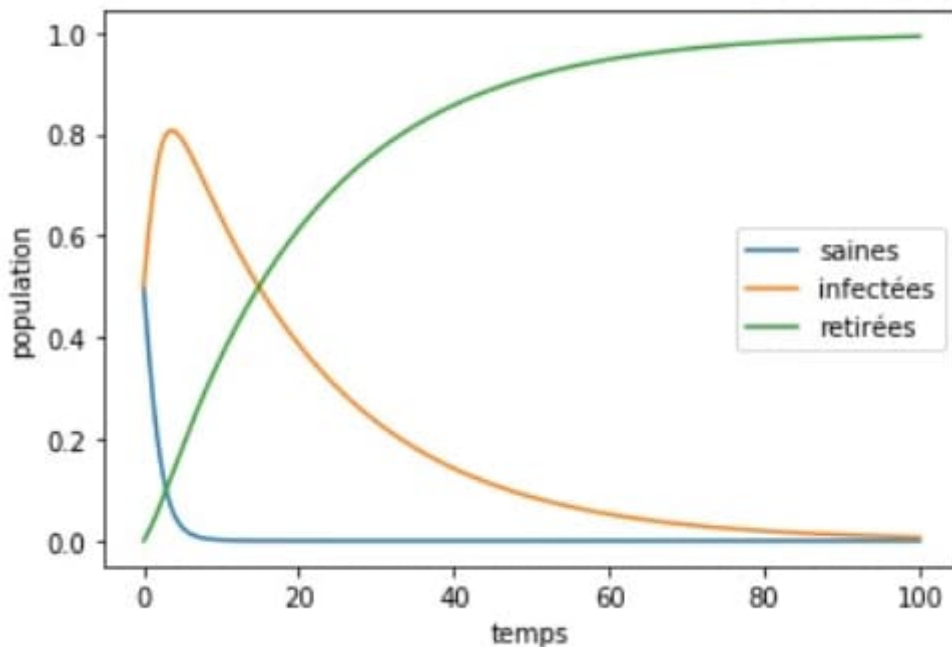
De plus, on vérifie bien que $\frac{dN}{dt} = 0$ car on a :

$$\begin{aligned} \frac{dN}{dt} &= \frac{dS_t}{dt} + \frac{dI_t}{dt} + \frac{dR_t}{dt} \\ &= -\beta S_t I_t + \beta S_t I_t - \gamma I_t + \gamma I_t = 0 \end{aligned}$$

La résolution du système d'équations différentielles donne l'expression explicite des différents états.

Au vu de ces éléments, le nombre d'infectés évolue à travers le temps comme une combinaison du taux de contact β et du taux de rétablissement γ . En effet, le nombre d'infectés décroît avec le taux de rétablissement car les personnes infectées ne peuvent plus infecter d'autres personnes quand elles sont rétablies mais il croît avec le taux de contamination.

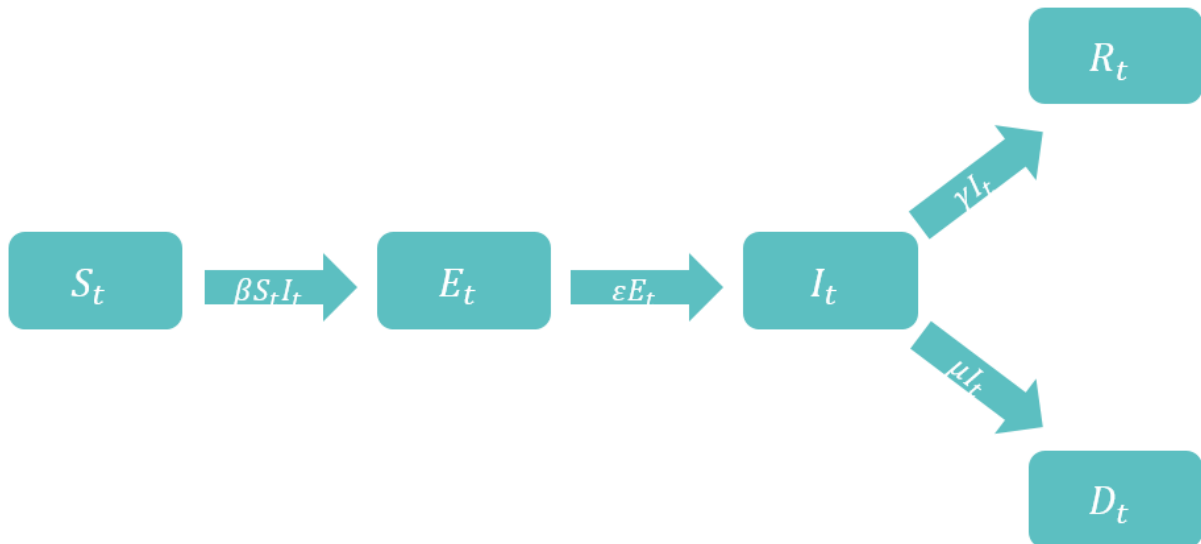
Ci-dessous, nous avons une représentation graphique de l'évolution d'un modèle SIR :



Solution du modèle SIR

1.4 Le modèle SEIRD

Le modèle SEIRD est lui aussi un modèle qui repose sur les passages d'un état à un autre. Ici, on retrouve les trois états déjà présents dans le SIR qui sont l'état sain, l'état infecté et l'état rétabli. A ces derniers se rajoutent l'état exposé et le décès. Il s'agit donc d'un modèle adapté pour des maladies létales. Schématiquement, il est illustré comme suit :



Modèle SEIRD

A chaque pas de temps, on a :

$$S_{t+1} = S_t - \beta S_t I_t$$

$$E_{t+1} = \beta S_t I_t + E_t - \varepsilon E_t$$

$$I_{t+1} = \varepsilon E_t + I_t - (\gamma + \mu) I_t$$

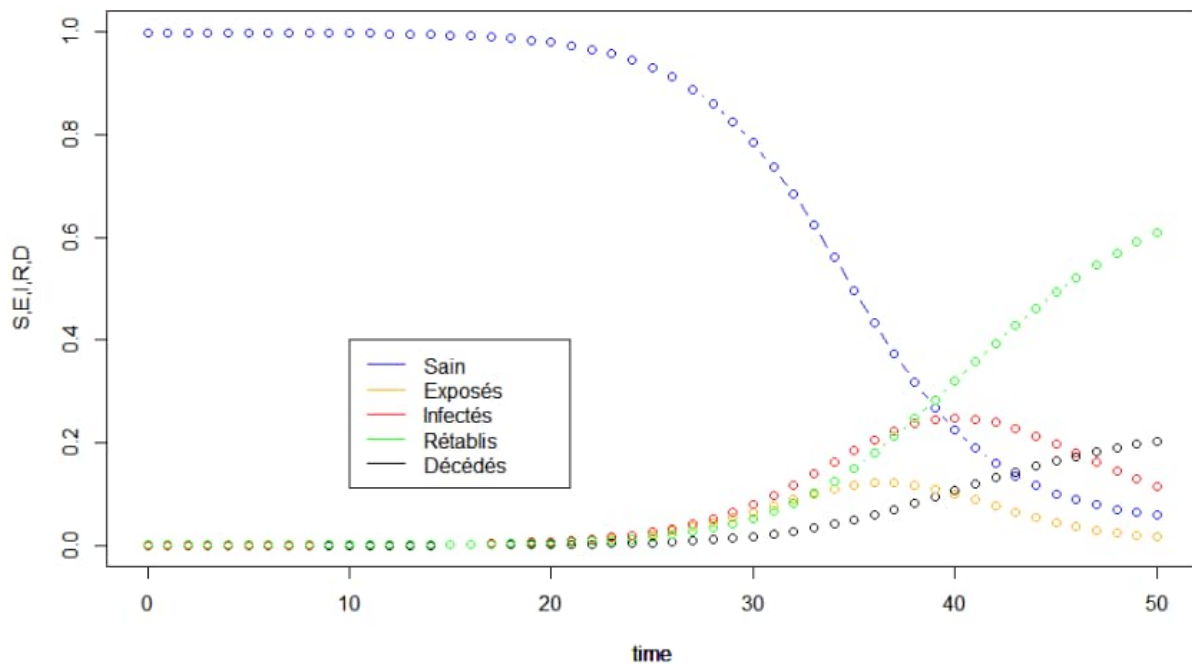
$$R_{t+1} = \gamma I_t + R_t$$

$$D_{t+1} = \mu I_t + D_t$$

Le système d'équations différentielles associé est le suivant :

$$\left\{ \begin{array}{l} \frac{dS_t}{dt} = -\beta S_t I_t \\ \frac{dE_t}{dt} = \beta S_t I_t - \varepsilon E_t \\ \frac{dI_t}{dt} = \varepsilon E_t - \gamma I_t - \mu I_t \\ \frac{dR_t}{dt} = \gamma I_t \\ \frac{dD_t}{dt} = \mu I_t \end{array} \right.$$

En additionnant les équations différentielles, on vérifie assez facilement que la population est bien constante. Graphiquement, l'évolution de la dynamique d'une épidémie de ce modèle est illustrée ci-dessous :



Solution du modèle SEIRD

On remarque que les sains diminuent à mesure que les populations des autres états augmentent ce qui est normal car c'est cet état qui alimente les autres. Les rétablis et les décédés représentent un stock de cumul.

2. Les paramètres des modèles

Dans cette section, nous allons traiter plus en détail des différents paramètres qui permettent à une épidémie d'évoluer au sein d'une population et accessoirement d'expliquer les comportements observés dans les modèles.

2.1 Taux de reproduction et contact

Il existe 3 indicateurs importants pour le suivi d'une épidémie : le taux de reproduction de base noté \mathcal{R}_0 , le nombre de contact σ et le taux de reproduction \mathcal{R} .

Le plus commun et aussi le plus connu est le taux de reproduction de base \mathcal{R}_0 qui se définit comme le nombre moyen de secondes infections qui ont lieu lorsqu'un individu infecté est introduit dans une population composée entièrement d'individus susceptibles. Kermack et McKendrick sont les premiers à étudier cet indicateur, ils sont arrivés au postulat suivant :

Si

$$\mathcal{R}_0 < 1$$

Alors la dynamique de la maladie avorte très rapidement au sein de la population. Le temps d'avortement de l'épidémie va beaucoup dépendre de la valeur de \mathcal{R}_0 : plus cette valeur est proche de 0 et plus l'épidémie peut disparaître en quelques jours et plus elle est proche de 1 plus cela peut prendre des semaines ou des mois.

Si

$$\mathcal{R}_0 > 1$$

Dans ce cas, la population s'expose à une propagation qui peut prendre une ampleur significative selon le niveau auquel \mathcal{R}_0 est au-dessus de 1 : si cette valeur reste proche de 1, la maladie pénètre seulement dans une partie de la population mais si cette valeur est plus grande que 2, une grande partie de la population peut être contaminée.

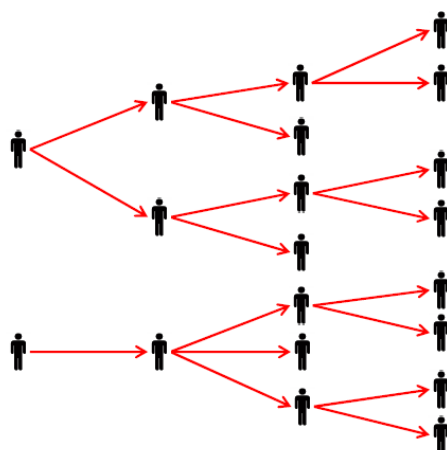
Dans le graphique en 2.2, les taux de reproduction de plusieurs virus sont mis en évidence et on peut y observer de grandes différences existantes.

Le nombre de contact σ est le nombre moyen de contact adéquat d'un infectieux typique au cours de sa période infectieuse. Un contact adéquat est celui qui mène à la transmission si l'individu en contact avec le susceptible est un infectieux. Il est implicitement supposé que le malade est dans la population pendant toute la période infectieuse et se mélange avec les autres individus.

Le taux de reproduction \mathcal{R} à ne pas confondre avec le taux de reproduction de base \mathcal{R}_0 est le nombre moyen de secondes infections causées par un infectieux au cours de sa période d'infection. Il faut noter que contrairement au taux de reproduction de base, le taux de reproduction change au cours du temps et suit l'évolution de la maladie au sein de la population. Plus simplement, si on réussit à reconstruire la chaîne de contamination au cours de la vie de l'épidémie, on a :

$$\mathcal{R} = \frac{\text{nombre de transmission}}{\text{nombre d'infectieux}}$$

L'exemple simple ci-dessous permet de se faire une meilleure idée de cet indicateur :



Exemple d'une chaîne de transmission

Dans cet exemple, une flèche représente une transmission, on compte 18 transmissions et on compte 9 individus qui ont infecté d'autres individus. On a donc un taux de reproduction de 2.

Ces trois quantités \mathcal{R}_0 , σ et \mathcal{R} sont toutes égales au début d'une épidémie quand toute la population est encore saine et qu'un seul individu est infecté. \mathcal{R}_0 est défini uniquement au commencement de l'invasion alors que σ et \mathcal{R} sont définis par la suite à chaque pas de temps. \mathcal{R} est toujours inférieur ou égale à \mathcal{R}_0 . De plus, après le début de la propagation, la proportion de susceptible dans la population totale est inférieure à 1 et tous les contacts ne donnent pas lieu à une infection. Ainsi le taux de reproduction est inférieur ou égal au nombre de contact. Ces résultats nous emmènent à cette inéquation :

$$\mathcal{R} \leq \sigma \leq \mathcal{R}_0$$

2.2 Le taux de létalité

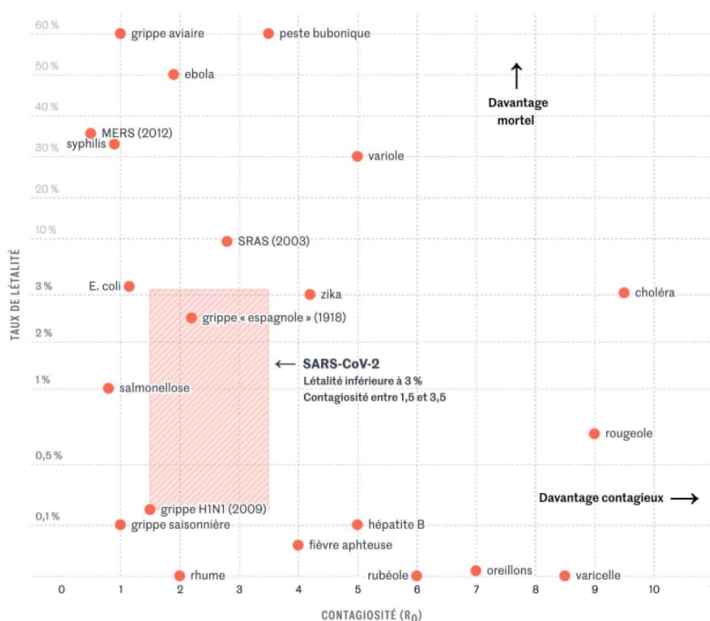
Le taux de létalité μ correspond au nombre de décès rapporté au nombre de personnes infectées. Il n'est pas à confondre avec le taux de mortalité qui est le nombre de décès rapporté à la population totale. On a donc :

$$\mu = \frac{\text{nombre de décès}}{\text{nombre d'infectés}}$$

Cependant, ce calcul présente des biais :

- Généralement les infectés asymptomatiques ne sont pas détectés contrairement aux personnes symptomatiques. Cela mène à une surestimation de la létalité car l'exposition liée au nombre d'infectés est sous-estimée.
- Le moment auquel le taux de létalité est évalué constitue aussi un biais non négligeable car parmi les cas infectieux observés, certains pourront potentiellement venir à décéder or cette information n'est pas disponible si la létalité est évaluée avant. C'est un biais classique dans l'analyse de survie connu sous le nom de censure mais ce n'est pas l'objet de notre mémoire.

Les différents taux de létalité de virus sont disponibles ci-dessous :



Taux de létalité et \mathcal{R}_0 de différents virus

3. Modèle Stochastique

Les modèles stochastiques en épidémiologie gardent le même principe que les modèles déterministes car ils fonctionnent aussi sur le passage entre les différents états mais ils permettent de prendre en compte les différents contacts entre les personnes. Le schéma des contacts au sein de la population est modélisé comme un réseau statique et la taille de la population est connue (de sorte que le nombre de nœuds dans le réseau est connu et fixe). La propagation du virus au sein d'un réseau est modélisée comme un modèle à compartiment stochastique. Ainsi le modèle va donner l'information sur les individus qui sont infectés au cours de l'épidémie le long des contacts et les différents moments auxquels ils passent d'un état à un autre.

Le réseau de contact est un réseau statique et non orienté dans lequel les individus sont des nœuds et une arête existe entre deux nœuds quand il y a un contact. Il est important de noter que dans le modèle, une arête (a, b) peut exister entre deux individus a infecté d'une part et b sain d'autre part traduisant un contact entre les deux sans pour autant qu'il y ait transmission du virus. En effet, la transmission est un risque possible lors d'un contact entre individu contaminé et individu infecté mais la transmission n'est pas systématiquement la résultante d'un contact et le modèle stochastique le prend en compte.

L'utilisation d'un modèle à compartiment stochastique pour la propagation d'un virus est largement répandue et non controversée. Bien que dans la suite nous parlerons uniquement du SEIR, le cadre utilisé peut largement être adapté à d'autres type de modèle à compartiments.

Le modèle prend en entrée le nombre d'individus à l'état susceptible et nous donne l'information sur le nombre de personnes infectées à la fin de l'épidémie et pour chaque personne infectée son temps d'exposition et son temps d'infection.

3.1 Réseau de contact

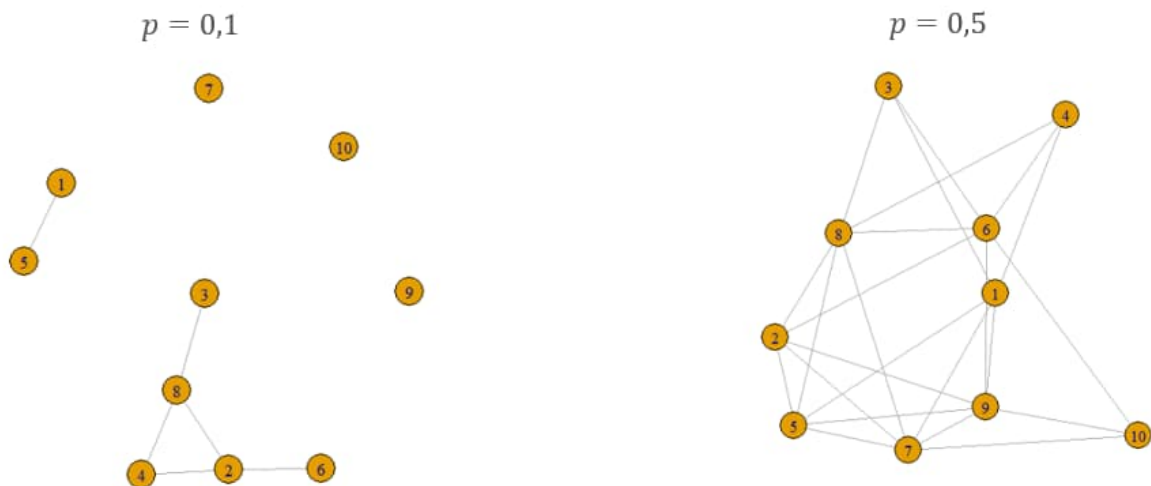
La propagation d'une épidémie tire sa principale « force » du nombre de personnes susceptibles de base autrement dit la population totale initiale mais aussi de l'existence des contacts au sein de cette population. Cela est compréhensible car si nous prenons un cas extrême où la population est grande mais dans laquelle aucun contact n'existe alors il n'y a pas de possibilité de propagation de virus car l'épidémie vit à travers les contacts entre individus alors que dans l'autre extrême où la population est grande et les contacts nombreux, on retrouve naturellement un environnement propice à la diffusion du virus.

Nous voyons donc l'intérêt de la construction d'un réseau de contact, ce dernier étant une sorte de matrice de diffusion du virus au sein de la population. Mais pour que ce réseau soit fiable, il faut qu'il se rapproche le plus possible de la réalité c'est-à-dire que, par exemple, tous les individus au sein d'une même population ne vont pas nécessairement rentrer en contact entre eux ou encore moins se contaminer systématiquement au cours de tous les contacts.

Groendyke a développé un modèle qui génère un graphe d'Erdős-Rényi. Le principe de ce graphe est plutôt simple et consiste à considérer une population de taille N avec chacun des N individus représentant un sommet puis de considérer des arêtes entre un individu i et un individu j pouvant exister avec une probabilité p_{ij} indépendamment de tous les autres sommets. S'il n'y a pas d'arêtes

entre i et j alors toute transmission est impossible entre ces deux individus même si l'un des deux est contaminé et l'autre sain.

Si pour tout $i, j \in \{1, \dots, N\}^2$ et $i \neq j$, on a $p_{ij} = p$ alors cela signifie qu'il existe une arête entre deux individus avec une probabilité p . Il y a une correspondance évidente entre ce procédé et des tirages de Bernoulli indépendants de paramètres p . On retrouve ci-dessous une illustration dans une population de $N = 10$ personnes avec d'une part $p = 0,1$ et d'autre part $p = 0,5$



Graphe d'Erdős-Rényi d'une population de 10 personnes avec $p=0,1$ (gauche) et $p=0,5$ (droite)

De manière plus générale, on a :

$$\ln\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \sum_m v_m X_{(i,j),m}$$

En passant à l'exponentielle, on obtient que l'expression de p_{ij} est :

$$p_{ij} = \frac{e^{\sum_m v_m X_{(i,j),m}}}{1 + e^{\sum_m v_m X_{(i,j),m}}}$$

Où :

- Les v_m sont des paramètres réels
- Les $X_{(i,j),m}$ sont les termes de m matrices dyadiques

On voit donc que la probabilité p_{ij} d'avoir un contact entre i et j va dépendre de m jeux de données sur les caractéristiques du couple (i, j) .

Quand on a $m = 1$ (cas simple) et pour tout i, j , $X_{(i,j),1} = 1$ alors on se ramène au cas illustratif que nous avons vu plus haut et $v_1 = 1$.

Il faut voir les $X_{(i,j),m}$ comme toutes sortes de caractéristiques possibles permettant de savoir si deux individus sont dans des situations favorisant la possibilité d'un contact entre les deux.

Par exemple, si au sein d'une ville, on s'intéresse au lieu d'habitation et au lieu de travail pour les contacts, on pourrait avoir :

$$X_{(i,j),m} = 0,5\mathbf{1}_A + 0,5\mathbf{1}_B$$

Avec

$$\mathbf{1}_A = \begin{cases} 1 & \text{si } i \text{ et } j \text{ vivent dans le même quartier} \\ 0 & \text{sinon} \end{cases}$$

$$\mathbf{1}_B = \begin{cases} 1 & \text{si } i \text{ et } j \text{ exercent un travail dans le même quartier} \\ 0 & \text{sinon} \end{cases}$$

Au vu de ces éléments, nous voyons déjà un avantage d'une modélisation épidémiologique sur une base stochastique par rapport au modèle déterministe qui est donc la prise en compte de la complexité des contacts entre les individus. Là où les contacts sont uniquement causés par un taux fixe dans le modèle déterministe, ils prennent place sous les traits d'un schéma d'Erdős-Rényi plus proche de la réalité dans le modèle stochastique.

Ce n'est pas le seul reproche qu'on peut faire au modèle déterministe. Bien qu'il soit simple à modéliser et très compréhensible, sa dynamique qui évolue avec des paramètres fixes ne donne qu'un unique scénario. L'aléas sur la transmission, les temps d'incubation et de rétablissement n'est pas adapté au modèle déterministe. Le modèle stochastique le permet en simulant des variables aléatoires sur les différents paramètres constitutifs du modèle.

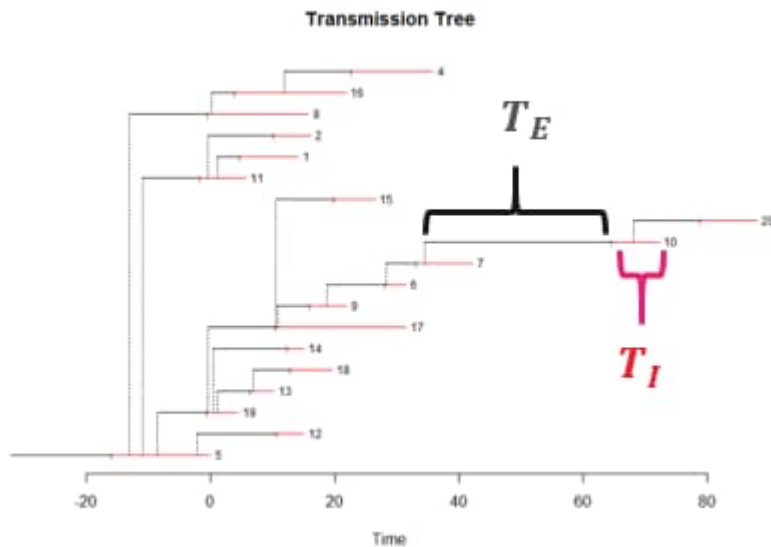
3.2 Les temps d'incubation et d'infection

Tout comme le SEIRD déterministe, nous avons aussi les temps d'incubation et d'infection qui désormais vont être modélisés par des variables aléatoires.

- On a d'une part T_E correspondant à la période d'incubation pendant laquelle l'individu est resté dans l'état E .
- D'autre part, on a T_I correspondant à la période d'infection pendant laquelle l'individu est resté dans l'état I .

Selon Groendyke, ces deux variables suivent des lois gamma respectivement de paramètre (k_E, θ_E) et (k_I, θ_I) . Ce choix de variable nous semble assez cohérent car la loi gamma est très souvent bien adaptée aux phénomènes temporels or ici on s'intéresse bien à des périodes d'incubation et d'infection.

Si on considère une population de $N = 20$ personnes, le modèle SEIR stochastique va graphiquement évoluer de la sorte :



Voici comment il faut lire ce graphique :

- L'individu 5 est le premier contaminé au sein de la population des 20 individus
- Les lignes noires représentent les périodes d'incubation. C'est donc la période pendant laquelle un individu qui a été contaminé ne peut pas encore contaminer d'autres personnes en cas de contact.
- Les lignes rouges représentent les périodes d'infection. C'est au cours de cette période que l'individu devient contagieux. D'ailleurs, les lignes verticales représentant la transmission ne peuvent croiser que les lignes rouges pour cette raison.
- L'épidémie aurait pu avorter si l'individu 5 n'avait eu aucun contact contaminant mais ici cet individu a contaminé l'individu 8 puis l'individu 11 ensuite l'individu 19 et sa dernière contamination fut l'individu 12.

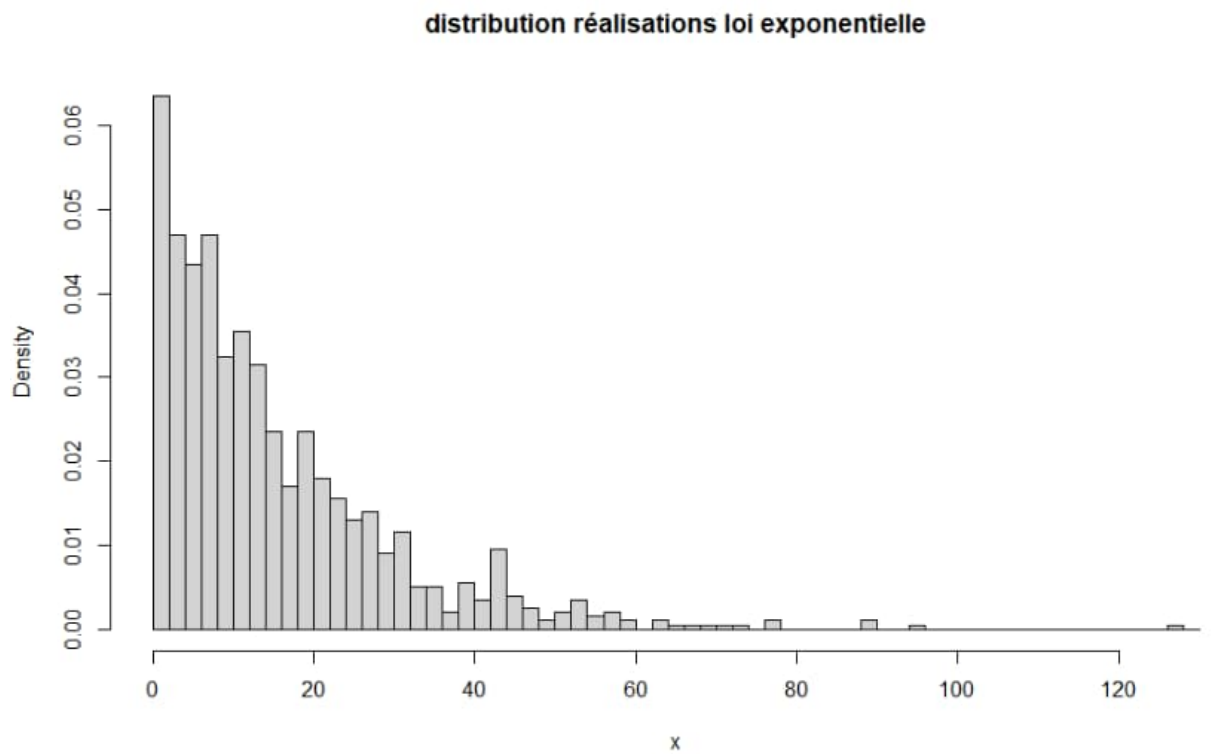
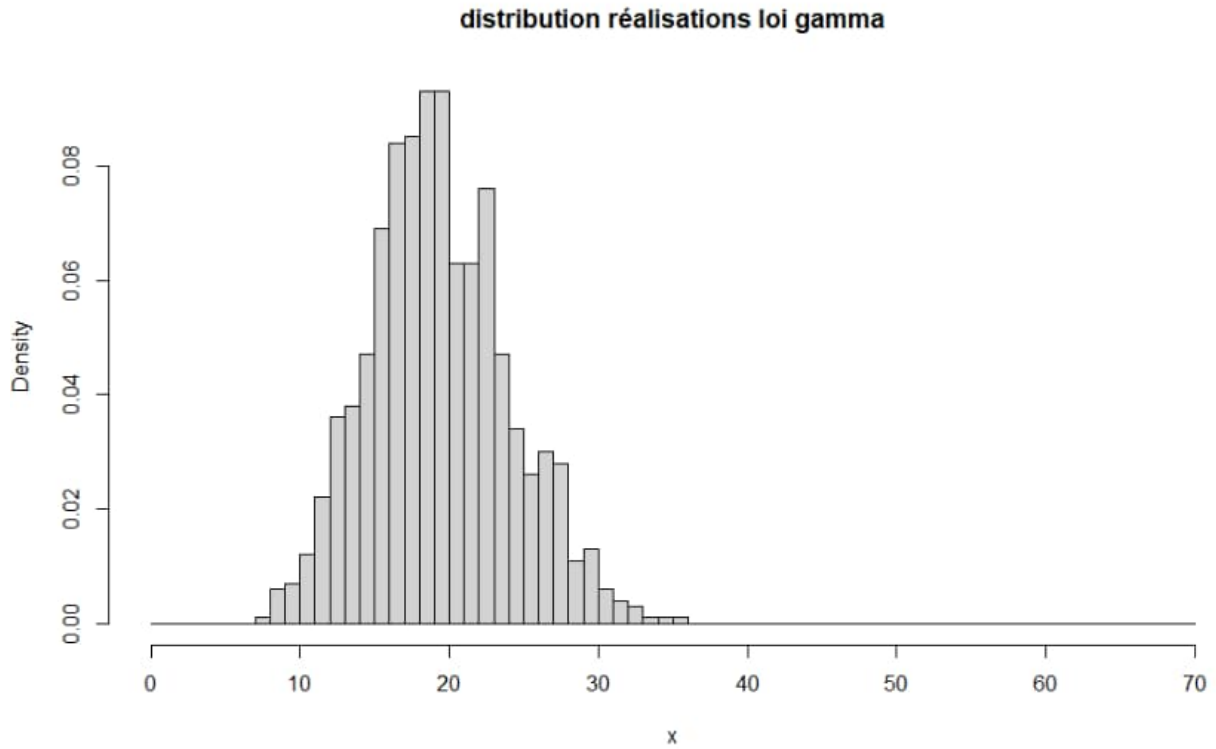
Cette petite simulation dans une population de 20 personnes a abouti à 19 personnes infectées excepter l'individu 3. On peut l'expliquer de plusieurs façons :

- Lors des tirages de Bernoulli, aucune liaison n'a été faite entre l'individu 9 et les autres individus. A partir de là, toute contamination de ce dernier devient totalement impossible. Cela peut arriver si l'individu s'isole du début à la fin de l'épidémie.
- L'individu rentre en contact avec des personnes qui ont été infectieuses mais sont désormais rétablies. Ainsi, les personnes infectieuses n'ont pas eu le temps de contaminer l'individu 9 au cours d'un contact.
- L'individu rentre en contact avec des personnes infectieuses pendant le période d'infection mais aucun contact n'aboutit à une contamination.

Pour la modélisation des temps, on pourrait aussi penser à la loi exponentielle qui est un cas particulier de la loi gamma. Mais nous visons des temps moyens pour nos distributions qui font que les réalisations de la loi exponentielle sont beaucoup trop irréalistes pour des temps liés à une maladie. En effet, nous préférons une loi qui se concentre autour de sa moyenne plutôt qu'une loi

très éparpillée autour de celle-ci car les temps d'infection et d'incubation sont généralement très proches entre les individus.

Si par exemple nous considérons une période d'infection moyenne de 15 jours alors nous avons les distributions simulées suivantes :



Dans les deux cas, la moyenne est de 15 jours mais nous voyons bien que les réalisations de la loi gamma sont concentrées autour de sa moyenne tandis que les réalisations de la loi exponentielle sont majoritairement très petites et il existe des réalisations beaucoup trop grandes qui peuvent être considérées comme aberrantes dans le contexte d'une maladie.

De plus, la loi gamma a l'avantage d'offrir plus de flexibilité grâce à ces deux paramètres contrairement à l'unique paramètre de la loi exponentielle. Ce nombre de paramètres est plutôt intéressant à exploiter et suffisamment petit pour éviter trop de complexité.

Pour simplifier, dans la suite nous considérerons que :

$$\theta_E = \theta_I$$

Nous rappelons que pour une variable aléatoire $T \sim \text{Gamma}(k, \theta)$, on a :

$$\begin{cases} k, \theta > 0 \\ \mathbb{E}[T] = k\theta \\ \mathbb{V}[T] = k\theta^2 \end{cases}$$

Le choix de k et θ est très important car c'est à partir de ces derniers qu'on décide de la forme de la distribution recherchée. Nous verrons par la suite le choix à faire sur les paramètres k et θ .

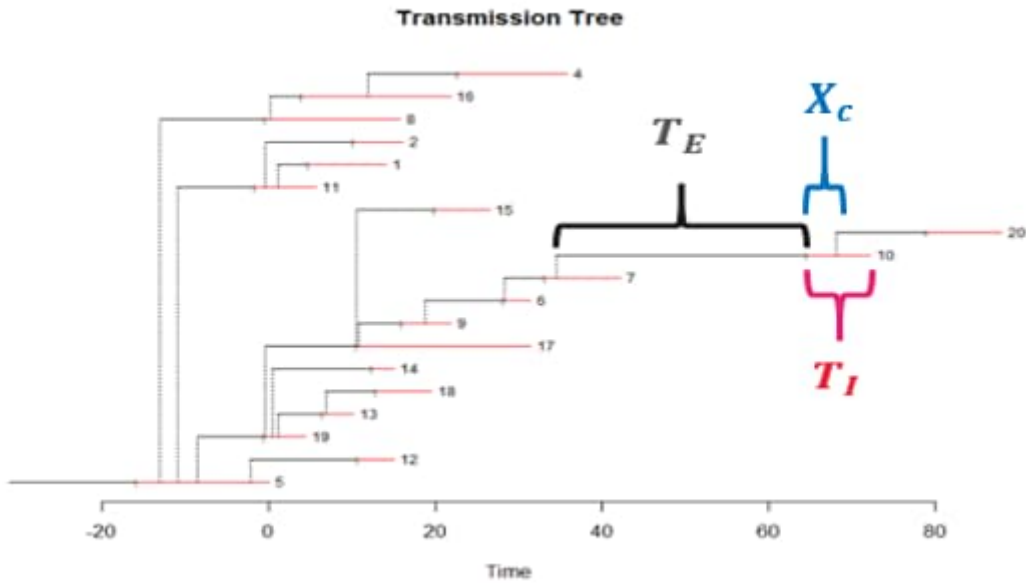
En amont des temps d'incubation et d'infection, il est primordial de connaître le temps de transmission d'un individu à un autre car c'est par la vitesse de transmission au sein d'une population qu'une épidémie va plus au moins vite se propager.

3.3 La transmission entre deux individus

Comme nous l'avons vu, pour qu'il y ait transmission entre deux individus, il faut au préalable l'existence d'un contact qui lie les deux.

Pour chaque contact c , on a un temps X_c correspondant au temps entre le début de la période d'infection du premier des deux individus et l'instant auquel le second sort de l'état susceptible pour entrer dans sa période d'incubation. On modélise X_c par une loi exponentielle de paramètre β . Comme le parcours viral d'un individu à un autre est totalement indépendant de son parcours précédent, la loi exponentielle semble être en adéquation avec le phénomène car c'est une loi qui modélise bien les durées de vie de phénomènes sans mémoire.

En partant du graphique de transmission précédent et en ajoutant le temps de transmission, nous avons :



Pour chaque contact c , il va s'agir de savoir si le temps de transmission X_c est inférieur à la période d'infection T_I . Autrement dit, une personne infectieuse contamine une personne saine au cours d'un contact c si et seulement si $X_c \leq T_I$. Il est donc clair qu'au départ c'est un contact social qui conditionne l'existence de l'épidémie puis sa vitesse de transmission selon la comparaison entre X_c et T_I .

A titre de rappel, pour une variable aléatoire $X_c \sim \text{Exp}(\beta)$, on a :

$$\begin{cases} \beta > 0 \\ \mathbb{E}[X_c] = \frac{1}{\beta} \\ \mathbb{V}[X_c] = \frac{1}{\beta^2} \end{cases}$$

3.4 Calcul du taux de reproduction de base \mathcal{R}_0

Comme nous l'avons vu à la section 2.1, le taux de reproduction \mathcal{R}_0 est le nombre moyen de secondes infections qui ont lieu lorsqu'un individu infecté est introduit dans une population composée entièrement d'individus susceptibles. Nous pouvons interpréter le nombre moyen de secondes infections comme l'espérance du nombre d'infections secondaires produit par une personne infectieuse. Notre réseau de contact est généré sur une population de N individus par un tirage de Bernoulli avec une probabilité p . En utilisant le temps de transmission X_c et la période d'infection T_I . Nous avons par (Britton & O'Neill, 2002) :

$$\mathcal{R}_0 = Np\mathbb{P}(X_c < T_I)$$

Nous avons posé que X_c suit une loi exponentielle de paramètre β et que T_I suit une loi gamma de paramètre (k_I, θ_I) . Nous avons donc l'expression explicite de \mathcal{R}_0 ci-dessous :

$$\mathcal{R}_0 = Np \left(1 - \left(\frac{1}{1 + \beta\theta_I} \right)^{k_I} \right)$$

Cette expression va nous aider à nous donner une idée sur le choix de p .

4. Données et construction de scénario

Dans cette section, nous allons mettre en évidence les différentes données qui vont servir pour la construction de nos scénarios d'épidémies notamment par le calcul des probabilités d'hospitalisation qui ne sont pas directement disponibles en sortie du modèle SEIR stochastique de base. En effet, avec le modèle épidémiologique, nous tentons de faire ressortir l'évaluation du risque des hospitalisations or le modèle qui a été présenté jusqu'à maintenant ne le prend pas en compte. Il va donc falloir déterminer des probabilités d'hospitalisation pour les personnes infectées au sein du modèle.

Dans l'idéal, il aurait été judicieux d'utiliser les données de Pacifica entité non-vie du Crédit Agricole Assurances qui a un portefeuille d'assurances santé collectives directement exposé au risque des hospitalisations. Mais le mémoire étant réalisé au sein de l'entité vie, nous n'avons pas accès aux données non-vie de la compagnie. C'est pourquoi nous allons utiliser des données publiques et tenter de les rapprocher avec un portefeuille fictif. Au vu des données disponibles que nous allons détailler plus bas, seules les hypothèses sur la classe d'âge du portefeuille et le niveau de risque d'hospitalisation seront nécessaires pour déterminer l'impact des chocs issus du modèle. Nous considérons donc que le portefeuille est celui des assurances santé collectives de Pacifica, les assurés sont actifs et âgés de 20 à 64 ans, ils ont donc en moyenne un risque d'hospitalisation réduit par rapport à la population générale. Nous avons conscience des simplifications faites sur le portefeuille car les impacts du modèle ne peuvent être interprétés que selon celles-ci.

Nous allons utiliser les données de l'Institut National de Veille Sanitaire (2005). Ces données sont issues des travaux de Aoife Doyle, Isabelle Bonmarin, Daniel Lévy-Bruhl, Yann Le Strat, Jean-Claude Desenclos. Cette étude avait pour objectif de mesurer l'impact d'une pandémie en déterminant potentiellement le nombre de cas et le nombre d'hospitalisations.

4.1 Les cas par âge

La population est répartie sous forme de trois classes d'âge : 0-19 ans, 20-64 ans et 65+ ans qui peuvent respectivement être interprétées comme la population jeune, la population des actifs et la population des retraités.

Age	Répartition des cas par classe d'âge
0-19	40%
20-64	50%
65+	10%

Pour estimer ces pourcentages, l'INVS s'est basée sur les informations des pandémies connues qui ont eu lieu au cours du 20^{ème} siècle. L'hypothèse du taux d'infection est entre 15 et 35% c'est-à-dire que si on considérait la taille totale de la pandémie comme étant le cumul du nombre de personnes qui ont été infectées au cours de la pandémie alors ce nombre cumulé d'infectés correspondrait à un intervalle de 15 à 35% de la population saine initiale.

4.2 Les cas à risque par âge

Pour faire une étude plus ciblée, l'INVS a fait la distinction parmi les cas entre ceux qui sont à faible risque de développer une forme grave et ceux qui peuvent en développer une assez facilement. Pour se donner une meilleure idée de ce dont il s'agit, les personnes à complications graves sont généralement des personnes atteintes d'une affection longue durée (ALD) comme le diabète ou les maladies cardiovasculaires.

Age	Pourcentage de la population à risque par classe d'âge
0-19	14%
20-64	5%
65+	50%

4.3 Taux d'hospitalisation par âge

Le taux d'hospitalisation au cours d'une pandémie dépend beaucoup de la gravité de la maladie et des moyens qui sont mis en œuvre pour la freiner. Les données sont issues des pandémies passées et comme pour la répartition des cas, elles sont fournies sous trois classes d'âge. De plus, comme tous les profils ne courent pas le même niveau de risque, une distinction est faite entre les personnes à faible risque et les personnes à plus fort risque.

		Age	Taux hospitalisation
Hospitalisation pour 1000 cas	Faible risque	0-19	1-20-80
		20-64	2-30-80
		65+	10-50-80
	Haut risque	0-19	5-100-300
		20-64	5-125-300
		65+	50-150-300

Telles qu'ils sont donnés, ces taux correspondent à des taux en excès de ce qui est observé en temps normal pour 1000 habitants. Par exemple le coefficient d'hospitalisation de 125 pour 1000 pour les personnes à risque élevé âgées de 20 à 64 ans est un excès d'hospitalisation dû à la pandémie qui n'existait pas dans les années précédentes sans pandémie.

On observe aussi que les taux sont donnés avec trois seuils pour chaque classe d'âge. C'est notamment une conséquence de l'incertitude autour de ces derniers. En lisant les coefficients horizontalement, nous voyons qu'ils augmentent et peuvent être assimilés de gauche à droite à un scénario faible, moyen (probable) et dur (critique).

4.4 Modélisation des hospitalisations dans le modèle SEIR stochastique

Comme ses différents états l'indiquent, le modèle SEIR stochastique ne prend pas en compte les éventuelles hospitalisations et la durée de ces hospitalisations or une pandémie impacte un portefeuille d'assurance en partie par les hospitalisations car les assurés sont indemnisés d'une part par la sécurité sociale et d'autre part le reste à charge concerne directement l'assureur.

Pour prendre en compte ces différents éléments au sein du modèle, nous allons utiliser les données de l'INVS présentées précédemment. A partir de la répartition des cas par âge, de la répartition de la population à risque par âge, des taux d'hospitalisation par âge selon que la population soit à risque ou non, nous allons établir des probabilités d'hospitalisation à appliquer à chaque simulation du modèle SEIR stochastique dans le but d'obtenir une distribution du nombre de personnes hospitalisées. Nous verrons ensuite comment considérer le nombre de jours d'hospitalisation.

Les taux d'hospitalisation de l'INVS vus en 4.3 sont donnés avec trois coefficients évaluant la criticité de la pandémie pour chaque classe d'âge et chaque niveau de risque, ces coefficients serviront de sensibilités pour l'évaluation du SCR Santé CAT Pandémie.

Enfin, nous omettons volontairement l'incidence dû aux arrêts de travail car bien qu'étant exposé à ce risque, les garanties PREDICA ont une franchise de 90 jours qui rendent obsolètes voire impossible un impact sur le portefeuille d'une sinistralité supplémentaire au cours de la pandémie car comme nous allons le voir par la suite, les périodes d'infection dépassent rarement ne serait-ce que 60 jours.

4.4.1 Les probabilités

Nous introduisons les évènements suivants :

- M l'évènement « tomber malade » / « être contaminé »
- A_i l'évènement « appartenir à la classe d'âge i » avec A_1 correspondant à la classe d'âge 0-19 ans, A_2 à la classe d'âge 20-64 ans et enfin A_3 à la classe d'âge 65+ ans.
- R l'évènement « être à risque »
- H l'évènement « être hospitalisé »

Avec ces évènements, plusieurs probabilités sont déjà connues. En effet, pour tout i , nous pouvons déterminer $\mathbb{P}(A_i|M)$ par le tableau de la répartition des cas par âge en section 4.1

La probabilité $\mathbb{P}(M)$ correspond à la probabilité d'être contaminé au cours de la pandémie. Cette probabilité détermine ainsi la taille finale de la pandémie que nous choisissons à 35%.

Il nous est aussi aisé de déterminer $\mathbb{P}(R|A_i)$ avec le tableau de la répartition des cas à risque par âge et donc a fortiori nous avons $\mathbb{P}(R^c|A_i)$.

Nous avons aussi $\mathbb{P}(H|M \cap A_i \cap R)$ et $\mathbb{P}(H|M \cap A_i \cap R^c)$ par le tableau sur les niveaux d'hospitalisation par niveau de risque et âge. Enfin, $\mathbb{P}(A_i)$ est déterminée par la répartition des classes d'âge au sein de la population française que l'on peut obtenir via des sources comme l'INSEE.

Nous faisons l'hypothèse que les évènements R et M sont indépendants. En effet, d'une part une personne se retrouve contaminée indépendamment du fait qu'elle soit à risque ou non car elle entre en contact avec d'autres individus. Il y a donc un lien inexistant entre contamination et risque de ce point de vue. D'autre part, une personne à risque ne l'est pas parce qu'elle est contaminée mais elle est à risque antérieurement à la contamination car comme nous l'avons mis en évidence en section 4.2, les personnes à risque sont des personnes qui de base ont une santé fragile en étant par exemple atteintes d'affection longue durée.

4.4.2 Probabilité d'hospitalisation

Les probabilités d'être hospitalisé conditionnellement au fait d'être malade ou contaminé sont des taux d'hospitalisation par âge dû aux infections de la pandémie. Nous calculons $\mathbb{P}(H|M)$:

$$\begin{aligned}\mathbb{P}(H|M) &= \frac{\mathbb{P}(H \cap M)}{\mathbb{P}(M)} \\ &= \frac{\sum_{i=1}^3 \mathbb{P}(H \cap M \cap A_i \cap R) + \mathbb{P}(H \cap M \cap A_i \cap R^c)}{\mathbb{P}(M)}\end{aligned}$$

Or,

$$\begin{aligned}
\mathbb{P}(H \cap M \cap A_i \cap R) &= \mathbb{P}(H|M \cap A_i \cap R) \cdot \mathbb{P}(M \cap A_i \cap R) \\
&= \mathbb{P}(H|M \cap A_i \cap R) \cdot \mathbb{P}(M|A_i \cap R) \cdot \mathbb{P}(A_i \cap R) \\
&= \mathbb{P}(H|M \cap A_i \cap R) \cdot \mathbb{P}(M|A_i \cap R) \cdot \mathbb{P}(R|A_i) \cdot \mathbb{P}(A_i)
\end{aligned}$$

On en déduit que $\mathbb{P}(H \cap M \cap A_i \cap R^c) = \mathbb{P}(H|M \cap A_i \cap R^c) \cdot \mathbb{P}(M|A_i \cap R^c) \cdot \mathbb{P}(R^c|A_i) \cdot \mathbb{P}(A_i)$

Ainsi,

$$\mathbb{P}(H|M) = \frac{\mathbb{P}(H|M \cap A_i \cap R) \cdot \mathbb{P}(M|A_i \cap R) \cdot \mathbb{P}(R|A_i) \cdot \mathbb{P}(A_i) + \mathbb{P}(H|M \cap A_i \cap R^c) \cdot \mathbb{P}(M|A_i \cap R^c) \cdot \mathbb{P}(R^c|A_i) \cdot \mathbb{P}(A_i)}{\mathbb{P}(M)}$$

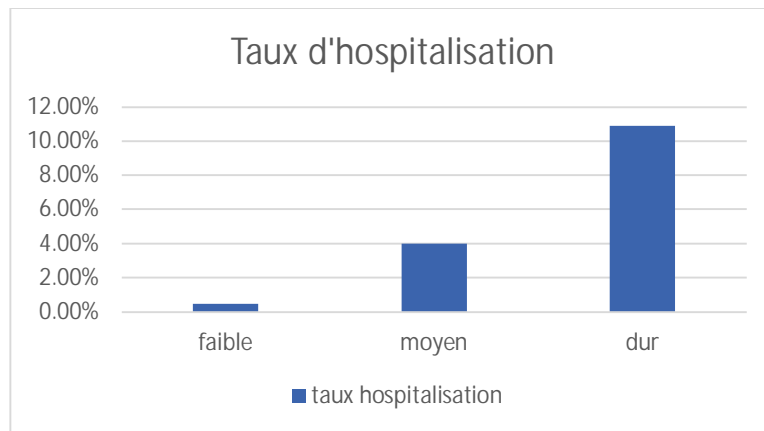
Comme M et R sont indépendants, alors, on a :

$$\mathbb{P}(M|A_i \cap R) = \mathbb{P}(M|A_i) = \frac{\mathbb{P}(M \cap A_i)}{\mathbb{P}(A_i)} = \frac{\mathbb{P}(A_i|M) \cdot \mathbb{P}(M)}{\mathbb{P}(A_i)}$$

En simplifiant les termes dans $\mathbb{P}(H|M)$, nous avons :

$$\mathbb{P}(H|M) = \sum_{i=1}^3 \mathbb{P}(H|M \cap A_i \cap R) \cdot \mathbb{P}(A_i|M) \cdot \mathbb{P}(R|A_i) + \mathbb{P}(H|M \cap A_i \cap R^c) \cdot \mathbb{P}(A_i|M) \cdot \mathbb{P}(R^c|A_i)$$

Nous observons les résultats suivant pour les trois scénarios:



Sans surprise, les résultats sont logiques avec ce à quoi on peut s'attendre. En effet, avec les coefficients du scénario faible, on obtient des probabilités d'hospitalisations très faibles et avec le scénario dur (ou critique), on obtient une probabilité d'hospitalisation d'un peu plus de 10% ce qui représente plus du double du scénario faible ou du scénario moyen.

4.5 Nombre de jours d'hospitalisation

Nous avons vu qu'un des soucis majeurs du SEIR stochastique est qu'il ne nous donne pas d'informations sur le nombre de jours d'hospitalisations or l'assureur est impacté par le coût journalier d'une hospitalisation. Nous allons donc élaborer une méthode à partir du modèle pour obtenir un nombre de jour moyen d'hospitalisation.

En partant du postulat que plus un individu va avoir un temps d'infection long, plus il y a de grandes chances qu'il s'agisse d'un individu qui a développé une forme grave de la maladie et est donc assujéti à une hospitalisation, nous allons utiliser la variable aléatoire T_I qui simule la période d'infection dans nos pandémies. Nous rappelons que cette variable aléatoire suit une loi gamma (k_I, θ_I) et comme toutes les personnes infectées vont avoir une période d'infection, il est donc question de déterminer un quantile sur notre distribution à partir duquel une personne infectée est considérée comme hospitalisée.

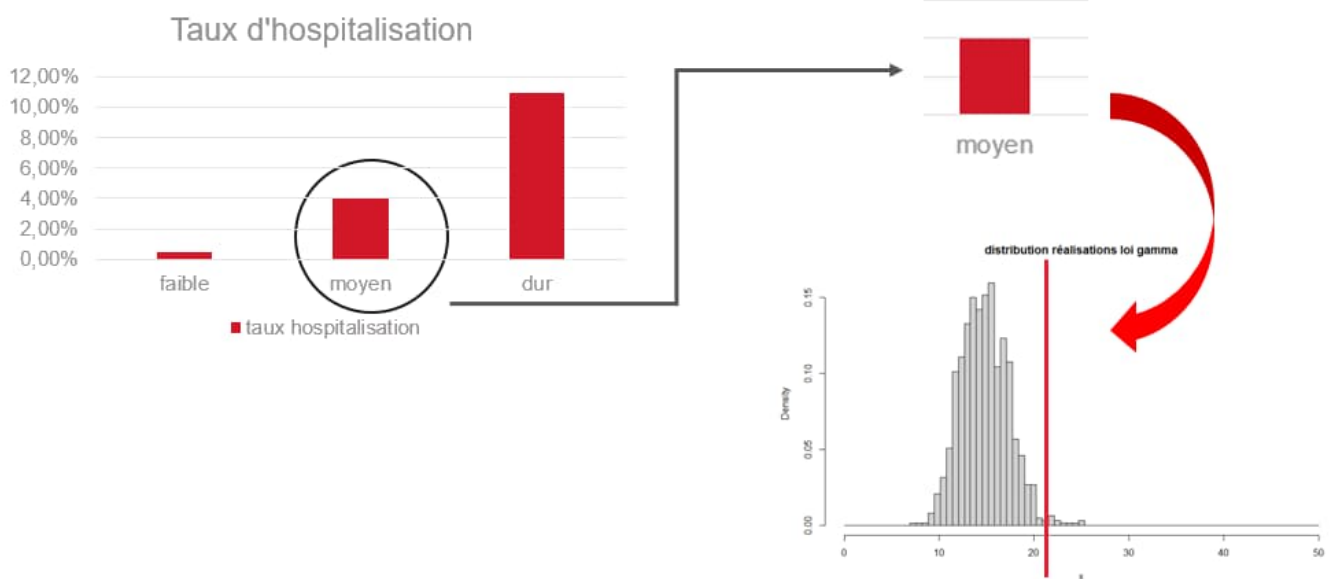
Comme $\mathbb{P}(H|M)$ est la probabilité d'être hospitalisé, il est intuitif de considérer $1 - \mathbb{P}(H|M)$ comme étant le quantile adéquat sur notre distribution gamma de la période d'infection. Comme $\mathbb{P}(H|M)$ est relativement faible, le quantile va donc être élevé et va faire directement la part entre la proportion des infectés hospitalisés et celle qui ne l'est pas.

Il reste donc à déterminer les paramètres de la loi gamma. Nous voulons simuler une pandémie générique et dans la littérature des différentes maladies existantes, nous trouvons qu'une période d'incubation de 5 jours et une période d'infection de 15 jours semblent tout à fait standard. Nous paramétrons donc la loi gamma qui modélise la période d'infection de telle sorte à avoir une moyenne de 15 jours c'est-à-dire $k_I \theta_I = 15$. Un nombre infini de combinaison est alors possible pour déterminer k_I et θ_I mais dans un contexte de maladie, les périodes d'infection sont globalement proches de la moyenne ainsi, il est préférable d'avoir $\theta_I < 1$. Finalement, nous décidons de prendre $\theta_I = 0,9$ ce qui nous amène à prendre $k_I = 16,66$. De même pour la période d'incubation, nous fixons $\theta_E = 0,9$ et donc $k_E = 5,55$.

Les résultats sont ci-dessous:

	Scénario		
	faible	moyen	dur
quantile hospitalisation	99,53%	96,01%	89,12%
nombre de jours correspondant	26,24	22	19,63

Schématiquement, si nous faisons, une simulation à partir des paramètres d'une pandémie moyenne, nous avons :



La probabilité d'hospitalisation nous permet de déterminer le nombre de jours d'hospitalisation d'un cas grave associé à la distribution des temps d'infection simulés au cours d'une pandémie puis on prend le quantile correspondant. Ainsi, dans l'exemple d'une pandémie moyenne illustré ci-dessus, la probabilité d'hospitalisation est de 3,99% et le quantile associé est donc le quantile d'ordre 96,01% correspondant à 22 jours. Dans cette simulation, on obtient donc que toutes les personnes qui n'ont pas atteint 22 jours d'infection sont des personnes non hospitalisées et que les autres le sont.

5. Choix du paramètre \mathcal{R}_0

Chaque simulation de pandémie va se faire au sein d'une population de $N = 1000$ individus et nous avons vu que dans le modèle stochastique c'est par des contacts que le virus se propage. Ces contacts sont modélisés par un graphe d'Erdős-Rényi qui s'assimile à un tirage de Bernoulli où un contact entre deux personnes a une probabilité p de se réaliser. Cette probabilité p est très importante dans la quantification du taux de reproduction de base \mathcal{R}_0 car nous rappelons que la formule de ce dernier dans le modèle stochastique est :

$$\mathcal{R}_0 = Np \left(1 - \left(\frac{1}{1 + \beta\theta_I} \right)^{k_I} \right)$$

A cela, nous ajoutons aussi la contrainte d'avoir une taille finale de pandémie ou encore nombre d'infectés cumulé d'environ 20 à 35% de la population saine de base.

5.1 Calibrage de β

Dans le modèle, lorsqu'un contact social concerne un infecté et une personne qui n'a pas encore été contaminée, une variable de temps de contamination X_c est modélisée pour savoir si ce contact va aboutir à la contamination de l'individu sain ou non. Nous rappelons que ce temps de contamination suit une loi exponentielle de paramètre β et donc sa moyenne est de $\frac{1}{\beta}$. Ainsi, le temps de

contamination est inversement proportionnel à β . Comme la vitesse de propagation d'un virus est très largement dépendante de la vitesse de transmission, ce paramètre est d'autant plus important. Un temps de transmission d'un jour nous semble raisonnable donc nous fixons $\beta = 1$.

5.2 Calibrage de p

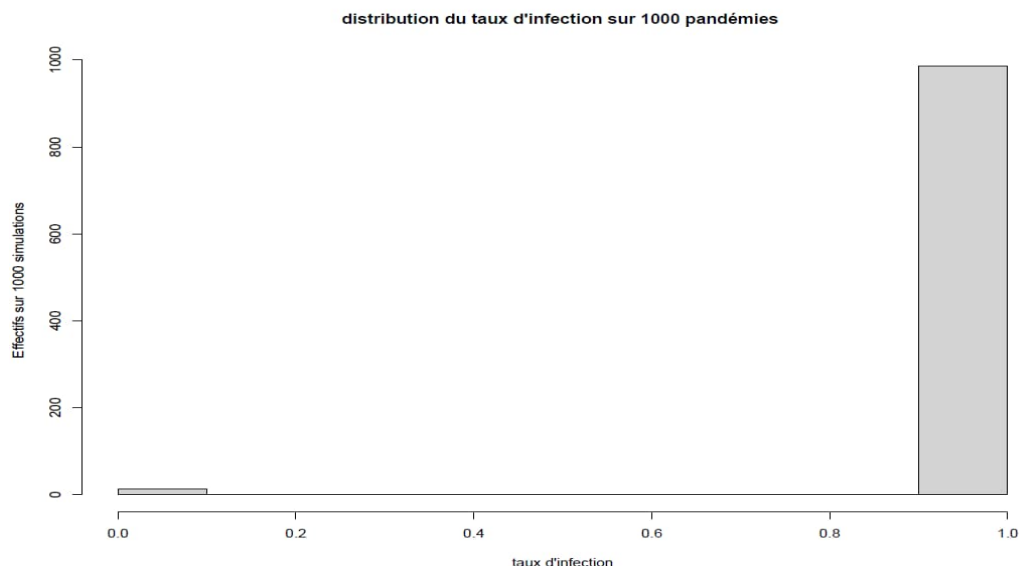
Comme nous l'avons vu dans le graphique des taux de létalité en fonction des taux de transmission en section 2.2, on peut considérer que le taux de reproduction de base \mathcal{R}_0 s'évalue généralement entre 0 et 4 or le coefficient $1 - \left(\frac{1}{1+\beta\theta_I}\right)^{k_I}$ qui compose la formule du \mathcal{R}_0 est compris entre 0 et 1. Ainsi, pour que \mathcal{R}_0 soit entre 0 et 4, il aurait suffi que le second coefficient Np soit égal à 4 et donc, on aurait :

$$p = \frac{4}{N} = 0,004$$

Comme on sait que

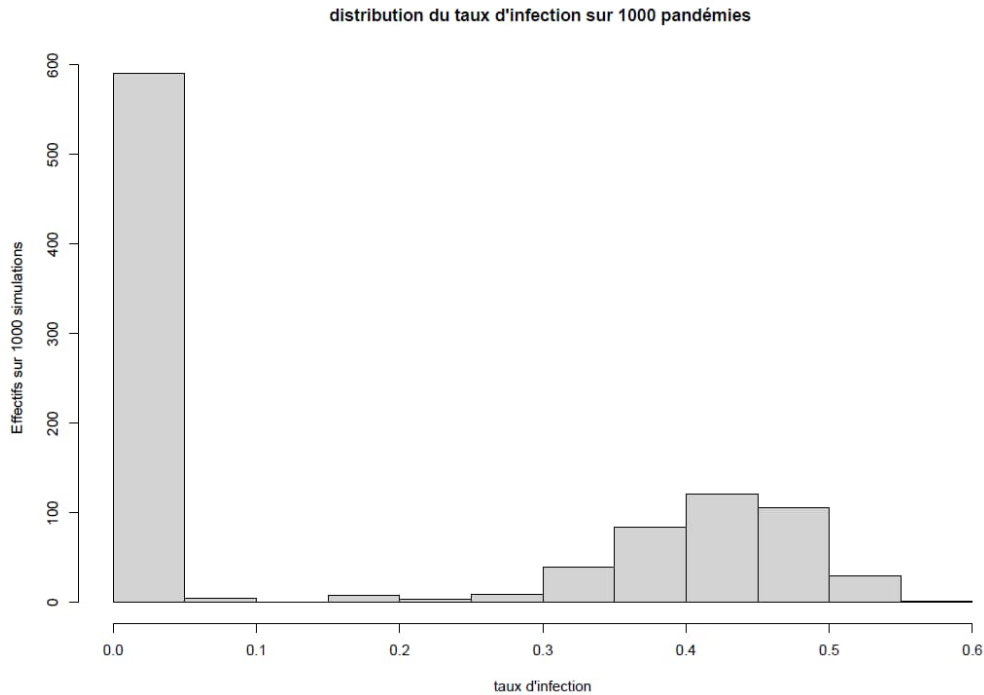
$$p = \frac{e^v}{1 + e^v} \text{ alors } v = -5,517$$

Cependant, en voulant une taille finale d'infectés entre 200 et 350 pour 1000 personnes, la pratique ne donne pas ce résultat si bien que le v adéquat doit être cherché de manière minutieuse. En effet, avec une population de 1000 personnes et $v = -5,517$, nous obtenons la distribution des taux d'infection suivante sur 1000 pandémies :



Nous obtenons un effet polarisant avec soit des pandémies qui ne durent pas car le virus ne se transmet pratiquement pas (taux d'infection à 0 dans l'histogramme) et soit des pandémies dans lesquelles le cumul du nombre d'infectés est égal à la population saine initiale (taux d'infection égal à 1 dans l'histogramme) autrement dit, toute le monde a été infecté au cours de la vie de la pandémie. On remarque d'ailleurs que la proportion des pandémies ayant un taux d'infection à 1 est largement plus significative que celles ayant un taux d'infection à 0. Cette valeur de v représente un maximum parmi les valeurs à tester car l'augmenter reviendrait à faire croître les contacts et donc réduire les pandémies à taux d'infection nul pour avoir uniquement des pandémies à taux d'infection égale à 1 ce qui est absurde.

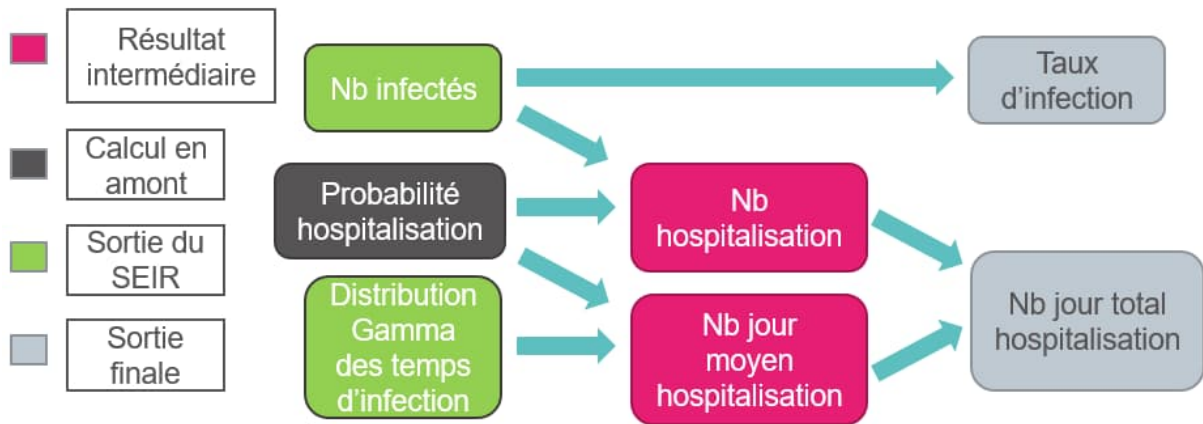
Après plusieurs tests, nous finissons par retenir $\nu = -6,644$ ce qui donne $p = 0,0013$. Bien qu'il soit difficile d'interpréter cette valeur de p , on peut en dire qu'au sein d'une large population, il est peu probable qu'une personne entre en contact avec toutes les personnes du reste de la population. De plus, cette valeur de p nous donne un taux de reproduction de base \mathcal{R}_0 qui est bien dans entre 0 et 4. Nous obtenons la distribution des taux d'infection suivante :



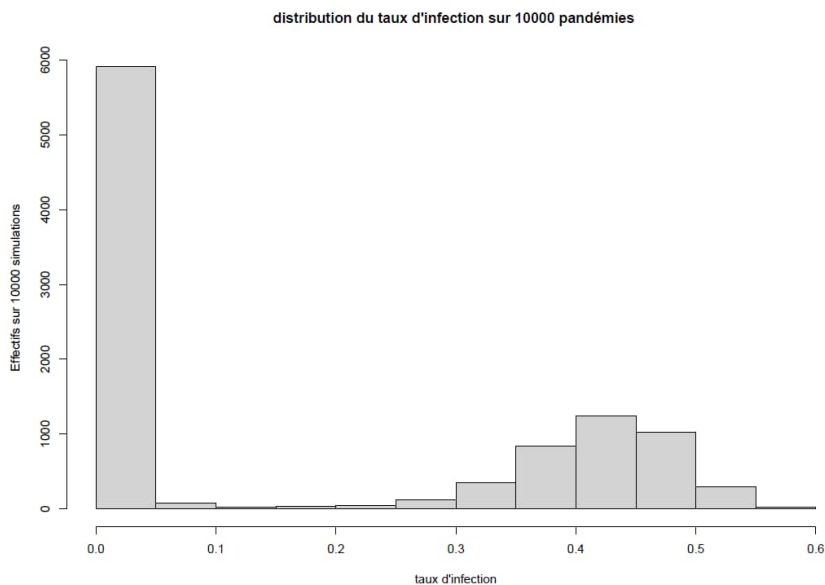
Avec cette nouvelle valeur de ν , nous n'observons plus d'effet polarisant. Cependant la proportion des taux d'infection nuls est largement supérieure au reste. Par ailleurs, il existe des taux d'infection très élevés sortant du cadre qu'on a fixé. Pour remédier à ce problème, nous sélectionnons uniquement les pandémies rentrons dans notre cadre d'étude c'est-à-dire celles qui ont un taux d'infection allant de 20 à 35%. Comme nous rencontrons beaucoup de pandémies avec un taux d'infection entre 35 et 40%, nous élargissons notre spectre de sélection à ces pandémies.

6. Modélisation des pandémies

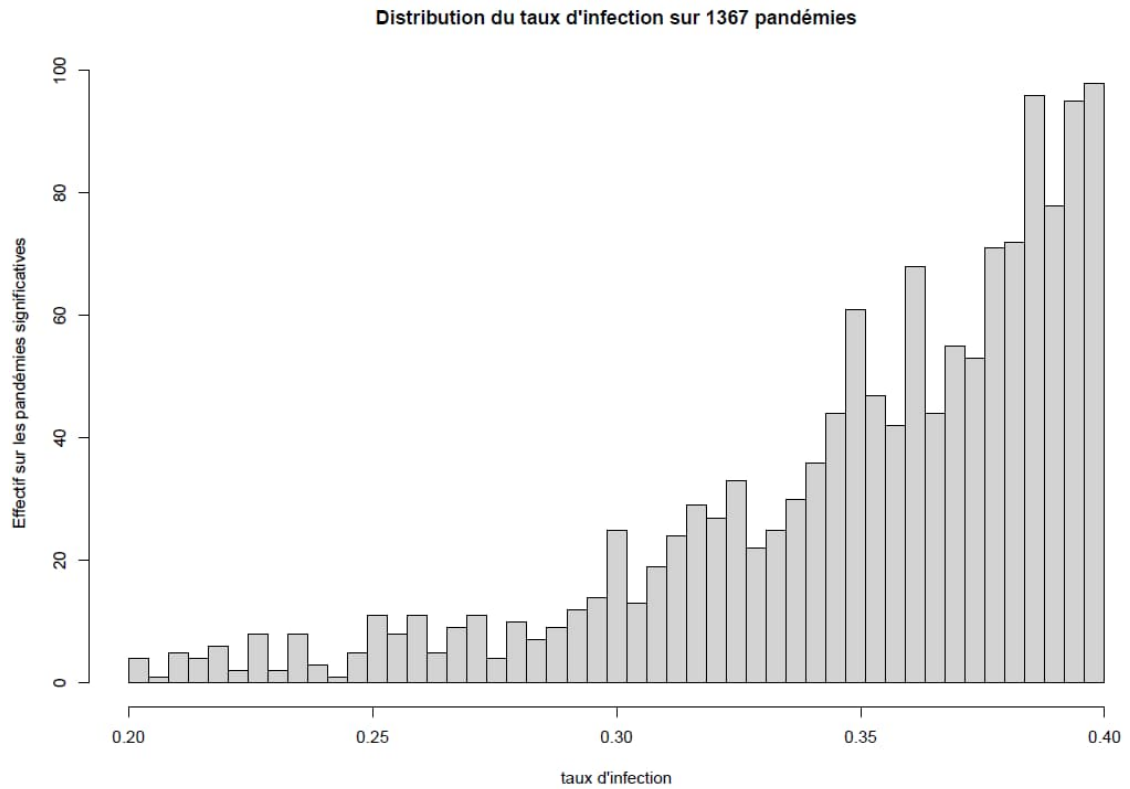
A ce stade, pour une simulation nous avons le taux d'infection et le nombre de jour total d'hospitalisation comme suit :



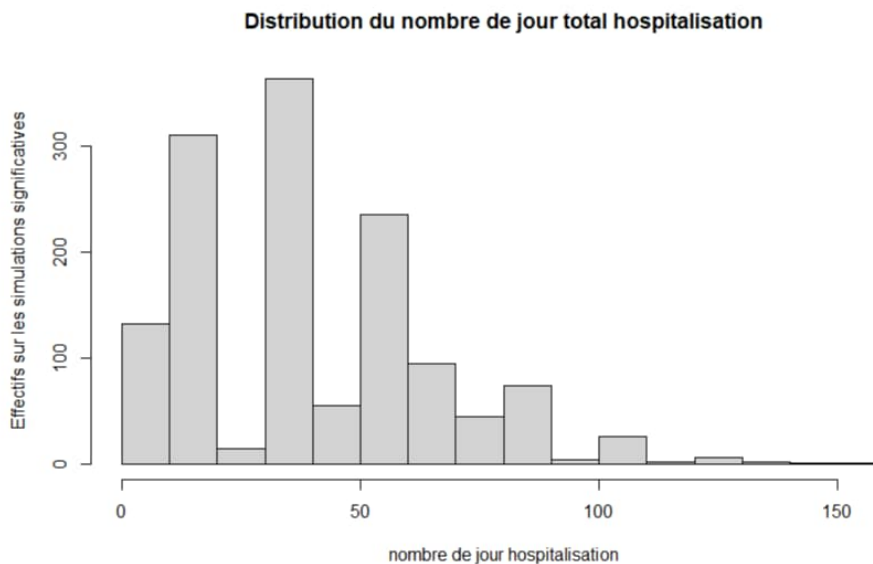
En faisant la simulation de 10000 pandémies et en retenant uniquement celles qui sont significatives sans être dans l'excès, nous obtenons donc en particulier une distribution du nombre de jour total d'hospitalisation. Avec 10000 pandémies, pour les taux d'infection, nous avons :



Bien qu'il y ait 10 fois plus de pandémie, on retrouve des résultats similaires à ceux de la section précédente car les paramètres sont les mêmes. En sélectionnant uniquement les pandémies qui ont un taux d'infection de 20 à 40%, nous obtenons :



Chacune de ces épidémies significatives produisent un nombre de jour total d'hospitalisation. Quand on utilise les coefficients du scénario moyen de l'INVS, la distribution issue des pandémies significatives est la suivante :



Une fois que nous obtenons cette distribution, nous en tirons le quantile bicentenaire.

Pour déterminer le quantile d'une pandémie correspondant à un scénario bicentenaire, on se base sur la fréquence annuelle de pandémie avancée par Swiss Re (2007) qui est de 3,33% soit une pandémie toutes les 30 années.

Soit X la mortalité supplémentaire annuelle dû à une pandémie alors

$$\mathbb{P}(X > x_{99,5\%}) = \mathbb{P}(Y > q_\alpha) \mathbb{P}(I = 1) = 0,5\%$$

Avec Y le taux de choc de mortalité supplémentaire I l'évènement « une pandémie a lieu ».

$$\text{Comme } \mathbb{P}(Y > q_\alpha) = 1 - \alpha$$

Alors, on a :

$$\alpha = 1 - \frac{0,5\%}{3,33\%} = 0,85$$

On obtient donc un quantile bicentenaire de 85% pour une pandémie basée sur une fréquence de 3,33%.

Nous pouvons désormais passer à l'évaluation du besoin en capital.

7. Approche pour l'évaluation du besoin en capital

Dans la section précédente, nous avons abouti au nombre de jour total d'hospitalisation d'une pandémie bicentenaire. Il est donc intuitif de considérer que le capital requis en cas de pandémie bicentenaire, pour l'assureur, serait ce nombre de jour total multiplié par le coût moyen d'une hospitalisation qui lui reste à charge après la sécurité sociale. Cependant, on peut appliquer une autre formule développée par [Guilbert \(2010\)](#) pour l'évaluation du besoin en capital dans le cadre des garanties d'assurance de personne. C'est cette approche que nous allons présenter dans cette section.

Cette approche se base sur la convergence conditionnellement à un facteur de risque systémique des coûts engendrés par une pandémie pour l'assureur qui a asymptotiquement un comportement gaussien.

Soit I un ensemble de contrat, soit X le montant des prestations découlant de ces contrats en cas de pandémie et soit γ un facteur de risque systémique et mutualisable. On suppose que X est uniformément borné alors $X|\gamma$ suit asymptotiquement une loi gaussienne :

$$\frac{X|\gamma - \mathbb{E}[X|\gamma]}{\sigma(X|\gamma)} \xrightarrow{|I| \rightarrow \infty} N(0,1)$$

Le facteur de risque γ est l'élément qui déclenche le risque de morbidité qui touche l'assureur au cours d'une pandémie. Etant donné que les garanties hospitalisations de l'assureur ne sont touchées que si les assurés passent par une étape d'infection, c'est donc l'infection qui est à l'origine des morbidités. Il est impossible de déterminer quels assurés exactement seront contaminés au cours de la pandémie, l'aléa est donc ici caractérisé par la probabilité qu'un assuré passe par l'étape d'infection pendant la pandémie qui n'est rien d'autre que le taux d'infection. Le taux d'infection correspond parfaitement au facteur de risque systémique.

Comme γ est aussi un risque mutualisable car il décrit le risque de morbidité alors $X|\gamma$ est asymptotique gaussienne. Ainsi, on a :

$$F_X(x) = \mathbb{P}(X < x) = \mathbb{E}[\mathbb{1}_{X < x}] = \mathbb{E}[\mathbb{E}[\mathbb{1}_{X < x}]] = \mathbb{E}_\delta[\mathbb{P}(X < x | \gamma = \theta)] \xrightarrow{|\Gamma| \rightarrow \infty} \int \Phi\left(\frac{x - \mu(\theta)}{\sigma(\theta)}\right) F_\gamma(d\theta)$$

Avec, $-\mu(\theta)$ espérance de $X|\gamma = \theta$ et $\sigma(\theta)$ variance de $X|\gamma = \theta$
 - Φ fonction de répartition d'une loi normale centrée réduite.

On peut donc caractériser X en calculant $\mu(\theta)$ et $\sigma(\theta)$

L'expression de la loi asymptotique est approchée par :

$$F_M(x) = \frac{1}{M} \sum_{m=1}^M \Phi\left(\frac{x - \mu(\theta_m)}{\sigma(\theta_m)}\right)$$

Où la suite $(\theta_m)_{1 \leq m \leq M}$ est la suite des réalisations du facteur systémique c'est-à-dire du taux d'infection.

Le SCR correspond au quantile d'ordre 99,5% de F_M . Autrement dit, il s'agit de résoudre numériquement $F_M(x_q) = q$ avec $q = 0,995$.

8. Application de la formule de Guilbert

Nous sommes dans un cas où la garantie portant sur les frais d'hospitalisation a été souscrite pour un ensemble de contrat. Nous disposons du nombre de jour d'hospitalisation total d'une pandémie bicentenaire permettant de caractériser le coût du risque de morbidité. Ce nombre de jour total d'hospitalisation a été déterminé sur une population de 1000 assurés. Il est donc possible de l'appliquer proportionnellement sur une population d'une autre taille.

La démarche s'articule autour de deux étapes :

- Nous déterminons tout d'abord $\mu(\theta_m)$ et $\sigma(\theta_m)$ qui sont les deux premiers moments conditionnels à chaque taux d'infection θ_m issu de la modélisation des pandémies significatives.

Soit N le nombre d'assurés et $n b j_{hosp}$ le nombre de jour d'hospitalisation total issu d'une épidémie bicentenaire.

$$\begin{cases} \mu(\theta_m) = N * \theta_m * n b j_{hosp} * Coût_{hosp} \\ \sigma(\theta_m) = \theta_m * (1 - \theta_m) * N * n b j_{hosp} * (Coût_{hosp})^2 \end{cases}$$

- On fixe le coût journalier moyen d'une hospitalisation à charge de l'assureur à 1 euro puis on résout numériquement l'équation :

$$F_M(x) = 0,995$$

Où M est le nombre de pandémies significatives.

En appliquant la formule du cadre de Guilbert, on obtient :

scénario bas	scénario moyen	scénario critique
13.8	74.45	252.5

Sans surprise, nous observons que le coût des hospitalisations est croissant avec le taux d'hospitalisation appliqué au scénario. Pour rappel, nous avons lié la gravité de l'hospitalisation à la durée d'infection sur la base d'un quantile et ce quantile est naturellement plus grand pour le scénario critique. Etant donné que le coût journalier est de 1 euro, on peut lire ces résultats directement comme étant le nombre de jour total d'hospitalisation généré par la pandémie au sein du portefeuille.

On peut se demander ce qui aurait été obtenu sans utiliser la formule du cadre de Guilbert mais en utilisant directement le quantile bicentenaire sur la distribution obtenue :

scénario bas	scénario moyen	scénario critique
10.31	67.8	229

Ces résultats sont proches de ceux obtenus précédemment. Nous conservons les résultats du quantile bicentenaire.

9. Le choc du SCR

Dans la formule standard, le choc appliqué aux frais d'hospitalisation est de 40%. Nous trouvons ce choc assez conséquent et notre objectif est de calibrer un autre choc par les résultats du modèle.

Pour continuer dans la même logique que précédemment, si on simplifie en considérant un coût journalier hospitalier à charge de l'assureur de 1 euro alors par les recommandations de l'EIOPA, l'exposition de l'assureur aux frais d'hospitalisations en cas de pandémie est :

$$F = \text{nb de jour moyen d'hospitalisation attendu par assuré} \times \text{nb assuré} \times \text{proportion d'assurés attendue en hospitalisation}$$

Le choc de SCR est donc

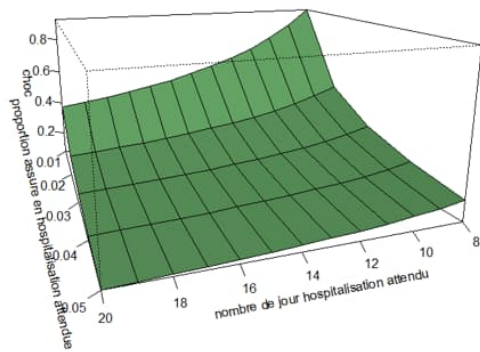
$$\text{Choc} = \frac{\text{montant en sorti du modèle}}{F}$$

Où le *montant en sorti du modèle* correspond directement au nombre de jour total d'hospitalisation bicentenaire, si on ne passe pas par la formule de Guilbert, car on a posé que le coût journalier est de 1 euro.

Le *nb de jour moyen d'hospitalisation attendu par assuré* et la *proportion d'assurés attendue en hospitalisation* étant arbitraires, il serait intéressant de voir comment évolue le choc selon les attentes de l'assureur. Nous obtenons les résultats suivants :

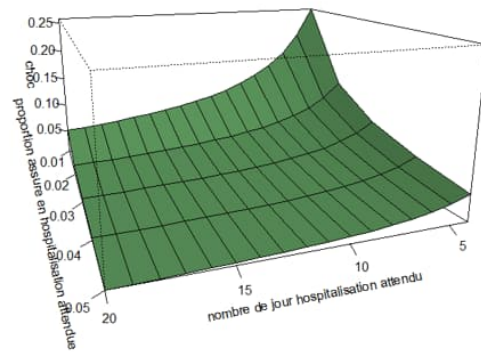
Scénario moyen

Niveau de choc en fonction des variables d'exposition attendues



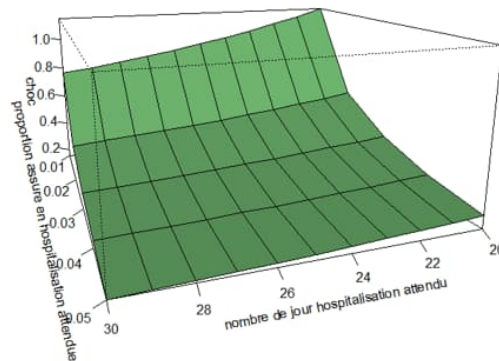
Scénario faible

Niveau de choc en fonction des variables d'exposition attendues



Scénario dur

Niveau de choc en fonction des variables d'exposition attendues



Dans tous les scénarios, plus le nombre de jour moyen d'hospitalisation attendu est grand plus le choc de SCR va inversement décroître, idem pour la proportion d'assurés en hospitalisation. En effet, un principe prudentiel beaucoup trop grand va s'éloigner de ce qui se produit lors d'une pandémie bicentenaire estimée par le modèle. Un assureur qui s'attend à un nombre de jour moyen d'hospitalisation grand (environ 15 jours) est un assureur qui considère que son exposition à la garantie hospitalisation est grande. Néanmoins ce qui ressort de la plupart des scénarios c'est qu'une pandémie ne semble pas être coûteuse en besoin de capital pour l'hospitalisation.

Ainsi, on observe que, dans le scénario moyen, il faudrait que l'assureur s'attende à environ 8 jours d'hospitalisation en moyenne pour avoir un choc en-dessous du choc fixé à 40% alors que ce nombre de jour est déjà relativement bas et ceci en prévoyant 1% des assurés en hospitalisation.

Dans le scénario faible, à partir de 4 journées en moyenne d'hospitalisation, le choc serait déjà calibré à 25%.

Seul le scénario dur semble critique car même en prévoyant 20 journées d'hospitalisations en moyenne, le choc à encaisser pour les fonds propres serait au-dessus de 40% mais cela est tout à fait compréhensible compte tenu des 10% d'assurés infectés finissant en hospitalisation.

PARTIE III : Modélisation du choc SCR Vie CAT

Pour le SCR Vie CAT, nous allons tenter une approche en nous basant sur l'historique des variations de mortalité des épidémies annuelles de grippe. De cet historique, nous allons faire l'évaluation de la surmortalité que pourrait causer un scénario bicentenaire. Cette modélisation de quantile dans un historique de données va nous emmener à utiliser des notions de théories de valeurs extrêmes pour mieux capter le comportement de la queue d'une distribution. Dans un premier temps, nous allons donc introduire des notions de théories de valeurs extrêmes avec les lois usuelles de celle-ci et les méthodes d'estimation de paramètres. Ensuite, il sera question d'utiliser un outil composé de plusieurs lois plus performant que les outils usuels de valeurs extrêmes tout en restant dans ce cadre et en utilisant les différentes connaissances mises en évidence. Enfin, on appliquera cet outil aux données pour en tirer des conclusions.

1. Théorie des valeurs extrêmes

Nous allons maintenant introduire les concepts fondamentaux de la théorie des valeurs extrêmes.

Lorsqu'on s'intéresse à un phénomène aléatoire, on étudie principalement la partie dite centrale de la loi modélisant au mieux le phénomène considéré (espérance, variance, loi des grands nombres, théorème centrale limite, etc). Cependant, l'étude des valeurs extrêmes du phénomène est parfois essentielle lorsqu'il va s'agir de quantifier le risque pour une compagnie d'assurance.

Une augmentation de mortalité causée par une pandémie est un phénomène qui peut conduire à des pertes et requiert un certain niveau de capital chez les assureurs. La théorie des valeurs extrêmes fournit une base mathématique et probabiliste rigoureuse sur laquelle il est possible de construire des modèles statistiques pour prévoir l'ampleur d'une éventuelle surmortalité.

Deux approches principales construisent la théorie des valeurs extrêmes. La première approche va se baser sur le maximum d'une suite de variables aléatoires ayant pour limite une loi extrême se déclinant en trois familles : Weibull, Gumbel, Fréchet. Concernant la seconde approche, il s'agit de fixer un seuil et de modéliser les observations au-delà de ce seuil par une loi de Pareto Généralisée.

1.1 Loi GEV

Si on considère $X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n,n}$ les statistiques d'ordre de n variables aléatoires indépendantes et identiquement distribuées de fonction de répartition $F : F(x) = \mathbb{P}(X_i \leq x)$

Alors $M_n = \max(X_1, \dots, X_n) = X_{n,n}$

Et $M_1 = \min(X_1, \dots, X_n) = X_{1,n}$

On a :

$$\mathbb{P}(M_n \leq x) = F^n(x) \text{ et } \mathbb{P}(M_1 \leq x) = 1 - (1 - F(x))^n$$

En regardant le comportement asymptotique des extrêmes c'est-à-dire quand $n \rightarrow \infty$, la limite de la fonction de répartition du maximum et aussi du minimum est dégénéré puisque le résultat est soit 0 soit 1. En effet,

$$\lim_{n \rightarrow \infty} F^n(x) = \begin{cases} 1 & \text{si } F(x) = 1 \\ 0 & \text{si } F(x) < 1 \end{cases}$$

Par le théorème centrale limite, nous savons que le comportement moyen d'une suite de variable aléatoire $(X_i)_{1 \leq i \leq n}$ iid qui a pour espérance μ et pour variance σ^2 est asymptotiquement gaussien avec

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow[n \rightarrow \infty]{} N(0,1)$$

L'intérêt de la théorie des valeurs extrêmes est d'étudier M_n en calquant le théorème centrale limite avec une transformation linéaire qui permet d'éviter la dégénérescence et une convergence vers une variable aléatoire Z qui a une fonction de répartition H telle que :

$$\frac{M_n - b_n}{a_n} \xrightarrow[n \rightarrow \infty]{} Z$$

De manière plus formelle, si H est une fonction de répartition non dégénérée, une fonction de répartition F appartient au domaine d'attraction de H noté $MDA(H)$ si et seulement s'il existe deux suites $(a_n) > 0$ et (b_n) telle que

$$\lim_{n \rightarrow \infty} F^n(a_n x + b_n) = H(x)$$

Le Théorème de Fisher-Tippet (1928) nous dit que si on a la convergence ci-dessus alors H est la fonction de répartition d'une loi Generalized Extremes Values (GEV).

La théorie des valeurs extrêmes nous apprend qu'il n'existe que 3 formes possibles pour la fonction de répartition H qui sont :

- Loi de Fréchet : $H_{1,a}(x) = e^{-x^{-a}} \mathbb{1}_{x>0}$ ($a > 0$)
- Loi de Weibull : $H_{2,a}(x) = \mathbb{1}_{x \geq 0} + e^{-(-x)^a} \mathbb{1}_{x < 0}$ ($a > 0$)
- Loi de Gumbel : $H_{3,a}(x) = e^{-e^{-x}}$ $\forall x \in \mathbb{R}$

Ces trois types de distribution extrêmes ont été combinées en une seule comportant trois paramètres (Jenkinson-Von Mises, 1955). L'expression empirique de la GEV est :

$$G_{\mu,\sigma,\xi}(x) = \begin{cases} \exp \left[- \left(1 + \xi \frac{x - \mu}{\sigma} \right)^{-\frac{1}{\xi}} \right] & \text{si } \xi \neq 0 \\ \exp \left(-e^{-\frac{x - \mu}{\sigma}} \right) & \text{si } \xi = 0 \end{cases} \quad \text{avec } 1 + \xi \frac{x - \mu}{\sigma} > 0$$

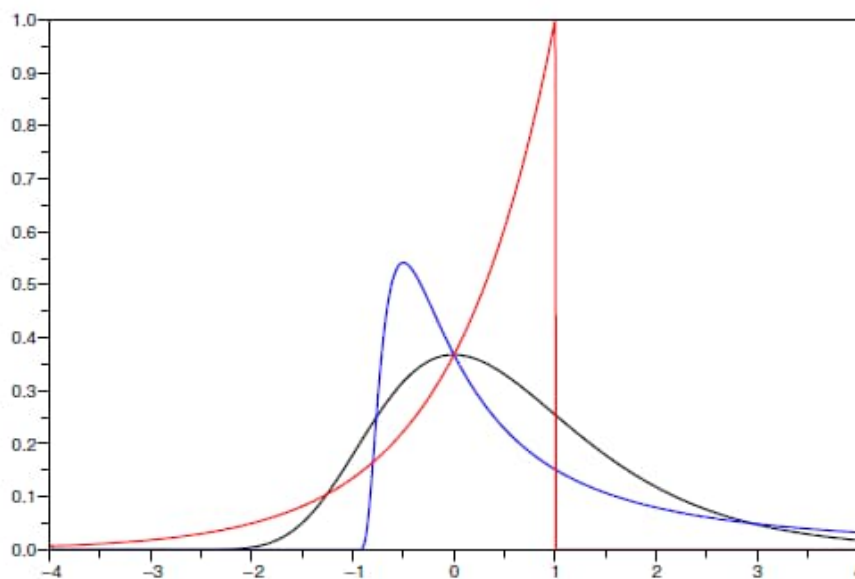
On a le cas $\xi = a^{-1} > 0$ qui correspond à une loi de distribution de Fréchet, le cas $\xi = 0$ qui correspond à une loi de distribution de Gumbel et enfin le cas $\xi = -a^{-1} < 0$ correspondant à une loi de distribution de Weibull.

Le paramètre ξ est celui qui permet de donner une idée sur la queue de la distribution. En effet, par la valeur de ce dernier, nous savons si la queue est épaisse ou non. Il est aussi appelé paramètre de forme et σ est le paramètre d'échelle.

C'est ainsi que nous avons les cas suivants :

- Si $F_X \in MDA(G_{\mu,\sigma,\xi})$ avec $\xi > 0$ alors X suit une loi de Fréchet et a une queue de distribution épaisse.
- Si $F_X \in MDA(G_{\mu,\sigma,\xi})$ avec $\xi = 0$ alors X suit une loi de Gumbel et a une queue de distribution dite fine.
- Si $F_X \in MDA(G_{\mu,\sigma,\xi})$ avec $\xi < 0$ alors X suit une loi de Weibull et est bornée à droite.

Voici ci-dessous un exemple de densités associées à des distributions de valeurs extrêmes :



$\xi = -1$ (Weibull) en rouge, $\xi = 0$ (Gumbel) en noir et $\xi = 1$ (Fréchet) en bleu

1.2 Blocs Maxima

La méthode des Blocs Maxima en théorie des valeurs extrêmes consiste comme son nom l'indique à diviser les observations par blocs temporels de même taille et ne considérer que les maximums de chacun des blocs. Sous l'hypothèse que les observations ont une distribution appartenant à un domaine d'attraction d'une loi extrême, les maximums obtenus convergent vers une loi de distribution extrême.

Autrement dit, soit $(X_i)_{1 \leq i \leq n}$ une suite de variables aléatoires iid ayant F pour fonction de répartition. On définit pour $m = 1, 2, \dots$ et pour $i = 1, 2, \dots, k$ les k blocs maxima :

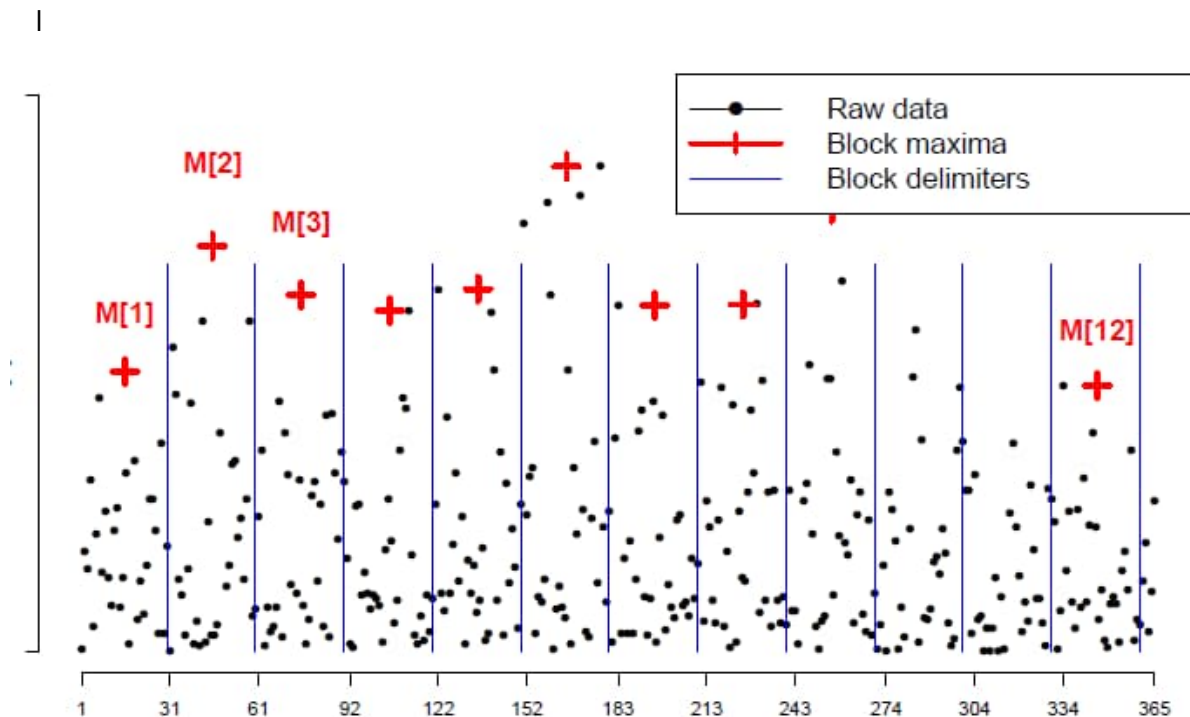
$$\tilde{X}_i = \max_{(i-1)m < j \leq im} X_j$$

Par conséquent, les $n = m \times k$ observations sont divisées en k blocs de taille m . Pour avoir des résultats significatifs, on considère que $k \rightarrow \infty$, $m \rightarrow \infty$ et donc $n \rightarrow \infty$. On a :

$$\frac{\tilde{X}_i - b_n}{a_n} \sim G_{\mu, \sigma, \xi} \text{ avec } \mu = b_n \text{ et } \sigma = a_n$$

L'estimation des paramètres par des méthodes statistiques se fait par la suite en se basant sur ces nouvelles observations.

Le graphique ci-dessous illustre le principe de la méthode :



Méthode Blocs Maxima sur données simulées

Il y a énormément de perte d'information avec cette méthode étant donné qu'on ne prend que des maximums sur une poignée de blocs. Un élément de la queue de distribution peut être omis par cette méthode si ce dernier est proche d'un maximum dans le même bloc. De plus, si on a déjà peu d'historique, ce qui est notamment le cas avec l'historique des pandémies, diviser les données par blocs pour ne prendre que les maximums devient un gros désavantage. Nous n'allons donc pas utiliser cette méthode.

1.3 Loi GPD et POT

Dans cette section, nous allons voir une autre méthode très utilisée dans la théorie des valeurs extrêmes mais qui est similaire à celle des Blocs Maxima. En effet, nous cherchons à modéliser la distribution des extrêmes. Cependant, cette méthode se différencie par :

- La définition des extrêmes : on considère une observation comme extrême dès lors que celle-ci dépasse un certain seuil défini.
- La loi de distribution qui permet de modéliser les extrêmes qui est ici une distribution de Pareto généralisée possédant deux paramètres : σ (un paramètre d'échelle) et ξ (un paramètre de forme/indice de queue).

A noter que dans toute la suite, le seuil à déterminer sera noté u .

1.3.1 GPD

Plusieurs concepts mathématiques permettent d'approximer la distribution des extrêmes par une distribution de Pareto généralisée. Introduisons ces concepts :

Définition :

Soit X une variable aléatoire avec une fonction de répartition F . La fonction de répartition des excès est définie par :

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u), \quad x \geq 0$$

Théorème de Pickands :

Soit F_u la fonction de répartition des excès. Nous avons lorsque u est suffisamment grand :

$$\lim_{u \rightarrow x^+} \sup_{x \in]0; x^+ - u[} |F_u(x) - G_{\xi, \sigma(u)}(x)| = 0$$

Avec $G_{\xi, \sigma(u)}$ la fonction de répartition d'une loi de Pareto Généralisée encore dite GPD, définie comme suit :

$$G_{\xi, \sigma(u)}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\sigma(u)}\right)^{-\frac{1}{\xi}} & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma(u)}\right) & \text{si } \xi = 0 \end{cases}$$

Donc pour u assez grand, on peut trouver $\sigma(u)$ et ξ tels que $F_u(x) \approx G_{\xi, \sigma(u)}$. On peut donc approcher la loi des excès par une loi de Pareto Généralisée et, par la suite, estimer les quantiles extrêmes.

En effet, par le théorème de Fischer et Tippet, nous considérons que pour n grand, on a :

$$\mathbb{P}(\max(X_i) \leq x) = F^n(x) \approx \exp\left(-\left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-\frac{1}{\xi}}\right)$$

Donc,

$$n \log(F(x)) \approx -\left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-\frac{1}{\xi}}$$

Etant donné que ce qui nous intéresse ici est la queue de distribution, nous considérons $x \geq u$ et donc pour ces observations ayant de grandes statistiques d'ordre, nous avons $F(x) \approx 1$, ce qui nous permet d'écrire :

$$\log(F(x)) \approx -(1 - F(x))$$

Ce qui nous amène à :

$$F(x) \approx 1 - \frac{1}{n} \left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-\frac{1}{\xi}}$$

Nous obtenons finalement le résultat suivant :

$$\begin{aligned} F_u(x) &= \mathbb{P}(X - u \leq x | X > u) \\ &= \frac{F(u + x) - F(u)}{1 - F(u)} \\ &= 1 - \left(1 + \xi \frac{x}{\sigma}\right)^{-\frac{1}{\xi}} \end{aligned}$$

1.3.2 POT

Nous rappelons que la loi de Pareto généralisée dépend de trois paramètres :

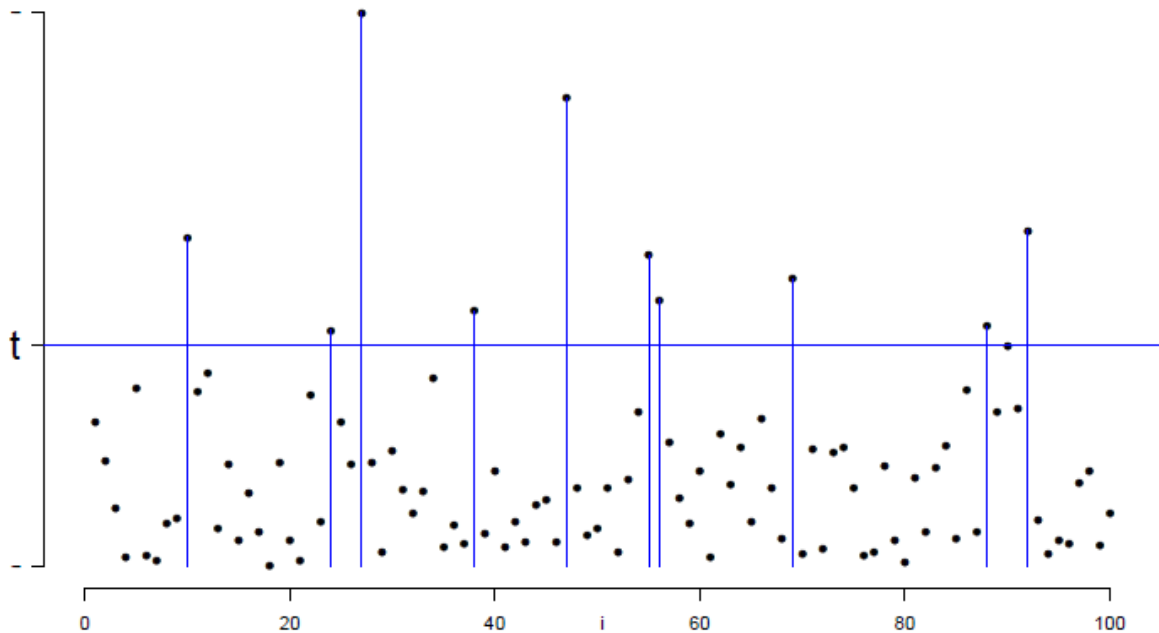
- Le paramètre de queue ξ qui décrit l'appartenance à un domaine d'attraction du maximum et donc l'épaisseur de la queue.
- Le paramètre u qui correspond au seuil.
- Le paramètre d'échelle $\sigma(u)$ qui dépend du seuil u .

L'objectif est de déterminer un seuil u optimal, et par la suite d'estimer les paramètres du modèle. La détermination du seuil u optimal est crucial afin d'avoir un compromis biais/variance raisonnable.

Notons également que le seuil choisi ne doit pas être trop petit afin de ne pas considérer des

observations qui ne seraient pas extrêmes, tout en n'étant pas trop élevé afin d'avoir suffisamment d'observations.

Le graphique ci-dessous illustre le principe de la méthode :



Méthode excès de seuil sur données simulées

2. Méthodes statistiques pour l'estimation de l'indice de queue

Comme nous l'avons vu, les méthodes des Blocs Maxima et Peak Over Threshold donnent lieu à l'utilisation de distributions. D'une part, nous avons la GEV pour la méthode Bloc Maxima et d'autre part la GPD pour le Peak Over Threshold. Ces distributions dépendent de paramètres et nous allons voir comment les estimer par maximum de vraisemblance.

2.1 Rappel sur le maximum de vraisemblance

Le maximum de vraisemblance est une méthode relativement simple qui permet de déterminer l'estimateur d'un paramètre inconnu θ qui peut être soit un réel soit un vecteur. Cette méthode a été introduite par R.A.Fisher un mathématicien anglais en 1912. L'estimation du maximum de vraisemblance peut être utilisée dans plusieurs problèmes statistiques et en particulier dans l'estimation des paramètres des lois GEV et GPD. Elle a un fort attrait intuitif et mène généralement à

un bon estimateur de θ . Pour ces raisons, la méthode du maximum de vraisemblance est largement une des plus utilisées par les méthodes d'estimation statistiques.

On suppose que x_1, \dots, x_n sont les n réalisations indépendantes d'une variable aléatoire X qui admet pour densité $f(x|\theta)$.

$$f(x|\theta) = \begin{cases} f_\theta(x) & \text{si } X \text{ est continue} \\ \mathbb{P}_\theta(X = x) & \text{si } X \text{ est discret} \end{cases}$$

Le symbole du conditionnement sert à représenter le fait que la distribution dépend du paramètre θ . Pour toutes les observations de l'échantillon x_1, \dots, x_n , on définit :

$$f(x_1, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$$

Si X est discret, $f(x_1, \dots, x_n|\theta)$ est la fonction de probabilité jointe et si X est continue, $f(x_1, \dots, x_n|\theta)$ est la fonction de densité jointe. Dans tous les cas, $f(x_1, \dots, x_n|\theta)$ est aussi la fonction de vraisemblance. Comme nous pouvons le voir, la fonction de vraisemblance dépend du paramètre θ et est toujours noté $L(\theta)$.

Si l'échantillon x_1, \dots, x_n est issu d'une distribution discrète et que l'on cherche à estimer θ , on évite de choisir toute valeur de θ pour laquelle il serait impossible d'obtenir l'échantillon observé. En outre, on suppose que la probabilité $f(x_1, \dots, x_n|\theta)$ d'obtenir l'échantillon observé x_1, \dots, x_n est assez haute quand θ a une valeur particulière, disons par exemple $\theta = \theta_0$, et cette probabilité est petite pour toute valeur différente de ce θ_0 . A partir de là, on estime naturellement la valeur de θ par θ_0 . Quand l'échantillon est issu d'une distribution continue, il est là aussi naturel de chercher une valeur de θ pour laquelle la densité de probabilité $f(x_1, \dots, x_n|\theta)$ est grande et d'utiliser cette valeur comme estimateur de θ . Ainsi, pour tout échantillon x_1, \dots, x_n , on est emmené par ce raisonnement à considérer la valeur de θ pour laquelle la fonction de vraisemblance $L(\theta)$ atteint son maximum et utiliser cette valeur comme l'estimateur de θ .

Théoriquement, si on ne possède pas de données observées, maximiser la fonction de vraisemblance conduirait à une fonction de n variables aléatoires X_1, \dots, X_n . Quand on a un échantillon de données observées, l'estimation prend une forme numérique qui est l'estimateur du maximum de vraisemblance.

L'estimation du maximum de vraisemblance demande de maximiser la fonction de vraisemblance $L(\theta)$ selon le paramètre inconnu θ . Nous avons vu que $L(\theta)$ est définie comme un produit de n termes et elle n'est donc pas simple à maximiser. Maximiser $L(\theta)$ est équivalent à maximiser $\log L(\theta)$ car \log est une fonction croissante. On définit $\log L(\theta)$ comme la fonction de log vraisemblance que l'on note $l(\theta)$ et on a :

$$l(\theta) = \log L(\theta) = \log \prod_{i=1}^n f(x_i|\theta) = \sum_{i=1}^n \log f(x_i|\theta)$$

Maximiser $l(\theta)$ selon θ donne l'estimateur du maximum de vraisemblance. Pour que θ_0 soit une valeur pour laquelle $l(\theta)$ est maximum, il faut que celle-ci remplisse deux conditions qui sont :

$$\begin{cases} \frac{\partial l(\theta_0)}{\partial \theta_0} = 0 \\ \frac{\partial^2 l(\theta_0)}{\partial \theta_0^2} < 0 \end{cases}$$

Si θ est un vecteur de plusieurs paramètres, la condition ci-dessus se généralise car alors, il faudrait que chaque dérivée partielle selon les éléments constituant le vecteur θ soit nulle et que la matrice hessienne soit définie négative.

2.2 Maximum de vraisemblance pour la loi GPD

L'approximation de la distribution GPD se fait à partir des observations qui se trouvent au-delà d'un certain seuil u . Si nous avons les observations x_1, \dots, x_n et que le seuil u vérifie $x_{(m-1)} < u \leq x_{(m)}$ avec $m \ll n$ et où $x_{(i)}$ correspond à la i ème statistique d'ordre alors les observations utilisées dans le calcul de la log vraisemblance sont les $m - n + 1$ observations suivantes : $x_{(m)}, x_{(m+1)}, \dots, x_{(n)}$. Il s'agit des excès et nous avons vu que ces derniers suivent une loi GPD. La GPD étant continue, nous devons déterminer sa densité pour obtenir sa log vraisemblance.

Nous rappelons que la fonction de répartition de loi GPD est la suivante :

$$G_{\xi, \sigma(u)}(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\sigma}\right)^{-\frac{1}{\xi}} & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma}\right) & \text{si } \xi = 0 \end{cases}$$

A noter que dans la GPD, $\sigma = \sigma(u)$ et donc ce paramètre dépend du seuil u mais pour simplifier, nous le notons juste σ .

En dérivant, nous obtenons la densité :

$$f(x|\xi, \sigma) = \begin{cases} \frac{1}{\sigma} \left(1 + \xi \frac{x}{\sigma}\right)^{-\frac{1}{\xi}-1} & \text{si } \xi \neq 0 \\ \frac{1}{\sigma} \exp\left(-\frac{x}{\sigma}\right) & \text{si } \xi = 0 \end{cases}$$

La log vraisemblance de la GPD est donc :

$$l(\xi, \sigma) = \begin{cases} \sum_{i=m}^n \log \left[\frac{1}{\sigma} \left(1 + \xi \frac{x_{(i)}}{\sigma}\right)^{-\frac{1}{\xi}-1} \right] & \text{si } \xi \neq 0 \\ \sum_{i=m}^n \log \left[\frac{1}{\sigma} \exp\left(-\frac{x_{(i)}}{\sigma}\right) \right] & \text{si } \xi = 0 \end{cases}$$

$$= \begin{cases} -n \log(\sigma) - (1 + \xi^{-1}) \sum_{i=m}^n \log\left(1 + \xi \frac{x^{(i)}}{\sigma}\right) & \text{si } \xi \neq 0 \\ -n \log(\sigma) - \frac{1}{\sigma} \sum_{i=m}^n x^{(i)} & \text{si } \xi = 0 \end{cases}$$

3. Estimateurs spécifiques de la queue dans la GPD

Dans l'approximation de la loi GPD notamment de son paramètre de queue, nous avons vu que les statistiques d'ordre jouent un rôle primordial car nous considérons uniquement les observations au-dessus d'un certain seuil. A partir de ce constat, plusieurs estimateurs spécifiques du paramètre de queue de la GPD ont été mis en évidence. Les plus utilisés sont : l'estimateur de Hill uniquement pour $\xi > 0$ (domaine d'attraction de Fréchet) datant de 1975, l'estimateur de Pickands datant aussi de 1975 et enfin l'estimateur de Dekkers, Einmalh et De Haan datant de 1990. Si nous avons n observations de X au total et m observations au-dessus du seuil, ces estimateurs sont définis comme suit :

$$\begin{aligned} \xi_{n,m}^{Hill} &= \frac{1}{m} \sum_{i=0}^{m-1} \log\left(\frac{X_{(n-i)}}{X_{(n-m)}}\right) \\ \xi_{n,m}^{Pickands} &= \frac{1}{\log 2} \log \frac{X_{(n-m)} - X_{(n-2m)}}{X_{(n-2m)} - X_{(n-4m)}} \\ \xi_{n,m}^{DEDE} &= \xi_{n,m}^{H(1)} + 1 - \frac{1}{2} \left[1 - \frac{(\xi_{n,m}^{H(1)})^2}{\xi_{n,m}^{H(2)}} \right]^{-1} \end{aligned}$$

Où $\xi_{n,m}^{H(r)} = \frac{1}{m} \sum_{i=1}^{m-1} (\log X_{(n-i)} - X_{(n-m)})^r$, $r = 1, 2, \dots$

Ces estimateurs ont aussi l'avantage d'être convergents quand $n \rightarrow \infty$ et $m \rightarrow \infty$ avec $\frac{m}{n} \rightarrow 0$ (n doit être largement plus grand que m). Nous avons la normalité asymptotique suivante :

$$\sqrt{m}(\xi_{n,m}^{Hill} - \xi) \xrightarrow{loi} N(0, \xi^2) \quad \text{pour } \xi > 0$$

$$\sqrt{m}(\xi_{n,m}^{Pickands} - \xi) \xrightarrow{loi} N\left(0, \frac{\xi^2(2^{\xi+1} + 1)}{(2(2^\xi - 1)\log 2)^2}\right)$$

$$\sqrt{m}(\xi_{n,m}^{DEDE} - \xi) \xrightarrow{loi} N(0, 1 + \xi^2) \quad \text{pour } \xi \geq 0$$

On remarque que l'estimateur de Hill est celui qui présente la plus petite variance d'où le fait qu'il soit le plus connu et le plus utilisé des trois.

Le paramètre m semble important dans chacun de ces estimateurs or il correspond au nombre d'excès au-dessus du seuil. Le choix du seuil doit donc être regardé de plus près.

4. Méthodes pour le choix du seuil

La distribution de Pareto Généralisée nous oblige à prendre en compte un seuil u dans l'établissement de sa loi. Ce seuil sert à caractériser les valeurs qui l'excèdent et sont dès lors considérées comme des valeurs extrêmes. En particulier, pour u assez grand, ces excédents peuvent être modélisés par une des trois lois correspondant à la famille des lois GPD c'est-à-dire soit Weibull, soit Gumbel, soit Fréchet. Dans ce contexte, le choix du seuil adéquat doit se faire avant l'estimation des paramètres et cette sélection implique une balance entre biais et variance. En effet, un seuil beaucoup trop élevé va aboutir à un modèle caractérisé par une très grande variance étant donné que très peu d'observations seront utilisées pour l'obtention des estimateurs tandis qu'un seuil trop bas va augmenter le biais car alors on se retrouverait dans une situation où le nombre d'excès serait équivalent au nombre d'observations total.

Plusieurs méthodes permettent de déterminer le choix du seuil et de voir si ce choix est adéquat. Ce sont toutes des méthodes qui reposent sur une analyse graphique.

4.1 Mean excess plot

La première méthode est basée sur un graphique appelé la mean excess plot. Ce graphique est un outil qui permet de distinguer si une queue est lourde ou pas. Elle va consister à tracer la mean excess function. Cette mean excess function se définit comme suit :

$$e(u) = \mathbb{E}[X - u | X > u]$$

Si X est une variable aléatoire qui suit une loi de Pareto Généralisée de paramètre σ et ξ alors on a :

$$e(u) = \frac{\sigma + u\xi}{1 - \xi} \mathbb{1}_{\{\sigma + u\xi > 0\}}$$

Cette fonction a la particularité d'être linéaire en fonction de u et de coefficient directeur $\xi/(1 - \xi)$. En combinant cette particularité avec le théorème de Pickands on peut déterminer le seuil à partir duquel les excès vont suivre une loi de Pareto Généralisée dans le but par la suite de déterminer les paramètres de cette loi, en particulier le paramètre de queue ξ .

En pratique, nous procédons comme suit :

(a) On suppose que la GPD est une loi valide pour modéliser les excès au-delà d'un seuil u . On considère la mean excess function $e(v) = \mathbb{E}[X - v | X > v]$ pour $v > u$ dont le graphique est donné par

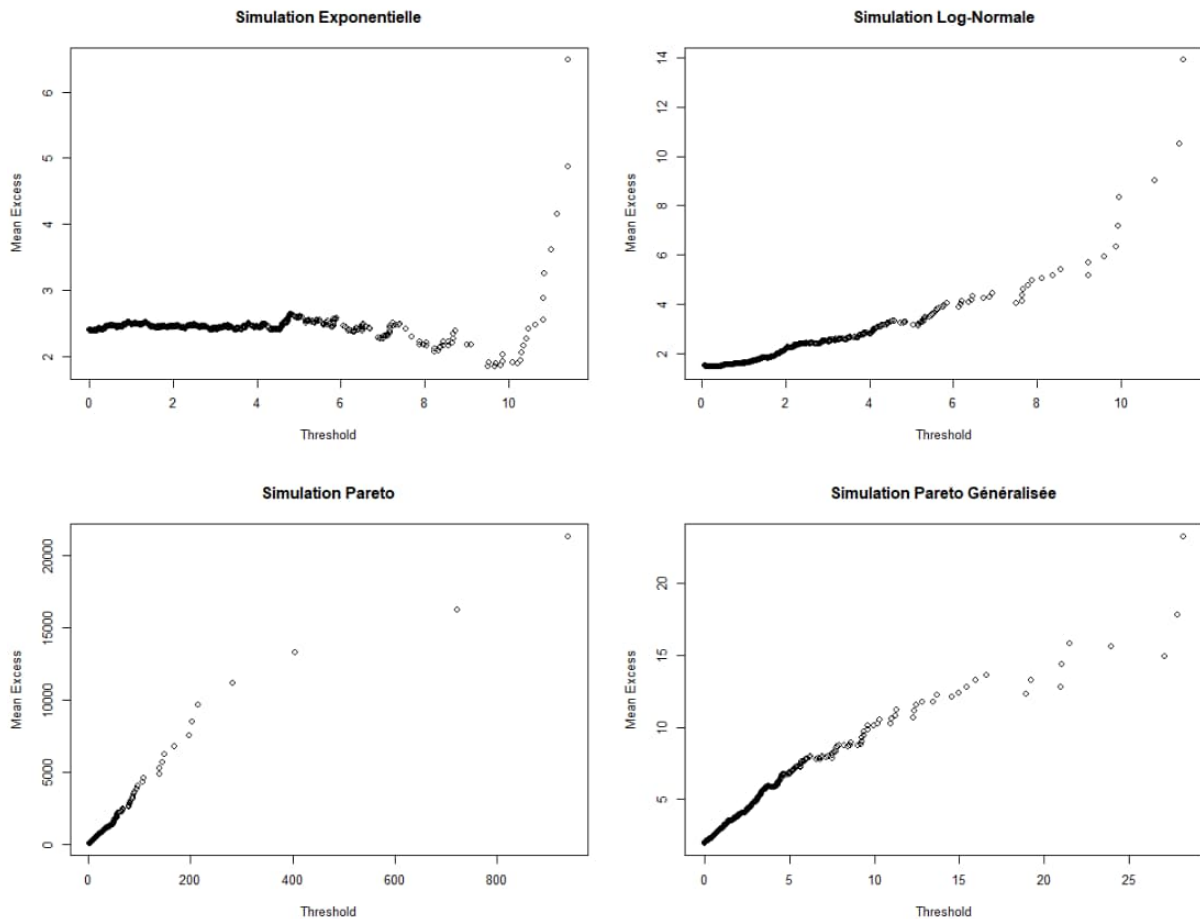
$$\left(v, \frac{1}{n_v} \sum_{i=1}^{n_v} (x_{(i)} - v) : v < x_{(n)} \right)$$

Où les $x_{(i)}$ correspondent aux n_v observations qui excèdent v .

(b) Au-dessus du seuil u pour lequel la GPD modélise assez bien la distribution des extrêmes, le graphique devrait avoir un comportement linéaire suivant v .

(c) Ensuite, par la méthode du maximum de vraisemblance ou celle des moments, nous déterminons les paramètres de la GPD et donc en particulier le paramètre de queue ξ qui va nous donner une idée sur l'épaisseur de la queue.

Voici ci-dessous des illustrations de mean excess plot pour différentes lois simulées :



Mean excess function de lois usuelles

Dans le premier graphique en haut à gauche correspondant au mean excess plot d'une variable aléatoire exponentielle de paramètre 2.5, on observe un comportement linéaire horizontale dans les premières abscisses. Ceci est totalement attendu car en calculant la mean excess fonction d'une loi exponentielle de paramètre γ , on trouve $e(u) = \gamma$. Le graphique empirique résultant des observations simulées doit donc avoir cette forme et avoir pour ordonnée γ .

Les graphiques sont aussi concluants pour la loi de Pareto et la loi de Pareto Généralisée. En effet, la mean excess fonction théorique de ces deux lois est une fonction affine en u et nous retrouvons un comportement affine là où il y a une concentration de points.

Par ces illustrations, on se rend compte de la pertinence de cet outil graphique. Il suffit de voir visuellement à partir de quelle abscisse, on observe une tendance linéaire. Cette abscisse servira de seuil u . C'est exactement la même façon de faire pour des données observées. Cependant, il n'y a

pas d'interprétation possible pour les plus grandes observations car par manque de données à ce niveau, il n'y a pas de tendance apparente.

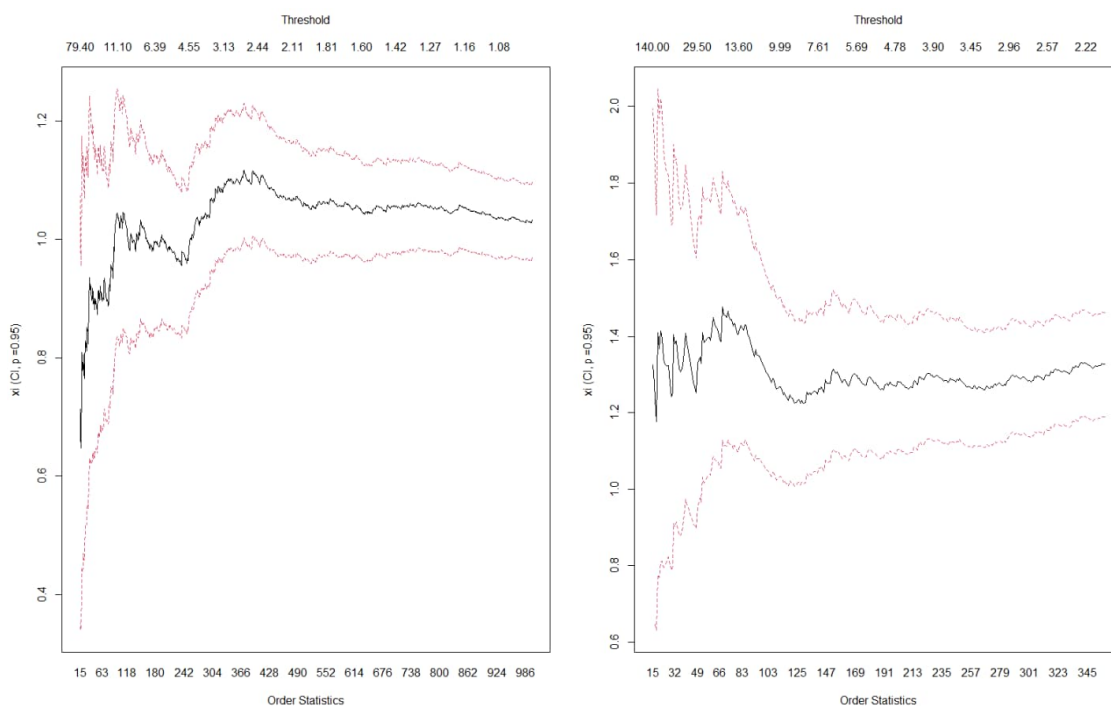
4.2 Hill plot

La méthode du Hill plot est une méthode graphique qui utilise le fait que la distribution de Pareto Généralisée est une distribution stable. Plus formellement, cette stabilité se caractérise de la manière suivante : si $X - u | X > u \sim GPD(\xi, \sigma)$ alors $X - u' | X > u' \sim GPD(\xi, \sigma)$. Cela traduit notamment le fait que pour tout $u > 0$, le paramètre de queue ξ va rester constant. La méthode va donc consister à tracer la courbe de ξ en fonction du seuil u ou encore en fonction du nombre d'excès car le nombre d'excès et le seuil u sont directement liés. Soit u le seuil qui fait état de m observations en excès parmi n observations au total alors l'estimateur de Hill qui approche ξ est donné par :

$$\xi_m^{Hill} = \frac{1}{m} \sum_{i=0}^{m-1} \log \left(\frac{X_{(n-i)}}{X_{(n-m)}} \right)$$

Il est clair que si m est trop faible, conséquence d'un seuil u très élevé, il va y avoir peu d'informations pour l'évaluation de l'estimateur de Hill ce qui va donner lieu à un intervalle de confiance large dû à la volatilité. L'objectif de l'analyse graphique est donc de trouver un seuil laissant place à un nombre suffisant d'excès qui réduit la volatilité et donc stabilise l'estimateur de Hill.

Pour illustrer cette méthode graphique, nous avons simulé d'une part un échantillon d'une loi de Pareto de paramètre $\alpha = 1$ et donc de paramètre de queue $\xi = 1$ et d'autre part un échantillon d'une loi de Pareto Généralisée de paramètre de queue $\xi = 1,2$. Les résultats sont ci-dessous :



Hill Plot d'une loi de Pareto(gauche) et d'une loi de Pareto Généralisée

Pour les deux graphiques, nous observons que les valeurs simulées les plus élevées et donc avec des seuils très hauts sont sujettes à un intervalle de confiance large pour le paramètre de queue ξ car il y a très peu d'excès pour ces seuils. Par ailleurs, dans les deux graphiques, on visualise assez bien une stabilisation de la courbe au fur et à mesure que le seuil baisse et cette stabilisation est bien en accord avec ce qui a été simulé. En effet, pour la loi de Pareto, la courbe se stabilise vers l'ordonnée $\xi = 1$ et pour la loi de Pareto Généralisée, la stabilisation a lieu vers l'ordonnée $\xi = 1,2$.

5. G-E-GPD

Un des principaux objectifs visés par la théorie des valeurs extrêmes est l'estimation de la queue de distribution très liée avec la détermination d'un seuil à partir duquel une loi de Pareto Généralisée peut être modélisée. Les différentes approches pour l'estimation peuvent être réparties en deux classes. D'une part, une classe de méthode qui utilise la méthode du Peak Over Threshold dans laquelle le seuil qui permet d'estimer le paramètre de queue est choisi graphiquement, c'est notamment ce dont nous avons parlé dans les sections précédentes. D'autre part, il y a une autre classe de méthode qui utilise un calibrage des paramètres où le seuil est déterminé par des algorithmes. Nous allons donc nous concentrer désormais sur cette seconde classe en proposant une distribution hybride pour les queues lourdes qui lie une loi gaussienne (ou log-normale) à une loi de Pareto Généralisée par une loi exponentielle servant de pont entre le comportement moyen et le comportement dans la queue de distribution. Un algorithme est développé pour estimer les paramètres du modèle. Nous évaluerons la pertinence de cet algorithme par rapport aux méthodes standards vues précédemment avec des données simulées. Et enfin, il sera appliqué sur les données de variation de mortalité dû aux épidémies de grippe.

5.1 Intérêt et fonctionnement d'un modèle hybride

La modélisation de données hétérogènes à plusieurs composantes est une problématique qui se retrouve dans beaucoup de domaines d'études. Généralement, ce n'est pas évident de trouver un modèle qui décrit assez bien les données observées. C'est pourquoi bien souvent, on a recours à des approches non paramétriques comme le noyau de densité (Parzen, 1962). Cependant, quand plusieurs composantes sont séparables, les modèles paramétriques deviennent malléables. De nombreux modèles hybrides ont été proposés dans ce contexte en combinant deux ou plus de densités (Carreau, 2009). Dans ce mémoire, nous allons nous parler d'un cas spécifique qui est celui où les données présentent une queue de distribution lourde.

Bien que la théorie des valeurs extrêmes s'intéresse à l'étude et la modélisation des observations extrêmes en utilisant uniquement l'information détenue dans la queue de distribution (ce qui est la force de cette théorie même si parfois c'est une faiblesse en pratique tant les données peuvent être rares), elle est aussi très utile pour combiner la queue de distribution avec le reste de l'information contenue notamment autour de la moyenne. Pour extraire les informations importantes contenues dans les extrêmes et mettre en évidence celles qui sont dans le reste de la distribution, il est important de prendre en compte la dissymétrie entre le poids que représentent les données au-delà d'un seuil haut (la queue) et en dessous de celui-ci (environ la moyenne). Différentes méthodes ont été présentées dans Nadarajah, 2014.

Notre objectif ici est de développer une méthode de calibration pour modéliser les données à queue lourde. Le choix de cette classe de méthode est clairement motivé par le fait que ça facilite les implémentations pratiques. En effet, le principal problème auquel on fait face lorsqu'on utilise les méthodes usuelles de la théorie des valeurs extrêmes, c'est qu'on se base sur des méthodes graphiques comme le Peak Over Threshold.

Cette méthode de calibrage alternative a deux gros avantages :

- Quand on regarde toutes les données et qu'on se rend compte qu'elles sont hétérogènes, à plusieurs composantes et avec une queue de distribution lourde, il est possible de faire une modélisation sur l'ensemble des observations.
- Quand on ne s'intéresse qu'à la queue de distribution, on peut obtenir par un algorithme le seuil à partir duquel la modélisation s'effectue.

De manière plus précise, on introduit un modèle hybride avec plusieurs composantes dont notamment une loi de Pareto Généralisée pour prendre en compte l'épaisseur de la queue de distribution ressortant des données observées. Sans perte de généralité, nous supposons une queue de distribution continue et asymétrique. Puisque ce qui nous intéresse est de modéliser l'ensemble de la distribution, l'idée est de considérer à la fois le comportement moyen et le comportement à la queue pour utiliser les théorèmes limite de chacun d'eux (comme c'est le cas dans Kratz, 2014) afin de construire un modèle aussi général que possible. Ainsi, nous introduisons une distribution Gaussienne pour le comportement moyen, justifiée par le théorème central limite et une distribution de Pareto Généralisée à partir d'un certain seuil pour l'épaisseur de la queue justifiée par le théorème de Pickands. Pour lier la moyenne et la queue de distribution, on utilise une distribution exponentielle. Un poids différent est assigné à chacune des composantes dans le but d'avoir une meilleure manipulation des extrêmes. Le modèle hybride qui résulte de tous ces éléments est un modèle à trois composantes et est appelé modèle G-E-GPD (G pour Gaussienne, E pour Exponentielle et GPD pour Generalized Pareto Distribution). Concernant le nombre de composantes, nous signalons que le modèle a besoin d'au moins deux composantes, incluant nécessairement la distribution de Pareto Généralisée pour être réalisable. En effet, le seuil au-dessus duquel la Pareto Généralisée est modélisée est déterminé dans un algorithme comme un point de jonction entre la Pareto Généralisée et la loi exponentielle. Contrairement aux méthodes usuelles de la théorie des valeurs extrêmes, cela signifie qu'on a besoin d'information se trouvant au-dessous du seuil pour bénéficier pleinement de cette méthode de calibrage.

De plus, la distribution intermédiaire (ici la loi exponentielle) est utilisée comme une « force » pour donner tout son sens au seuil de queue au dernier point de jonction entre la Pareto Généralisée et la distribution intermédiaire. La distance entre deux points de jonction consécutifs tendra automatiquement vers 0.

Un algorithme itératif est développé pour estimer les paramètres du modèle hybride à trois composantes. Il commence par imposer la continuité et la régularité des trois composantes aux points de jonctions, puis il procède de manière itérative pour déterminer les seuils successifs et les paramètres des distributions impliquées. Il offre aussi une judicieuse pondération pour les trois composantes. Cet algorithme est basé, pour chaque itération, sur la résolution numérique d'un problème d'optimisation dans le sens des moindres carrés, en utilisant la méthode de Levenberg Marquardt (Levenberg, 1944 et Marquardt, 1963).

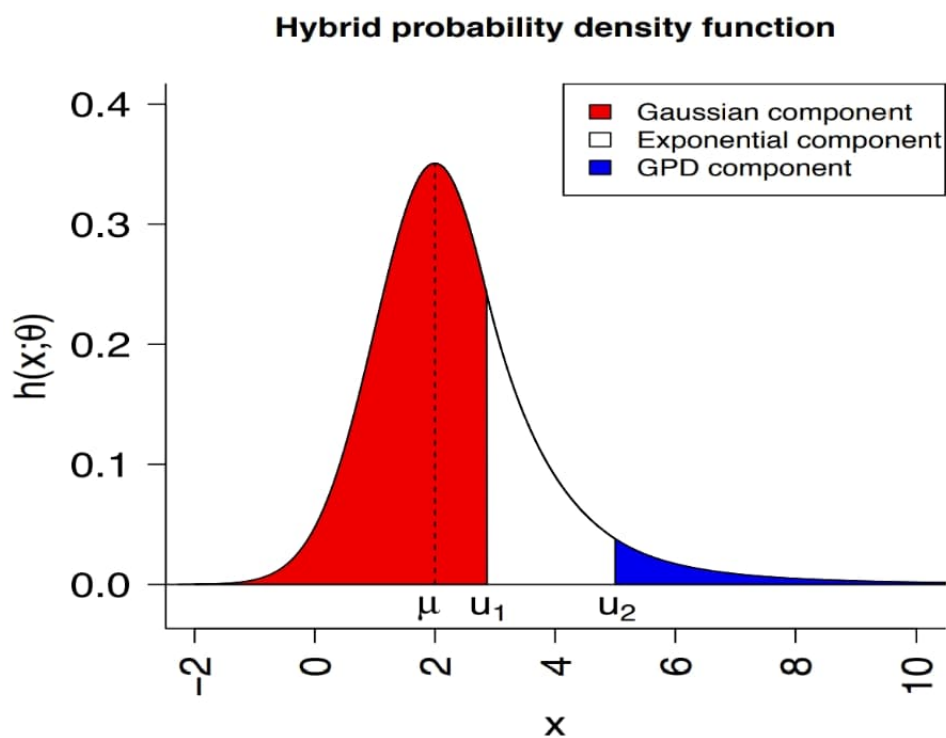
La performance de la méthode de calibrage est étudiée en terme de qualité de modélisation sur des données simulées d'une G-E-GPD. Nous verrons par la suite que ce modèle est plutôt performant et fort de ce constat, on l'appliquera sur les données de variation de mortalité annuelle dû à une épidémie.

5.2 Formalisation du modèle hybride à trois composantes

On considère un modèle pièce par pièce où chaque composante représente un comportement différent au sein des données qui peuvent être hétérogènes ou non. On suppose que les données admettent une distribution continue et donc, nous pouvons introduire une densité hybride de probabilité avec quelques contraintes de régularité. Sans perte de généralité, on considère un modèle à trois composantes. Plus précisément, le modèle hybride proposé lie trois distributions différentes les unes aux autres par deux points de jonction notés u_1 et u_2 : une distribution Gaussienne pour modéliser le comportement moyen, une distribution de Pareto Généralisée pour la queue de distribution et une distribution exponentielle pour relier la moyenne et la queue. Ce modèle noté G-E-GPD est caractérisé par sa densité h qui s'exprime explicitement comme suit :

$$h(x; \theta) = \begin{cases} \gamma_1 f(x; \mu; \theta) & \text{si } x \leq u_1 \\ \gamma_2 e(x; \lambda) & \text{si } u_1 \leq x \leq u_2 \\ \gamma_3 g(x - u_2; \xi; \beta) & \text{si } u_2 \leq x \end{cases}$$

Pour illustrer cette densité, voici sa courbe ci-dessous :



Densité d'une distribution G-E-GPD

Le paramètre θ est un vecteur paramètre qui contient l'ensemble des 10 paramètres du modèle. Pour commencer, les γ_i , $i = 1,2,3$ correspondent aux poids associés à chaque composante. Les paramètres $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}_+^* = \mathbb{R} \setminus \{0\}$ représentent, respectivement, la moyenne et l'écart-type de la distribution Gaussienne ayant pour densité :

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \forall x \in \mathbb{R}$$

Nous rappelons que la densité de loi GPD est la suivante :

$$g(x; \xi, \beta) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-\frac{1}{\xi}} & \text{si } \xi \neq 0 \\ 1 - \exp\left(-\frac{x}{\beta}\right) & \text{si } \xi = 0 \end{cases}, \forall x \in D(\xi, \beta)$$

$$\text{Où } D(\xi, \beta) = \begin{cases} [0, \infty) & \text{si } \xi \geq 0 \\ \left[0, -\frac{\beta}{\xi}\right] & \text{si } \xi < 0 \end{cases}$$

Notons qu'auparavant, nous notions le paramètre de forme de la Pareto Généralisée par σ mais pour éviter la confusion avec l'écart-type de la gaussienne dans ce contexte de modèle hybride, le paramètre de d'échelle de la Pareto Généralisée est désormais noté β .

Enfin, nous avons $\lambda \in \mathbb{R}_+^*$ qui est le paramètre de loi exponentielle de densité :

$$e(x; \lambda) = \lambda e^{-\lambda x}, \quad \forall x \geq 0$$

Dans la suite, nous considérons que le passage d'un comportement à un autre est régulier. Et donc en conséquence, on contraint la densité hybride h à être C^1 . Notons qu'en combinant cette contrainte et l'hypothèse de données ayant une queue lourde, le nombre de paramètres et par conséquent la taille de θ se réduit.

5.3 Les hypothèses du modèle

Les deux premières hypothèses font partie de la construction du modèle G-E-GPD.

1. Tout d'abord, on suppose, par construction, que les données admettent une densité h . Cela signifie que h est positive et vérifie $\int h(x; \theta) dx = 1$ c'est-à-dire :

$$\gamma_1 F(u_1; \mu, \theta) + \gamma_2 (e^{-\lambda u_1} - e^{-\lambda u_2}) + \gamma_3 = 1,$$

Où F est la fonction de répartition de la distribution Gaussienne.

2. Comme les queues lourdes font partie intégrante du modèle, cela implique que h appartient au domaine d'attraction de Fréchet ($\xi > 0$) et donc $\beta = \xi u_2$ (Embrechts, 1997)

La principale contrainte est la régularité de la densité et donc la réduction de paramètres, c'est pourquoi nous avons imposé une densité h de classe C^1 .

3. h est continue et régulière avec deux points de jonction u_1 et u_2 .

Les hypothèses 1 et 3 conduisent à six équations liant tous les paramètres du modèle :

$$\begin{cases} \beta = \xi u_2 \\ \lambda = \frac{1 + \xi}{\beta} \\ u_1 = \mu + \lambda \sigma^2 \end{cases} \quad \begin{cases} \gamma_1 = \gamma_2 \frac{e(u_1; \lambda)}{f(u_1; \mu; \sigma)} \\ \gamma_2 = \left[(\lambda \beta - 1) e^{-\lambda u_2} + \left(1 + \lambda \frac{F(u_1; \mu; \sigma)}{f(u_1; \mu; \sigma)} \right) e^{-\lambda u_1} \right]^{-1} \\ \gamma_3 = \beta \gamma_2 e(u_2; \lambda) \end{cases}$$

Par conséquent, le vecteur paramètre θ contient moins de paramètres et on a $\theta = [\mu, \sigma, u_2, \xi]$. La fonction de répartition de la G-E-GPD est :

$$H(x; \theta) = \begin{cases} \gamma_1 F(x; \mu; \sigma) & \text{si } x \leq u_1 \\ \gamma_1 F(x; \mu; \sigma) + \gamma_2 (e^{-\lambda u_1} - e^{-\lambda x}) & \text{si } u_1 \leq x \leq u_2 \\ 1 - \gamma_3 \left(1 + \frac{\xi}{\beta} (x - u_2) \right)^{-\frac{1}{\xi}} & \text{si } x \geq u_2 \end{cases}$$

La fonction de répartition inverse est donc

$$H^{-1}(x; \theta) = \begin{cases} F^{-1}\left(\frac{x}{\gamma_1}; \mu; \sigma\right) & \text{si } x \leq p_1 = \gamma_1 F(x; \mu; \sigma) \\ \lambda^{-1} \log\left(\frac{\gamma_2}{p_1 - x + \gamma_2 e^{-\lambda u_1}}\right) & \text{si } p_1 \leq x \leq p_2 = 1 - \gamma_3 \\ \frac{\beta}{\xi} \left(\left(1 - \frac{x - p_2}{\gamma_3} \right)^{-\xi} - 1 \right) + u_2 & \text{si } p \geq p_2 \end{cases}$$

Un problème classique mis en évidence quand on utilise des modèles paramétriques concerne la manière à laquelle on va estimer les paramètres. Pour répondre à cela, un algorithme itératif a été développé pour l'estimation du paramètre θ (Kratz, 2016). Cet algorithme se base sur la résolution d'un problème d'optimisation utilisant la méthode de Levenberg Marquardt. Nous allons décrire cette méthode par la suite.

5.4 Algorithme itératif pour le modèle hybride

L'algorithme itératif permet de décomposer le problème d'optimisation des paramètres du vecteur θ en deux problèmes d'optimisation. Les paramètres $p = [\mu, \sigma, u_2]$ et ξ sont estimés alternativement. En effet, pour chaque itération, on estime les paramètres de p en minimisant le carré des écarts entre la fonction de répartition empirique et la fonction de répartition issue de la modélisation en considérant l'estimation de ξ venant de l'itération précédente. Ensuite, ξ est à nouveau estimé en

réduisant de même le carré des écarts entre la fonction de répartition empirique et la fonction de répartition issue de la modélisation en utilisant cette fois-ci le vecteur p qui a été estimé précédemment. Cette procédure commence en fixant des paramètres initiaux et s'arrêtent lorsqu'une condition est vérifiée.

Plus précisément, si on considère un échantillon $X = (X_i)_{1 \leq i \leq n}$ ayant une distribution G-E-GPD, on note $x = (x_i)_{1 \leq i \leq n}$ les réalisations associées. Pour la suite, on note $\tilde{a}^{(0)}$ et $\tilde{a}^{(k)}$ l'initialisation et l'estimation du paramètre a à la $k^{\text{ème}}$ itération.

Pour commencer le processus itératif, l'algorithme a besoin d'une valeur initiale $\tilde{\xi}_0$ de l'indice de queue ξ . Cependant, la seule information que nous avons au sujet de ξ est qu'il est positif ce qui rend son initialisation assez difficile. Pour cette raison, nous commençons par initialiser $\tilde{p}^{(0)} = [\tilde{\mu}^{(0)}, \tilde{\sigma}^{(0)}, \tilde{u}_2^{(0)}]$. Pour le faire, on choisit $\tilde{\mu}^{(0)}$ comme le mode des données. En partant du fait qu'environ 16% des réalisations d'une loi gaussienne sont en-dessous de $\mu - \sigma$, on prend $\tilde{\sigma}^{(0)} = \tilde{\mu}^{(0)} + q_{16\%}$ où $q_{16\%}$ est le quantile d'ordre 16% associé à H . Enfin, $\tilde{u}_2^{(0)}$ est un quantile suffisamment élevé (par exemple 90%). Ensuite, nous utilisons $\tilde{p}^{(0)}$ pour déterminer $\tilde{\xi}^{(0)}$ en minimisant le carré des erreurs entre la fonction de répartition de la loi hybride sachant $p = \tilde{p}^{(0)}$ (p est donc fixé à ce stade) et la fonction de répartition empirique H_n associée à l'échantillon $X = (X_i)_{1 \leq i \leq n}$ défini par

$$H_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{X_i \leq t} \quad , \quad t \in \mathbb{R}$$

Pour le faire, on n'évalue pas seulement le carré des écarts sur les réalisations de x_i uniquement (comme il y a peu d'observations à la queue de distribution) mais sur une suite $y = (y_j)_{1 \leq j \leq m}$ de taille m (m peut être différent de n) dans le but d'accroître le nombre de points au-dessus de u_2 . Plus précisément, pour $1 \leq j \leq m$, y_j est exprimé comme suit :

$$y_j = \min_{1 \leq i \leq n} (x_i) + \left(\max_{1 \leq i \leq n} (x_i) - \min_{1 \leq i \leq n} (x_i) \right) \log_{10} \left(1 + \frac{9(j-1)}{m-1} \right)$$

A noter que l'introduction de nouveaux points entre les observations de X a un impact sur H car on l'évalue sur plus de points mais ça n'a pas de conséquence sur H_n .

Ainsi, $\tilde{p}^{(0)}$ est déterminé en résolvant le problème d'optimisation ci-dessous par l'algorithme de Levenberg Marquardt :

$$\tilde{\xi}^{(0)} \leftarrow \underset{\xi > 0}{\operatorname{argmin}} \| H(y; \theta | \tilde{p}^{(0)}) - H_n(y) \|_2^2$$

Où $\theta | \tilde{p}^{(0)}$ correspond à θ pour $p = \tilde{p}^{(0)}$

Une fois que $\tilde{\xi}^{(0)}$ est déterminé, on peut appliquer le processus itérativement. Pour $k \geq 1$, la $k^{\text{ème}}$ itération est divisée en deux problèmes de minimisation qui sont résolus alternativement comme suit :

Etape 1 : Détermination de $\tilde{p}^{(k)} = [\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{u}_2^{(k)}]$ en minimisant le carré des erreurs entre la fonction de répartition hybride sachant $\tilde{\xi}^{(k-1)}$ et la fonction de répartition empirique.

$$\tilde{p}^{(k)} \leftarrow \underset{\substack{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^* \\ u_2 \in \mathbb{R}_+}}{\operatorname{argmin}} \| H(y; \theta | \tilde{p}^{(k)}) - H_n(y) \|_2^2$$

Où $\theta|\xi^{(k)}$ correspond à θ pour $\xi = \xi^{(k-1)}$ (fixé)

Etape 2 : Détermination de $\xi^{(k)}$ en minimisant le carré des erreurs entre la fonction de répartition hybride sachant $\tilde{p}^{(k)}$ et la fonction de répartition empirique.

$$\xi^{(k)} \leftarrow \underset{\xi > 0}{\operatorname{argmin}} \| H(y; \theta|\tilde{p}^{(k)}) - H_n(y) \|_2^2$$

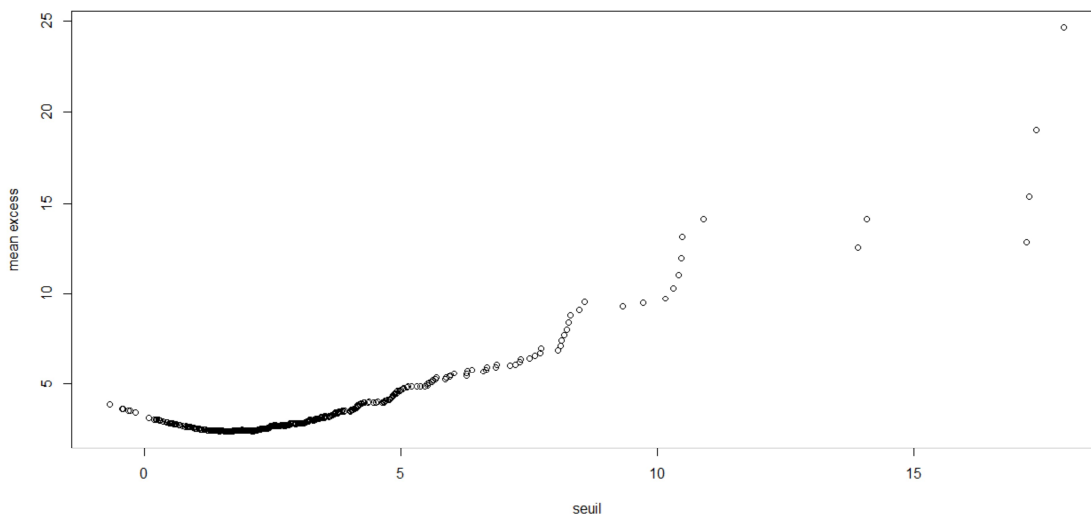
Où $\theta|\tilde{p}^{(k)}$ correspond à θ pour $p = \tilde{p}^{(k)}$ (fixé)

Condition d'arrêt :

$$|H_n(y)| < (1 + 10^{-k})|H(y, \theta^{(k)})| \quad \text{ou} \quad k = k_{\max} = 20 \text{ itérations}$$

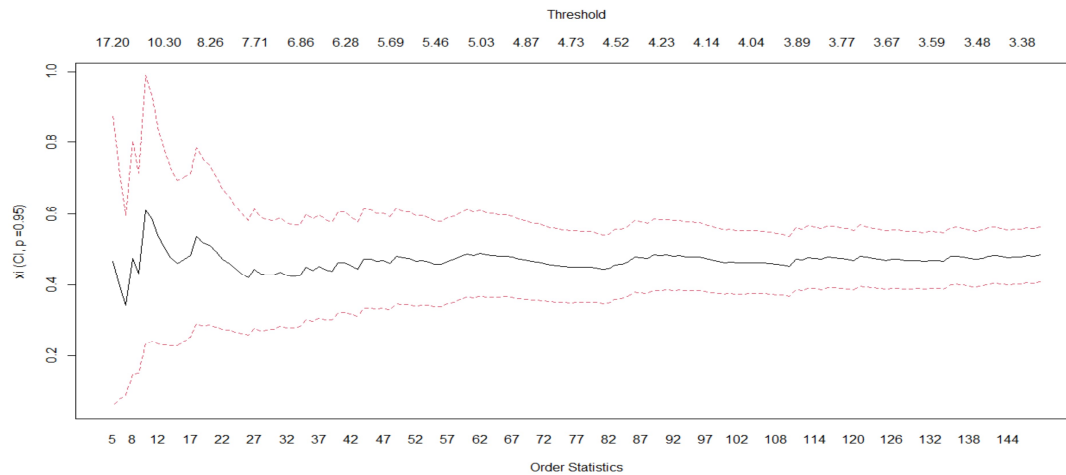
5.5 Performance du G-E-GPD avec des données simulées

Nous allons tenter de voir l'utilité de l'algorithme vu précédemment par rapport aux méthodes graphiques de recherche de seuil et de l'estimation des paramètres qui en découle. Nous simulons 1000 observations issues d'une loi G-E-GPD de paramètre $\theta = [2, 1, 5, 0.5]$. On commence par le mean excess plot pour rechercher un comportement linéaire :



Nous voyons qu'il n'est pas évident de déterminer le seuil à partir duquel on a un comportement linéaire tant il semble y avoir plusieurs seuils pour la linéarité dont en particulier les valeurs autour de 5.

Avec le hill plot, on obtient :



Le hill plot est beaucoup plus clair et on observe une stabilité vers un seuil de 4,04. On estime le paramètre de queue en partant de ce constat par le maximum de vraisemblance. Nous déterminons ensuite le paramètre de queue avec l'algorithme. Tous les résultats sont donnés dans le tableau ci-dessous où N_{u_2} représente le nombre d'observations au-dessus du seuil et la distance donne la moyenne des carrés des erreurs dans la queue entre la distribution empirique et la distribution modélisée par les différentes méthodes.

Modèle	indice de queue	seuil	N_{u_2}	distance (queue)
GPD-MLE	0,41	4,04 = $q_{78\%}$	108	1,97E-05
G-E-GPD	0,47	5,2 = $q_{88\%}$	55	9,02E-06

Non seulement l'algorithme donne un estimateur de l'indice de queue proche de celui qui a été fixé mais en plus, il réduit considérablement les écarts par rapport aux autres méthodes alors qu'il s'agit déjà d'un échantillon de données assez réduit. Sachant que nous aurons moins de données avec l'historique des pandémies, il est donc judicieux de préférer l'algorithme du G-E-GPD plutôt que les autres méthodes.

Pour étudier la performance de l'algorithme permettant la détermination des paramètres de loi G-E-GPD, on utilise des simulations de Monte-Carlo. On procède en 4 étapes :

Etape 1 : On considère N ensembles d'entraînements $\{x^q = (x_p^q)_{1 \leq p \leq n}\}_{1 \leq q \leq N}$ de taille n et N ensemble de tests $\{y^q = (y_p^q)_{1 \leq p \leq n}\}_{1 \leq q \leq N}$ de taille l suivant une loi G-E-GPD de paramètre θ fixé.

Etape 2 : Pour chaque ensemble de test x^q , $1 \leq q \leq N$, on estime θ par $\tilde{\theta}^q = [\tilde{\mu}^q, \tilde{\sigma}^q, \tilde{u}_2^q, \tilde{\xi}^q]$ en utilisant l'algorithme précédent. On note par \tilde{a}^q l'estimateur du paramètre a relatif au q^{eme} ensemble d'entraînement.

Etape 3 : On calcule la moyenne et la variance empirique des estimateurs de chaque paramètre des N ensembles d'entraînement. Pour un paramètre a , on note par \tilde{a} et \tilde{S}_a sa moyenne et sa variance empirique définies par :

$$\tilde{a} = \frac{1}{N} \sum_{q=1}^N \tilde{a}^q \quad \text{et} \quad \tilde{S}_a = \frac{1}{N-1} \sum_{q=1}^N (\tilde{a}^q - \tilde{a})^2$$

On peut vérifier la précision de \tilde{a} en utilisant le critère du mean squared error (MSE) : Le MSE d'un paramètre a est $MSE_a = \frac{1}{N} \sum_{q=1}^N (\tilde{a}^q - a)^2$. Une valeur proche de 0 du MSE témoigne d'une bonne précision de l'estimation des paramètres par l'algorithme.

Etape 4 : On compare la densité hybride h sachant θ avec la densité estimée \tilde{h} sur chaque ensemble de test y^q sachant θ^q . Pour le faire, on calcule la moyenne du log des ratios de vraisemblance sur N simulations :

$$D = \frac{1}{Nl} \sum_{q=1}^N \sum_{p=1}^l \log \left(\frac{h(y_p^q; \theta)}{h(y_p^q; \tilde{\theta}^q)} \right)$$

		n=500	n=1000
$\mu = 2$	$\tilde{\mu}$	1,96	1,98
	\tilde{S}_N^μ	7,50E-02	6,20E-03
	MSE_μ	7,40E-02	6,18E-03
$\sigma = 1$	$\tilde{\sigma}$	1,06	1,01
	\tilde{S}_N^σ	8,33E-02	5,40E-03
	MSE_σ	8,31E-02	5,38E-03
$u_2 = 5$	\tilde{u}_2	4,97	5,01
	$\tilde{S}_N^{u_2}$	9,01E-02	1,50E-03
	MSE_{u_2}	8,99E-02	1,40E-03
$\xi = 0.5$	$\tilde{\xi}$	0,47	0,49
	\tilde{S}_N^ξ	5,77E-02	9,87E-03
	MSE_ξ	5,74E-02	9,83E-03
D		3,72E-02	3,01E-03

Les résultats sont satisfaisants de par le peu d'écart entre les paramètres estimés par l'algorithme et les paramètres réels ainsi que l'écart très réduit entre la densité estimée et la densité réelle.

5.6 Application à des taux bruts de mortalité dû à une épidémie

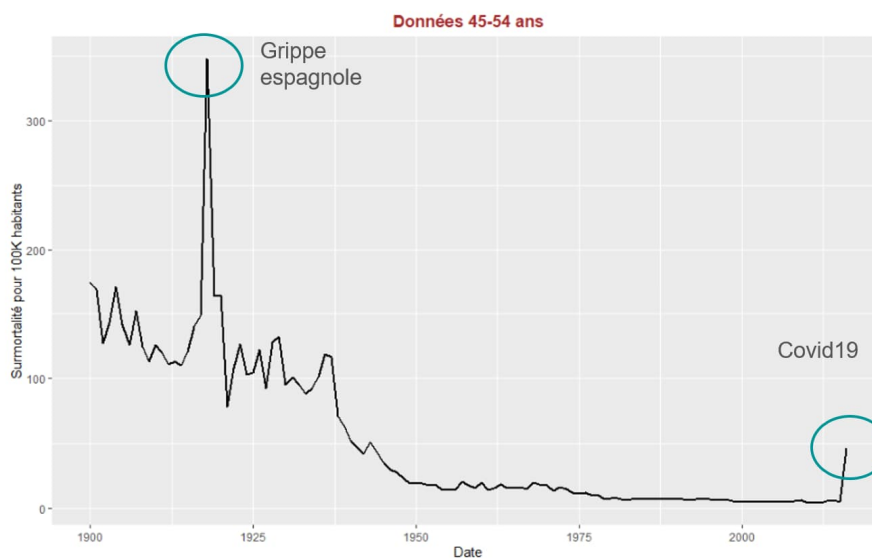
Pour déterminer le choc de mortalité dû à une épidémie, nous allons utiliser l'historique des taux bruts de mortalité dû à une épidémie. La plupart des historiques disponibles sur les épidémies concernent les épidémies grippales qui sont annuelles avec notamment les gros excès dû aux épidémies ayant eu lieu au cours du 20ème siècle comme la grippe espagnole. Sur un portefeuille d'assurés, il est intéressant d'avoir une segmentation de ce choc de mortalité par âge tant l'âge des personnes est un réel critère de différenciation en cas de sévérite. L'idéal serait donc d'avoir le nombre de décès causés par une épidémie grippale par âge rapporté à la population de chaque

classe d'âge pour obtenir un taux brut de mortalité dû à l'épidémie sur un historique suffisamment long en France. Cependant, nous n'avons pas trouvé d'informations de ce type pour les classes d'âge françaises bien qu'on ait l'information pour la population générale. Nous avons donc utilisé les données américaines fournies par le CDC qui répertorie l'ensemble des taux bruts de mortalité pour cause de grippe de 1900 jusqu'à 2015 (les données de 2016 à 2020 ne semblent pas encore consolidées). Nous avons artificiellement rajouté une année 2016 supplémentaire qui a les données 2020 du covid19. Cette manipulation n'a pas d'énormes conséquences étant donné que pour toutes catégories d'âge, les taux bruts de mortalité dû à une épidémie de grippe ont peu varié dans les années 2000. Nous ne considérons pas le covid19 comme une épidémie de grippe mais nous voulons surtout capter la surmortalité de cette épidémie par rapport à mortalité annuelle apportée par une épidémie de grippe. On peut aussi noter que la grippe et le covid19 ont des caractéristiques similaires d'un point de vue épidémiologique comme la mutation de leur virus, la transmission directe par voie interhumaine, des caractères symptomatiques et asymptomatiques.

5.6.1 Démarche

Pour obtenir les résultats, nous prenons les variations de mortalité annuelle d'épidémies que nous modélisons par une G-E-GPD. Une fois que notre G-E-GPD est modélisée, nous déterminons le quantile bicentenaire correspondant au choc de mortalité dû à une pandémie.

Prenons par exemple les données de la classe d'âge 45-54 ans représentées par l'évolution de leurs taux bruts de mortalité grippale ci-dessous :



Les taux bruts de mortalité dû à une épidémie de grippe sont donnés pour 100000 habitants. Nous observons un fort pic à plus de 300 décès correspondant à la grippe espagnole de 1918 puis le taux brut de cette classe d'âge va au fur et à mesure décroître pour se stabiliser dans les années 2000 grâce notamment à l'amélioration des soins et un système immunitaire plus solide. Malgré cette stabilité, un sursaut du taux brut est observé avec le covid19 pour cette classe d'âge. A l'échelle, ce sursaut semble beaucoup moins important que celui de la grippe espagnole de 1918 mais quand nous regardons l'effet de la grippe espagnole et du covid19 par rapport aux années qui les précèdent, on a :

Année	AllAges	<1year	1-4years	5-14years	15-24years	25-34years	35-44years	45-54years	55-64years	65-74years	75-84years	>85
1914	132.4	1439.8	214.80	21.90	25.50	38.70	67.50	110.20	211.5	460.1	1096.3	2244.60
1915	145.9	1489.4	205.50	21.80	28.00	43.00	71.00	121.10	240.0	567.9	1371.5	2930.40
1916	163.3	1414.0	221.70	23.00	35.00	51.40	87.70	140.40	277.3	661.3	1631.5	3698.80
1917	164.5	1474.5	211.50	24.00	38.90	59.30	98.10	148.80	281.4	614.6	1503.0	3187.40
1918	588.5	2273.3	718.00	176.20	580.50	992.60	554.80	347.80	381.9	646.3	1179.0	2230.60

Taux brut de mortalité causé par la grippe par âge de 1914 à 1918

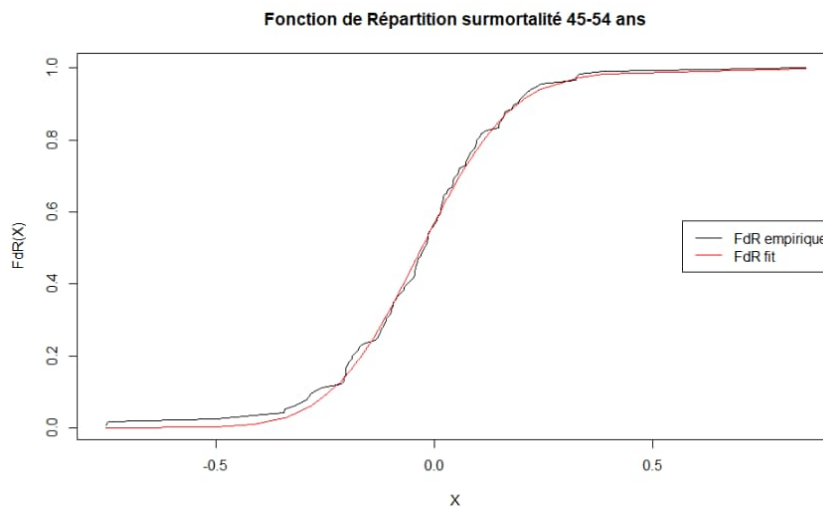
Année	AllAges	<1year	1-4years	5-14years	15-24years	25-34years	35-44years	45-54years	55-64years	65-74years	75-84years	>85
2012	16.1	NA	0.60	0.20	0.30	0.80	1.70	4.10	10.2	26.1	98.2	408.40
2013	18.0	NA	0.60	0.30	0.40	1.00	2.20	5.10	12.2	29.5	103.7	441.00
2014	17.3	NA	0.70	0.20	0.50	1.30	2.80	6.30	13.4	29.8	96.4	385.90
2015	17.8	NA	0.60	0.20	0.40	0.90	1.70	4.70	11.3	29.5	101.6	421.40
2016	NA	NA	0.16	0.12	1.47	5.79	16.32	46.15	109.3	257.6	657.5	1876.95

Taux brut de mortalité causé par la grippe par âge de 2012 à 2016 (2016 avec des données 2020)

La classe d'âge 45-54 ans a vu son taux de mortalité brut de mortalité causé par une épidémie de grippe doublé avec la grippe espagnole de 1918 alors que ce même taux a été multiplié par 10 environ pendant le covid19.

5.6.2 Résultats

En utilisant toujours les données de la classe d'âge 45-54 ans et en y appliquant le modèle sur ces données, on obtient :



Nous aboutissons à une fonction de répartition assez bien modélisée malgré le peu de données. On prend le quantile d'une épidémie bicentenaire. On détermine un choc de 0,198 point sur le SCR CAT Vie pour les 45-54 ans.

Si on refait la même chose sans considérer la grippe espagnole de 1918, on trouve un choc de 0,201 point. Cela peut sembler étonnant au premier abord mais comme nous l'avons vu, aux Etats-Unis, la classe d'âge 45-54 ans faisait partie des classes d'âge les moins touchées par la grippe avec un taux brut de mortalité qui a doublé alors que cette même classe d'âge a vu son taux brut de mortalité être

multiplié par 10 avec le covid19 comme on peut le voir avec les données ci-dessus qui comparent les 4 années précédant la grippe espagnole et le covid19.

Pour l'ensemble des classes d'âge avec et sans prise en compte de la grippe 1918 (grippe espagnole), on a :

Choc de mortalité		
Classe d'âge	Avec grippe 1918	Sans grippe 1918
15-24	0.12	0.1
25-34	0.15	0.16
35-44	0.17	0.16
45-54	0.19	0.2
55-64	0.21	0.20
65-74	0.27	0.28

A ces différents chocs de mortalité, on applique un coefficient de 60% comme l'a fait L'European Actuarial Consultative Group sur les taux européens pour obtenir un choc spécifique sur les assurés.

Choc de mortalité		
Classe d'âge	Avec grippe 1918	Sans grippe 1918
15-24	0.072	0.06
25-34	0.09	0.096
35-44	0.1	0.096
45-54	0.11	0.12
55-64	0.13	0.12
65-74	0.16	0.17

Nous observons donc qu'avec ou sans considération de la grippe espagnole dans les données, le choc de mortalité sur les taux de mortalité va de 0,06 point pour les plus jeunes à 0,17 point pour les plus âgés en pourcentage en cas de pandémie. Cela donne donc un choc en-dessous du choc de la formule standard à 0,15 point pour toutes les classes d'âge sauf pour la classe d'âge 65-74.

PARTIE IV : Impact des chocs estimés

1. Indicateurs et traitements

Dans cette section, nous allons mettre en évidence les indicateurs du Crédit Agricole qui vont nous servir de référence pour apprécier l'impact des chocs obtenus au cours du mémoire et nous ferons un traitement des résultats du modèle épidémiologique afin de pouvoir les appliquer à une population assurée.

1.1 Les indicateurs de base

Nous rappelons que le capital de solvabilité requis du groupe Crédit Agricole Assurances est de 14 milliards d'euros au 31 décembre 2020. Le capital de solvabilité requis de base du groupe est de 28.8 milliards d'euros.

Le capital de solvabilité requis du module de risque de souscription en vie est de 2.7 milliards d'euros. Il s'évalue en agrégeant les sous-modules de risque mortalité, longévité, incapacité/invalidité, dépenses, révision, cessation et CAT via la matrice de corrélation ci-dessous dont les coefficients sont rendus publics au sein de la directive solvabilité 2. Le SCR Vie CAT est de 31.4 millions d'euros.

Corrélations (i,j)	Mortalité	Longévité	Incap/Inval	Dépenses	Révision	Cessation	CAT
Mortalité	1	-0.25	0.25	0.25	0	0	0.25
Longévité	-0.25	1	0	0.25	0.25	0.25	0
Incap/Inval	0.25	0	1	0.5	0	0	0.25
Dépenses	0.25	0.25	0.5	1	0.5	0.5	0.25
Révision	0	0.25	0	0.5	1	0	0
Cessation	0	0.25	0	0.5	0	1	0.25
CAT	0.25	0	0.25	0.25	0	0.25	1

Le SCR Vie est donnée par la formule :

$$SCR_{Vie} = \sqrt{\sum_{i,j} Corr_{i,j} \cdot SCR_i \cdot SCR_j}$$

Le capital de solvabilité requis du module de risque de souscription en santé est égal à 967 millions d'euros. Il s'évalue en agrégeant les sous-modules de risque en santé assimilable à la vie, risque en santé assimilable à la non-vie et risque CAT via une matrice de corrélation dont les coefficients sont rendus publics au sein de la directive solvabilité 2. Ainsi le SCR Santé CAT contribue au SCR Santé de manière non proportionnelle.

Cependant, comme ça a été dit en 6. , au niveau du profil de risque du groupe Crédit Agricole, c'est le risque de marché qui est largement prédominant avec une exposition à 64% du BSCR reléguant le risque de souscription santé au rang de risque secondaire avec une simple exposition à 5%. C'est pourquoi le Crédit Agricole Assurances utilise un proxy pour l'évaluation de son besoin en capital sur un sous-module comme le sous-module CAT en Santé. Nous allons donc d'abord estimer le besoin en

capital de ce sous-module pour mieux évaluer l'impact des chocs du modèle épidémiologique que nous tentons d'appliquer.

Si nous considérons que le sous-module santé assimilable à la vie a un besoin en capital de 800 millions d'euros et que celui du sous-module santé assimilable à la non-vie est nul alors il est possible de déterminer le montant implicite du besoin en capital du sous-module CAT. La matrice d'agrégation du module santé est :

Corrélations (i,j)	SLT	Non SLT	CAT
SLT	1	0.5	0.25
Non SLT	0.5	1	0.25
CAT	0.25	0.25	1

Le SCR santé est donc donné par la formule :

$$SCR_{sante} = \sqrt{\sum_{i,j} Corr_{i,j} \cdot SCR_i \cdot SCR_j}$$

Sachant que le SCR santé est de 967 millions d'euros, nous déterminons aisément que le SCR santé CAT est de 379 millions d'euros.

Le SCR santé CAT est composé du risque de concentration de masse, du risque de concentration d'accident et du risque pandémie. Comme les risques de concentration de masse et d'accident ne sont pas significatifs nous considérons que le SCR santé CAT est entièrement déterminé par le SCR santé CAT Pandémie. Ainsi, dans notre analyse d'impact, la valeur centrale du SCR est 14 milliards, celle du SCR Vie CAT est de 31.4 millions et celle du SCR santé CAT Pandémie est de 379 millions d'euros.

1.2 De la population générale à la population assurée

Le modèle épidémiologique que nous avons présenté se base sur une population quelconque qui n'est pas nécessairement la population assurée. Cette population quelconque nous permet de mettre en évidence les indicateurs de taux d'infection et nombre de jour d'hospitalisation qui pourraient être générés si elle est formée de 1000 individus puis le coefficient à appliquer à une exposition aux frais d'hospitalisation qui serait aussi basée sur cette population. Nous n'avons pas cette problématique avec les taux de surmortalité obtenus pour le SCR Vie CAT car nous avons appliqué un coefficient de 60% pour faire la distinction entre les deux types de population.

De plus, les 1000 individus qui sont utilisés au sein de chacune des simulations représentent un nombre assez limité pour le nombre d'agents constituant une population. Cela est dû au fait de la puissance de calcul qui était à disposition pour générer les 10000 pandémies du modèle. En effet, l'idéal aurait été de pouvoir simuler des pandémies sur des populations constituées de 1 million, 10 millions voire encore plus d'habitants mais les temps de simulation étaient déjà conséquents à partir de 3000 habitants pour 10000 simulations d'où la limitation à 1000 individus.

Il est difficile d'évaluer un impact pour le portefeuille assuré directement avec ces simulations de 1000 individus car les pandémies retenues ont un taux d'infection de 20 à 40% ce qui veut dire que le

nombre de personnes infectées varient entre 200 et 400 personnes. Or sur ce nombre d'infectés, très peu vont en hospitalisation selon les probabilités qui ont été déterminées. Ainsi, si on considère que la population assurée constitue une proportion de la population générale (5% par exemple), le nombre d'assurés infectés serait considérablement réduit ce qui ferait ressortir pour le modèle un nombre de jour total d'hospitalisation généré par une pandémie très proche de 0 dans la majorité des cas voire nul.

Pour mesurer l'impact avec le portefeuille assuré c'est-à-dire en considérant les âges des assurés et le niveau de risque lié à l'hospitalisation de chacun d'entre eux, nous allons utiliser un ratio entre la probabilité de régler des prestations aux assurés et la probabilité de régler une prestation aux individus de la population générale. Ce ratio se calcule comme suit :

$$\text{ratio portefeuille} = \frac{\text{probabilité de prestation sur le portefeuille}}{\text{probabilité de prestation sur la population générale}}$$

Typiquement si ce ratio est de 0.33 alors le SCR Santé CAT Pandémie serait 3 fois moins important que celui de la population générale.

Comme le portefeuille est composé d'actifs âgés de 20 à 64 ans dont le contrat a été souscrit par leur entreprise, ils appartiennent à la classe d'âge A_2 qui a été fixée en 4.4.1 . De plus, nous avons fait l'hypothèse que les assurés vont globalement avoir un niveau de risque moindre par rapport à la population générale sur les hospitalisations. Ces hypothèses, nous permettent de calculer le ratio selon la formule suivante :

$$\text{ratio portefeuille} = \frac{\mathbb{P}(H|A_2 \cap R^c)}{\mathbb{P}(H)}$$

Par le fait que M et M^c forment un système complet d'évènement, nous avons d'une part :

$$\mathbb{P}(H|A_2 \cap R^c) = \mathbb{P}(H|A_2 \cap R^c \cap M) \cdot \mathbb{P}(M) + \mathbb{P}(H|A_2 \cap R^c \cap M^c) \cdot \mathbb{P}(M^c)$$

D'autre part, on a :

$$\mathbb{P}(H) = \mathbb{P}(H|M) \cdot \mathbb{P}(M) + \mathbb{P}(H|M^c) \cdot \mathbb{P}(M^c)$$

Mais comme un individu ne peut pas être hospitalisé s'il n'est pas préalablement tombé malade alors $\mathbb{P}(H|A_2 \cap R^c \cap M^c)$ et $\mathbb{P}(H|M^c)$ sont nuls.

D'où,

$$\mathbb{P}(H|A_2 \cap R^c) = \mathbb{P}(H|A_2 \cap R^c \cap M) \cdot \mathbb{P}(M)$$

Et,

$$\mathbb{P}(H) = \mathbb{P}(H|M) \cdot \mathbb{P}(M)$$

Ainsi, nous avons :

$$\text{ratio portefeuille} = \frac{\mathbb{P}(H|A_2 \cap R^c \cap M)}{\mathbb{P}(H|M)}$$

Selon le type de scénario présenté en 4.3 , nous obtenons les ratios ci-dessous :

Type de scénario	ratio portefeuille
scenarion faible	42.6%
scenarion moyen	75.3%
scenarion dur	73.5%

Le ratio du scenarion faible pénalise de manière importante la population générale pour passer à la population assurée. La pénalité est moins importante dans les scénarios moyen et dur car elle est d'environ 3/4 dans les deux cas.

Le SCR Santé CAT Pandémie résultant de cet impact est :

$$SCR\ Sante\ CAT\ Pandemie_{impact} = Exposition\ aux\ frais\ d'hospitalisation * Choc * ratio\ portefeuille$$

Mais comme l'exposition aux frais d'hospitalisation représente 250% du SCR Santé CAT Pandémie, on a :

$$SCR\ Sante\ CAT\ Pandemie_{impact} = \frac{SCR\ Sante\ CAT\ Pandemie}{0.4} * Choc * ratio\ portefeuille$$

Il nous reste à déterminer le choc à utiliser. En 9. , nous avons vu comment évolue le choc selon le nombre de jour moyen d'hospitalisation et selon la proportion d'assurés allant en hospitalisation. Comme cette proportion est fixée à 1% dans la directive Solvabilité 2, nous allons prendre la moyenne des chocs pour une proportion d'assurés allant en hospitalisation de 1% avec entre 10 et 24 jours d'hospitalisation attendus en moyenne pour les assurés. On obtient les chocs suivants :

Type de scénario	choc
scenarion faible	6.5%
scenarion moyen	42.8%
scenarion dur	144.5%

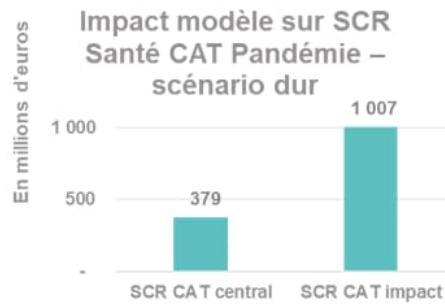
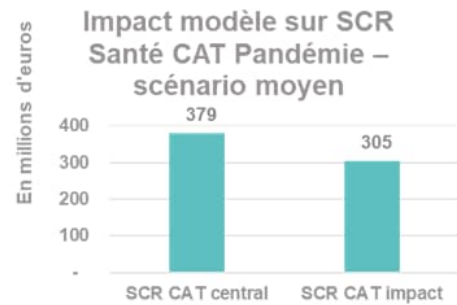
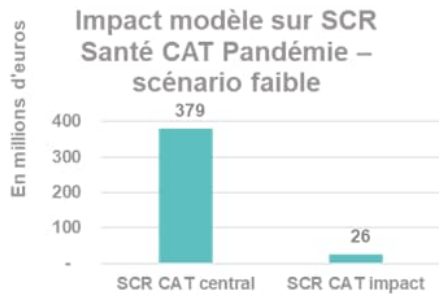
Nous sommes désormais en mesure de déterminer l'impact sur le SCR.

2. Analyse des résultats

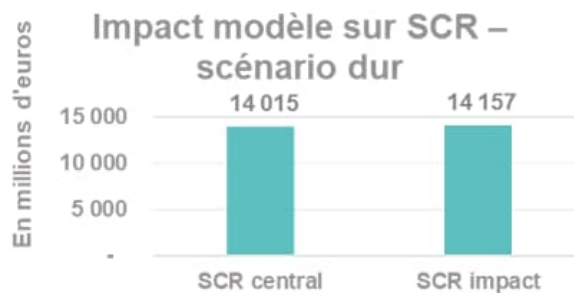
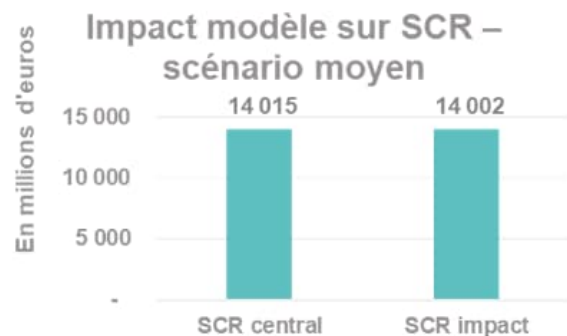
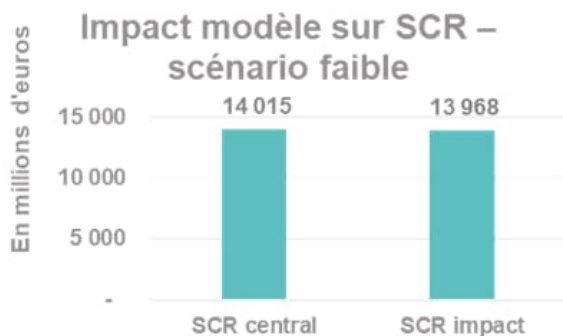
Dans cette section, nous allons évaluer l'impact que peut avoir les chocs en sortie des modèles présentés dans les sections précédentes sur le SCR de la compagnie.

2.1 Résultats impact Santé CAT

Pour déterminer l'impact de notre étude, nous utilisons la nouvelle valeur de SCR Sante CAT Pandémie obtenue par les différents calculs de la section précédente. Il correspond d'emblée à notre SCR Santé CAT :



Puis nous l'agrégeons par matrice des corrélations avec les sous-modules du risque Santé afin d'obtenir le SCR Santé impacté par notre étude. Ensuite, il suffit d'agréger cette fois-ci le module de risque Santé obtenu avec les autres modules de risque principaux de la formule standard et enfin nous ajoutons le SCR du risque opérationnel et retranchons les capacités d'absorption de pertes des impôts différés et des provisions techniques pour obtenir le SCR impacté. Ainsi nous observons les résultant suivants :



De manière plus synthétique, nous avons ces variations :

Type de scénario	Δ SCR
scénario faible	-0.3%
scénario moyen	-0.1%
scénario dur	1.0%

Le constat est assez clair car nous voyons que peu importe le scénario pandémie qui nous permet d'avoir de nouveaux chocs à appliquer à l'exposition aux frais d'hospitalisation, le SCR de la compagnie va en être peu impacté. La plus large variation relative que nous obtenons provient du scénario le plus critique et même dans ce cas le SCR de la compagnie augmenterait de seulement 1% alors que ce scénario est particulièrement extrême avec une probabilité d'hospitalisation de 10% parmi les personnes infectées.

2.2 Résultats impact Vie CAT

Pour l'impact du SCR Vie CAT, nous utilisons la formule simplifiée donnée par la directive solvabilité 2 que nous adaptons à notre étude. En effet, cette formule se base sur un unique taux de surmortalité fixé à 0.15% à appliquer à l'ensemble des capitaux sous-risque or notre étude par approche avec la théorie des valeurs extrêmes nous a permis d'avoir une hétérogénéité pour les taux de surmortalité à appliquer selon la classe d'âge.

Nous avons aussi fait un test de sensibilité pour avoir une idée de l'impact de la grippe espagnole de 1918 sur les taux de surmortalité. Ce test a fait ressortir que les taux étaient peu sensibles donc les taux de surmortalité que nous utilisons pour avoir l'impact sur le SCR sont ceux qui sont obtenus sans restriction.

Ainsi le SCR Vie CAT s'évalue comme suit :

$$SCR\ Vie\ CAT = \sum_i (tx\ mortalite_i * \sum_{j \in i} CSR_j)$$

Avec

- $tx\ mortalite_i$: le taux de mortalité dû à la pandémie pour la classe d'âge i

- CSR_j : le capital sous risque total des assurés d'âge j

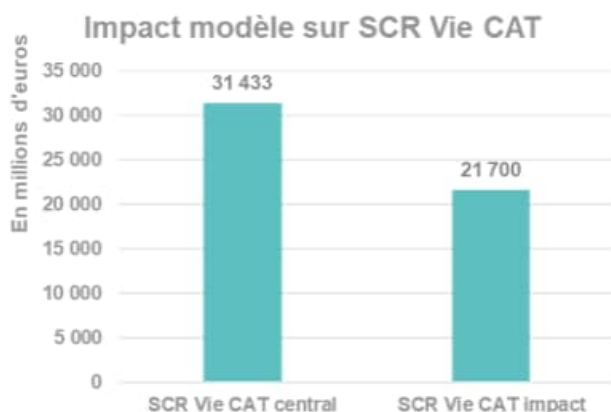
Après mécanismes d'atténuations des risques tels que la réassurance et commissions variables, les capitaux sous risque de l'entité Vie du Crédit Agricole Assurances s'élèvent à 20 milliards répartis de la manière suivante :

Classe d'âge	Capital sous risque
18-24	324 673 840
25-34	5 419 162 691
35-44	7 245 099 411
45-54	5 110 541 839
55-64	2 444 714 104
65+	340 126 114

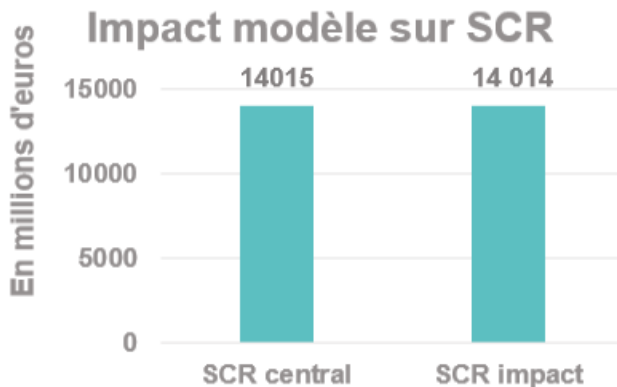
Comme les capitaux sous risque des personnes âgées de 75 ans et plus ne représentent que 0.01% des capitaux en portefeuille et que la classe d'âge la plus âgée ayant un coefficient de surmortalité en sortie du modèle est la classe d'âge 65-74 ans, nous allons utiliser la surmortalité des personnes âgées de 65 à 74 ans sur l'ensemble de la classe d'âge 65+.

Le reste du calcul d'impact est identique à ce qui a été fait à la section précédente.

On détermine l'impact sur le sous-module SCR Vie CAT :



Puis nous faisons une première couche d'agrégation avec les sous-module de risque puis une seconde couche d'agrégation avec les modules de risque principaux suivi d'un ajout du module de risque opérationnel et d'un retranchement des capacités d'absorption de pertes des impôts différés et des provisions techniques. Nous obtenons le résultat suivant :



La variation de SCR après les effets de diversification est de -0.0086% ce qui n'est pas très impactant. Ainsi même en utilisant une surmortalité hétérogène par classe d'âge, le SCR obtenu reste peu variable par rapport au SCR central.

CONCLUSION

Notre travail s'est articulé sur deux axes : tout d'abord, on s'est intéressé à la genèse et l'historique épidémiologique. Nous avons vu qu'un changement de l'environnement des espèces les poussant à migrer vers des zones géographiques où vivent des populations non immunisées contre eux pourrait être une source de pandémie à l'avenir en plus de la mondialisation qui favorise la diffusion d'un virus à une grande échelle très rapidement comme ça a été le cas avec le covid19. Nous avons établi la manière à laquelle une pandémie pourrait avoir un impact sur le secteur de l'assurance et nous avons fait état des outils dont disposent les compagnies d'assurance en France notamment pour pouvoir y faire face.

Deuxièmement, on a regardé des méthodes de modélisation pouvant donner un meilleur aperçu des impacts d'une pandémie. Il y a beaucoup de prévisions mais peu d'informations au sujet des pandémies. Les calculs effectués ont été faits sur divers hypothèses et utilisant le peu d'information disponible qui sont en général des informations globales. Typiquement, pour l'entité vie du Crédit Agricole Assurances, Predica, il est impossible d'interroger sa base de données d'assurés pour avoir un historique de sinistralité épidémiologique ainsi que la situation sanitaire des assurés en portefeuille ce qui romprait le secret médical en plus de ne pas être en accord avec le RGPD. Cela limite donc une analyse fine. Cependant, certains éléments peuvent quand même être mis en évidence : bien qu'avant le risque de pandémie pouvait être considéré comme un risque apportant des conséquences économiques sévères pour les compagnies d'assurance, ce risque semble un peu plus atténuer avec notamment, comme on peut le voir chez Prédica, des garanties sur le risque incapacité prévoyant 3 mois de franchise ce qui réduit considérablement l'exposition au risque de pertes de revenus en cas de pandémie. De plus, on peut aussi considérer la conjecture qui voudrait qu'avec le temps, le système immunitaire des individus s'améliore.

Nous avons fait une partie de notre travail sur le niveau de risque lié aux hospitalisations en cas de pandémie car le choc à 40% appliqué au sein de la formule standard pour ce risque nous semblait très en décalage de la réalité épidémiologique. Il en est sorti que le choc des hospitalisations est très lié aux attentes de l'assureur concernant la durée moyenne d'hospitalisation attendue. Nous avons construit trois scénarios différents en nous basant sur une étude externe car nous ne pouvions pas utiliser les données non-vie du Crédit Agricole Assurances. Ces scénarios dont notamment le plus critique ce sont révélés peu impactant pour le besoin en capital de la pandémie car le Crédit Agricole Assurances est surtout exposé en risque de marché. Cependant, les nouveaux chocs qui nous permettent d'évaluer l'impact sur le SCR sont obtenus avec un modèle qui a des limites donc l'interprétation de ces résultats doit se faire en connaissance des limites du modèle.

Par ailleurs, en partant du principe qu'une pandémie bicentenaire représenterait un quantile de 85%, nous avons déterminé des chocs de mortalité par âge pour le besoin en capital sur le sous module catastrophe en vie afin de pouvoir avoir une surmortalité en cas de pandémie qui soit hétérogène et adapté à l'âge de chaque individu au lieu d'une surmortalité uniforme sans distinction de classe d'âge. Ces chocs ramenés à la population assurée se sont révélés globalement en-dessous du choc de la formule standard à 0,15%. L'impact sur le SCR s'est là encore montré peu significatif.

Ainsi, ce mémoire a permis de montrer que même dans un contexte de pandémie le Crédit Agricole Assurances est peu impacté pour son besoin en capital sur les souscriptions liées aux risques biométriques et ce même en tentant de redéfinir les chocs utilisés dans la formule standard. Dans ce mémoire, nous avons fait un focus sur les risques biométriques mais la pandémie du covid19 nous a

montré que le risque de marché est très important à étudier car les marchés financiers ont connu une chute très rapide en mars 2020. C'est un axe qui pourrait être approfondi pour voir l'impact d'une pandémie sur le besoin en capital d'un acteur comme le Crédit Agricole Assurances fortement exposé au risque de marché.

BIBLIOGRAPHIE

Actuarial reflections on pandemic risk and its consequence (2006), [pandemics_web.doc \(actuary.eu\)](#)

Carreau, J., and Bengio, Y. A hybrid pareto model for asymmetric fat-tailed data: the univariate case. *Extremes* 12, 1 (2009)

DNB (De Nederlandsche Bank), PBL Netherlands Environmental Assessment Agency (2020), "Indebted to nature, Exploring biodiversity risks for the Dutch financial sector". Available at: <https://www.dnb.nl/en/news/news-and-archive/dnbulletin-2020/dnb389169.js>

Doyle, A., Bonmarin, I., Lévy-Bruhl, D., Le Strat, Y., & Desenclos, J. (2005). Estimation de l'impact d'une pandémie grippale et analyse de stratégies. Institut National de Veille Sanitaire.

Embrechts, P., Kluppelberg, C., and Mikosch, T. "Modelling Extremal Events for Insurance and Finance. Springer-Verlag, 1997

FRB (Fondation pour la Recherche sur la Biodiversité) (2020), « Mobilisation de la FRB par les pouvoirs publics français sur les liens entre Covid-19 et biodiversité ». Available at: <https://www.fondationbiodiversite.fr/wp-content/uploads/2020/05/Mobilisation-FRB-Covid-19-15-05-2020-1.pdf>

Grippe espagnole, 2685_PASTEUR_LIP91_v12.indd

Groendyke, C., Welch, D., & Hunter, D. (2010). Bayesian inference for contact networks given epidemic data. Penn State Department of Statistics.

GUIBERT, Q. , PLANCHET, F. , JUILLARD, M. (2010). Un cadre de référence pour un modèle interne partiel en assurance de personnes : application a un contrat de rentes viagères. Cahier de recherche de l'ISFA n°2126.

Kratz, M. Normex, a new method for evaluating the distribution of aggregated heavy tailed risks. *Extremes*, 4 (2014). Special issue on Extremes in Finance (Guess Ed. P. Embrechts)

Kratz, M. , Debbabi, N. , Mboup, M. , A self-calibrating method for heavy tailed data modeling : Application in neuroscience and finance 2016

LeMonde (2021) https://www.lemonde.fr/sciences/article/2020/12/22/a-l-origine-de-la-pandemie-de-covid-19-un-virus-sars-cov-2-aux-sources-toujours-enigmatiques_6064168_1650684.html

Levenberg, K. A method for the solution of certain nonlinear problems in least squares quart. *Applied Math* 2 (1944)

Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics* 11 (1963)

Munich Re (2021), Jan. 7th 2021, "Record hurricane season and major wildfires – The natural disaster figures for 2020". Available at: <https://www.munichre.com/en/company/media-relations/media-information-and-corporate-news/media-information/2021/2020-natural-disasters-balance.html>

Nadarajah, S., and Bakar, S. New composite models for the danish fire insurance data. *Scandinavian Actuarial Journal* 2014, 2 (2014)

New York Times (2018), Jan. 10th 2018, "To fight Climate change New York City takes on Oil companies". Available at: <https://www.nytimes.com/2018/01/10/nyregion/new-york-city-fossil-fuel-divestment.html#:~:text=Seeking%20to%20position%20himself%20as,in%20the%20fossil%20fuel%20business>

OMS (Organisation Mondiale de la Santé) ,
https://www.google.fr/aclk?sa=l&ai=DChcSEwi6vbSWwNjyAhWEutUKHe09BsAYABAAGgJ3cw&ae=2&sig=AOD64_06kISADnlcOr5IA7qZFckwp8yrfw&q&adurl&ved=2ahUKEwj462WwNjyAhVix4UKHcGwBDEQ0Qx6BAgCEAE

One Health Commission (2020), One Health Commission's website, "What is One Health" webpage, available at: https://www.onehealthcommission.org/en/why_one_health/what_is_one_health/ (consulted on Dec. 1st 2020)

Pandémie — Wikipédia (wikipedia.org)

Parzen, E. On estimation of a probability density function and mode. The annals of mathematical statistics 33, 3 (1962)

PwC (PricewaterhouseCoopers), WWF (2020), "Nature is too big to fail". Available at: https://wwfint.awsassets.panda.org/downloads/nature_is_too_big_to_fail_en_web.pdf

RÈGLEMENT DÉLÉGUÉ (UE) 2015/ 35 DE LA COMMISSION - du 10 octobre 2014 - complétant la directive 2009/ 138/ CE du Parlement européen et du Conseil sur l'accès aux activités de l'assurance et de la réassurance et leur exercice (solvabilité II) (europa.eu)

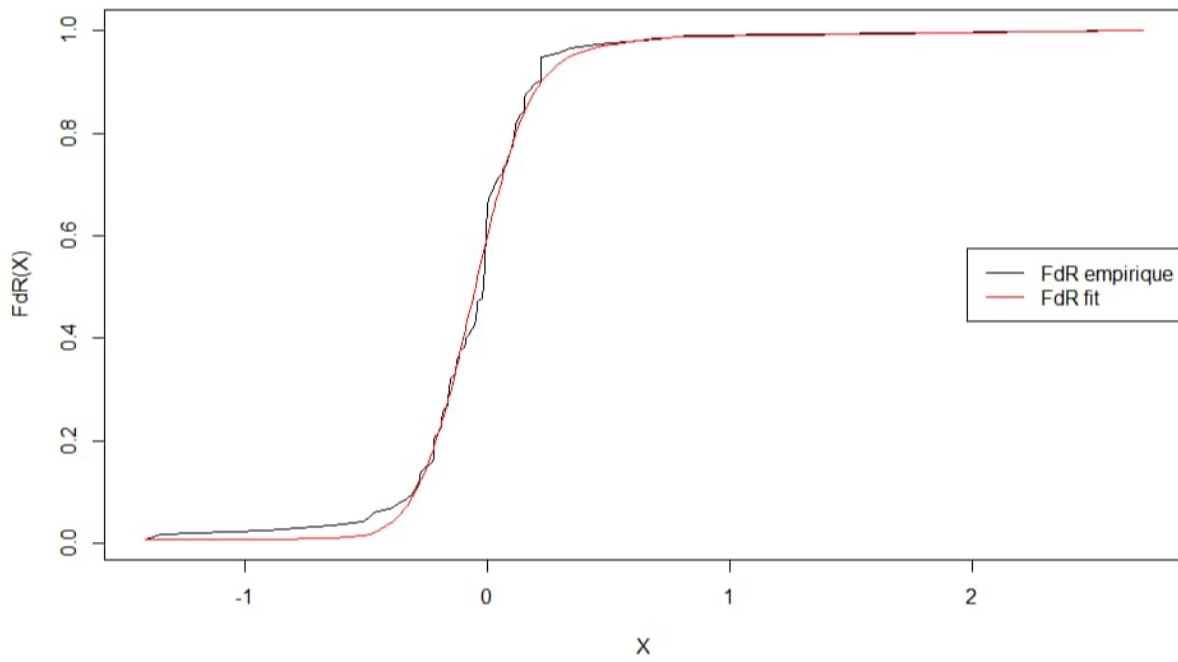
RSI(2005),<https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjDgLKMwdjyAhVC4YUKHYK2DB0QFnoECAIQAO&url=https%3A%2F%2Fapps.who.int%2Firis%2Fbitstream%2Fhandle%2F10665%2F246187%2F9789242580495-fre.pdf&usq=AOvVaw0rGNqF72-zHcFNGIFE1enl>

Schmidt K.A., Ostfeld R.S. (2001), "Biodiversity and the dilution effect in disease ecology", Ecology Vol. 82, No. 3, Mar., 2001. DOI: 10.2307/2680183

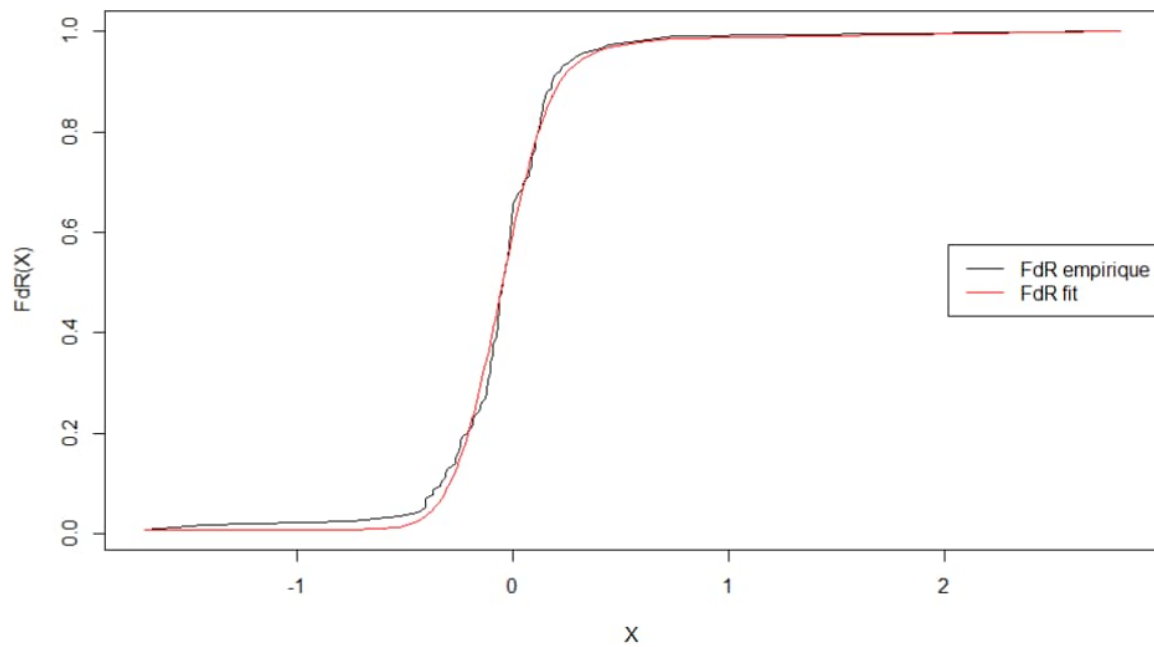
Swiss Re. (2007). Pandemic influenza : A 21st century model for mortality shocks

ANNEXE

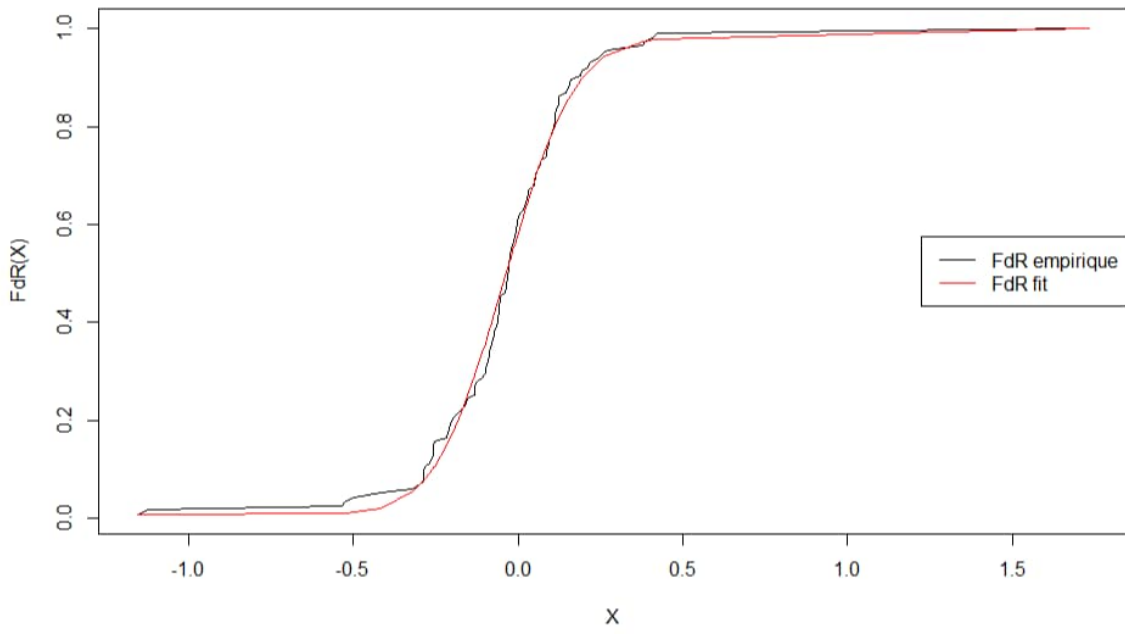
Fonction de Répartition surmortalité 15-24 ans



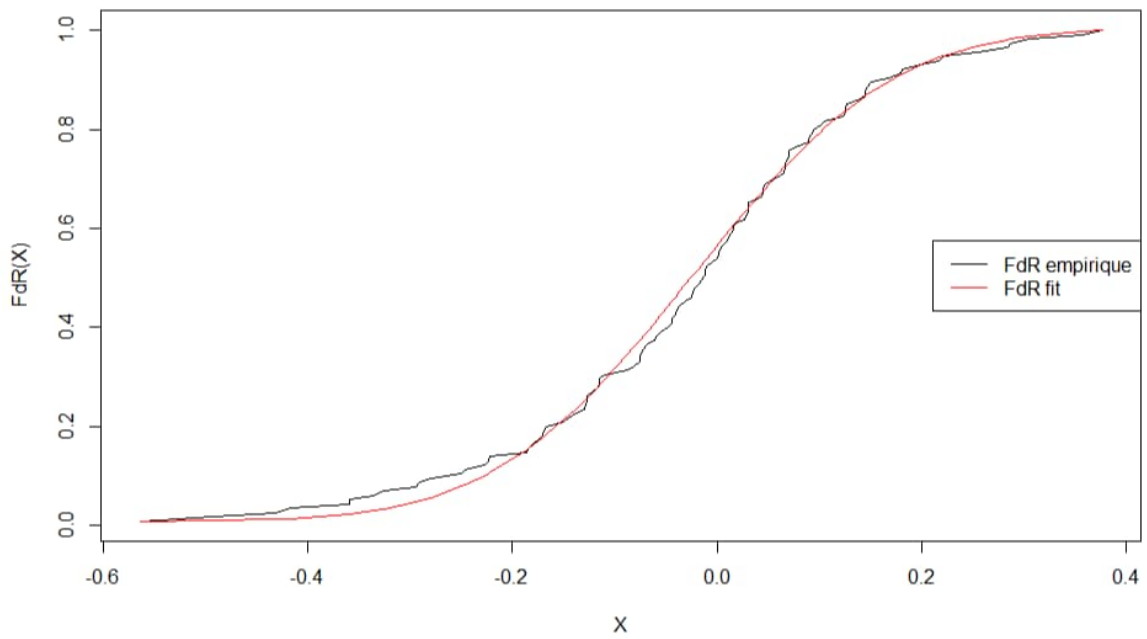
Fonction de Répartition surmortalité 25-34 ans



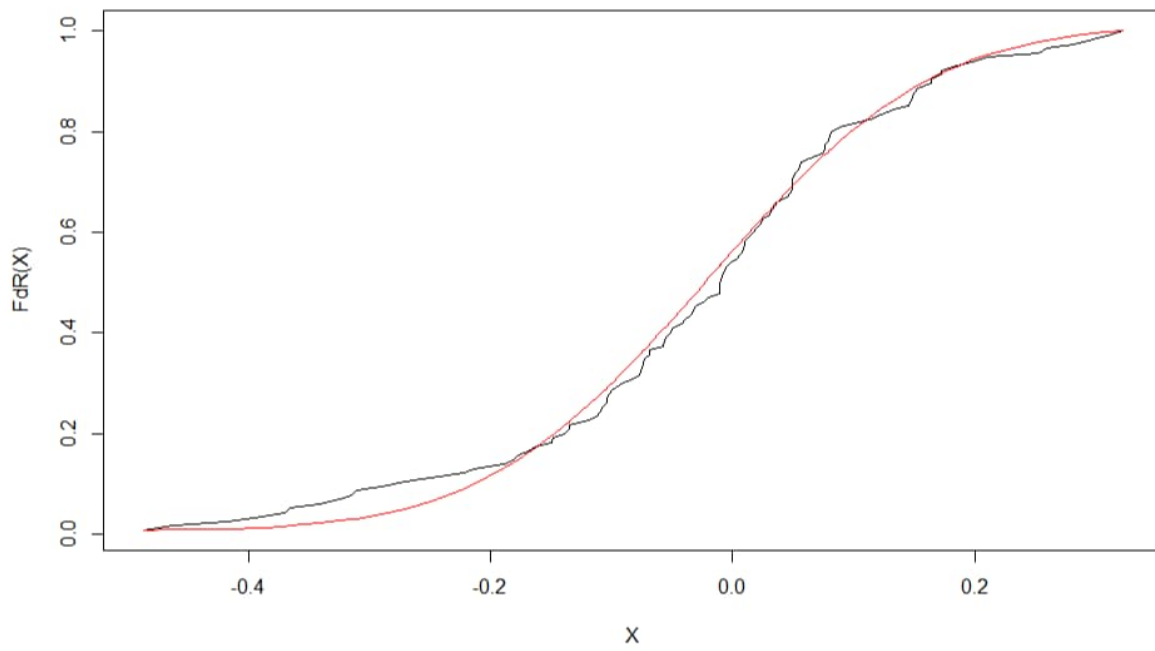
Fonction de Répartition surmortalité 35-44 ans



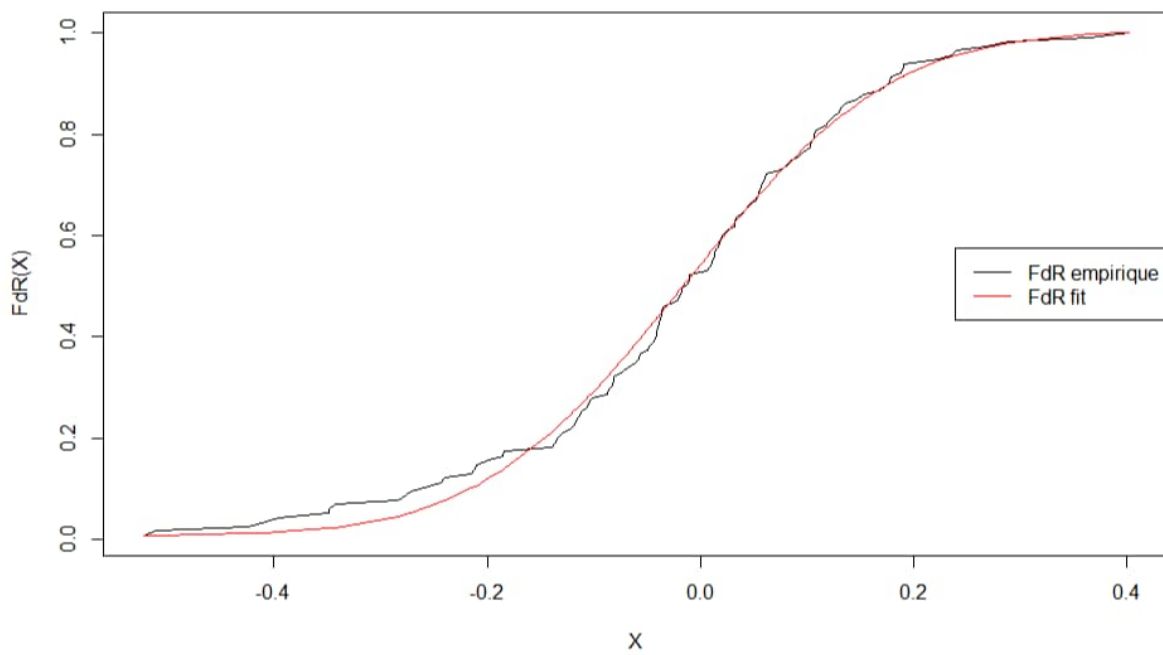
Fonction de Répartition surmortalité 55-64 ans



Fonction de Répartition surmortalité 65-74 ans



Fonction de Répartition surmortalité 75-84 ans



Fonction de Répartition surmortalité 85+ ans

