Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 24 septembre 2021

Par : Geoffroy DALLOZ
Titre : Modèles dynamiques de tarification santé en Italie dans un contexte d'épidémie de Covid

Confidentialité : Oui Durée : 2 ans

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

Membres présents du jury de l'Institut des Actuaires :
Swan BROUTARD
Marine HABART
Léonard FONTAINE
Signature :

Entreprise :
Europ Assistance
Signature :

Membre présent du jury de l'EURIA :
Philippe LENCA

Directeur de mémoire en entreprise :
Etienne BONNET
Signature :

Tuteur académique
Anthony NAHELOU
Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Europ Assistance a développé en Italie un produit original d'assurance santé composé de prestations de services et d'indemnisation pour des faits générateurs aussi variés que le diagnostic d'une maladie grave ou un accident. La première partie de ce mémoire présente en détail la modélisation de la fréquence de ce produit ainsi que la nouvelle segmentation proposée. L'analyse et la sélection des variables explicatives les plus discriminantes pour la tarification révèle l'impact important des méthodes de vente sur le comportement des assurés et la sinistralité.

Face à ce produit pour lequel nous avons un historique de données sur une dizaine d'années, nous avons modélisé dans la deuxième partie de ce mémoire la fréquence d'un nouveau produit lancé en 2020 et couvrant spécifiquement les hospitalisations liées à l'épidémie de Covid-19. Nous avons pour cela adapté et calibré un modèle compartimental permettant de suivre et de projeter l'évolution de l'épidémie dans le temps. L'adaptabilité du modèle à l'apparition de nouveaux variants de l'épidémie le rend particulièrement simple d'utilisation et permet de modéliser des scénarios variés en jouant avec ses différents paramètres. La précision de nos estimations, du moins sur un horizon de temps où la variabilité des paramètres est suffisamment prévisible, valide ce modèle qui a par ailleurs été plébiscité par la communauté scientifique et les pouvoirs publics pendant la crise sanitaire.

Enfin, la troisième partie de ce mémoire imagine la conception et la modélisation d'un produit d'assurance combinant les garanties des produits analysés dans les deux premières parties. Le réseau bayésien dynamique construit permet de faire la synthèse entre le modèle régressif du produit historique vu en première partie et le modèle prédictif du nouveau produit vu en deuxième partie. Ce type de modèle permettant de faire interagir des variables aléatoires dont le paramètres évoluent dans le temps présente de nombreux champs d'application possibles en assurance, et notamment sur les produits commercialisés par Europ Assistance.

Mots clefs: Assurance santé, segmentation, CART, Covid-19, modèle compartimental SIR, réseaux bayésiense

Abstract

Europ Assistance has developed in Italy a unique health insurance product consisting of services and compensation for events as diverse as the diagnosis of a serious illness or an accident. The first part of this thesis presents in detail the frequency modeling of this product as well as the proposed new segmentation. The analysis and selection of the most discriminating explanatory variables for pricing reveals the significant impact of sales methods on policyholder behavior and claims experience.

In contrast to this product for which we have historical data over the past ten years, we modeled in the second part of this thesis the frequency of a new product launched in 2020 covering specifically hospitalizations linked to the epidemic of Covid-19. For that purpose, we have adapted and calibrated a compartmental model enabling us to monitor and project the evolution of the epidemic over time. The adaptability of the model to the outbreak of new variants of the epidemic allows various scenarios to be modeled by playing with the dynamic parameters of its differential equations. The accuracy of our estimates, at least over a time horizon where the variability of the parameters is sufficiently predictable, validates this model which was also acclaimed by the scientific community and the public authorities during the health crisis.

Finally, the third part of this thesis foresees the design and modeling of an insurance product combining the guarantees of the products analyzed in the first two parts. The dynamic Bayesian network developed here makes it possible to synthesize the regressive model of the historical product seen in the first part and the predictive model of the new product seen in the second part. This type of model, allowing random variables whose parameters change over time to interact, presents many possible fields of application in insurance, particularly on products marketed by Europ Assistance.

Keywords: health insurance, segmentation, CART, Covid-19, compartmental model, SIR, Bayesian networks

Remerciements

À ma chère épouse, ingénieure de son état, pour son soutien moral et logistique tout au long de la rédaction de mon mémoire.
À mes enfants, pour leur patience pendant que leur papa planchait encore sur son mémoire pendant les vacances..
À mes parents, pour m'avoir donné le goût de l'effort.
À mes beaux-parents pour leur accueil et leur soutien de baby-sitting pendant cet été studieux.
À la météo bretonne qui m'a donné envie de rester travailler à la maison pendant ces vacances.
À mes collègues d'Europ Assistance, notamment Edoardo d'EA Italie, pour leurs données, leurs conseils et leur amitié.
À Étienne, Chiara et Olivier de m'avoir soutenu dans ma démarche de VAE pour devenir un actuaire en bonne et due forme.
À l'équipe pédagogique de l'EURIA, notamment Dominique, Franck, Romain et Anthony, pour leurs enseignements, leurs conseils et leur sympathie.
A Auguste et tout le Groupe de travail sur la Microassurance de l'IA qui m'accueille en son sein depuis 5 ans bien que n'étant pas encore officiellement membre de l'Institut.

Table des matières

Introduction

Le système de santé italien est un service public de santé (Servizio Sanitario Nazionale (SSN)) organisé au niveau régional qui offre une couverture universelle et des soins largement gratuits. Au niveau national, le Ministère de la santé, appuyé par plusieurs institutions spécialisées, fixe les principes et objectifs fondamentaux du système de santé, détermine l'ensemble des prestations de santé de base remboursables dans tout le pays et alloue des fonds nationaux aux régions, en utilisant les ressources collectées par la fiscalité générale. Les régions sont responsables de l'organisation et de la prestation des soins de santé. Au niveau local, les autorités sanitaires locales fournissent des services de santé publique et des services de soins primaires directement, ainsi que des soins secondaires et spécialisés, soit directement, soit par le biais d'hôpitaux publics ou de prestataires privés conventionnés [1].

La prise en charge des dépenses pour les soins de longue durée échoit aux municipalités pour les services sociaux et aux unités sanitaires locales pour le service des soins de santé et certains services sociaux. Dans le contexte du vieillissement rapide de la population, les soins de longue durée représentent l'un des principaux domaines où une intégration efficace entre les soins de santé et les services sociaux est plus que jamais nécessaire. Toutefois, les soins de longue durée sont encore caractérisés par la multiplicité des prestataires et des modes de prestation et on constate historiquement dans la population italienne beaucoup de dépenses santé «out of the pocket». En conséquence, le marché de l'assurance privée est en forte croissance depuis plusieurs années.

Dans ce contexte, Europ Assistance (EA) Italie a développé un produit modulaire d'assurance et assistance visant à couvrir les besoins quotidiens de santé et les événements plus graves (accidents de la vie, hospitalisation, rééducation, entrée en dépendance). Ce produit *EURA Salute 360°* a été tarifé historiquement sur un modèle assez simple de «burning cost». La première partie de ce mémoire a pour objectif d'explorer différents modèles actuariels afin de proposer une nouvelle segmentation et une tarification plus fine, notamment des garanties d'indemnisation forfaitaire en cas de diagnostic d'affection longue durée ou d'accident grave.

1. Source : Centre des Liaisons Européennes et Internationales de Sécurité Sociale (Cleiss)

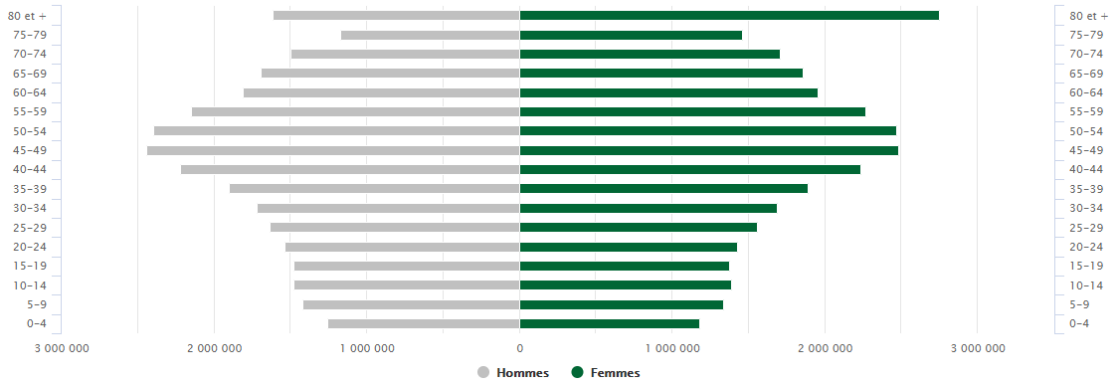

FIGURE 1 – Pyramide des âges de la population italienne recensée en 2018 - source Banque mondiale

En parallèle de ces travaux structurels de modélisation et de tarification sur le portefeuille historique *Eura Salute 360°* en Italie, l'épidémie de Corona Virus Disease apparue fin 2019 (Covid-19) a suscité une nouvelle demande conjoncturelle des entreprises clientes d'Europ Assistance pour la protection de leurs salariés. Nous avons donc cherché à créer dès avril 2020 un nouveau produit santé spécifique, appelé *Covid-19 Protezione*, offrant des garanties d'assistance et d'indemnisation forfaitaire en cas d'hospitalisation de personnes malades du Covid-19. Le fort risque de concentration sur ce produit nous a conduit à modéliser la dynamique de l'épidémie, en adaptant un modèle compartimental de type Susceptible Infected Removed (SIR).

La dernière partie de mon mémoire cherchera à faire le lien entre la sinistralité du produit santé italien historique (*EURA Salute 360°*) et celle du nouveau produit Covid-19 afin d'envisager un nouveau produit exhaustif couvrant les maladies chroniques traditionnelles aussi bien que les épidémies de Corona Virus Disease (Covid) avec tous leurs variants potentiels. Nous déterminerons à cette fin des probabilités conditionnelles entre les différents faits générateurs à partir d'une autre source de données venant des déclarations de sinistres d'annulation de voyages durant toute la période de l'épidémie sur notre portefeuille d'assurance voyages. Enfin, nous chercherons à faire interagir les deux modèles de fréquences respectifs du produit *EURA Salute 360°* et *Covid protection* en utilisant des méthodes de réseaux bayésiens.

Europ Assistance a un rôle pionnier et une agilité qui lui permettent de développer et de mettre sur le marché rapidement des produits de niche, dont la prime et la matérialité demeurent faibles grâce à des limites assez basses de montants couverts et d'exposition temporelle comme en nombre de polices. Néanmoins, dans ce contexte de pandémie d'une ampleur sans précédent et dont les développements sont largement imprévisibles étant donnés le nombre de paramètres - épidémiologiques, sociaux, médicaux et politiques - entrant en jeu, il est important de rappeler la politique de risques et de souscription du

groupe Generali dont Europ Assistance fait partie intégrante. Une étude d'impact de la pandémie sur l'ensemble de notre portefeuille d'assurance et d'assistance voyage a d'abord été réalisée (cf. partie 3.2 de ce mémoire) avant de considérer le développement potentiel d'un nouveau produit *Covid protection*. Puis, une première limite de souscription d'1 million de polices *Covid protection* individuelles a été validée avec Generali Head Office (GHO). Etant distribuées essentiellement comme des polices de groupes de type *employee benefits*, le risque de dépasser le seuil à chaque nouvelle souscription par une entreprise poue l'ensemble de ses employés a dû être surveillé et contrôlé de près.

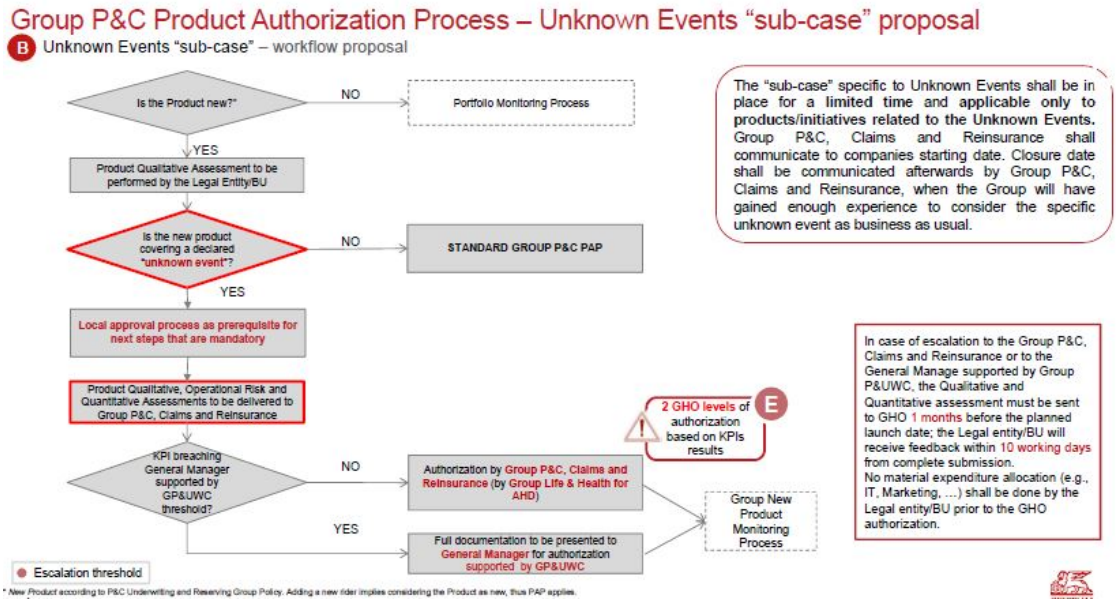

FIGURE 2 – Diagramme de décision pour le lancement et le suivi de nouveaux produits dans un contexte inconnu

Enfin, les formes dites *longues* de Covid ne sont pas prises en compte dans cette étude puisque le produit d'assurance proposé est très limité dans le temps (durée d'hospitalisation de maximum 10 jours) mais celles-ci feront certainement l'objet de nouvelles problématiques de santé publique auxquelles les assureurs risquent d'êtres confrontés dans les prochaines années.

Chapitre 1

Sophistication de la tarification du produit santé italien *EURA Salute 360°*

1.1 Design du produit et caractéristiques techniques des garanties indemnitaires

EURA Salute 360° est à l'origine un produit simple d'assistance médicale qui a été enrichi au fil des années par des garanties multiples et variées : remboursement de frais médicaux et de rééducation, forfait hospitalier, versement d'un capital en cas de diagnostic d'affection de longue durée (ALD) ou d'accident de la vie, aide à l'entrée en dépendance.

La nature du risque et les prestations associées ont donc progressivement évolué d'un service de mise à disposition d'un réseau médical et d'une plateforme de chargés d'assistance (les 2 actifs principaux d'Europ Assistance) vers de l'assurance indemnitaire plus traditionnelle, certaines garanties se rapprochant même de produits d'assurance vie mais avec un fait générateur bien identifié pour quand même rester dans le contexte et le mandat d'Europ Assistance dont la raison d'être d'entreprise est *From distress to relief, anytime, anywhere*. D'ailleurs, ce produit a parfaitement trouvé sa place sur le marché de l'assurance italien, dominé par Generali dont les entités spécialisées en assurance santé et en prévoyance n'ont pas vu en ce produit labellisé Europ Assistance un concurrent direct de leurs produits traditionnels. On pourrait se poser la question de la pertinence d'un tel produit sur le marché français par exemple où le reste à charge des ALD reste élevé pour les ménages concernés. J'avais d'ailleurs participé en 2016-2017 aux travaux de la commission santé de l'Institut des Actuaires (cf. [A.Dieng, 2017]) sur ce sujet qui avait été sollicité par la commission des affaires sociales du Sénat.

Le produit *EURA Salute 360°* d'Europ Assistance Italie est composé de 5 familles de couvertures :

- Assistance médicale dans la vie de tous les jours
- Remboursement de frais médicaux
- Versement d'un capital en cas de diagnostic d'ALD
- Versement d'un capital en cas d'accident grave
- Aide à l'entrée en dépendance

La sophistication en couches successives du produit avec un catalogue d'options pour chaque famille de couvertures a été accompagnée par la production de grilles de prix de plus en plus difficiles à lire et perdant progressivement le fondement actuariel qui avait prévalu à ses origines, notamment sur la modélisation de la fréquence. J'ai donc fait le choix de concentrer mon étude sur les couvertures de versement d'un capital (en cas d'ALD et/ou accident grave), celles-ci étant non seulement les moins courantes chez Europ Assistance mais aussi perçues comme les plus risquées car les montants d'indemnisation en cas d'accident grave entraînant une invalidité totale ou un décès peuvent atteindre jusqu'à 250 000€, montant largement supérieur aux sinistres individuels généralement observés sur les portefeuilles traditionnels d'Europ Assistance.

Les conditions générales du produit prévoient une liste de maladies ou d'états graves résultant d'un accident pouvant faire l'objet d'un versement de capital, dont le montant dépendra, d'une part, des niveaux de garanties souscrites (faible, moyen ou élevé) et, d'autre part, de la gravité du diagnostic :

- Les couvertures pour invalidité résultant d'un accident sont rangées selon 2 classes de gravité (voir détails dans la figure 1.1)

LESIONI PARTICOLARI				
STATO DI COMA VEGETATIVO PERSISTENTE POST TRAUMA CRANICO	V	€ 25.000,00	€ 50.000,00	€ 100.000,00
ASPORTAZIONE CHIRURGICA DI PARTE DI TECA CRANICA (indipendentemente dall'estensione della breccia)	II	€ 1.250,00	€ 2.500,00	€ 5.000,00
ROTTURA MILZA CON SPLENECTOMIA	II	€ 2.500,00	€ 5.000,00	€ 10.000,00
ROTTURA RENE CON NEFRECTOMIA	II	€ 3.000,00	€ 6.000,00	€ 12.000,00
ESITI EPATECTOMIA (oltre un terzo del parenchima)	II	€ 2.500,00	€ 5.000,00	€ 10.000,00
EPATITI TOSSICHE O INFETTIVE (con test enzimatici e sieroproteici alterati e con bilirubinemia oltre i valori normali)	II	€ 3.750,00	€ 7.500,00	€ 15.000,00
ERNIA CRURALE O IPOIEPIGASTRICA O OMBELICALE O DIAFRAMMATICA (trattate chirurgicamente)	I	€ 200,00	€ 400,00	€ 800,00
PERDITA ANATOMICA DI UN GLOBO OCULARE	IV	€ 7.600,00	€ 15.200,00	€ 30.400,00
CECITA' MONOLATERALE (Perdita irreversibile non inferiore a 9/10 di visus)	III	€ 6.300,00	€ 12.600,00	€ 25.200,00
PERDITA TOTALE DELLA FACOLTA' VISIVA DI AMBEDUE GLI OCCHI	V	€ 25.000,00	€ 50.000,00	€ 100.000,00
SORDITA' COMPLETA UNILATERALE	II	€ 3.000,00	€ 6.000,00	€ 12.000,00
SORDITA' COMPLETA BILATERALE	III	€ 6.300,00	€ 12.600,00	€ 25.200,00
PERDITA NASO (oltre i due terzi)	II	€ 3.750,00	€ 7.500,00	€ 15.000,00
PERDITA LINGUA (oltre i due terzi)	III	€ 6.300,00	€ 12.600,00	€ 25.200,00

FIGURE 1.1 – Exemples d'indemnisation forfaitaire résultant d'un accident grave

- Les couvertures pour ALD résultant d'un diagnostic de maladie sont rangées selon 2 classes de gravité (voir détails dans la figure 1.2)

ELENCO GRAVI MALATTIE E LORO CLASSE DI APPARTENENZA		
1.	Diabete I e II tipo con retinopatia	I
2.	Diabete complicato con neuropatia, vasculopatia e retinopatia periferica	II
3.	Insufficienza cardiaca iniziale (I-II classe NYHA)	I
4.	Insufficienza cardiaca con edemi e/o aritmie (II-III classe NYHA)	II
5.	Insufficienza cardiaca con stasi polmonare o cuore polmonare (IV classe NYHA)	II
6.	Infarto miocardico acuto	I
7.	Angina instabile	II
8.	Fibrillazione atriale cronica (in trattamento)	I
9.	Anomalie della conduzione del ritmo	I
10.	Insufficienza respiratoria, caratterizzata da trattamento farmacologico e/o riabilitativo	II
11.	Insufficienza respiratoria, caratterizzata da ossigenoterapia h24 e/ventilazione domiciliare	II
12.	Insufficienza renale acuta	I
13.	Insufficienza renale cronica in trattamento dialitico (emodialitico o peritoneale)	II
14.	Neoplasie maligne: accertamenti e cure	II
	se con interessamento linfonodale o metastatico a distanza	II
15.	Gravi osteoartropatie e collagenosi con gravi limitazioni funzionali che comportino anchilosi o rigidità articolari superiori al 50%	I
16.	Tetraplegia	II
17.	Sclerosi multipla	I
18.	Malattia del I e II motoneurone	II
19.	Ictus e/o emorragie cerebrali con gravi reliquati neurologici	II
20.	Morbo di Parkinson, purchè caratterizzato da marcata riduzione dell'attività motoria e dal mantenimento con difficoltà della stazione eretta	II
21.	Stato di coma	II
22.	Morbo di Alzheimer (diagnosticato clinicamente)	II

FIGURE 1.2 – Liste exhaustive des ALD couvertes classées selon 2 niveaux

1.2 Qualité des données, retraitement de la base d'étude et premières analyses

Les bases de données utilisées pour mon étude m'ont été fournies par la direction actuarielle d'EA Italie et résultent de 11 années d'historique (de 2011 à 2021) de ventes et de sinistres sur le produit italien *EURA Salute 360°*. Comme décrit précédemment, la nature des couvertures et des prestations a évolué au fil des ans pour arriver aujourd'hui à un produit assez exhaustif, avec de nombreuses garanties en option. J'ai donc créé une nouvelle maille *nature de la garantie* plus agrégée en identifiant toutes les garanties indemnitaires (par opposition aux prestations de service) qui pouvaient se trouver dans différents sous-produits. En effet, le nombre de modalités possibles en combinant toutes les options était trop grand par rapport au nombre de polices et de sinistres de notre base d'étude pour pouvoir établir un modèle de fréquence différencié pour chacune de ces modalités. En revanche, les différentes options de limites de montants d'indemnisation permettront quand même de différencier les primes en combinant avec notre modèle unique de fréquence différents modèles de coûts en fonction des options proposées dans le produit *EURA Salute 360°*.

Base des ventes : Elle est constituée d'un historique d'environ 50 000 polices d'assurance qu'il a fallu découper par exercices pour lesquels j'ai dû faire un calcul d'exposition ainsi que le calcul, à partir de la date de naissance de chaque assuré, de son âge moyen pour les exercices où sa police d'assurance était active. J'ai ensuite regroupé les modalités de la variable âge en 5 tranches d'âges qui me paraissaient optiquement relativement homogènes (cf. figure 1.3) quand on considère tous les sinistres de façon indifférenciée entre faits générateurs.

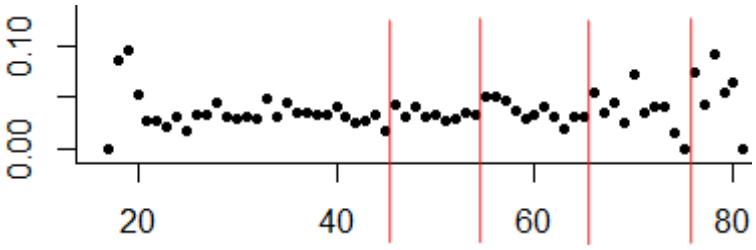

FIGURE 1.3 – Fréquences empiriques par âge des assurés et seuils de regroupement retenus.

En raison de la nature variée des couvertures, il a fallu procéder à la segmentation de la base des assurés en regroupant ceux-ci entre 3 macro-couvertures possibles :

- C1 : versement d'un forfait hospitalisation pour maladie et d'un capital en cas d'ALD
- C2 : versement d'un forfait hospitalisation pour accident et d'un capital en cas d'invalidité ou de mort

- C3 : l'assuré a souscrit simultanément dans sa police les 2 couvertures C1 et C2.

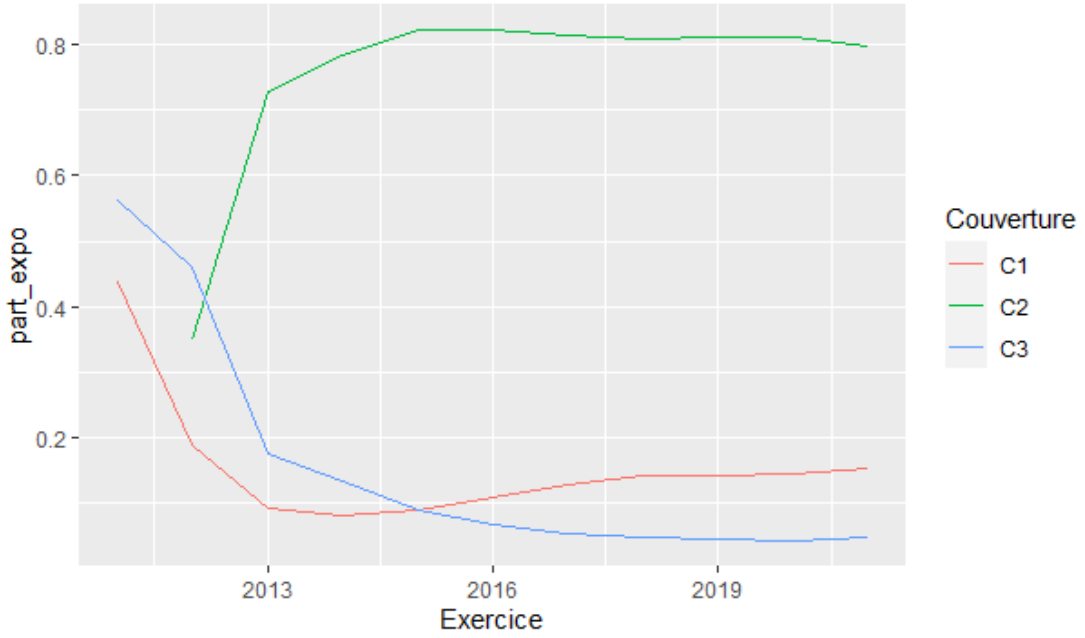

FIGURE 1.4 – Evolution de l'exposition relative de chaque couverture par années de souscription

Il est intéressant de constater que lorsque le produit a été lancé en 2011, les couvertures vendues étaient soit uniquement des couvertures pour maladie (type *C1*) soit des couvertures complètes maladie et accident (type *C3*) mais pas des couvertures pour versement d'un capital en cas d'accident uniquement (type *C2*), celles-ci étant probablement les plus lointaines de l'appétit au risque d'Europ Assistance à l'époque. Puis, face à la demande du marché, la tendance s'est progressivement inversée pour arriver à une grande majorité ($\simeq$ 80%) de couvertures de type *C2* souscrites, tendance stable depuis 2015 (figure 1.4).

Base des sinistres : Une étape essentielle de l'étude que nous cherchons à réaliser est le regroupement des sinistres à la maille Police x Exercice pour pouvoir établir pour chaque police à chaque exercice actif un compteur N_sini du nombre de sinistres encourus, ainsi que la somme des coûts de l'ensemble des prestations de l'assurance, et enfin le calcul de la variable $Ante\_sini$, indicatrice égale à 1 si un assuré a déjà déclaré un sinistre les années précédentes.

L'objet de ce mémoire se concentre sur l'analyse et la modélisation des fréquences de nos portefeuilles d'assurés mais devra être complété à terme par une analyse similaire de

la sévérité puis de l'interaction entre ces deux facteurs de risques dont nous pouvons voir l'évolution dans le temps sur le graphique en figure 1.5.

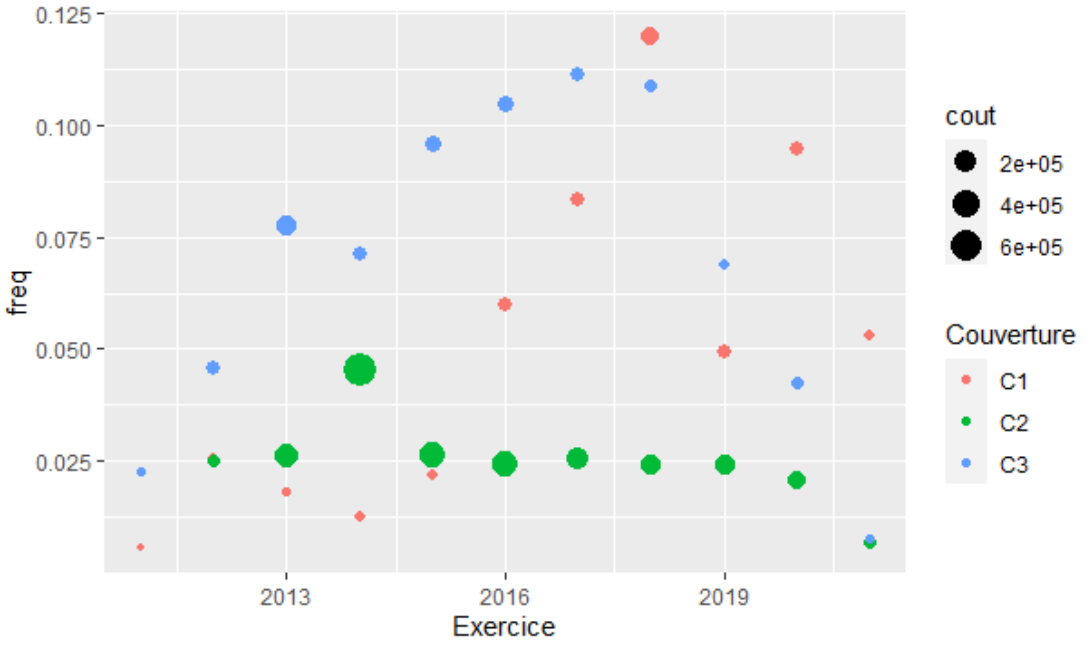

FIGURE 1.5 – Evolution dans le temps des fréquences et coûts moyens de chaque couverture

Sous réserve d'indépendance entre les coûts individuels des sinistres et leur nombre d'occurrence, la théorie du modèle collectif permet de modéliser la sinistralité en espérance comme le produit des espérances respectives de la fréquence et du coût moyen, ce qui est en pratique le modèle le plus répandu aujourd'hui en tarification non-vie, mais d'autres méthodes statistiques existent pour sortir du cadre de l'hypothèse forte d'indépendance mentionnée ci-dessus, comme nous le verrons en troisième partie de ce mémoire.

Rappelons ici que notre produit ne comporte aucune garantie à développements longs comme cela peut être le cas dans des portefeuilles d'assurance santé plus classiques. En effet, ce produit, certes atypique dans le paysage d'Europ Assistance, est de rester relativement proche du concept d'assistance et donc d'apporter un service ou un secours financier ponctuel en cas d'événement inattendu. L'exercice de provisionnement sur ce produit est donc relativement simple car le coût potentiel du sinistre est souvent entièrement connu lors de sa déclaration. Nous avons d'ailleurs ces deux variables bien distinctes (coût à l'ouverture du dossier et coût réel) dans notre base de données et j'ai pu observer très peu d'écart entre les deux. Les éventuels écarts proviennent de sinistres déclarés mais jugés non recevables par nos chargés de sinistres et médecins. Nous ne retiendrons donc

que les coûts réels pour déterminer si un sinistre a eu lieu ou non (recensé par le compteur N_sini défini précédemment) et nous intéresserons dans ce mémoire à modéliser cette variable N_sini.

Quelques statistiques simples des coûts de sinistres nous révèlent néanmoins que nous ne sommes pas confrontés à des valeurs extrêmes (en raison de la nature des couvertures avec des limites de garantie plafonnées) et que la distribution de ces coûts de sinistres peut être approchée par une loi Gamma dont la fonction de densité suit bien l'histogramme de distribution empirique des coûts de sinistres de notre portefeuille comme on peut le voir dans la figure 1.6 :

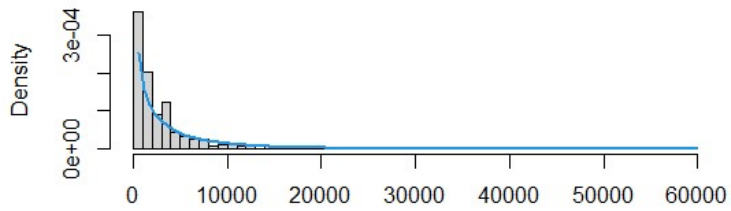

FIGURE 1.6 – Adéquation de la distribution des coûts de sinistres à la densité d'une loi Gamma

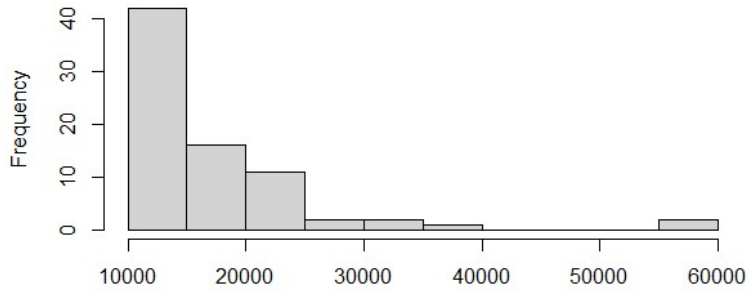

FIGURE 1.7 – Zoom sur la distribution des coûts de sinistres supérieurs à 10k€

Enfin, la fusion des bases retraitées des ventes et des sinistres nous donne notre base d'étude, sur laquelle nous allons opérer un ultime retraitement pour prendre en compte la disparité géographique du système de santé italien. Ce retraitement consiste à regrouper la variable *province de souscription* en 3 zones afin de pouvoir prendre en compte cette variable dans notre modélisation. Si les prestations octroyées par le Service national de santé sont les mêmes sur l'ensemble du territoire italien, les modalités de prise en charge varient d'une région à l'autre. La législation nationale prévoit, en effet, un certain nombre de tarifs de référence tout en laissant aux régions la possibilité d'adapter les dispositifs

prévus en fonction de leurs particularités, par exemple, en termes de population, de structures médicales présentes sur le territoire ou de fonds disponibles.

On peut voir la disparité des fréquences empiriques de notre portefeuille entre les différentes provinces sur la représentation graphique de la carte d'Italie figure 1.8 :

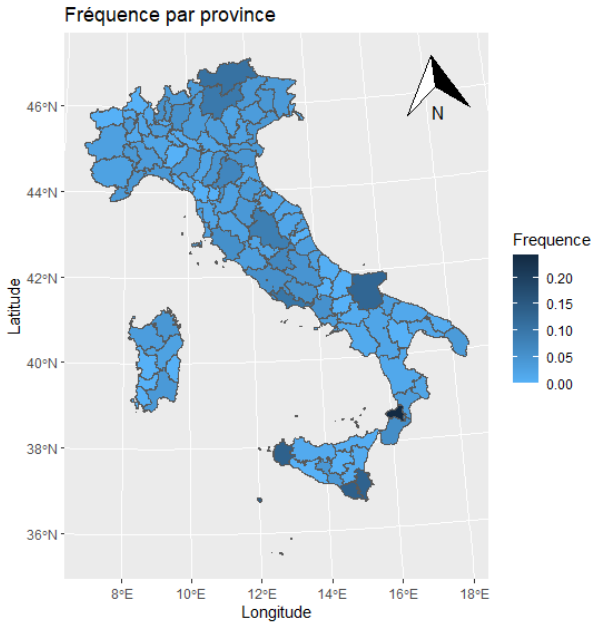

FIGURE 1.8 – Représentation graphique des fréquences empiriques par province

Il y a 112 provinces en Italie donc il était inenvisageable de garder une maille aussi fine pour la tarification que nous cherchons à sophistiquer en étant simple et lisible. J'ai donc décidé de créer un zonier pour répartir les 112 provinces en 3 zones sur la base des fréquences observées sur notre portefeuille en cherchant à maximiser la variance inter-classes et minimiser la variance intra-classe de ces zones. J'ai arbitrairement défini le nombre de zones à 3 au regard de la distribution des fréquences par province et afin d'avoir une homogénéité du nombre de modalités entre les variables catégorielles du modèle. La méthode de classification non hiérarchique par regroupement autour des centres mobiles m'a paru la plus appropriée dans ce contexte. La relation de Huygens permet de montrer que maximiser la variance interclasses afin d'avoir une séparation la plus claire possible de nos zones est équivalent à minimiser la variance intra-classe puisque leur somme est constante. L'algorithme de regroupement autour des centres mobiles procède par des améliorations successives d'une partition initiale arbitraire : $S^{(0)} = \{Z_1^{(0)}, Z_2^{(0)}, Z_3^{(0)}\}$.

Soit $j = 0$, nous effectuons alors une nouvelle partition $S^{(j+1)} = \{Z_1^{(j+1)}, Z_2^{(j+1)}, Z_3^{(j+1)}\}$ en regroupant les provinces dont les fréquences sont les plus *proches* (au sens de la dis-

tance définie comme la norme euclidienne sur $\mathbb{R}$) de chaque moyenne intra-classe. Ainsi la province p est affectée à $Z_i^{(j+1)}$ si $|p, p_i^{(j)}| = min\{|p, p_k^{(j)}|, k = 1,\ 2,\ 3\}$. Nous recommençons ce processus de partition de façon itérative pour $j := j+1$ jusqu'à ce que la variance de notre partition (égale à la somme des 3 variances intra-classes) soit suffisamment faible pour que nous puissions considérer que les 3 zones sont homogènes, c'est-à-dire dont les fréquences empiriques sont peu étendues autour de leurs moyennes respectives.

Nous obtenons alors la nouvelle représentation géographique avec les 112 provinces d'Italie réparties selon 3 zones à respectivement faible, moyenne et forte fréquence en figure 1.9 :

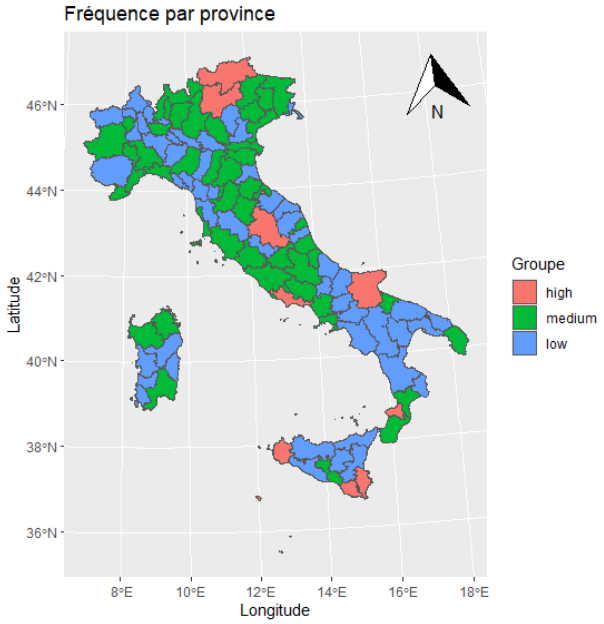

FIGURE 1.9 – Zonier calibré en fonction des fréquences empiriques par province

Une hypothèse forte sous-tend la création de ce zonier sur la base des fréquences empiriques observées dans notre portefeuille : une distribution identique et indépendante entre les provinces des autres variables explicative de la fréquence, tels que l'âge ou le type de couverture choisie. La taille limitée de notre portefeuille et du nombre d'assurés respectifs dans chacune des 112 provinces ne nous permet pas de tester statistiquement cette hypothèse de populations *iid* entre les provinces mais nous savons dans les faits que la réalité démographique et l'accès à l'offre de santé sont hétérogènes entre les différentes provinces d'Italie (cf Annexe A).

Avant de créer notre modèle statistique de fréquence, il est nécessaire de tester la corrélation entre les différentes variables explicatives et de décider si ces niveaux sont

acceptables pour ne pas introduire de biais statistique. Nous serons donc particulièrement attentifs à la corrélation entre la variable *Zonier* et les autres variables étant donnée la réflexion menée au paragraphe précédent.

La matrice des corrélations empiriques représentée en figure 1.10 résulte du calcul du V de Cramer pour chaque couple possible de variables selon la formule suivante :

Soit un couple de variables discrètes A et B aux valeurs respectives $A_1, ..., A_r$ et $B_1, ..., B_k$ représenté par un échantillon de taille n dont on note respectivement n_i, n_j et n_{ij} le nombre d'observations des valeurs A_i, B_j et (A_i, B_j).

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - \frac{n_i n_j}{n})^2}{\frac{n_i n_j}{n}}$$

$$\text{et } V \text{ de Cramer} = \sqrt{\frac{\frac{\chi^2}{n}}{min(k-1, r-1)}}$$

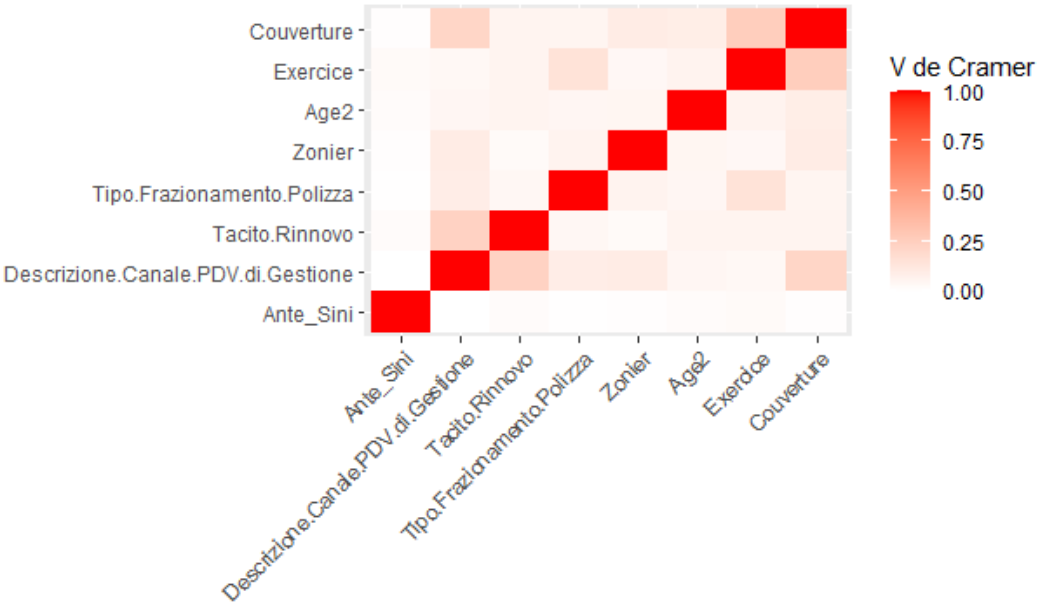

FIGURE 1.10 – Matrice des corrélations empiriques entre les variables explicatives

L'ensemble des variables étudiées précédemment seront donc retenues comme variables explicatives potentielles de notre modèle de fréquence (figure 1.11) :

```
$ N_sini                              : int  0 0 0 0 0 0 0 0 0 1 ...
$ Expo                                : num  1 1 1 1 1 ...
$ Exercice                            : int  2018 2020 2019 2020 2017 2021 2016 2021 2016 2015 ...
$ Couverture                          : chr  "C2" "C2" "C2" "C2" ...
$ Descrizione.Canale.PDV.di.Gestione: chr  "AGENTI" "AGENTI" "DIRETTO" "AGENTI" ...
$ Tacito.Rinnovo                      : chr  "SI" "SI" "SI" "SI" ...
$ Tipo.Frazionamento.Polizza          : chr  "Annuale" "Annuale" "Annuale" "Annuale" ...
$ Age2                                : chr  "a. <=45" "a. <=45" "d. 66-75" "#" ...
$ Ante_Sini                           : int  0 0 0 0 0 0 0 0 0 0 ...
$ Zonier                              : chr  "Z2" "Z1" "Z1" "Z1" ...
```

FIGURE 1.11 – Liste des variables du modèle et détail des premières lignes de la base de données

En sus des variables déjà présentées précédemment (et parfois même retraitées), il convient de décrire les 3 variables aux noms d'origine en italien :

Descrizione.Canale.PDV.di.Gestione **:** cette variable renseigne le canal de distribution par lequel la police d'assurance a été vendue, soit via de la vente en direct sur le site d'Europ Assistance (*DIRETTO*), soit via des agents généraux (*AGENTI*) ou bien des courtiers (*BROKER*). Il existe une quatrième modalité de cette variable (*SPORTELLI* ou accord de branche) qui n'est quasiment pas représentée dans notre échantillon, aussi l'ai-je regroupée avec la classe des courtiers dans la mesure où ces accords de branche sont des assurances de groupe souvent intermédiées et négociées par des courtiers.

Tacito.Rinnovo **:** cette variable est binaire et décrit si oui (*SI*) ou non (*NO*) la police d'assurance fait l'objet d'un renouvellement tacite.

Tipo.Frazionamento.Polizza **:** cette variable décrit si le paiement de la prime de la police d'assurrance se fait une fois par an à la date anniversaire de la souscription (*ANNUALE*) ou si elle est fractionnée par semestre (*SEMESTRALE*).

1.3 Sélection des variables par une recherche exhaustive et un arbre de régression

Afin de guider la construction du modèle de fréquence le plus efficace possible, c'est-à-dire à le juste compromis entre simplicité et précision de l'estimation des fréquences que nous cherchons à modéliser, nous cherchons au préalable dans cette section à déterminer, au sein de toutes ces variables explicatives, celles qui sont réellement discriminantes pour la prédiction de la fréquence.

J'ai choisi pour cela deux méthodes à base de régressions successives à la fois distinctes et complémentaires :

- Une première méthode de recherche dite « exhaustive » puisqu'elle va tester pour les N modalités de variables explicatives potentielles l'ensemble des 2^{N-1} modèles de régression linéaire possibles et les classer selon leur qualité de prédiction (un critère objectif comme le R^2 ajusté sera retenu).

- Une seconde méthode, issue de l'algorithme Classification And Regression Tree (CART), repose sur le partitionnement récursif et dyadique de l'espace des observations, ce qui se représente par un arbre binaire de décision ou arbre de segmentation.

La fonction « regsubsets » dans R applique une méthode de recherche exhaustive des meilleures variables explicatives à notre ensemble de données en appliquant la technique appelée «Subset Selection» (cf. [Hastie *et al.*, 2001]) qui fournit, pour chaque $k \in \{1, ..., n\}$, le meilleur sous-ensemble de variables de taille k, c'est-à-dire le sous-ensemble de taille k qui est associé au résidu le plus faible issu de la méthode des moindres carrés (figure 1.12) :

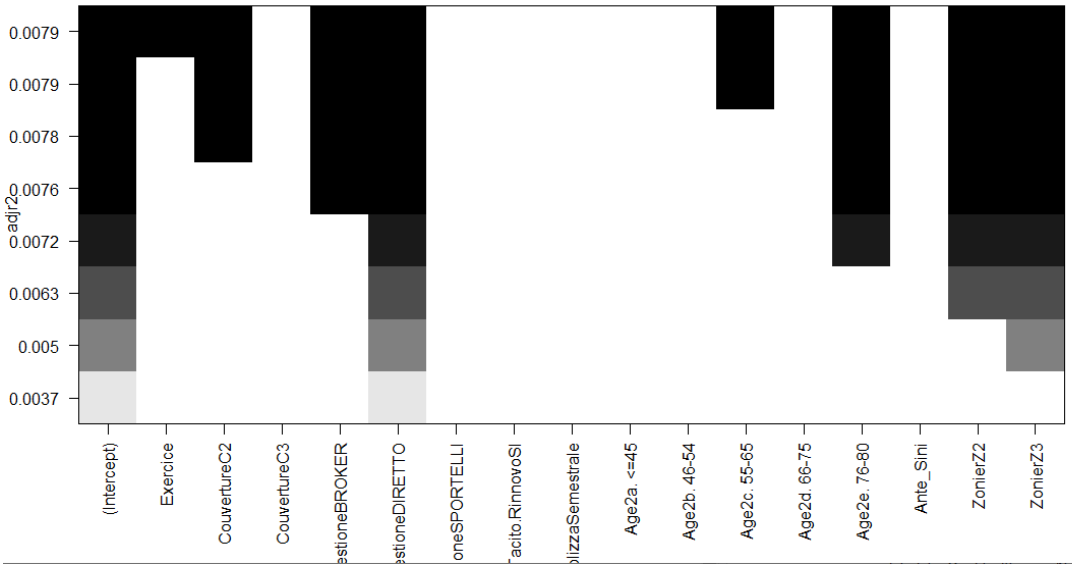

FIGURE 1.12 – Résultat de la fonction « regsubsets » dans R appliquée à notre base d'étude

L'algorithme CART a été introduit par Breiman, Friedman, Olshen et Stone en 1983 (cf. [Breiman *et al.*, 1983]). En partant de la racine, notée t_1, qui regroupe l'espace tout entier des observations et constitue le point de départ d'un arbre, l'algorithme consiste à partitionner cet espace en deux en affectant toutes les observations ayant une valeur répondant à un critère de segmentation pour une variable donnée (de type $\{X^j < s\}$ avec $s \in \mathbb{R}$ pour les variables quantitatives, et de type $\{X^j = C^l\}$ pour les variables de classes avec un nombre fini de modalités) au sous-arbre de gauche et les autres au sous-arbre de droite. La méthode sélectionne la meilleure coupure possible (i.e. le choix de la variable discriminante et le niveau de seuil ou la classe) en minimisant la variance intra-classe, c'est-à-dire la somme des distances totales respectives pour les 2 sous-arbres créés de la fréquence de chaque observation à la fréquence moyenne du sous-arbre (de la

même façon que nous l'avons vu précédemment pour la création du zonier). Une fois la racine de l'arbre ainsi partitionnée, on itère le processus sur chacun des deux sous-arbres obtenus, en recherchant de nouveau la coupure optimale à chaque nœud jusqu'aux nœuds terminaux appelés les feuilles de l'arbre et qui ne seront plus découpés car répondant à un certain critère d'arrêt.

Dans la pratique, nous faisons face à un problème d'optimisation sous contrainte dans la mesure où nous cherchons à sophistiquer le modèle de tarification actuel tout en simplifiant la lecture du modèle. L'étape d'élagage de l'algorithme CART va donc nous permettre de sélectionner le meilleur sous-arbre élagué à partir de l'arbre maximal en construisant une suite de sous-arbres élagués les uns des autres correspondant à une famille de partitions emboitées dont nous minimisons à chaque étape la variance intra-classe pénalisée par le nombre de feuilles du sous-arbre. Le graphique 1.13 représente l'évolution de l'erreur de prédiction (relative à l'erreur quadratique moyenne du modèle trivial consistant à affecter la fréquence moyenne à l'ensemble des polices) en fonction de la taille de l'arbre retenue, c'est-à-dire pour chaque sous-arbre élagué de la suite optimale définie précédemment. Il est obtenu dans R à partir de la fonction *rpart* (pour *recursive partitioning*).

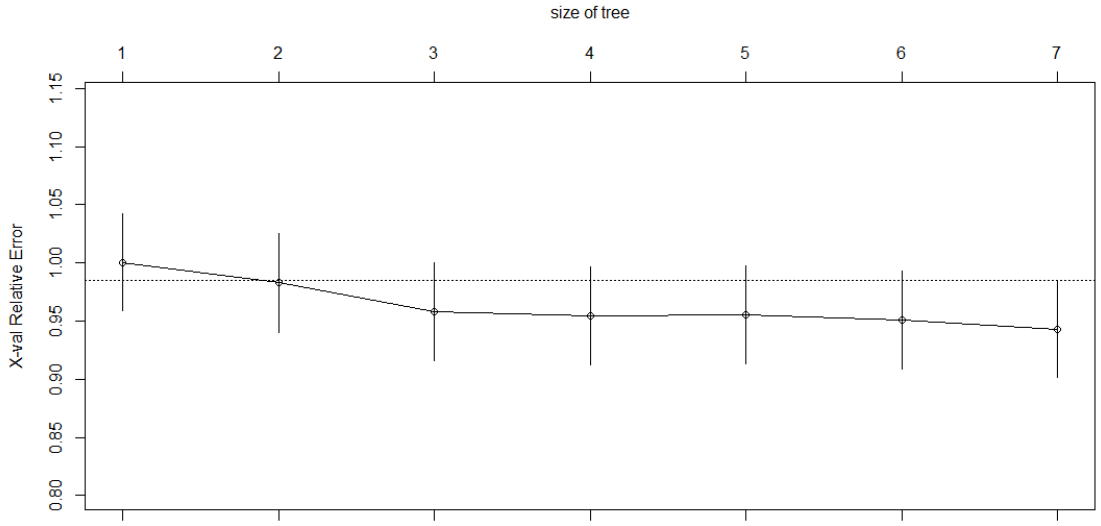

FIGURE 1.13 – Evolution de l'erreur de prédiction en fonction de la taille de l'arbre retenue

On voit graphiquement sur la figure 1.13 qu'un arbre à 3 niveaux semble être le bon compromis entre simplicité du modèle et minimisation de l'erreur quadratique moyenne. Voici en figure 1.14 l'arbre optimal de segmentation à 3 niveaux en sortie de l'algorithme CART appliqué à notre base de données. A chaque nœud (y compris la racine de l'arbre) les valeurs suivantes sont représentées dans un encadré : la valeur prédite pour la fréquence

du sous-groupe, le nombre de sinistres observés / le nombre de polices sans sinistres, le poids que représente le sous-groupe en termes d'effectif par rapport l'ensemble des polices du portefeuille.

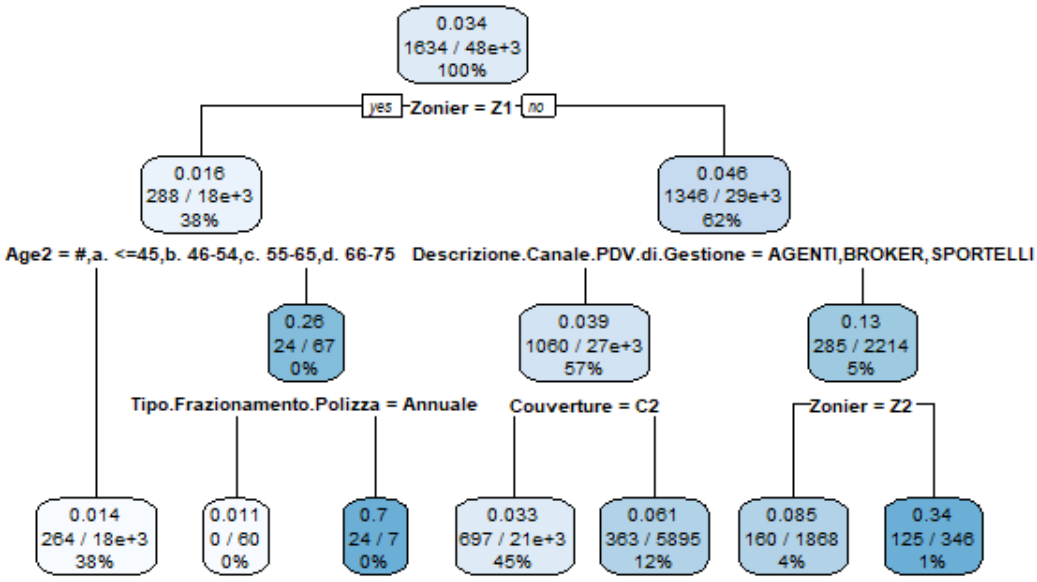

FIGURE 1.14 – Arbre optimal de segmentation à 3 niveaux en sortie de l'algorithme CART

Un des avantages des arbres de décision est la lisibilité et la facilité à interpréter les résultats, grâce au graphique sous forme d'arbre que nous pouvons voir figure 1.14 et qui offre une compréhension naturelle du modèle. De plus, la liaison entre la variable expliquée et les variables explicatives peut être non linéaire contrairement à l'algorithme « regsubsets » vu précédemment. En revanche, il est important de noter que les arbres de type CART sont instables, dans le sens où une légère fluctuation de l'échantillon d'apprentissage peut provoquer une grande variation dans le résultat de la procédure CART, d'où l'importance de comparer les résultats des 2 méthodes de sélection de variables retenues et de ne pas baser notre modèle d'estimation de fréquences uniquement sur cet algorithme.

On voit donc dans les deux méthodes de sélection de variables étudiées dans cette section que sans surprise l'âge et le zonier sont deux des variables les plus discriminantes, ainsi que le mode de distribution de l'assurance, ce qui reflète bien la réalité de notre pratique quotidienne de souscription chez Europ Assistance où nous avons une distribution multi-canaux (B2B, B2C, B2B2C) qui peut être intermédiée par des acteurs très

variés (agents généraux, courtiers, holdings ou captives de grands groupes). L'impact sur la sinistralité de la sélection du risque par le distributeur est d'autant plus fort dans un contexte d'information et de comportement des assurés en constante évolution.

L'importance prépondérante de la variable *Zonier*, qui ressort en tête de notre arbre de segmentation, doit également nous inciter à la prudence quant à la réplicabilité et la tarification du produit *EURA Salute 360°* dans d'autres géographies que le marché italien.

Enfin, bien qu'importante et ressortant parmi les variables explicatives à sélectionner, la variable Couverture n'apparaît pas comme la plus discriminante alors que les faits générateurs entre diagnostic de maladie chronique et accident grave sont nettement différents. Cela résulte de la nature hybride du produit construit autour d'une police d'assistance privée qui vient en complément du système d'assistance publique.

1.4 Calibrage d'un GLM, test et validation de la nouvelle segmentation fréquence

La variable que nous cherchons à modéliser est N_sini, équivalente à la fréquence lorsqu'elle est rapportée à l'exposition (variable $Expo$) à une maille plus agrégée que chaque police individuelle. Celle-ci étant typiquement modélisable par une loi de Poisson (cf. figure 1.15), nous pouvons nous rapporter au cadre plus simple et formel de la statistique paramétrique avec un modèle de régression de type Generalised Linear Model (GLM) dont la fonction de lien est le logarithme :

$$\log(\frac{N_{sini}}{Expo}) \sim \beta \cdot X + e$$

$$\log N_{sini} \sim \log Expo + \beta \cdot X + e$$

En passant le $log(Expo)$ en « offset » dans la régression, on obtient l'estimateur suivant :

$$\widehat{N_{sini}} = exp(\hat{\beta} \cdot X)$$

Nous allons calibrer et tester plusieurs modèles GLM en ajoutant successivement de nouvelles variables explicatives, en commençant par un premier GLM avec une seule variable explicative de type classe à 3 modalités : la variable discriminante du canal de distribution, dont l'interprétabilité commerciale est particulièrement forte comme expliqué précédemment (figure 1.16).

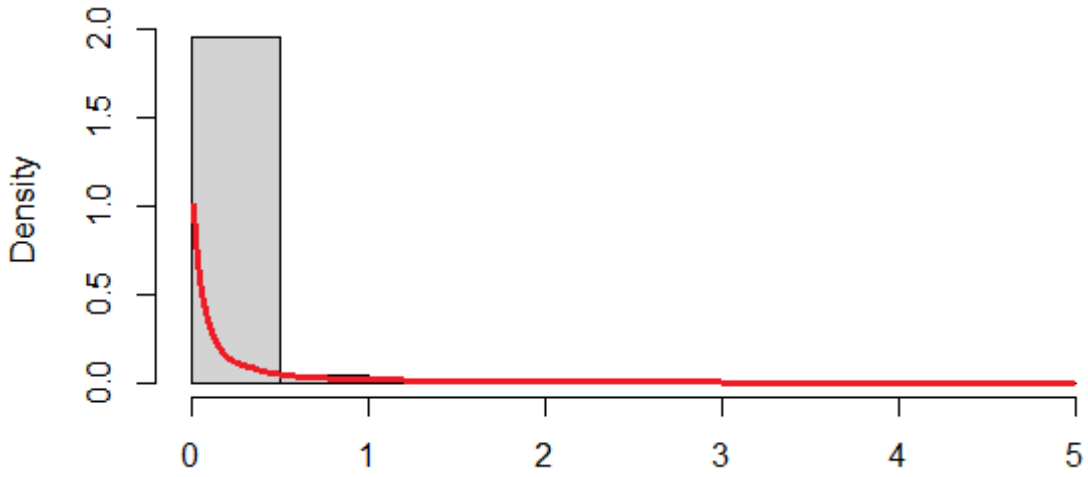

FIGURE 1.15 – Adéquation de la distribution des occurrences de sinistres à la densité d'une loi de Poisson

```
Call:
glm(formula = N_sini ~ offset(log(Expo)) + Descrizione.Canale.PDV.di.Gestione,
    family = poisson(link = "log"), data = bdd_select)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5766  -0.2313  -0.2265  -0.1588   6.5084

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               -3.62108    0.03714 -97.505  < 2e-16 ***
Descrizione.Canale.PDV.di.GestioneBROKER   0.50851    0.07506   6.774 1.25e-11 ***
Descrizione.Canale.PDV.di.GestioneDIRETTO  1.38300    0.07808  17.713  < 2e-16 ***
```

FIGURE 1.16 – Paramètres estimés du modèle GLM à une seule variable *canal de distribution*

L'intercept correspond au paramètre du canal de distribution *AGENTI*. Celui-ci est négatif alors que les paramètres respectifs des canaux *BROKER* et *DIRETTO* sont positifs, ce qui nous indique que les polices d'assurance distribuées par des agents généraux ont des fréquences en moyenne plus faibles que celles distribuées par des courtiers, et a fortiori par le canal de vente en ligne où les fréquences observées sont les plus élevées de l'ensemble du portefeuille (de fait, ces derniers sont des clients B2C qui vont de leur propre chef sur le site internet d'Europ Assistance Italie pour acheter ce type de polices sans intermédiation : leur connaissance de l'acte de souscription et de son coût n'est donc pas érodée par le discours d'une tierce personne ou fondue avec une multitude d'autres polices). L'arbre de segmentation réalisé dans la section précédente (figure 1.14 page 18) distingue d'ailleurs bien la vente directe des autres canaux de distribution.

Les fréquences estimées par ce GLM simple avec une seule variable explicative sont égales aux fréquences moyennes empiriques respectives de chaque sous-groupe correspondant aux 3 modalités de la variable *canal de distribution* (figure 1.17).

```
Descrizione.Canale.PDV.di.Gestione         V1  esti_fit2
                            AGENTI 0.02675387 0.02675387
                            BROKER 0.04448655 0.04448655
                           DIRETTO 0.10666316 0.10666316
```

FIGURE 1.17 – Fréquences observées et estimées pour chaque canal de distribution

Le GLM testé ensuite ajoute la variable *Couverture* au modèle en gardant toujours comme variable explicative le *canal de distribution*. L'intercept correspond cette fois-ci à la combinaison des 2 modalités suivantes : canal de distribution *AGENTI* et couverture *C1*. Toutes les autres combinaisons peuvent être réalisées à l'aide des 4 autres paramètres estimés sur la figure 1.18)

```
Call:  glm(formula = N_sini ~ offset(log(Expo)) + Descrizione.Canale.PDV.di.Gestione +
    Couverture, family = poisson(link = "log"), data = bdd_select)

Coefficients:
                              (Intercept)  Descrizione.Canale.PDV.di.GestioneBROKER
                                -3.2270479                                0.4148222
Descrizione.Canale.PDV.di.GestioneDIRETTO                              CouvertureC2
                                 1.0341096                               -0.5472289
                             CouvertureC3
                                 0.2539833
```

FIGURE 1.18 – Paramètres estimés du modèle GLM à 2 variables explicatives : *canal de distribution* et *Couverture*

Nous voyons ici que la fréquence de la couverture C2 (accident comme fait générateur) est plus faible que celle de la couverture C1 (diagnostic de maladie comme fait générateur) et que logiquement, la fréquence de la couverture C3 (couvrant au sein d'une même police les 2 types de faits générateurs) est la plus élevée, mais, comme nous le verrons ensuite, celle-ci demeure inférieure à la somme arithmétique des fréquences individuelles respectives des couvertures C1 et C2, ce qui signifie que les 2 types de faits générateurs ne sont pas indépendants d'un point de vue probabiliste. Intuitivement, on peut imaginer que le comportement d'un individu souffrant d'une affection de longue durée est sensiblement différent de celui d'une personne qui n'est pas malade, ce qui pourrait expliquer un risque plus faible d'occurrence d'un accident.

Notons que ce GLM avec 2 variables explicatives estime 5 paramètres (y compris l'intercept) pour calibrer un modèle de fréquence à 9 modalités. Le nombre restreint de degrés de liberté explique que les estimations de fréquence issues de ce GLM (cf. $esti\_fit3$ dans

la figure 1.19) ne correspondent plus parfaitement aux fréquences moyennes empiriques de chaque modalité croisée (cf. V1 figure 1.19).

Descrizione.Canale.PDV.di.Gestione	Couverture	V1	esti_fit3
AGENTI	C1	0.04730386	0.03967445
AGENTI	C2	0.02167467	0.02295369
AGENTI	C3	0.05553641	0.05114633
BROKER	C1	0.04925703	0.06007114
BROKER	C2	0.03596437	0.03475421
BROKER	C3	0.08608585	0.06593531
DIRETTO	C1	0.10091713	0.11158839
DIRETTO	C2	0.10590109	0.06455955
DIRETTO	C3	0.11637865	0.14385421

FIGURE 1.19 – Fréquences observées et estimées pour chaque combinaison entre les modalités de *canal de distribution* et de *Couverture*

Etant donnée l'importance de ces 2 variables explicatives et leur corrélation potentielle, comme le montrent les différentes études statistiques réalisées dans les sections précédentes, nous allons ajouter au modèle GLM l'interaction entre ces 2 variables *canal de distribution* et *couverture* comme variable explicative (figure 1.20) :

```
Call:  glm(formula = N_sini ~ offset(log(Expo)) + Descrizione.Canale.PDV.di.Gestione +
    Couverture + Descrizione.Canale.PDV.di.Gestione * Couverture,
    family = poisson(link = "log"), data = bdd_select)

Coefficients:
                                                  (Intercept)
                                                   -3.05116337
                                Descrizione.Canale.PDV.di.GestioneBROKER
                                                    0.04046008
                                Descrizione.Canale.PDV.di.GestioneDIRETTO
                                                    0.75770778
                                                  CouvertureC2
                                                   -0.78044768
                                                  CouvertureC3
                                                    0.16044698
                 Descrizione.Canale.PDV.di.GestioneBROKER:CouvertureC2
                                                    0.46592437
                Descrizione.Canale.PDV.di.GestioneDIRETTO:CouvertureC2
                                                    0.82865356
                 Descrizione.Canale.PDV.di.GestioneBROKER:CouvertureC3
                                                    0.39784605
                Descrizione.Canale.PDV.di.GestioneDIRETTO:CouvertureC3
                                                   -0.01789758
```

FIGURE 1.20 – Paramètres estimés du modèle GLM à 2 variables explicatives *canal de distribution* et *Couverture* avec leurs interactions

Nous retrouvons bien alors pour chacune des 9 modalités croisées des 2 variables *canal de distribution* et *couverture* une fréquence estimée correspondant à la fréquence moyenne empirique de chaque sous-groupe respectif (figure 1.21)

Descrizione.Canale.PDV.di.Gestione	Couverture	V1	esti_fit4
AGENTI	C1	0.04730386	0.04730386
AGENTI	C2	0.02167467	0.02167467
AGENTI	C3	0.05553641	0.05553641
BROKER	C1	0.04925703	0.04925703
BROKER	C2	0.03596437	0.03596437
BROKER	C3	0.08608585	0.09706025
DIRETTO	C1	0.10091713	0.10091713
DIRETTO	C2	0.10590109	0.10590109
DIRETTO	C3	0.11637865	0.11637865

FIGURE 1.21 – Fréquences observées et estimées pour chaque combinaison entre les modalités de canal de distribution et de Couverture dans ce nouveau GLM avec interaction croisée

En vertu de la sélection de variables réalisée dans la section précédente, notre GLM le plus complet comportera les variables explicatives suivantes : *Zonier*, *canal de distribution*, *Couverture*, *Age*2, *Exercice* et *Ante_sini*, auxquelles nous rajouterons l'interaction que nous venons d'étudier entre *Canal de distribution* et *Couverture*, ainsi que l'interaction entre *Couverture* et *Age*2 (les risques respectifs de maladie et d'accident grave n'étant pas les mêmes entre les différentes tranches d'âges) et enfin également l'interaction entre *Couverture* et *Exercice* (la tendance de la fréquence au fil des exercices pouvant varier d'une couverture à l'autre en fonction de l'évolution du comportement des assurés et de leur rapport à ce produit d'assurance).

Pour des raisons de confidentialité, nous ne montrerons pas ici le détail des coefficients d'ajustement en sortie de ce modèle exhaustif particulièrement sensible aux données propres au portefeuille d'Europ Assistance Italie mais nous pouvons néanmoins partager quelques résultats intéressants, comme le fait que le biais temporel (introduit par l'année d'Exercice comme variable explicative) soit positif pour la couverture C1 (maladie) alors qu'il est négatif pour la couverture C2 (accident) ou que la surprime pour la tranche d'âge la plus élevée (76 à 80 ans) soit quasi nulle pour la couverture C2 (accident) mais 3 fois plus élevée pour la couverture C3 (qui regroupe les mêmes faits générateurs que C1 et C2) qu'elle ne l'est pour la couverture C1 (maladie uniquement).

Afin de comparer la performance de ces différents GLMs, j'ai calculé pour chacun d'eux les critères suivants :

- La déviance, qui compare l'écart de log-vraisemblance entre un modèle théorique saturé qui aurait autant de paramètres que d'observations et le modèle calibré par notre GLM. Plus la déviance est faible, plus le modèle est bien ajusté aux données, ce qui est systématiquement le cas en rajoutant des variables, aussi peu explicative soit-elle (le simple fait de rajouter un paramètre augmente le nombre de degrés de liberté du modèle, indépendamment de la pertinence de la variable ajoutée). Dans le cadre d'une distribution de Poisson, comme c'est le cas pour notre famille de

GLMs, la formule explicite suivante permet de calculer la déviance :

$$\text{deviance(Poisson)} = 2 \times \sum_{i=1}^{\infty} \{ \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i) \}$$

- L'Akaike Information Criterion (AIC) cf. [Akaike, 1974]) pénalise la complexité du modèle en ajoutant à la déviance un terme propotionnel au nombre de paramètres du modèle. Contrairement à la déviance, l'AIC ne diminuera en ajoutant une variable explicative au modèle que si celle-ci a un impact suffisamment important sur l'ajustement du modèle aux données. C'est donc un critère discriminant dans notre cas où nous n'avons pas arrêté ex-ante le nombre de variables souhaitées dans notre modèle mais où nous sommes néanmoins vigilants à ne pas trop complexifier la grille de tarification qui découlera du modèle que nous aurons élaboré.
- La racine de l'erreur quadratique moyenne (RMSE) est calculée sur un échantillon-test de notre base de données après avoir calibré les estimateurs des paramètres des différents GLM sur un échantillon d'apprentissage (nous sélectionnons de façon aléatoire 80% des observations de notre base de données).

Au-delà des valeurs absolues de ces trois critères, nous allons surtout être attentifs pour chaque critère étudié individuellement à l'impact relatif qu'apporte l'ajout de nouvelles variables ou interactions de variables au modèle (cf. figure 1.22) :

	Variables ajoutées	Δ déviance	Δ AIC	Δ RMSE x 1000
fit2 vs fit1	*canal de distribution*	-258	-253	-64
fit3 vs fit2	*Couverture*	-108	-104	-39
fit4 vs fit3	*canal distrib * Couverture*	-26	-18	5
fit5 vs fit4	*Age2*	-58	-60	-9
fit6 vs fit5	*Age2 * Exercice*	-52	-21	16
fit7 vs fit6	*Ante_Sini, Zonier*	-181	-168	-107

FIGURE 1.22 – Comparaison des critères d'estimation des différents modèles GLMs créés

Les 3 critères étudiés valident le modèle le plus complet fit7 présenté précédemment comme étant celui aux meilleures qualités de prédiction sans trop complexifier l'interprétation de celui-ci. Nous pouvons voir aussi dans la figure 1.22 que les modèles fit4 et fit6 qui ajoutent des interactions de variables par rapport aux précédents (respectivement fit3 et fit5) peuvent conduire à des erreurs de prédiction plus importantes comme le révèlent les Root Mean Square Error (RMSE) sur la base de test. Cet effet de surapprentissage est mitigé par l'ajout de variables tierces au potentiel discriminant plus important.

Regardons à présent l'adéquation des fréquences estimées du modèle fit7 retenu aux fréquences empiriques observées sur notre échantillon de test pour chacune des modalités croisées des variables explicatives les plus importantes :

Canal de distribution x Couverture : La corrélation empirique du nuage de 9 points de la figure 1.23 est de 94,6%. Le biais induit par le canal de distribution dans la sinistralité de chacune des couvertures est donc bien capturé par le modèle GLM de fréquence que nous venons de valider.

Descrizione.Canale.PDV.di.Gestione	Couverture	Freq_obs	Freq_pred
AGENTI	C1	0.02669017	0.04291493
AGENTI	C2	0.01917049	0.01879572
AGENTI	C3	0.07432000	0.04966767
BROKER	C1	0.05280550	0.04713226
BROKER	C2	0.03369076	0.02839249
BROKER	C3	0.08808994	0.07062177
DIRETTO	C1	0.09297511	0.08475881
DIRETTO	C2	0.10053660	0.08552713
DIRETTO	C3	0.10971263	0.10495588

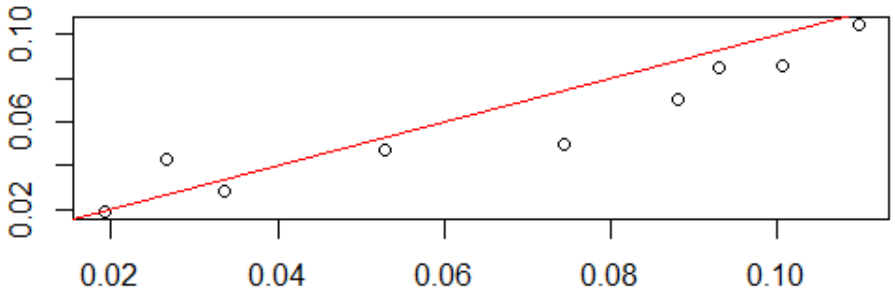

FIGURE 1.23 – Fréquences observées versus prédites pour Canal de distribution x Couverture

Zonier x Couverture : La corrélation empirique du nuage de 9 points de la figure 1.24 est de 91,1%. Le biais induit par le zonier dans la sinistralité de chacune des couvertures est donc relativement bien capturé par le modèle GLM de fréquence que nous venons de valider, même si la fréquence de la couverture C3 pour la zone Z3 est largement sous-estimée par le modèle. La convergence de la fréquence moyenne observée sur ce sous-groupe C3 x Z3 d'après le théorème centrale limite n'est pas assez significative en raison de la faible représentativité de cette modalité croisée sur notre échantillon de test de taille limitée (environ 10 000 observations).

Nous voyons néanmoins que la forte dispersion inter-classes de la variable *Zonier* justifie largement la prise en compte de cette variable dans notre modèle de tarification, ce qui n'était pas le cas jusqu'à présent, d'où le besoin de faire migrer notre portefeuille existant en même temps que les nouvelles souscriptions vers un nouveau modèle de tarification prenant en compte les variables discriminantes que nous avons mises en exergue

Zonier	Couverture	Freq_obs	Freq_pred
Z1	C1	0.004597245	0.02527185
Z1	C2	0.014716268	0.01172511
Z1	C3	0.066756507	0.03260086
Z2	C1	0.055013241	0.05648522
Z2	C2	0.026281424	0.02643032
Z2	C3	0.076464384	0.06745471
Z3	C1	0.066718454	0.08557466
Z3	C2	0.052180874	0.04775143
Z3	C3	0.186626712	0.12645018

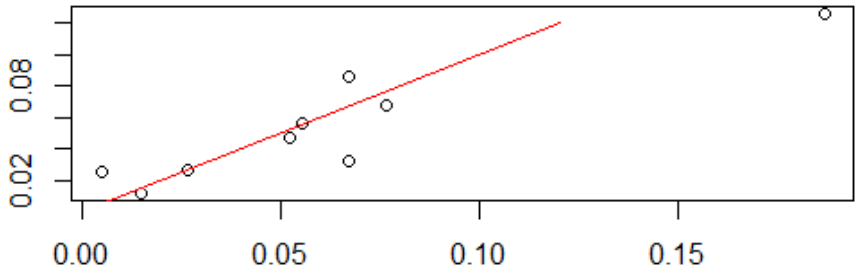

FIGURE 1.24 – Fréquences observées versus prédites pour Zonier x Couverture

dans cette étude. Nous allons pour cela avoir recours à la théorie de la crédibilité.

1.5 Migration du portefeuille existant avec recours à la théorie de la crédibilité

La grille de tarification actuelle de la couverture C1 (cf. figure 1.25) ne dépend que de la tranche d'âge de l'assuré (dont la segmentation est assez proche de celle que nous avons déterminée de façon empirique en construisant notre modèle) et du type de fractionnement de la police (*ANNUALE* ou *SEMESTRALE*). La grille de tarification actuelle de la couverture C2 est de structure similaire mais subdivisée en 3 sous-grilles en fonction des 3 niveaux possibles de montants de garanties (faible, moyen ou élevé).

INDENNITARIA DA MALATTIA	0-40	41-55	56-65	66-70	71-75	76-80
Premio Annuale	Pa1	Pa2	Pa3	Pa4	Pa5	Pa6
Premio Semestrale	Ps1	Ps2	Ps3	Ps4	Ps5	Ps6

FIGURE 1.25 – Structure de la grille de tarification actuelle du produit *EURA SALUTE 360°*

Les 2 variables explicatives les plus discriminantes d'après notre étude, *Zonier* et *canal de distribution*, ne sont donc pas du tout prises en compte aujourd'hui dans la

tarification. La problématique qui se pose est donc de les intégrer de façon dynamique dans notre tarification en faisant progressivement évoluer notre portefeuille de la tarification actuelle vers la tarification plus sophistiquée dont nous avons validé la pertinence dans notre étude.

La théorie de la crédibilité nous donne un cadre approprié pour cet exercice dans la mesure où nous avons constaté une forte hétérogénéité dans les fréquences de notre portefeuille (notre arbre de segmentation en figure 1.14 en est une bonne illustration avec un rapport de 1 à 3 rien qu'au premier nœud entre les fréquences des 2 sous-groupes homogènes de tailles relativement équivalentes) dont nous disposons d'un historique remontant jusqu'à 11 années. La théorie de la crédibilité propose une structure de prime pour l'année n+1 adaptée à l'expérience des années précédentes pour chaque classe de contrats existants c de la forme suivante :

$$p_{c,n+1} = \lambda \overline{X_c} = (1-\lambda)\overline{X}$$

où $\lambda \in [0,1]$ est le facteur de crédibilité qui accorde un poids plus ou moins important à l'historique des sinistres (dont l'information est synthétisée par l'estimation de prime issue de notre modèle de régression) et $\overline{X_c}$ la prime moyenne estimée par notre modèle de régression pour chaque classe de contrats c.

Dans le cadre de notre problématique de faire évoluer la tarification actuelle progressivement sans entraîner de variations de primes trop importantes pour un assuré d'une année sur l'autre, nous considèrerons X comme étant la prime souscrite actuellement par chaque tranche d'âges d'assurés, soit $\overline{X} = P_{a\ ou\ s}(k),\ k = 1, \ldots, 6$. Le loss-ratio de notre portefeuille étant historiquement bon et stable, nous savons que ce $\overline{X}$ nous permet de charger assez de primes pour couvrir tous les sinistres (mais que ces primes ne sont a priori pas réparties assez équitablement entre les assurés en raison des facteurs de risques discriminants que nous avons identifiés et qui ne sont pas pris en compte aujourd'hui, d'où l'exercice que nous sommes en train de mener qui devrait à terme permettre de limiter l'antisélection et améliorer la rentabilité du portefeuille).

Nous adaptons ici les résultats du modèle de [Bühlmann, 1967] en supposant le coût moyen des sinistres homogène sur l'ensemble du portefeuille des assurés et en nous ramenant donc à notre modèle GLM fréquence pour le calcul de λ et des $\overline{X_c}$.
Ainsi

$$\lambda = \frac{nM^2}{\sum^2 + nM^2}$$

avec $\sum^2$ la variance intra-classe (i.e. la moyenne des variances de chaque estimateur des fréquences de classes) et M^2 la variance inter-classe (i.e. la variance empirique des fréquences moyennes de classes estimées).

On voit que le facteur de crédibilité λ augmente avec n et tend vers 1 lorsque n tend vers l'infini : on donne donc plus de poids au fil des exercices à la sinistralité passée et au modèle de régression dont l'historique de sinistralité augmente. Cette nouvelle tarification dynamique fait conjointement évoluer le facteur de crédibilité et les facteurs d ?ajustement du GLM qu'on pourra recalibrer chaque année.

Nous verrons dans la troisième partie de ce mémoire un autre cadre de la théorie de la crédibilité avec une approche bayésienne.

Chapitre 2

Modèle dynamique de fréquence du nouveau produit *Covid-19 Protezione*

2.1 Design du produit et caractéristiques du sous-produit indemnitaire

L'épidémie de Covid-19 a suscité une attente forte de la part des entreprises clientes d'Europ Assistance pour proposer des solutions d'assistance et d'assurances adaptées à leurs nouveaux besoins qui ont émergé pendant cette crise d'une forme et d'une ampleur inattendues. Les médecins de nos plateaux d'assistance ont été fortement mobilisés pour répondre aux nombreuses questions des assurés dans les premiers mois de la crise sanitaire, notamment en Italie, un des premiers pays touchés en Europe. La téléconsultation médicale ainsi que l'assistance psychologique ont donc été largement plébiscitées par les entreprises comme avantages sociaux à mettre en avant auprès de leurs employés.

Devant le nombre accru d'hospitalisations, nécessitant pour les cas les plus graves des soins intensifs, de nouvelles dépenses inattendues ont dû être supportées par les personnes touchées par la maladie, notamment une fois de retour chez elles après un séjour à l'hôpital (parfois écourté en raison du manque de lits) : dépenses d'ordre médical comme certains frais optionnels d'hospitalisation, de transport en ambulance ou des consultations de suivi à domicile, l'achat de médicaments, ainsi que des dépenses d'aide à la personne pour faire ses courses et retrouver ses gestes du quotidien. C'est dans ce contexte qu'Europ Assistance Italie a lancé dès avril 2020 son nouveau produit *Covid-19 Protezione* proposant ses services d'assistance ainsi qu'une indemnité financière en cas d'hospitalisation pour faire face à ces dépenses inattendues.

Le produit *Covid-19 Protezione* ne s'adresse au départ qu'aux entreprises souhaitant offrir cette garantie à l'ensemble de leurs salariés dans le cadre d'une police d'assurance

collective, l'objectif étant de mutualiser au maximum le risque nouveau liée à l'épidémie dont l'évolution était (et est encore aujourd'hui) très incertaine et de limiter le risque d'antisélection. A ce titre, la politique de souscription mise en place pour ce nouveau produit avait pour effet d'éviter une trop forte concentration dans notre portefeuille de populations ou professions particulièrement exposées au virus (critère objectivé par le paramètre η que nous étudierons ci-après dans ce mémoire), la police d'assurance ne couvrant évidemment que les cas de Covid-19 diagnostiqués après la souscription de l'assurance par l'entreprise cliente. Quant aux services d'assistance, ceux-ci sont proposés pendant 14 jours après la sortie de l'hôpital (pour bien rester dans le cadre de la branche d'assistance, il ne s'agit pas de créer là de l'assurance dépendance, qui est une branche complètement différente d'assurance, avec des développements longs).

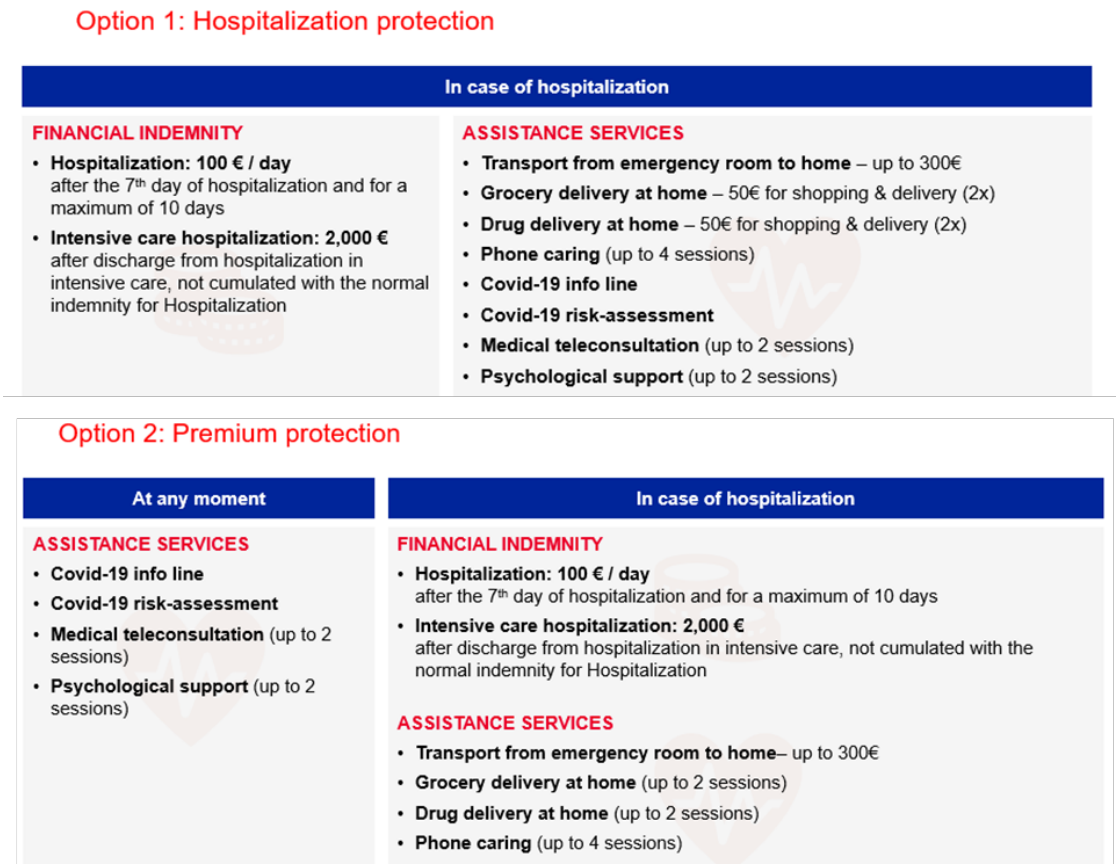

FIGURE 2.1 – Détail des garanties des couvertures standards du produit *Covid-19 Protection*

Deux couvertures standards ont été créées pour le produit *Covid-19 Protezione*, la première (cf. figure 2.1) étant restreinte au cadre des infections au Covid-19 entraînant une hospitalisation d'au moins 7 jours, alors que la deuxième offre en sus des services

auxiliaires d'assistance ne nécessitant pas de fait générateur particulier.

Un produit modulaire (cf. figure 2.2) a également été conçu afin de proposer aux entreprises demandeuses une offre sur mesure en choisissant parmi les couvertures standards vues précédemment les plus pertinentes pour leurs salariés et en y associant quelques garanties d'assistance supplémentaires, telles que le *phone caring* consistant à appeler de façon systématique et régulière pendant la période de crise sanitaire les employés bénéficiaires de la police de groupe pour s'assurer de leur état de santé et psychologique (stress potentiel lié à des proches malades, une période de chômage partiel ou des nouveaux modes de travail à distance).

Nous reviendrons sur ce concept de produit modulaire dans la troisième partie de ce mémoire où nous essaierons de construire un produit exhaustif aux faits générateurs aussi variés qu'un diagnostic de maladies chroniques, un accident ou une hospitalisation due à l'infection par une épidémie.

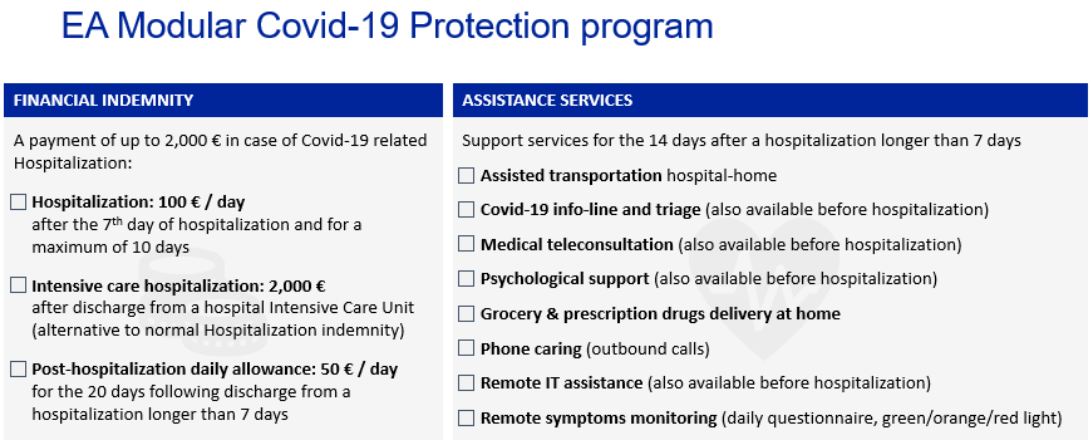

FIGURE 2.2 – Détail des garanties éligibles au produit modulaire *Covid-19 Protection program*

Nous nous concentrerons à nouveau ici sur le sous-produit indemnitaire aux garanties les plus coûteuses, et nous chercherons à modéliser sa fréquence, principal facteur de risque et particulièrement volatile. Celle-ci est aussi par construction directement lié l'évolution du nombre d'hospitalisions dues au Covid. Leur croissance à chaque nouvelle vague est exponentielle, comme n'importe quelle épidémie dont le phénomène se nourrit de lui-même. Les mesures de maîtrise comme les confinements ou la vaccination ont tout de même réussi jusqu'à présent à stopper les pics de croissance avec un décalage en moyenne de deux semaines entre l'introduction de la mesure et ses premiers effets (cf. figure 2.3 [1]).

1. Source : data.gouv.fr

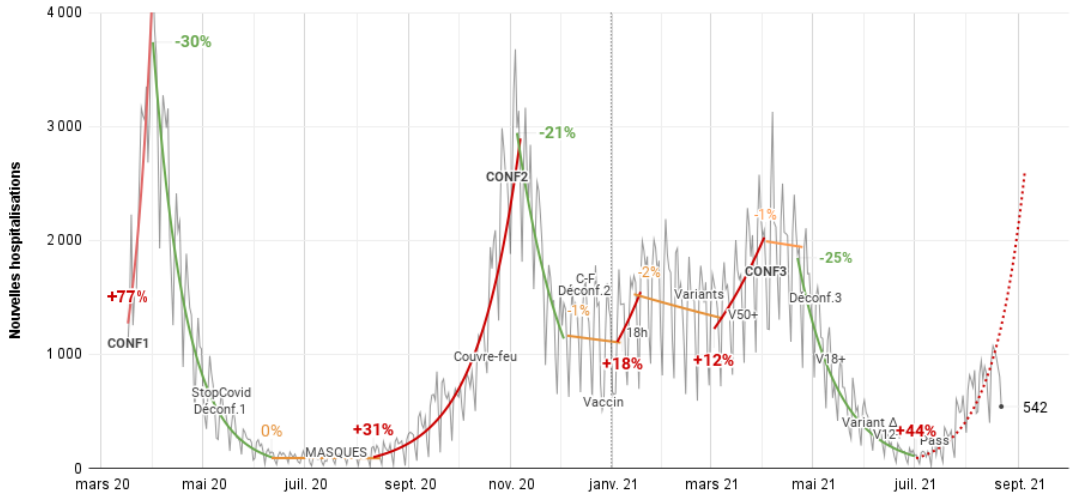

FIGURE 2.3 – Tendance hebdomadaire des nouvelles hospitalisations dues au Covid en France

La proportion de patients devant être admis en réanimation a également fortement fluctué au gré des évolutions de l'épidémie, de même que la mortalité liée à la maladie. La figure 2.4 représente l'évolution dans le temps du rapport de ces 2 indicateurs, ce qui permet de voir la tendance de la tension hospitalière depuis le début de la crise sanitaire (figure 2.4[2]) :

La nature particulière du risque induit par le produit *Covid-19 Protection* d'Europ Assistance a nécessité un suivi détaillé et quotidien de l'épidémie puis la mise en place d'un modèle de tarification dynamique permettant d'ajuster la prime à l'évolution de l'épidémie et des mesures sanitaires subséquentes, avec également une vision prospective, en particulier en Italie où nous avons lancé en premier ce produit au plus fort de la crise sanitaire.

2. Source : data.gouv.fr

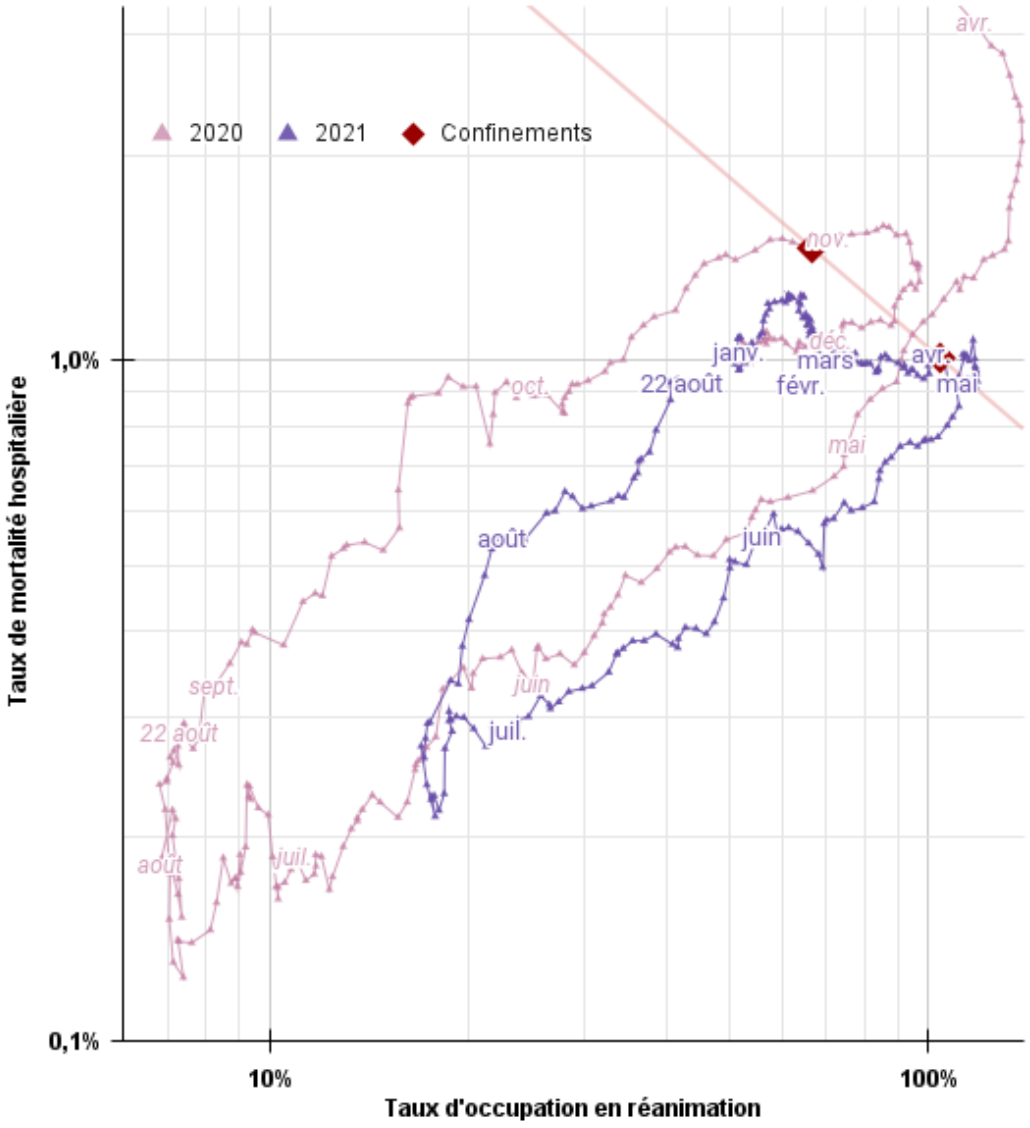

FIGURE 2.4 – Tendance de la tension hospitalière due au Covid en France

2.2 Calibrage du modèle SIR à l'épidémie de Covid-19 en Italie

Les modèles compartimentaux ont été mis en pratique pour faciliter les calculs de probabilité de contagion liées à un épidémie. Les compartiments divisent la population en divers états possibles par rapport à la maladie. Kermack et McKendrick (cf. [McKendrick, 1926]) proposèrent dans le contexte de la grippe espagnole un modèle fondateur dans lequel la population est divisée entre individus susceptibles de contracter la maladie (compartiment S) et individus infectés et contagieux (compartiment I). Ces derniers peuvent soit décéder de la maladie (et donc sortir de la population), soit en guérir et acquérir une immunité contre une la réinfection pendant une certaine durée (compartiment R pour *removed*). Ce modèle est connu sous le nom de SIR.

Le modèle SIR représente à l'aide d'équations différentielles l'évolution du nombre d'individus dans chaque compartiment au cours du temps :

- Notons p le paramètre de transmission de la maladie, c'est-à-dire la probabilité qu'une personne contagieuse transmette la maladie à une personne saine. Alors $\frac{dS(t)}{dt} = -p \cdot I(t) \cdot S(t)$, c'est-à-dire que l'effectif d'individus sains diminue à proportion du nombre d'individus déjà infectés et du nombre d'individus encore non-infectés.
- Notons γ le taux de convalescence et μ le taux de mortalité liés à la maladie. Alors $\frac{dI(t)}{dt} = p \cdot I(t) \cdot S(t) - (\gamma + \mu) \cdot I(t)$
- Notons σ le taux de déclin de l'immunité gagnée par les personnes guéries de la maladie. Alors $\frac{dR(t)}{dt} = \gamma \cdot I(t) - \sigma \cdot R(t)$
- Enfin, notons $N(t)$ l'effectif total de la population à l'instant t.
  Alors $N(t) = S(t) + I(t) + R(t)$ et $\frac{dN(t)}{dt} = -\mu \cdot I(t)$. Dans ce système fermé d'équations différentielles, nous faisons l'hypothèse d'une population stable dans le temps (hors décès liés à la maladie étudiée).

En appliquant ce modèle SIR à l'épidémie de Covid-19, nous pouvons décomposer le paramètre de transmission p pour prendre en compte les mesures sanitaires visant à restreindre le nombre de contacts quotidiens (confinement, fermeture d'écoles, télétravail généralisé, chômage partiel,...) ainsi que la campagne de vaccination lancée à la fin de l'année 2020 :

$$p = p_{t,vax} \cdot (1 - \theta) \cdot \eta \cdot \beta + (1 - p_{t,vax)} \cdot \eta \cdot \beta$$

où :

- $p_{t,vax}$ est la proportion de la population vaccinée à la date t
- θ est l'efficacité du vaccin, c'est-à-dire la proportion par laquelle la probabilité d'être infecté par un contact avec la maladie est réduite
- η est le nombre moyen de contacts quotidiens d'un individu dans la population

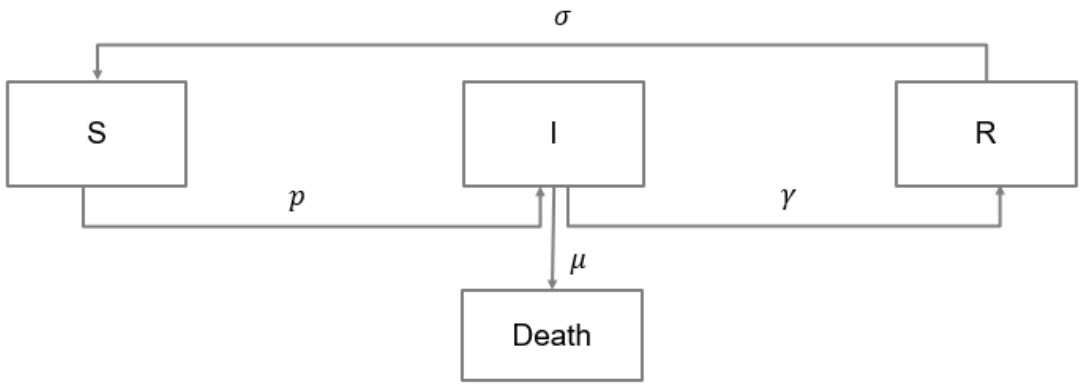

FIGURE 2.5 – Dynamique du modèle SIR et paramètres-clés

- β est la probabilité qu'une personne saine soit infectée au contact d'une personne contagieuse

Le pas de temps choisi pour notre modélisation est de 24 heures (en lien avec les publications quotidiennes des chiffres des autorités de santé). Le paramètre $p_{t,vax}$ varie quotidiennement au gré des campagnes de vaccination et doit être mis à jour à chaque pas de temps. Le paramètre η relève quant à lui des modes de vie et des habitudes comportementales de chaque sous-groupe de la population (cf. figure 2.6 [3]), il est donc a priori plus stable dans le temps mais il peut être fortement réduit de manière brutale par des mesures sanitaires restrictives comme les fermetures de crèches, d'écoles ou de commerces et l'instauration du télétravail de masse.

	Répartition de la population italienne			Paramètre η					
Tranches d'âge	Effectif	% du total	% actifs	Total	A la maison	A l'école	Au travail	Transports	Temps libre	Autre activité
0-4	2 367 686	3,92%		16,54						
5-9	2 722 796	4,51%		20,49						
10-14	2 871 733	4,76%		27,38						
15-19	2 897 141	4,80%		29,28						
20-24	2 990 245	4,95%	7,63%	22,15	3,51	1,17	4,49	0,96	7,23	4,79
25-29	3 211 025	5,32%	8,19%	21	3,47	2,23	5,21	1,13	6,3	2,66
30-34	3 369 346	5,58%	8,59%	18,03	3,55	0,85	3,92	0,76	5,24	3,71
35-39	3 704 872	6,14%	9,45%	21,25	4,38	0,68	7,78	1,05	3,92	3,44
40-44	4 418 357	7,32%	11,27%	22,35	3,88	2,53	7	0,67	4,48	3,79
45-49	4 824 297	7,99%	12,30%	19,27	2,99	2,61	8,24	0,88	1,93	2,62
50-54	4 934 336	8,17%	12,59%	22,3	2,75	5,54	8,05	0,52	2,02	3,42
55-59	4 417 895	7,32%	11,27%	18,27	2,88	1,41	4,6	0,68	3,62	5,08
60-64	3 846 237	6,37%	9,81%	18,43	3,28	1,07	6,05	0,87	3,53	3,63
65-69	3 490 973	5,78%	8,90%	12,74	3,1	0,55	0,48	0,95	3,33	4,33
70+	10 292 607	17,05%		10,55						
Population totale	**60 359 546**			**18,85**						
Population active	**39 207 583**			**19,66**	**3,34**	**2,03**	**5,81**	**0,83**	**3,94**	**3,72**

FIGURE 2.6 – Paramètre η moyen de la population italienne (totale et active) hors Covid

3. Source : Instituto Nazionale di Statistica italiano (ISTAT)

Nous ferons l'hypothèse suivante d'un retour progressif à la quasi-normale du paramètre η au cours de l'année 2021 que nous avons modélisée pour l'intégralité de la population italienne dès le mois de février : η est estimé à 5,6 en début d'année où la population est encore largement confinée avec de larges mesures restrictives sur tous types d'activités. Puis η augmente de façon linéaire jusqu'à s'établir à une valeur de 13 en fin d'année où nous ne nous attendons pas à de nouveaux confinements mais où les nouvelles habitudes de télétravail et de distanciation sociale resteront ancrées, entraînant une réduction structurelle du nombre de contacts quotidiens dans les transports, au travail, les stades, cinémas ou autres rassemblements.

Les paramètres purement épidémiologiques, tels que β, γ, σ ou l'efficacité du vaccin θ ont fait l'objet de nombreux débats et de publications scientifiques au coeur de la crise sanitaire pour tenter d'appréhender au mieux l'épidémie de Covid-19 et de proposer des mesures gouvernementales efficaces. Nous retiendrons la valeur moyenne de 95% estimée par l'Agence européenne des médicaments (EMA) pour l'efficacité du vaccin (paramètre θ). Le taux de convalescence γ mesure le temps moyen pour une personne infectée par le virus de passer de la catégorie de personne infectée à la catégorie de personne guérie et immunisée ne pouvant plus contaminer d'autres personnes avant un certain temps moyen mesuré par le paramètre σ. Le consensus des scientifiques italiens (cf. [Cereda *et al.*, 2020]) s'établit à 6,6 jours pour le paramètre γ mais nous retiendrons une hypothèse plus conservatrice de 9 jours dans notre modèle. Le paramètre σ fait encore largement débat aujourd'hui mais nous avions pour référence une valeur de 8 mois (240 jours) lorsque la modélisation a été faite au mois de février 2021 (cf. [Dan *et al.*, 2021]). Les taux journaliers de convalescence et de déclin de l'immunité de notre modèle seront donc respectivement 1/9 et 1/240. Quant au paramètre β, nous avons pu observer en 2020 un effet de saisonnalité, au même titre que l'épidémie de la grippe. La statistique officielle du paramètre β publiée par le gouvernement italien est représentée graphiquement dans la figure 2.7 [4].

Le paramètre $p_{t,vax}$,mesurant la proportion de la population vaccinée à chaque date t, a été calibré sur les projections du plan vaccinal du gouvernement italien (cf. figure 2.8 [5]), que nous avons jugé trop optimiste et par conséquent corrigé d'un facteur d'ajustement négatif croissant avec le temps pour prendre en compte les écueils possibles de la campagne de vaccination pas toujours bien accueillie par la population.

Enfin, le paramètre μ est calibré et projeté dans le temps par une régression simple à partir des statistiques empiriques de décès liés au Covid publiées quotidiennement par le gouvernement, de même que le nombre de personnes hospitalisées ou en soins intensifs à cause du Covid-19.

4. Source : Service national de la protection civile italienne
5. Source : https://www.trovanorme.salute.gov.it

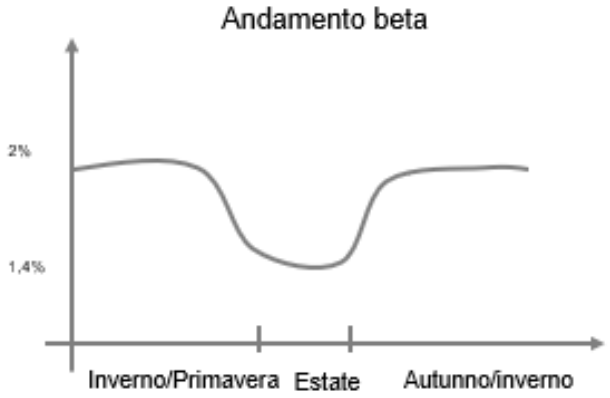

FIGURE 2.7 – Saisonnalité du paramètre β en Italie

Vaccini (azienda)	DIC 2020	Q1 2021	Q2 2021	Q3 2021	Q4 2021	Q1 2022	Q2 2022	TOTALI
Astra Zeneca*		8,028000	18,209000	13,929000	-	-	-	40,166000
PF/BT dosi iniziali	0,456000	7,352000	8,760000	10,792000	-	-	-	27,360000
PF/BT dosi aggiuntive		6,642991		6,642991				13,285982
J&J **		-	14,806000	32,304000	6,730000	-	-	53,840000
Sanofi/GSK		-	-	-	-	20,190000	20,190000	40,380000
Curevac		1,992000	5,312000	6,640000	7,968000	7,968000	-	29,880000
Moderna dosi iniziali		1,328000	4,650000	4,650700	-	-	-	10,628700
Moderna dosi aggiuntive				3,321000	7,307700			10,628700
TOTALE	**0,456000**	**20,360748**	**56,719243**	**74,958196**	**25,327196**	**28,158000**	**20,190000**	**226,169382**
media x mese	0,456	6,787	18,906	24,986	8,442	9,386	6,730	

agg.to 31/12/20 *in milioni di dosi*

** numero/mese provvisorio per i mesi successivi ad aprile*

*** se due dosi per regime vaccinale, altrimenti dimezzare*

FIGURE 2.8 – Plan vaccinal initial du Ministère de la santé italien

L'évolution dans le temps de notre population et des ses 3 compartiments se construit donc de façon itérative, chaque ligne représentant la répartition de la population et la valeur de chacun des paramètres à la date t (correspondant à 1 jour, cf. figure 2.9).

Date	η	Beta	pvax	Population	S	I	R	Death	Total hospitalised	Intensive Care
08/01/2021	**5,61**	**2,00%**	**0%**	**60 239 619**	58 079 640	570 389	1 589 590	77 381		
09/01/2021	5,61	1,90%	0,00%	60 239 114	58 024 548	568 222	1 646 343	77 886	26 129	2 610
10/01/2021	5,61	1,90%	0,06%	60 238 611	57 973 070	562 922	1 702 619	78 389	25 885	2 585
11/01/2021	5,62	1,90%	0,12%	60 238 112	57 922 406	557 634	1 758 072	78 888	25 642	2 561
12/01/2021	5,62	1,90%	0,18%	60 237 618	57 872 550	552 362	1 812 706	79 382	25 400	2 537
13/01/2021	5,63	1,90%	0,24%	60 237 129	57 823 497	547 105	1 866 527	79 871	25 158	2 513
14/01/2021	5,63	1,90%	0,30%	60 236 644	57 775 240	541 866	1 919 539	80 356	24 917	2 489
15/01/2021	5,64	1,90%	0,36%	60 236 164	57 727 772	536 644	1 971 748	80 836	24 677	2 465
16/01/2021	5,64	1,90%	0,42%	60 235 689	57 681 087	531 442	2 023 160	81 311	24 438	2 441
17/01/2021	5,65	1,90%	0,48%	60 235 218	57 635 178	526 261	2 073 779	81 782	24 199	2 417
18/01/2021	5,65	1,90%	0,54%	60 234 752	57 590 040	521 100	2 123 612	82 248	23 962	2 393
19/01/2021	5,66	1,90%	0,61%	60 234 290	57 545 664	515 963	2 172 663	82 710	23 726	2 370
20/01/2021	5,66	1,90%	0,67%	60 233 833	57 502 045	510 848	2 220 940	83 167	23 491	2 346
21/01/2021	5,67	1,90%	0,73%	60 233 380	57 459 175	505 758	2 268 447	83 620	23 257	2 323
22/01/2021	5,67	1,90%	0,79%	60 232 932	57 417 049	500 694	2 315 190	84 068	23 024	2 300
23/01/2021	5,67	1,90%	0,85%	60 232 489	57 375 658	495 655	2 361 176	84 511	22 792	2 277
24/01/2021	5,68	1,90%	0,91%	60 232 050	57 334 996	490 643	2 406 411	84 950	22 562	2 253
25/01/2021	6,00	1,90%	0,97%	60 231 615	57 295 056	485 659	2 450 900	85 385	22 332	2 231
26/01/2021	6,00	1,90%	1,03%	60 231 185	57 253 086	483 449	2 494 650	85 815	22 231	2 220
27/01/2021	6,01	1,90%	1,09%	60 230 757	57 211 563	481 222	2 537 972	86 243	22 128	2 210
28/01/2021	6,01	1,90%	1,15%	60 230 330	57 170 486	478 978	2 580 866	86 670	22 025	2 200
29/01/2021	6,02	1,90%	1,21%	60 229 906	57 129 856	476 717	2 623 332	87 094	21 921	2 190
30/01/2021	6,02	1,90%	1,27%	60 229 484	57 089 672	474 442	2 665 370	87 516	21 817	2 179
31/01/2021	6,03	1,90%	1,33%	60 229 063	57 049 932	472 151	2 706 980	87 937	21 711	2 169
01/02/2021	6,03	1,90%	1,39%	60 228 645	57 010 636	469 846	2 748 163	88 355	21 605	2 158
02/02/2021	6,04	1,90%	1,45%	60 228 229	56 971 582	467 730	2 788 917	88 771	21 508	2 148

FIGURE 2.9 – Projections du modèle SIR calibré sur la population italienne au 8 janvier 2021

2.3 Résultats du modèle prédictif appliqué au nombre de cas d'infections/hospitalisations

Nous avons ainsi projeté sur toute l'année 2021 l'évolution des différents compartiments de la population italienne à partir de notre modèle SIR adapté aux paramètres de l'épidémie de Covid-19 connus ou estimés à la date du 8 janvier 2021.

Une première macro-statistique de la modélisation qu'on peut regarder (figure 2.10) est le nombre cumulé de cas de Covid-19 recensés dans la population italienne, c'est-à-dire l'effectif de la population qui aura contracté le Covid-19 depuis l'apparition de l'épidémie.

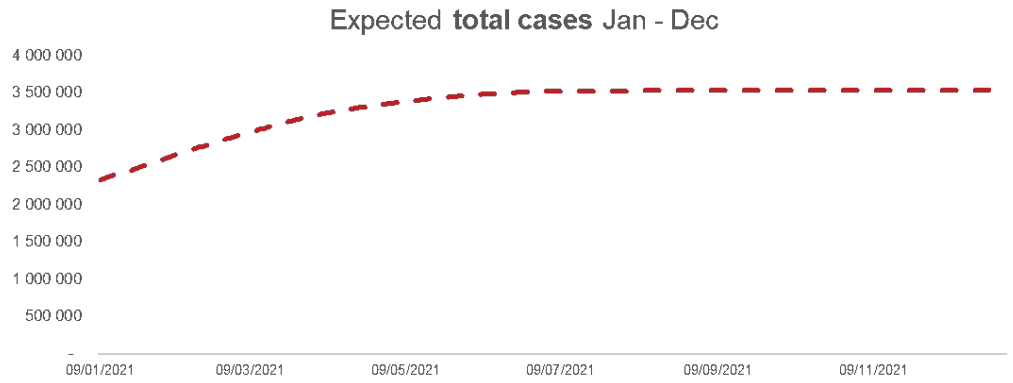

FIGURE 2.10 – Projection du nombre de cas cumulés de Covid-19 en Italie

Notre modèle estime que ce nombre converge de façon asymptotique vers 3 540 000 cas cumulés de Covid-19 sur l'ensemble de la population italienne.

Comparons cette projection avec le nombre officiel de cas de Covid-19 recensés par le gouvernement italien jusqu'à début février (figure 2.11) :

Nous voyons sur la figure 2.11 que notre modèle SIR surestime légèrement le nombre de cas cumulés de Covid-19 mais converge progressivement vers le nombre réel de cas observés, ce qui nous conforte quant à la performance du modèle.

Analysons à présent l'évolution du compartiment I des personnes contagieuses dont nous pouvons rappeler la dynamique : $\frac{dI(t)}{dt} = p \cdot I(t) \cdot S(t) - (\gamma + \mu) \cdot I(t)$, c'est-à-dire qu'il est alimenté par les nouveaux cas de Covid proportionnels au nombre d'individus déjà infectés et du nombre d'individus encore non-infectés et diminué simultanément des cas de Covid ayant entraîné un décès et des personnes convalescentes depuis 9 jours.

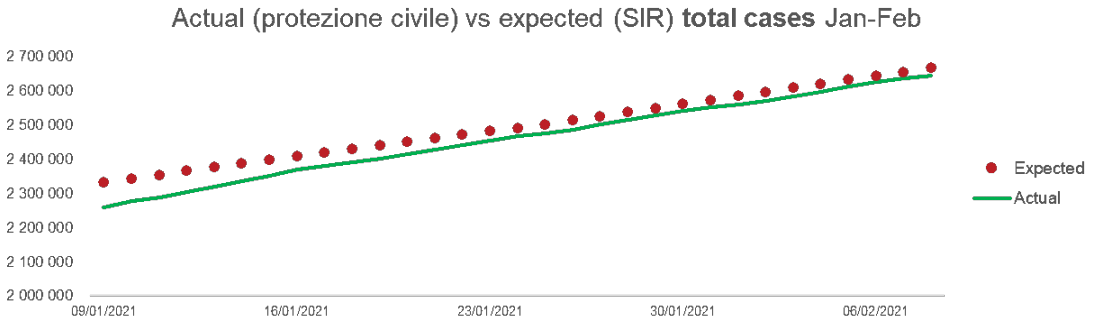

FIGURE 2.11 – Performance du modèle SIR pour estimer les nouveaux cas de Covid-19

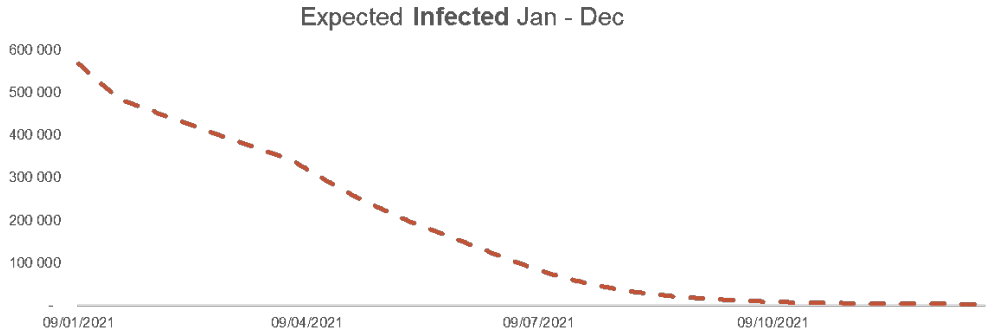

FIGURE 2.12 – Projection de l'effectif du compartiment *Infected* de la population italienne

Notre modèle prévoit que l'effectif des personnes infectées en Italie décroisse fortement les 6 premiers mois de l'année 2021 puis converge asymptotiquement vers 0 au deuxième semestre (cf. figure 2.12). L'effectif estimé du compartiment I au 31/12/2021 n'est plus que de 2840 individus.

Comparons à nouveau cette projection avec les chiffres officiels du gouvernement italien sur les 30 premiers jours modélisés du 8 janvier au 8 février 2021 (figure 2.13) :

Nous voyons sur la figure 2.13 que le modèle prévoit bien la tendance décroissante réelle de l'effectif des individus contagieux au cours du premier mois de l'année. Le modèle donne également de bons résultats de prévisions sur les projections des guérisons (figure 2.14) ou décès du Covid-19 (figure 2.15), ainsi que les effectifs de personnes hospitalisées (figure 2.16) ou en soins intensifs (figure 2.17), qui sont les deux variables critiques que nous cherchions à modéliser dans le cadre du suivi et de la tarification de notre produit *Covid-19 Protezione*.

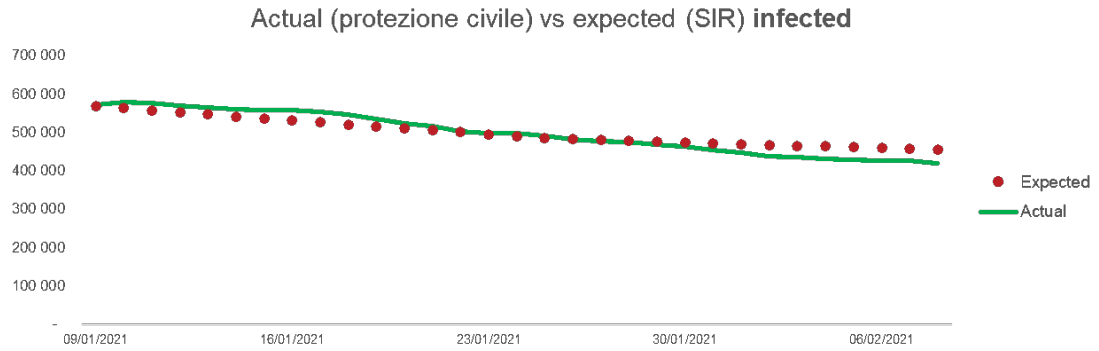

FIGURE 2.13 – Performance du modèle SIR pour estimer l'évolution de l'effectif des individus contagieux

Le calcul des erreurs de prédiction moyennes pour chaque variable modélisée (figure 2.18) confirme la précision des estimations du modèle développé et sa possible mise en application dans un outil de tarification dynamique.

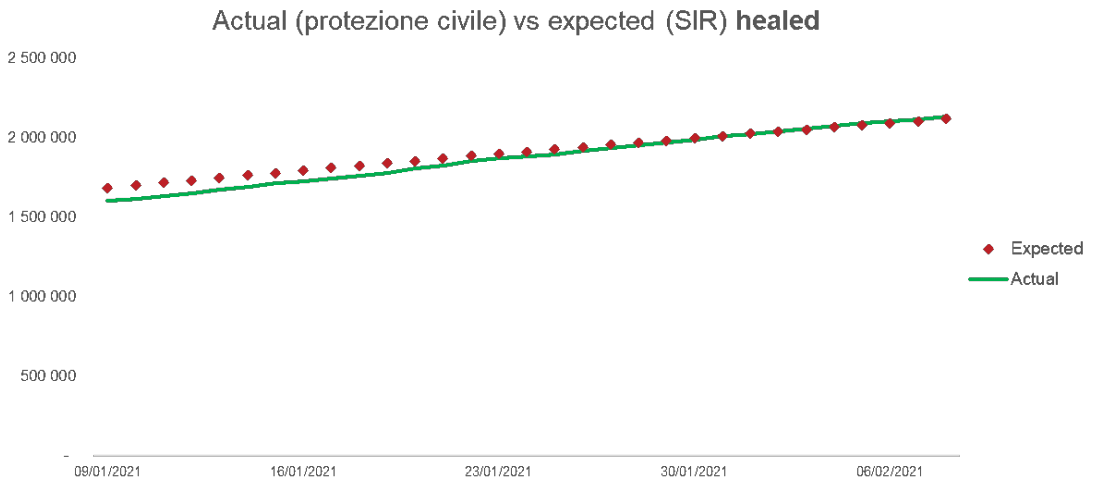

FIGURE 2.14 – Performance du modèle SIR pour estimer l'évolution de l'effectif des personnes guéries

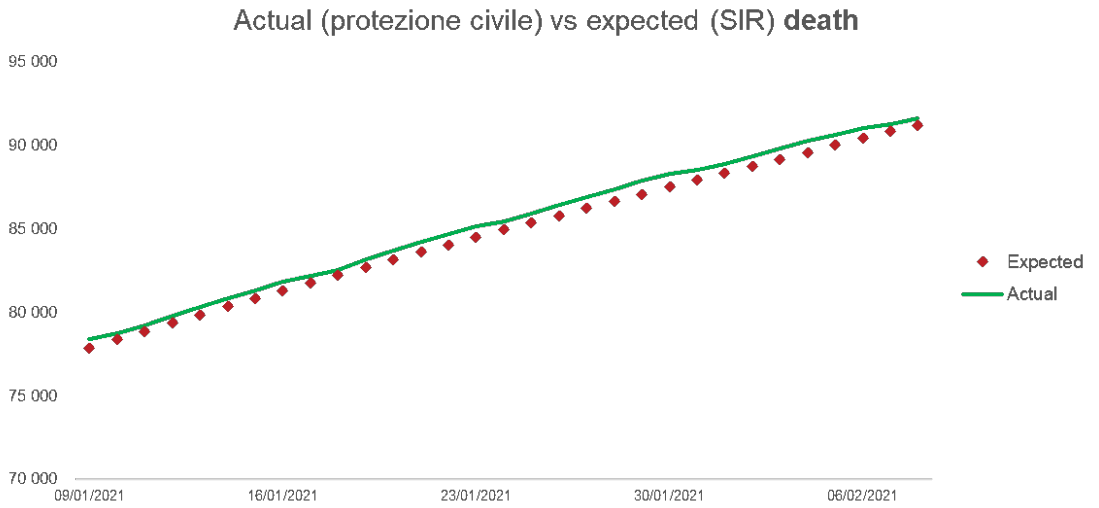

FIGURE 2.15 – Performance du modèle SIR pour estimer l'évolution des décès dus au Covid-19

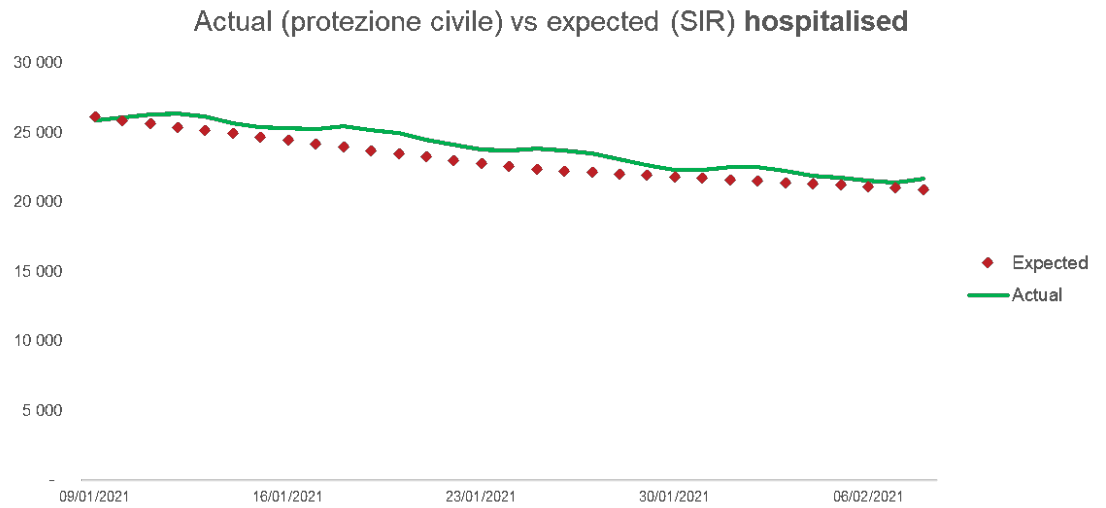

FIGURE 2.16 – Performance du modèle SIR pour estimer l'effectif des personnes hospitalisées

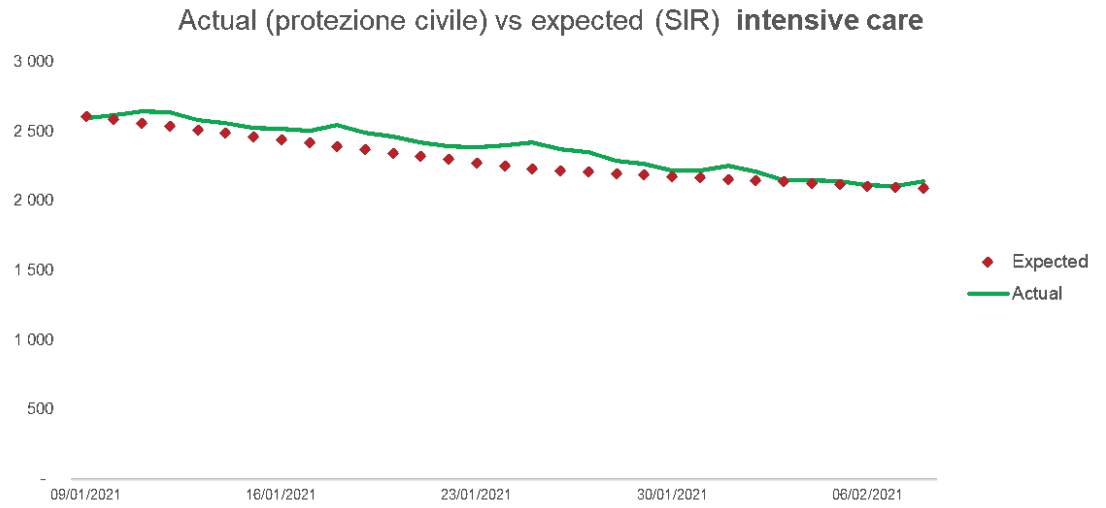

FIGURE 2.17 – Performance du modèle SIR pour estimer l'effectif des personnes en réanimation

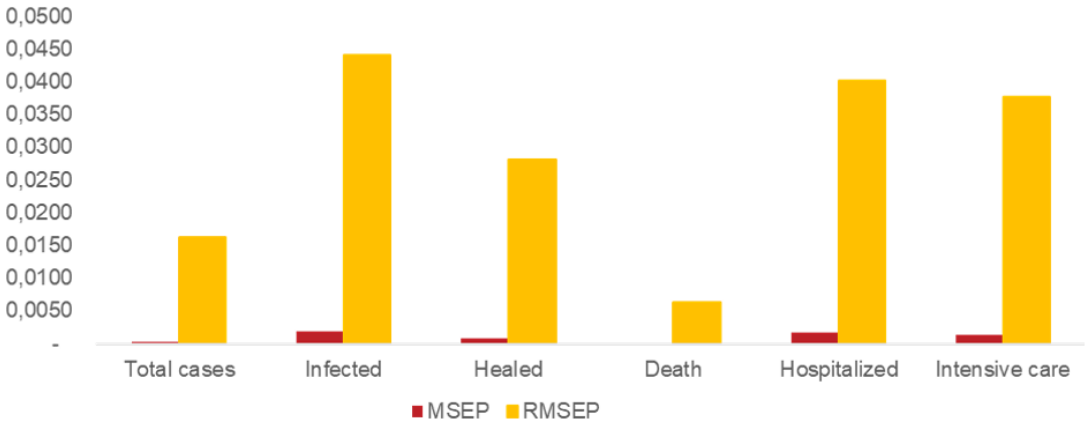

FIGURE 2.18 – Erreurs de prédiction moyennes de notre modèle SIR Covid-19

2.4 Application à l'outil de tarification dynamique développé pour ce nouveau produit

Le modèle compartimental permet une grande souplesse dans sa mise en oeuvre, ce qui le rend facilement adaptable à des contextes de pays variés ou à l'évolution de l'épidémie avec ses différents variants. Ainsi, le compartiment I peut être en segmenté en sous-compartiments, par exemple I_{eng} et $I_{no\ eng}$ pour distinguer la population infectée par le variant anglais de celle infectée par les autres variants. Chaque sous-compartiment a alors sa propre dynamique modélisée avec notamment un β qui lui est propre, de même qu'une efficacité vaccinale θ différente. Le système d'équations différentielles est alors adapté pour tenir compte de la distinction de catégories de personnes infectées :

- $N(t) = S(t) + I_{eng}(t) + I_{no\ eng}(t) + R(t)$
- $\frac{dS(t)}{dt} = -(p_{eng} \cdot I_{eng}(t) + p_{no\ eng} \cdot I_{no\ eng}(t)) \cdot S(t)$
  où $p_{eng} = p_{t,vax} \cdot (1 - \theta_{eng}) \cdot \eta \cdot \beta_{eng} + (1 - p_{t,vax}) \cdot \eta \cdot \beta_{eng}$
  et $p_{no\ eng} = p_{t,vax} \cdot (1 - \theta_{no\ eng}) \cdot \eta \cdot \beta_{no\ eng} + (1 - p_{t,vax}) \cdot \eta \cdot \beta_{no\ eng}$
- $\frac{dI_{eng}(t)}{dt} = p_{eng} \cdot I_{eng}(t) \cdot S(t) - (\gamma_{eng} + \mu_{eng}) \cdot I_{eng}(t)$
- $\frac{dI_{no\ eng}(t)}{dt} = p_{no\ eng} \cdot I_{no\ eng}(t) \cdot S(t) - (\gamma_{no\ eng} + \mu_{no\ eng}) \cdot I_{no\ eng}(t)$
- $\frac{dR(t)}{dt} = \gamma_{eng} \cdot I_{eng}(t) + \gamma_{no\ eng} \cdot I_{no\ eng}(t) - \sigma \cdot R(t)$

L'apparition du variant anglais et sa rapide propagation en Italie au début du mois de mars 2021 a entraîné un changement de régime dans les séries temporelles modélisées (à l'exception du taux de décès resté relativement stable, cf. figure 2.19). Il a donc fallu recalibrer le modèle pour tarifer correctement les nouvelles souscriptions, au risque de sous-estimer la fréquence du nombre d'infections et d'hospitalisations qui ont été fortement impactées par la force de contagion accrue de ce variant.

Devant le succès commercial du produit *Covid-19 Protezione* en Italie et sa bonne performance technique grâce à une modélisation et un suivi minutieux, nous avons souhaité développer ce produit de façon agile et rapide dans d'autres pays en appliquant un facteur d'ajustement égal au ratio des niveaux de risques respectifs du nouveau pays (cf. figure 2.20) et du pays de référence (l'Italie).

Le niveau de risque est défini selon une échelle logarithmique allant de 0 (très faible) à 5 (catastrophique) de telle sorte que $risque = 10^{Niveau}$. Un niveau à 5 pendant 1 an correspond au seuil maximal de risque pouvant décimer 100% de la population. Un niveau 4 correspond à un risque de 10% de décès dans la population, un niveau 3 à 1% de décès,... Le niveau de risque est calculé de la manière suivante à partir des statistiques empiriques nationales recensées pour chaque pays [6] :

$$NR^{(c,t)} = \log(P^{(c,t)}_{hospi} \cdot G^{(c,t)}_{hospi} + P^{(c,t)}_{rea} \cdot G^{(c,t)}_{rea} + P^{(c,t)}_{deces} \cdot G^{(c,t)}_{deces})$$

6. Sources : `https://www.worldometers.info/coronavirus/` et `https://github.com/owid/covid-19-data/tree/master/public/data`

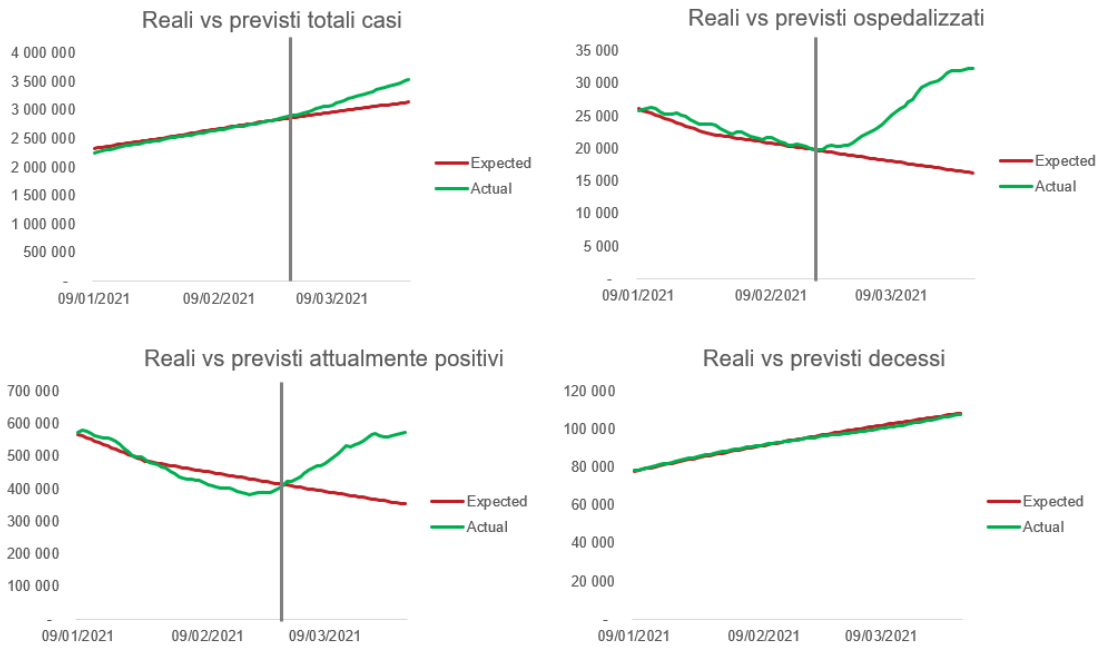

FIGURE 2.19 – Impact du variant anglais en Italie sur la performance du modèle SIR (avant recalibrage)

avec $P_i^{(c,t)}$ et $G_i^{(c,t)}$ les fréquences et sévérités respectives des différentes composantes du risque pour le pays c à la date t, estimées par les ratios ci-dessous :

- $P_{hospi}^{(c,t)}$= effectif des individus hospitalisés dans le pays c à la date t / effectif de la population
- $G_{hospi}^{(c,t)}$ = log (moyenne mobile 7j de (nombre de décès / effectif des hospitalisations) $\times 10^5$)
- $P_{rea}^{(c,t)}$= effectif des individus en réanimation dans le pays c à la date t / effectif de la population
- $G_{rea}^{(c,t)}$ = log(moyenne mobile 7j de (nombre de décès / effectif des réanimations) $\times 10^5$)
- $P_{deces}^{(c,t)}$= nombre de décès depuis le début de l'épidémie dans le pays c à la date t / effectif de la population
- $G_{deces}^{(c,t)} = 5$

L'outil de tarification dynamique développé spécifiquement pour le produit *Covid-19 Protection* a ensuite permis d'adapter de façon plus fine la prime d'assurance collective en fonction de l'évolution temporelle des variables exogènes du modèle SIR mais également de certains paramètres propres au groupe assuré : répartition des âges et nature de l'activité professionnelle exercée conditionnent le paramètre η et le taux de vaccination (cf.

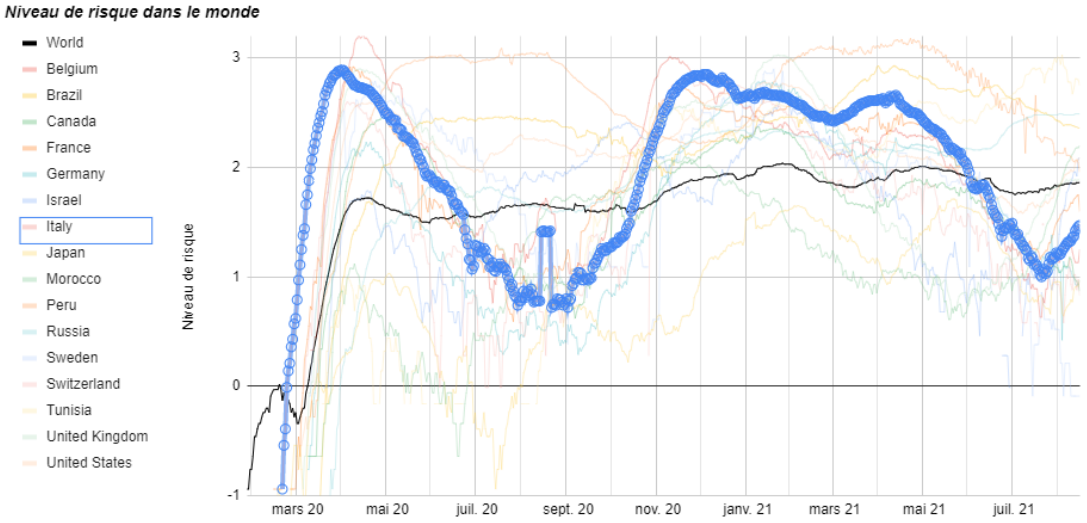

FIGURE 2.20 – Evolution dans le temps du niveau de risque de l'épidémie de Covid-19

figures 2.21 [7] et 2.22 [8] avec l'exemple de la France), et par conséquent la force d'infection de l'épidémie pour cette population spécifique.

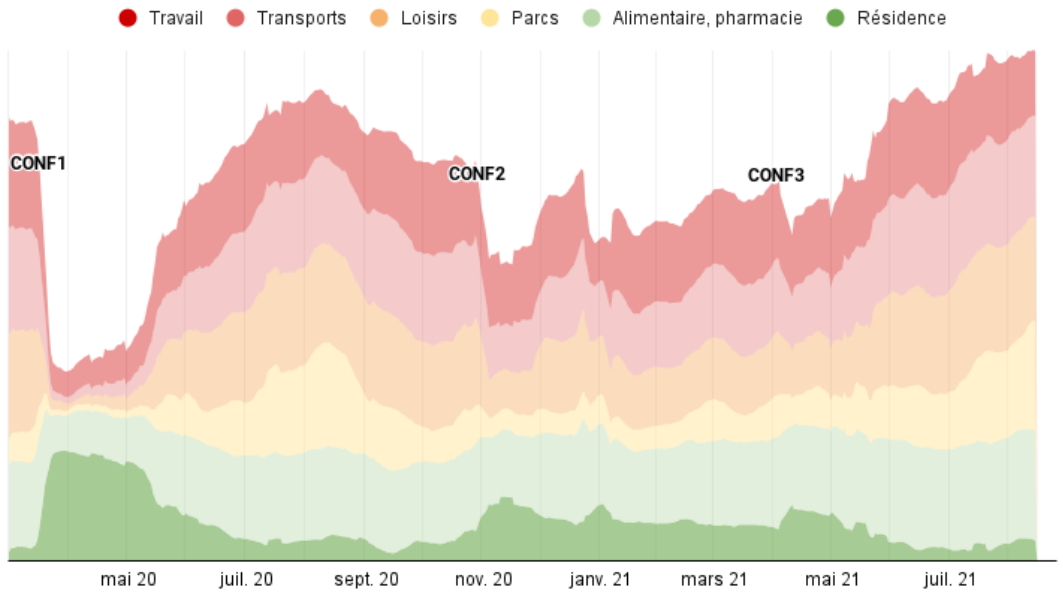

FIGURE 2.21 – Evolution dans le temps des activités quotidiennes exercées en France

7. Indicateur calculé à partir des données de mobilité Google
8. Données statistiques publiques en santé et social (DREES)

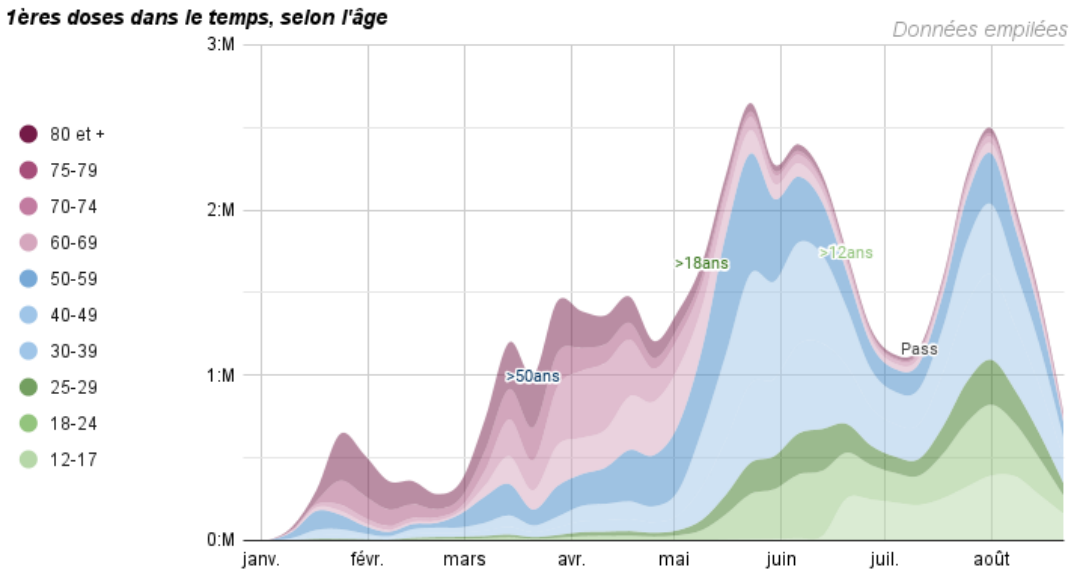

FIGURE 2.22 – Evolution de la campagne de vaccination en France par tranches d'âges

La modélisation des fréquences d'infections et d'hospitalisations dues au Covid-19 peut également être utile pour la projection de la sinistralité d'autres produits d'assurance que le produit spécifique *Covid-19 Protection*. En effet, la crise sanitaire a impacté tous les domaines de la vie courante et les besoins d'assistance à court et moyen termes ont clairement évolué pendant cette période, ce qui s'est traduit par des variations notables de fréquences de sinistres sur les portefeuilles d'Europ Assistance en auto, habitation et surtout voyages. Nous avons donc cherché à adapter nos outils de tarification en conséquence.

2.5 Autres applications possibles commes l'assistance médicale en voyage

Les produits d'assurance ou d'assistance voyages ont la particularité de couvrir un ou plusieurs risques pendant une fenêtre de temps limitée plus ou moins éloignée de la date de contractualisation. Ces risques sont également largement dépendants de la saison et de la destination du voyage. Dans le cadre de l'épidémie soudaine et violente de Covid-19, les besoins d'assistance médicale en voyage à l'étranger ont été nombreux et divers, allant de la simple demande d'informations au rapatriement sanitaire. Quant à la garantie d'annulation de voyage, la plupart des conditions générales d'assurances l'excluaient dans un contexte d'épidémie, mais de nombreux remboursements ont quand même pu avoir lieu dans la pratique, comme nous aurons l'occasion de l'analyser dans la troisième partie de ce mémoire.

En tout état de cause, il a fallu intégrer ce nouveau facteur de risque dans la modélisation et la tarification des produits d'assurance et d'assistance voyage. Un modèle déterministe possible consiste à projeter le modèle SIR sur une année complète pour les pays de départ et de destination les plus représentés dans nos portefeuilles et de calculer pour chaque voyage les fréquences d'infection, d'hospitalisation ou d'entrée en réanimation en fonction des caractéristiques du voyage (pays, dates, nature et durée du séjour, nombre de personnes,...) auxquelles on ajoutera une marge de risque proportionnelle à l'éloignement dans le temps de la date du voyage programmé.

Devant les nombreuses incertitudes de l'évolution de la crise sanitaire, a fortiori en matière de voyages, un modèle stochastique paraît plus approprié qu'une approche déterministe dans la mesure où il permet d'appréhender la variabilité accrue des erreurs de prédiction des modèles de tarification existants (GLM ou autres) à l'aune du Covid sans avoir besoin d'une caractérisation complète de l'aléa. Cela pourra faire l'objet de travaux ultérieurs à ce mémoire, en s'inspirant par exemple des travaux de Husson en la matière (cf. [F.Husson, 2001]).

Chapitre 3

Vers un futur produit santé couvrant aussi les risques d'épidémie type Covid-19

3.1 Contexte épidémique et changement de régime dû au comportement des assurés

L'épidémie de Covid-19 a bouleversé tous les modes de vie et de travail et entraîné une prise de conscience collective du risque épidémique. En effet, elle est à l'origine d'un profond changement de comportement des assurés dans leur rapport à l'assurance : une consommation accrue et de nouvelles exigences (cf. figure 3.1) en termes de garanties pour qu'elles intègrent la pandémie et de façon générale tout type d'épidémie.

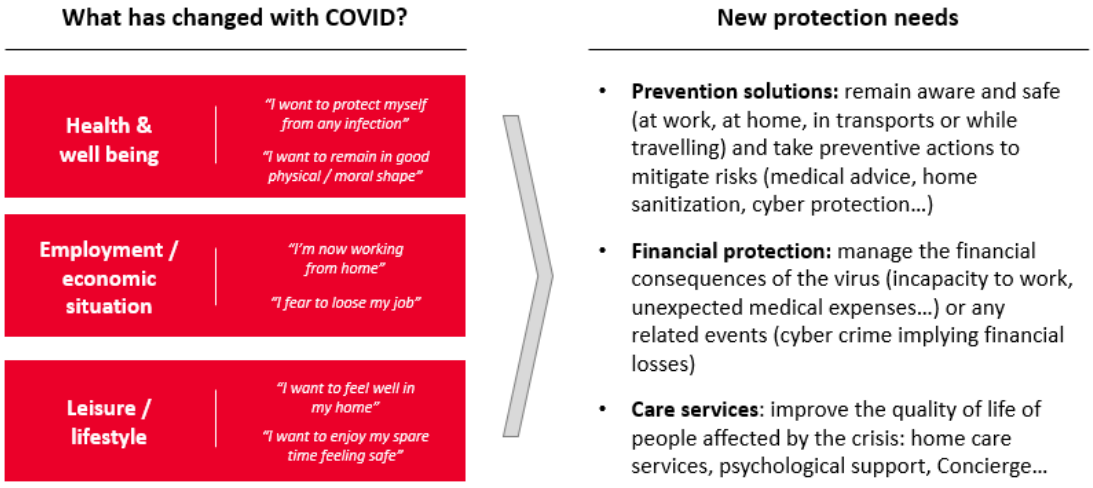

FIGURE 3.1 – Cartographie des nouveaux besoins des assurés dans un monde post-Covid

Dans ce nouveau contexte, j'ai cherché à combiner les 2 produits d'assurance respectivement étudiés dans les deux premières parties de ce mémoire, et commercialisés de façon totalement indépendantes aujourd'hui. Cela signifie qu'une personne cherchant à être couverte pour des faits générateurs aussi variés qu'un diagnostic de maladies chroniques, un accident ou une hospitalisation due à une épidémie, doit en l'état souscrire à plusieurs produits d'assurance. Elle ne peut d'ailleurs pas souscrire individuellement au produit *Covid-19 Protection* car il ne peut aujourd'hui être souscrit que par des entreprises dans le cadre d'une police d'assurance collective.

Si tant est que nous ouvrions la commercialisation du produit *Covid-19 Protection* à des polices individuelles qui pourront être distribuées via des agents, des courtiers ou en direct en ligne, nous serions réduits, sans autre modélisation faite à ce jour, à proposer les 2 produits indépendamment à un prix égal à la somme des primes respectives des 2 produits pour l'individu au regard de ces caractéristiques. L'objectif de cette troisième partie est donc d'établir une corrélation potentielle entre tous ces faits générateurs, ou tout du moins un biais sur le comportement des assurés à qui nous offririons un choix aussi large à la souscription. Nous avons déjà pu observer pendant la période du Covid-19 un certain nombre de biais comportementaux dans les déclarations de sinistres, notamment sur notre portefeuille d'assurance annulation de voyages.

3.2 Probablilités conditionnelles déduites de nos données empiriques d'assurance voyage

Nous avons observé sur le portefeuille d'assurance annulation de voyages d'Europ Assistance une flambée d'annulations pour causes médicales à partir du mois de mars 2020, certaines d'entre elles ayant un lien de causalité direct avec le Covid-19, sans pour autant être déclaré comme tel, les assurés ayant été progressivement bien informés au cours de la crise que les cas de pandémies étaient exclus de la plupart des polices d'assurance. Nous avons par conséquent classé les causes médicales d'annulation de voyages selon 2 catégories détaillées dans la figure 3.2 :

- I1 pour les maladies saisonnières ayant une probabilité forte d'être dues au Covid-19
- I2 pour les maladies chroniques ou n'ayant aucun lien direct évident avec le Covid-19

Pour chacune des catégories de maladies, nous pouvons calculer la fréquence moyenne d'annulation de voyages pour cause de diagnostic médical tombant explicitement dans la catégorie concernée. Les données observées sur notre portefeuille nous permettent de distinguer une fréquence saisonnière mais stable dans la durée (que nous nommerons f_{HC}) jusqu'à la matérialisation de la pandémie de Covid-19 au cours du premier trimestre 2020. Celle-ci a entraîné une montée en flèche de la fréquence d'annulations (que nous nommerons f_C), toutes causes comprises (y compris I2).

I1 :
I1_Medical Cause_Gastro enteritis
I1_Medical Cause_Neck pain - Back pain
I1_Medical Cause_Respiratory track infection - Bronchitis
I2 :
I2_Medical Cause_Cancer
I2_Medical Cause_Chronical Illness
I2_Medical Cause_Heart failure
I2_Medical Cause_Hematological pathology
I2_Medical Cause_Immunodeficiency
I2_Medical Cause_Kawasaki Illness
I2_Medical Cause_Pneumonia - Asthma
I2_Medical Cause_Vascular pathology

FIGURE 3.2 – Détail de la classification des causes médicales d'annulation de voyages en fonction de leur lien d'ordre 1 ou d'ordre 2 avec le Covid-19

Nous pouvons alors déduire de la différence entre ces deux fréquences moyennes (Hors Covid f_{HC} et en période de Covid f_C , cf. figure 3.3) la probabilité conditionnelle qu'un assuré soit infecté par le Covid-19 sachant qu'il a déclaré un sinistre d'annulation de voyage pour cause de type I1 ou I2 :

$$P[\Theta = 1 | I = 1] = \frac{f_C - f_{HC}}{f_C}$$

où Θ est une variable aléatoire binaire prenant la valeur 1 si l'assuré est infecté par le Covid-19
et I est une variable aléatoire binaire prenant la valeur 1 si l'assuré déclare une maladie de type I.

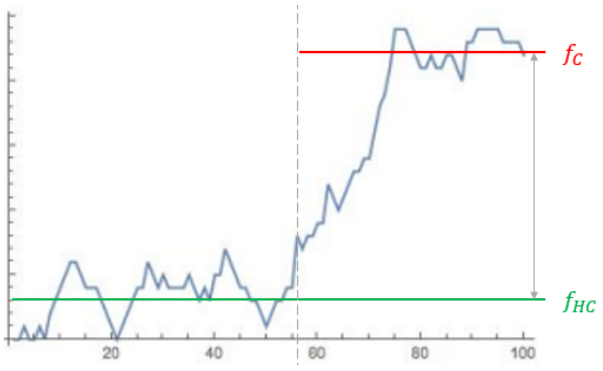

FIGURE 3.3 – Fréquences empiriques d'annulation de voyages pour causes médicales sur les 100 premiers jours de l'année 2020

Nous nous intéresserons particulièrement aux maladies de type I2 dans le cadre du produit que nous cherchons à construire en faisant la synthèse des produits *Eura Salute*

360° et *Covid-19 Protection*. Néanmoins les probabilités conditionnelles pour les maladies saisonnières de type I1 peuvent être intéressantes dans le cadre d'un produit d'assurance santé classique pour la Sécurité sociale, par exemple, dont le budget est fortement affecté par la pandémie.

3.3 Synthèse de la représentation de nos connaissances a priori et choix du modèle

Forts de nos connaissances a priori grâce aux travaux menés dans les deux premières parties de ce mémoire, nous allons pouvoir calculer la fréquence théorique f_{new} du nouveau produit que nous cherchons à construire, c'est-à-dire un produit indemnitaire couvrant aussi bien les faits générateurs du produit *Eura Salute 360°* qu'une infection par une épidémie (nous garderons l'exemple du Covid-19 déjà modélisé dans les développements ci-après).
Ainsi, $f_{new} = P([I = 1] \cup [\Theta = 1]) = P[I = 1] + P[\Theta = 1] - P([I = 1] \cap [\Theta = 1])$ Or, d'après la formule de Bayes, $P([I = 1] \cap [\Theta = 1]) = P[\Theta = 1 | I = 1] \cdot P[I = 1]$ Nous savons estimer a priori $P[I = 1]$ pour chaque groupe d'individus de la segmentation induite par la modélisation des fréquences du portefeuille *Eura Salute 360°*. De même, nous connaissons $P[\theta = 1]$ d'après le modèle SIR développé pour le produit *Covid Protection*. Enfin, nous pouvons utiliser les probabilités conditionnelles $P[\Theta = 1 | I = 1]$ déduites précédemment de notre portefeuille d'annulation de voyages. Le modèle probabiliste synthétisant le mieux l'ensemble des observations dont nous disposons est le réseau bayésien suivant (figure 3.4) :

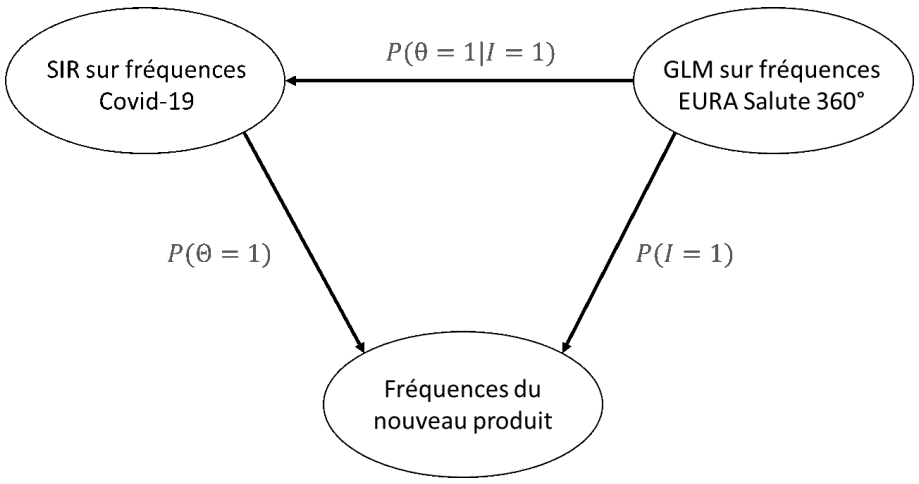

FIGURE 3.4 – Réseau bayésien issu du modèle de données étudiées

Revenons un instant sur l'importance de la variable *Zonier* que nous avons établie dans la première partie et qui est tout autant discriminante dans le cadre d'un modèle

épidémique comme celui développé dans le cadre du Covid-19. Le nombre de contacts quotidiens étant lié à la densité de population, nous observons logiquement des fréquences plus importantes de l'épidémie dans les zones urbaines dont l'activité économique attire également des populations non-résidentes venant d'autres provinces ou de l'étranger (cf. figure 3.5). Nous pourrons donc créer un zonier plus granulaire que celui à 3 modalités proposé dans le cadre restreint du produit *Eura Salute 360°* afin de prendre en compte la nouvelle variable exogène de la densité de population. Le zonier permet enfin une représentation géographique très visuelle de nos connaissances a priori.

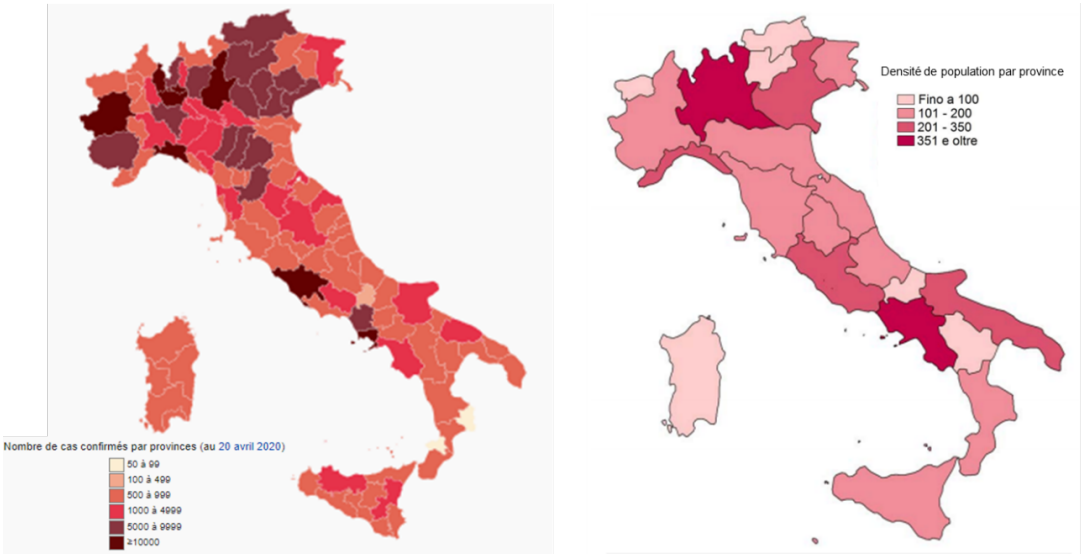

FIGURE 3.5 – Comparaison du nombre de cas de Covid-19 à la densité de population par province en Italie

3.4 Construction d'un réseau bayésien dynamique

Un réseau bayésien dynamique modélise l'évolution d'un réseau bayésien dont les probabilités ne sont pas figées dans le temps. Ce type de modèle permet d'appréhender le changement de comportement des assurés en engrangeant de l'expérience. La crise épidémique du Covid-19 en est un bon exemple : la connaissance a priori du virus et de la maladie n'ont rien à voir avec ce que nous en savons aujourd'hui, ni avec le recul supplémentaire que nous aurons dans les années à venir.

Le modèle de fréquence GLM du produit *EURA Salute 360°* étant relativement stable dans le temps par rapport au modèle épidémique SIR avec un pas de temps quotidien, nous pouvons postuler que la dynamique temporelle du réseau bayésien sera donnée avant tout par la variable indicatrice d'infection ou non de l'assuré par l'épidémie à chaque pas de temps t ; nous la noterons X_1^t. Toujours grâce à la formule de Bayes, nous pouvons

inférer du réseau bayésien construit en figure 3.4 le nouveau réseau bayésien statique ci-dessous (figure 3.6) :

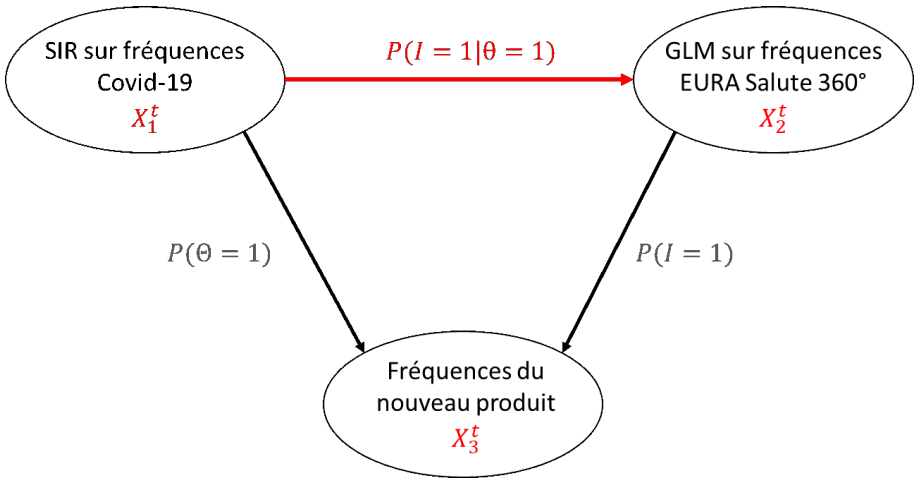

FIGURE 3.6 – Réseau bayésien de référence du modèle dynamique

Pour parachever notre modèle dynamique de tarification, nous introduirons la variable X_4^t que nous définirons comme la moyenne mobile des projections des X_3^j (avec j allant de t à $t + 365$ pour une prime annuelle). En fonction de la confiance que nous accordons aux projections sur des horizons de temps différents, nous pourrons faire le choix d'une moyenne mobile arithmétique ou géométrique. Cette dernière me semble plus adaptée au modèle stochastique (dont les intervalles de confiance s'élargissent avec l'horizon de projection) que nous évoquions à la fin de la deuxième partie de ce mémoire. La figure 3.7 ci-dessous schématise la structure du réseau bayésien dynamique proposé.

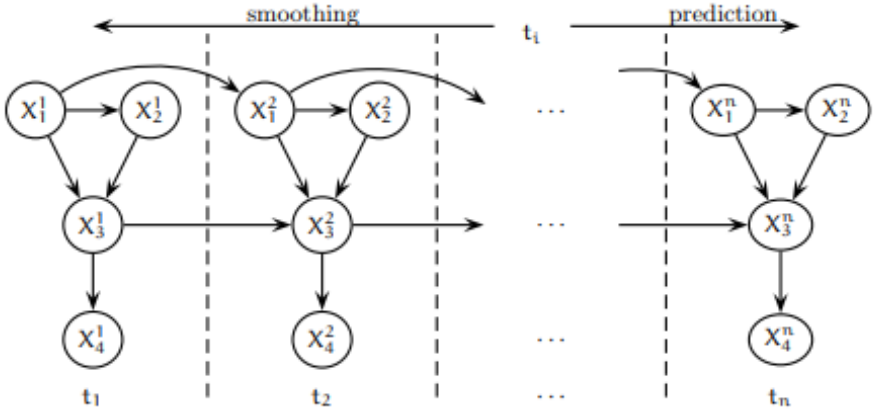

FIGURE 3.7 – Structure du modèle de réseau bayésien dynamique

3.5 Validation du modèle et champs d'application possibles

La nature probabiliste du modèle proposé complexifie son implémentation bien qu'elle nous donne une vision plus complète de la dynamique sous-jacente des différents risques inhérents au nouveau produit d'assurance que nous avons pour ambition de créer. Celui-ci reste aujourd'hui théorique mais le projet de ce mémoire a permis de faire avancer la réflexion et la demande du marché pour ce type de produit continue d'évoluer au gré des prises de conscience collectives. On notera par exemple que le produit *Covid-19 Protection* a connu un fort succès commercial pendant sa première année de souscription, concomitante aux débuts de la pandémie de Covid en 2020, mais que l'arrivée de la vaccination, les nouveaux modes de télétravail et l'habitude progressive de vivre avec le Covid ont nettement atténué la demande des entreprises pour cette assurance collective de leurs employés. En revanche, nous pouvons très bien imaginer un regain d'enthousiasme pour ce produit si les espoirs portés par la vaccination (sur laquelle tous les gouvernements ont largement misé) étaient éteints par un nouveau variant du virus plus résistant.

En effet, en reprenant l'ensemble des hypothèses sur les paramètres de notre modèle SIR Covid-19, dont la plupart sont issues des dires d'experts médicaux et scientifiques en épidémiologie et vaccination, nous avons simulé (cf. figure 3.8) l'impact d'une chute de l'efficacité du vaccin θ de 95% à 50% sur la trajectoire du taux de personnes infectées dans la population italienne sur l'année 2021. Nous pouvons voir que ce nouveau calibrage du modèle SIR ne prévoit plus une convergence asymptotique vers 0 du nombre de personnes infectées mais au contraire une nouvelle remontée exponentielle augurant d'une nouvelle vague tout aussi brutale que les précédentes. Le variant delta actuel laisse présager d'une moindre efficacité vaccinale, toutefois relativement moins importante que la simulation faite avec un taux de 50%. Mais il n'est pas exclu que de nouvelles mutations plus résistantes apparaissent à l'avenir.

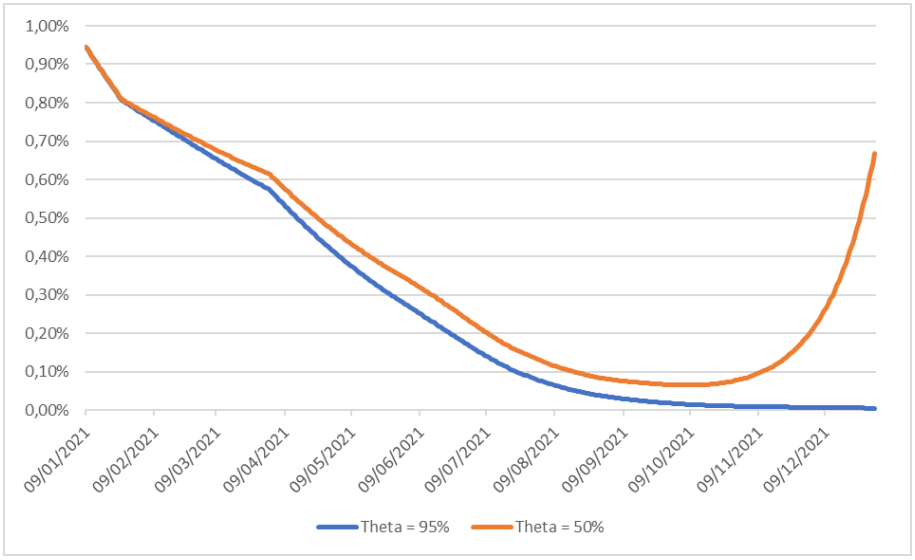

FIGURE 3.8 – Projections du taux de personnes infectées dans la population italienne selon 2 scénarios

Conclusion

Ce mémoire, écrit dans un contexte inédit de pandémie mondiale, nous fait prendre conscience des enjeux politiques et comportementaux de l'assurance santé, au-delà des aspects médicaux, économiques et sociaux traditionnellement connus comme moteurs principaux. La crise du Covid laisse déjà augurer qu'elle aura un impact structurel sur les politiques de santé publique. Aussi le modèle de réseau bayésien dynamique construit dans la dernière partie de ce mémoire permet-il de donner un cadre théorique pour faire évoluer nos modèles. Quand bien même le nouveau produit exhaustif imaginé ne verrait-il jamais le jour, sa modélisation pourra toujours être utile à la revalorisation des produits existants comme l'*Eura Salute 360°*. En effet, nous avons proposé à la fin de la première partie un modèle dynamique de crédibilité linéaire pour améliorer progressivement la qualité de sa tarification. Le réseau bayésien dynamique offre une alternative probabiliste des modèles de crédibilité.

La modélisation des coûts de sinistres des produits santé pourra également être envisagée dans un contexte de réseau bayésien dynamique lorsque nous commencerons à avoir un certain recul sur les formes longues du Covid et les facteurs de comorbidité. Quant au coût public de la pandémie, celui-ci pourra avoir un impact sur les politiques sanitaires à l'avenir avec des répercussions sur l'assurance privée. Combien de temps pourra encore durer la martingale du « quoi qu'il en coûte » ? Les échéances électorales dans ce contexte de crise sanitaire apporteront peut-être des réponses différentes aux stratégies gouvernementales adoptées jusqu'à présent. Mais notre modèle SIR montre bien que le nombre de paramètres réellement influençables est limité.

Enfin, l'hypothèse standard d'indépendance entre l'espérance de la fréquence et celle des coûts de sinistres peut elle aussi être remise en cause dans un contexte d'assurance où une pandémie d'ampleur mondiale n'est plus exclue. Nos modèles GLM traditionnels de tarification devront donc peut-être aussi évoluer vers ce type de modèles dynamiques avec une composante stochastique. En tout état de cause, le biais comportemental et la tendance consumériste des assurés doivent être pris en compte pour éviter des phénomènes d'anti-sélection. Nous avons bien vu dans la première partie de ce mémoire l'impact important du canal de distribution sur la sinistralité. Par ailleurs, le retour d'expérience des entreprises ayant souscrit le produit Covid-19 protection pour leurs employés témoigne de l'importance du taux d'usage de l'assurance. Celui-ci fut particulièrement faible par

rapport à nos estimations sur la base de la population active générale. Cela peut s'expliquer par une moyenne d'âge des assurés relativement basse, la nature des activités (principalement de service) et la taille critique des entreprises qui se virent imposer de nombreuses contraintes par les gouvernements. D'où l'importance de continuer à innover pour proposer de nouveaux produits d'assurance et d'assistance (dont les taux d'usage sont en général plus élevés en raison des nombreux services annexes proposés) en phase avec les attentes de nos clients.

Annexe A

Corrélation du zonier créé avec des variables macro-économiques

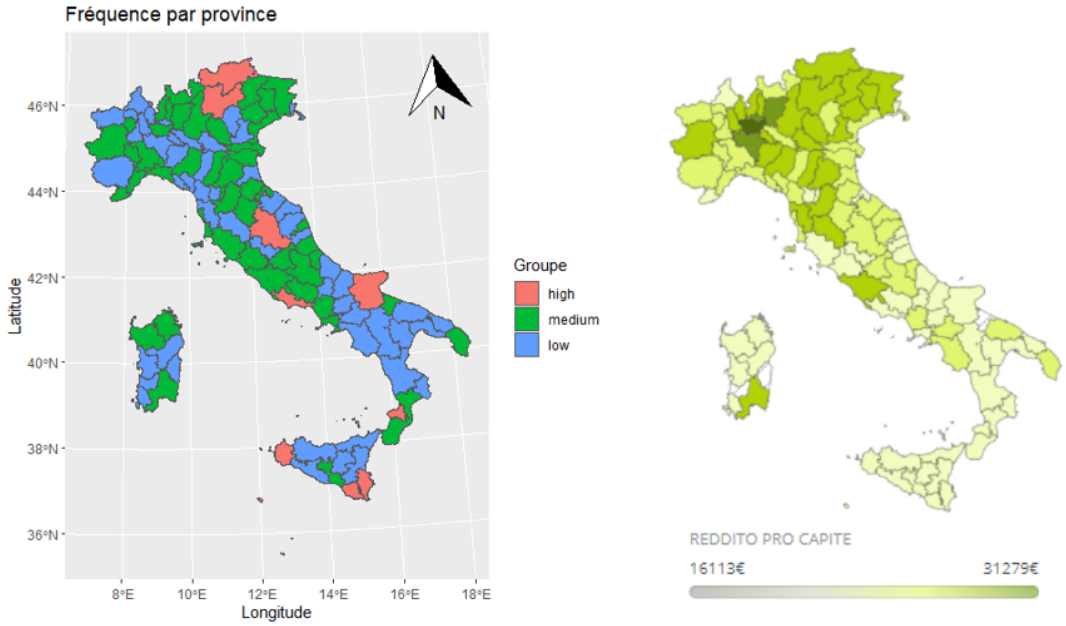

FIGURE A.1 – Comparaison des fréquences empiriques des garanties indemnitaires du produit *EURA Salute 360°* avec le revenu par habitant par province

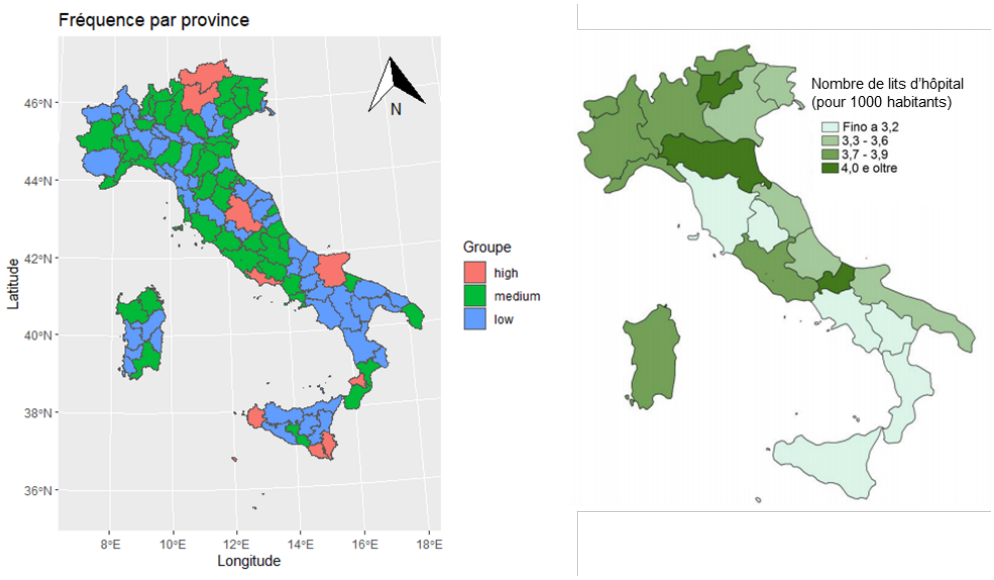

FIGURE A.2 – Comparaison des fréquences empiriques des garanties indemnitaires du produit *EURA Salute 360°* avec l'offre hospitalière par province

Annexe B

Sensibilité du modèle SIR aux paramètres β, σ, γ et η

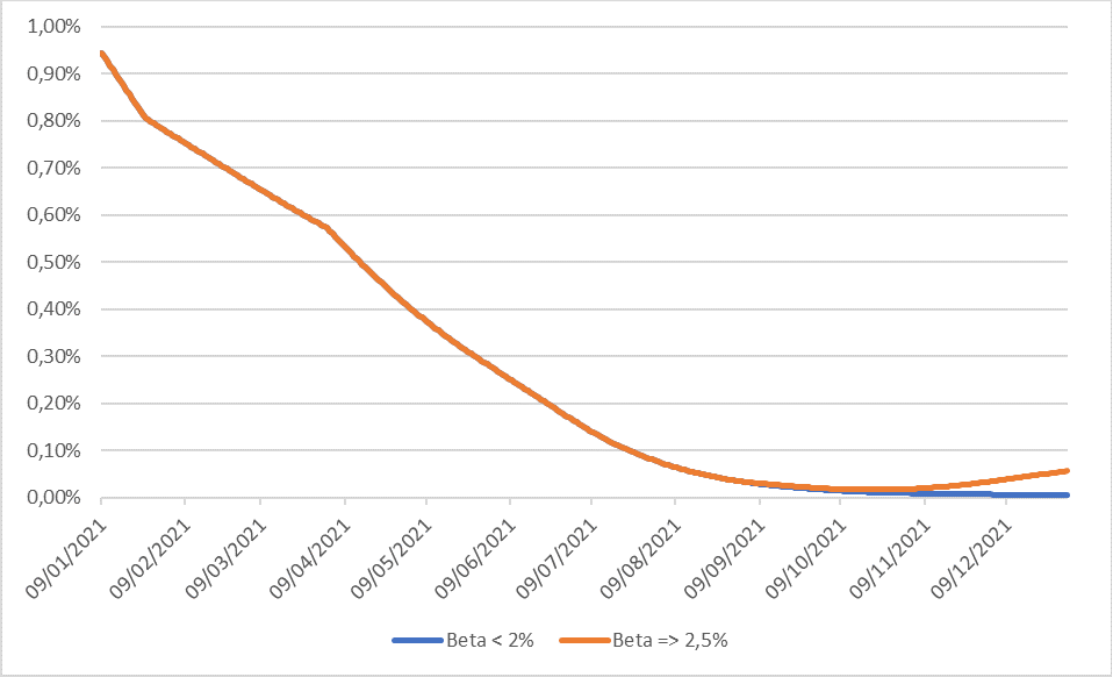

FIGURE B.1 – Projections du taux de personnes infectées dans la population italienne selon 2 niveaux du paramètre β (contagiosité du virus)

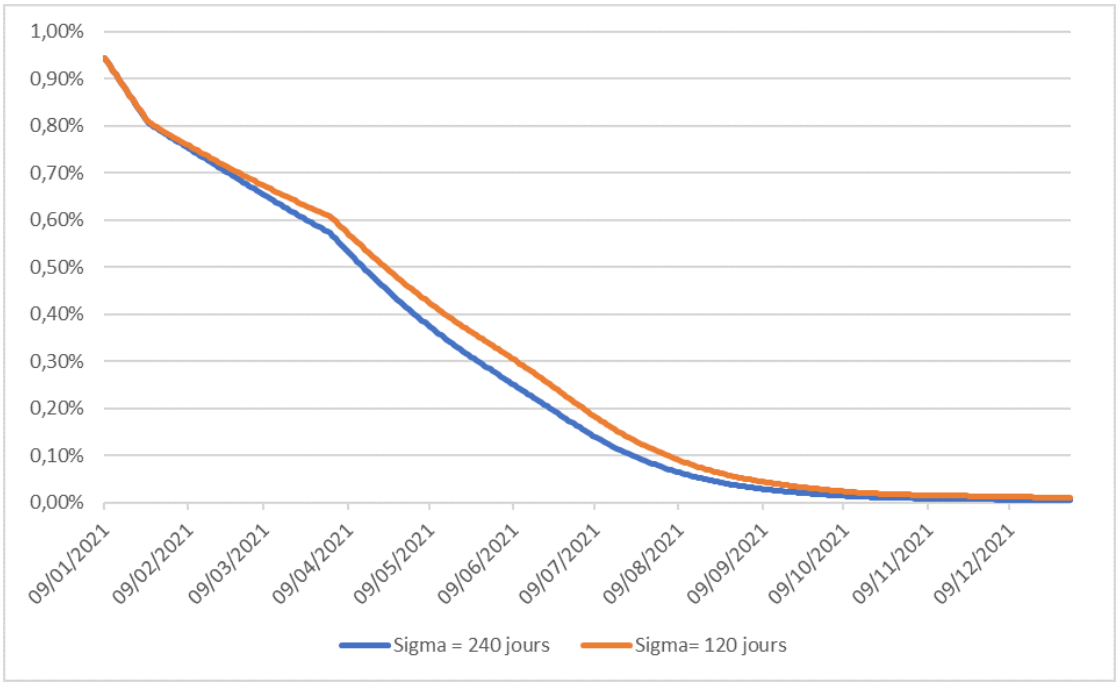

FIGURE B.2 – Projections du taux de personnes infectées dans la population italienne selon 2 vitesses de déclin de l'immunité σ

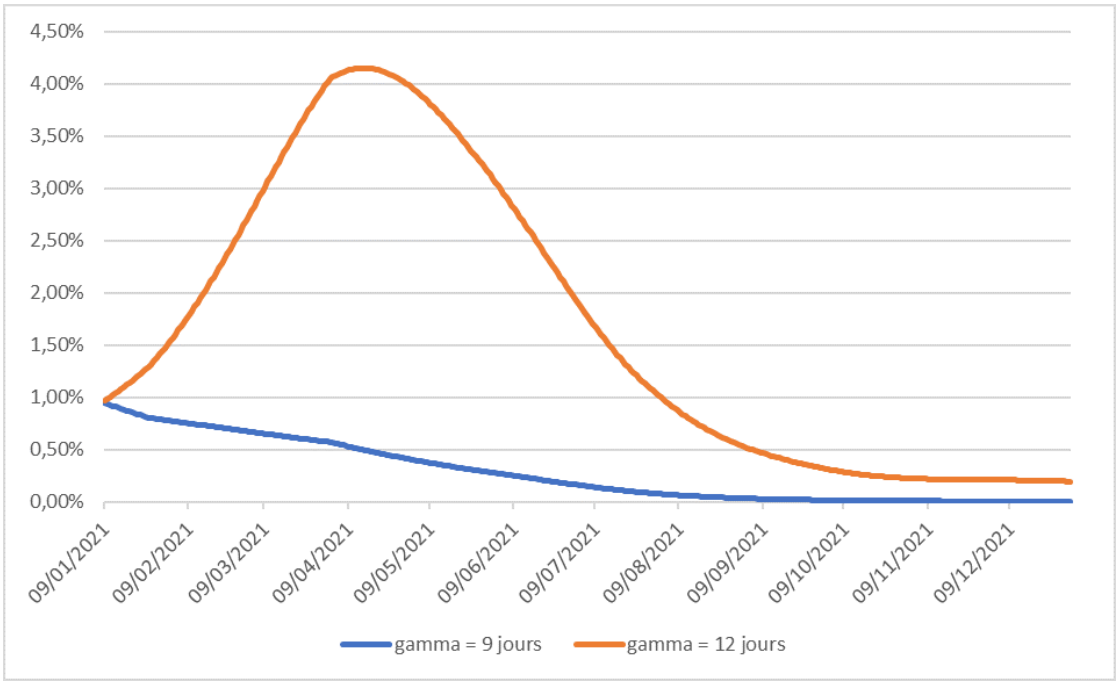

FIGURE B.3 – Projections du taux de personnes infectées dans la population italienne selon 2 vitesses de convalescence γ

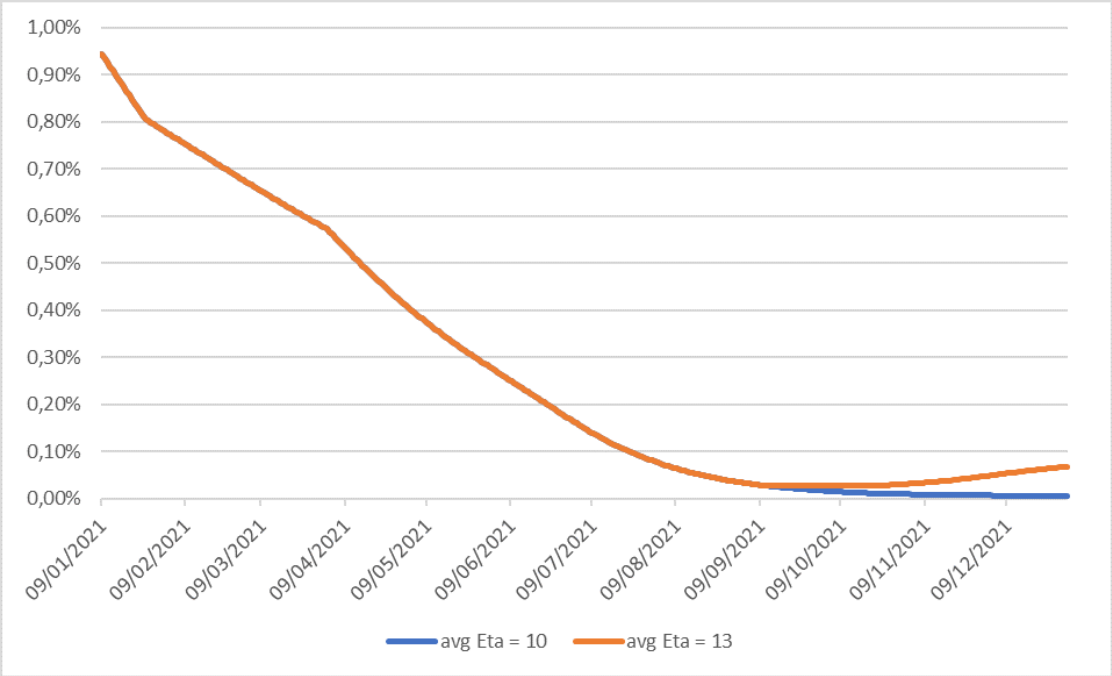

FIGURE B.4 – Projections du taux de personnes infectées dans la population italienne selon 2 niveaux de fréquence de contacts quotidiens η

Bibliographie

[A.Dieng, 2017] A.Dieng (2017). *Impact de l'évolution des dépenses de santé et de leur prise en charge sur les ménages.* Mémoire d'actuariat, Institut des actuaires.

[Ailliot, 2020] Ailliot, P. (2020). Théorie de la crédibilité. *In EURIA.*

[Akaike, 1974] Akaike, H. (1974). *A new look at the statistical model identification.* Institute of Electrical and Electronics Engineers.

[Bühlmann, 1967] Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin : The Journal of the IAA*, 4.

[Breiman *et al.*, 1983] Breiman, L. *et al.* (1983). *Classification And Regression Trees.* Wadsworth.

[Cereda *et al.*, 2020] Cereda, D. *et al.* (2020). The early phase of the COVID-19 outbreak in lombardy, italy. *Cornell University publications.*

[Dan *et al.*, 2021] Dan, J. *et al.* (2021). Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science magazine.*

[de Bordeaux et CRPV de Marseille, 2021] de Bordeaux et CRPV de Marseille, C. (2021). Enquête de pharmacovigilance du vaccin Pfizer BioNTech Comirnaty. Rapport n°16, ANSM.

[F.Husson, 2001] F.Husson (2001). Construire un modèle stochastique à partir d'un modèle déterministe. *Revue de statistique appliquée*, pages 5–27.

[Garnier, 2013] Garnier, S. (2013). *Application of Credibility Theory to Healt Pricing.* Mémoire d'actuariat, Institut des actuaires.

[Hastie *et al.*, 2001] Hastie, T. *et al.* (2001). *The Elements of Statistical Learning.* Springer.

[Kjaerulff et Madsen, 2008] Kjaerulff, U. et Madsen, A. (2008). *Bayesian networks and influence diagrams.* Springer.

[McKendrick, 1926] McKendrick, A. (1926). Applications of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, 14:98–130.

[Qualitiso, 2021] Qualitiso (2021). www.qualitiso.com. `https://www.qualitiso.com/`.

[Therneau et Atkinson, 2019] Therneau et Atkinson (2019). An introduction to recursive partitioning using the RPART routines. *R-CRAN.*

[Thérond, 2017] Thérond, P. (2017). Théorie de la crédibilité. *In ISFA*.

[Wikistat, 2021] Wikistat (2021). Arbres binaires de décision. `https://www.math.univtoulouse.fr/besse/Wikistat/pdf/st-m-app-cart.pdf`.

Liste des figures

Liste des acronymes

AIC Akaike Information Criterion
ALD affection de longue durée
CART Classification And Regression Tree
Covid Corona Virus Disease
Covid-19 Corona Virus Disease apparue fin 2019
EA Europ Assistance
EMA Agence européenne des médicaments
GHO Generali Head Office
GLM Generalised Linear Model
RMSE Root Mean Square Error
SIR Susceptible Infected Removed
SSN Servizio Sanitario Nazionale