Les apports de l'open data pour les assureurs

Remi BELLINA, IA
Antoine CAUTRU

Milliman

La capacité à accéder à de nouvelles sources de données offre de nombreuses perspectives concernant la gestion des risques des assureurs. L'évolution[1] du site data.gouv.fr, plateforme ouverte des données publiques françaises, constitue un bon exemple de ces évolutions. Les données se densifient, avec pour certaines, un renforcement de leur qualité au fil du temps.

Avec l'enrichissement du contenu et le développement du partage des informations, des concepts comme celui de l'*OSINT* (*Open Source Intelligence*) émergent. L'idée est d'acquérir (dans un cadre légal normé) des informations disponibles publiquement pour enrichir ses propres systèmes d'informations. Avec l'*OSINT*, l'objectif est de récupérer de l'information disponible sur la vie numérique des personnes et entreprises pour en déduire des informations sur la vie réelle (profilage). On parle aussi de *GeoInt* pour tenter de recontextualiser une photographie ou une vidéo, particulièrement utile en journalisme par exemple et dans le diagnostic de « *fake news* ».

Les données peuvent être issues du milieu académique et scientifique, il peut également s'agir de données commerciales, gouvernementales, et de données issues des médias, internet et réseaux sociaux. En général, l'*OSINT* est surtout présentée comme une approche permettant de lutter contre la fraude et les escroqueries dans les secteurs financiers.

Alors que les Insurtechs transforment le marché de l'assurance[2], tant sur le plan des services et de l'expérience utilisateur que de la création de produits innovants, les assureurs traditionnels misent également sur la donnée pour enrichir leur activité et la connaissance de leurs risques. La France confirme son statut de leader européen en matière d'Open Data[3].

Les enjeux en assurance

Un des enjeux des assureurs résulte en leur capacité à suivre leur exposition et ce, dans un contexte caractérisé par l'augmentation du risque cyber, les impacts du changement climatique, la pandémie. Ils bénéficient pour cela de données internes : localisation des risques soucrits, caractéristiques des biens et des individus, historiques des mouvements financiers et des estimations dossiers, dates clés et autres informations de classification interne. Des informations complémentaires peuvent aussi être utilisées : quelles étaient les conditions météorologiques au moment du sinistre et avaient-elles récemment évolué, quelles sont les caractéristiques du marché immobilier dans un quartier depuis les 5 dernières années dans un rayon de 10km, la zone est-elle factuellement plus risquée

[1] https://www.data.gouv.fr/fr/posts/retour-sur-les-activites-de-data-gouv-fr-en-2021/
[2] https://fr.milliman.com/fr-fr/insight/Developpement-et-perspectives-des-insurtechs-sur-le-marche-francais
[3] https://www.etalab.gouv.fr/retour-sur-lopen-data-maturity-index-2021-qualite-du-portail-de-donnees-3-4

en termes de criminalité ou d'accidents de la route, à quelle distance se situe l'hôpital le plus proche, etc. Toutes ces informations pourraient, à l'avenir, nourrir des modèles de Machine Learning pour affiner les modèles prédictifs, anticiper des dérives de sinistralité, modifier les structures tarifaires, etc. Ces évolutions nécessiteront bien entendu des investissements spécifiques.

En lien avec l'évolution des attentes des assurés, la relation client, le developpement de services au-delà de l'assurance font également partie des enjeux majeurs des assureurs. Et l'open data offre de nouvelles perspectives. Un questionnaire de souscription réduit au strict minimum, avec parfois la promesse d'une tarification à l'adresse, peut devenir peu à peu le standard à atteindre : il pourrait ne plus être demandé au prospect s'il dispose d'un appartement ou d'une maison, le nombre d'étages, la surface, l'année de construction du bâtiment, etc. Faciliter la souscription est souvent une première étape vers une digitalisation des échanges avec l'assuré.

Au-delà de la réduction des irritants ressentis par les clients, l'utilisation des données externes peut aussi avoir un intérêt mathématique lorsque l'on connait les « *proxies* » utilisés en tarification Automobile ou MRH par exemple. Certaines informations comme des zoniers de risque très précis pourront par exemple compléter ou remplacer des variables tarifaires.

Cette tendance à l'utilisation des données externes est sans nul doute à relier à l'essor des Data Labs[4] et la multiplication des profils formés à la data. Tantôt objectif à moyen terme, tantôt réalité concrète, souvent les deux à la fois, le développement d'approches data driven est certain. Cela s'accompagne de son lot inhérent de difficultés, de la disponibilité et qualité irrégulière de certaines données à l'implémentation dans les systèmes d'information non encore pleinement dimensionnés ou construits pour les intégrer. A titre d'exemple, les données ne sont jamais parfaitement géocodées (l'attribution de coordonnées GPS latitude et longitude à une adresse postale), alors même que la plupart des données externes sont reliées au risque par la localisation.

Cet essor de l'Open Data s'observe directement auprès des compagnies. Citons par exemple l'initiative de MAIF, visant à proposer des outils adaptés aux besoin de ses partenaires, et d'Open Assurance[5]. Ou encore le Generali Climate Lab créé en 2015 qui partage des informations à propos des risques climatiques sur LinkedIn[6]. Le partage et l'ouverture de certaines ressources des assureurs à des tiers (prestataires ou assurés) contribuent à entretenir ce vivier d'informations propice à la transformation du secteur.

Des données sous toutes leurs formes

Cette effervescence se traduit aussi dans la pluralité des données et des modalités d'accès : données disponibles « à plat » comme sur le site data.gouv, utilisation d'API (Google, OpenStreetMap, etc.), possibilités de web scrapping lorsque le cadre juridique le permet, certaines données peuvent également parfois être achetées.

Plusieurs types de données sont disponibles, et la liste n'est pas exhaustive :

- Données socio-démographiques (salaires, richesse, criminalité, chômage, habitudes de vie, tendances des consommateurs, tourisme, etc.)
- Données macroéconomiques
- Changements de lois / législation / réglementation (FGAO, arrêté Cat Nat sécheresse, flat tax, etc.)
- Données géologiques, météorologiques et prévisions climatiques long terme (humidité, précipitations, sécheresse, températures, nature des sols, arrêtés cat nat, etc.)
- Réseaux sociaux (Twitter, YouTube, Facebook, Instagram, LinkedIn, etc.)
- IoT ou Internet des Objets connectés (télématique, téléphones portables, montres connectées, health trackers, etc.)
- Données de Santé

Au-delà de l'utilisation de l'Open Data en assurance Dommages (auto, MRH), les possibilités en assurance vie sont aussi multiples, en lien notamment avec l'analyse de la santé ou de la mortalité. Dans deux papiers, nous présentons plus en détail les données externes pertinentes[7] et nous exposons des cas d'utilisation[8] qui ont été développés en assurance vie par les assureurs, les réassureurs et les Insurtechs. Tous ces cas ont prouvé leur capacité à générer de la valeur dans un large éventail d'applications.

En Santé, des analyses de la base DAMIR[9] ont également été publiées. La première étude vise à analyser l'impact de la COVID-19 sur les dépenses de santé et les indemnités journalières en France, en s'appuyant sur les données publiques de l'assurance maladie issues, d'une part des bases DAMIR et d'autre part des rapports mensuels AMELI de remboursements de la Sécurité Sociale. Elle montre notamment les impacts par grands postes de soins et les tendances 2021. La seconde étude est une analyse annuelle plus large en 2022

[4] https://fr.milliman.com/fr-fr/News-and-Events/Les-webinaires-Milliman : Retours d'expérience sur les challenges des Datalabs en Assurance
[5] https://entreprise.maif.fr/actualites/presse/2021/maif-deploie-portail-api ; https://github.com/MAIF
[6] https://www.argusdelassurance.com/green-assurance/generali-climate-lab-fer-de-lance-de-l-action-de-generali-dans-le-domaine-climatique.183259
[7] https://www.milliman.com/en/insight/potential-data-sources-for-life-insurance-ai-modelling
[8] https://www.milliman.com/en/insight/the-use-of-artificial-intelligence-and-data-analytics-in-life-insurance
[9] https://www.milliman.com/en/insight/impacts-sur-les-depenses-de-sante-de-la-covid19-en-france ; https://fr.milliman.com/fr-fr/insight/Barometre-2022-des-depenses-de-sante-en-France

des dépenses de santé et des indemnités journalières en France.

Partout, la question de l'éthique et de la légalité est un sujet crucial. A titre d'exemple, alors qu'il est possible d'acheter des données de santé aux US, la France est beaucoup plus protectrice. De même, l'utilisation des réseaux sociaux est source de controverses, tout comme d'autres discriminations liées au genre, la religion, etc.

Geo Analytics

Il est fréquent de relier des données externes de différentes sources par la localisation. Ainsi, les bases de données des assureurs sont complétées avec des données externes à la maille de la région, du département, du code postal, du code INSEE, voire de l'adresse. On peut alors parler de *Geo Analytics*. Nous présentons ci-dessous plusieurs sources.

La plateforme ouverte DataGouv des données publiques du Gouvernement français offre plus 40.000 jeux de données, divisés en ressources (plus de 200.000)[10] qui correspondent aux tables disponibles dans chaque jeu de données. Les bases de données (ressources) sont souvent téléchargeables au format CSV. Certaines ressources ont fait l'objet de réutilisations[11] : cartes interactives, recherche de données par filtres, apport d'informations. Voici quelques jeux de données géolocalisées (latitude / longitude ou alors par département / commune) à titre d'exemple :

- Indices annuels d'anomalies de précipitations issus du modèle Aladin-Climat, avec des données fournies par Météo France. Les prévisions incluent des valeurs jusqu'à l'horizon 2100, réparties sur trois horizons temporels : horizons proche (2021-2050), moyen (2051-2070) et lointain (2071-2100).
- Accidents corporels de la circulation routière (pour les années 2005 à 2020).
- Demande de Valeurs Foncières (DVF) : cette base contient les transactions immobilières intervenues au cours des cinq dernières années. Les données sont issues des actes notariés et des informations cadastrales.
- Crimes et délits enregistrés par les services de police et de gendarmerie : pour chaque poste de CGD (Compagnie de Gendarmerie nationale).

Le site Géorisques, créé par le ministère chargé de l'environnement afin d'évaluer les risques naturels et technologiques autour d'un point donné, constitue également une très bonne source, et peut être consulté de deux manières différentes ; via une API en ligne ou via le téléchargement de données.

L'INSEE, l'Institut National de la Statistique et des Etudes Economiques, collecte, produit, analyse et diffuse des informations sur l'économie et la société française. Les données sont souvent au format CSV. On peut par exemple citer la base SIRENE des entreprises et de leurs établissements (SIRET), combinée aux données sur les codes APE, et les géolocalisations de chaque établissement, largement exploitée par les assureurs.

Au delà du marché français, OpenStreetmap est une base de données mondiale collaborative, avec des données qui peuvent être extraites via l'API nommée *Overpass*. Google propose aussi de nombreuses API (la plupart payantes), dont une permet d'obtenir des images de *Streetview* et une autre des images satellites à partir d'une adresse postale ou d'un point (lat, lon). Certains services, par exemple l'obtention d'un historique des vues sur *Streetview*, ne sont toutefois pas disponibles par API.

Des sources spécifiques sur des sujets précis peuvent également être intéressantes. A titre d'exemple, le site de la CCR recense la liste des arrêtés Cat Nat, par publication au Journal Officiel, avec la décision retenue pour chaque commune ayant réalisé une demande de reconnaissance, depuis les années 1980 à aujourd'hui, et pour différents périls (dont les inondations et la sécheresse).

Illustrations de quelques cas d'usage

Risque

La récupération des données, automatisée et transparente pour l'assuré, peut permettre d'améliorer la connaissance du risque et la qualité prédictive des modèles. En assurance auto, il est possible par exemple de récupérer les limitations de vitesses environnantes, de faire un diagnostic de l'accidentologie, etc. Des statistiques à différentes mailles géographiques peuvent alors être établies. À une maille plus fine, en l'occurrence les coordonnées GPS, il est possible de dresser une analyse de points accidentogènes.

REPRESENTATION DES ACCIDENTS CORPORELS (BAAC)

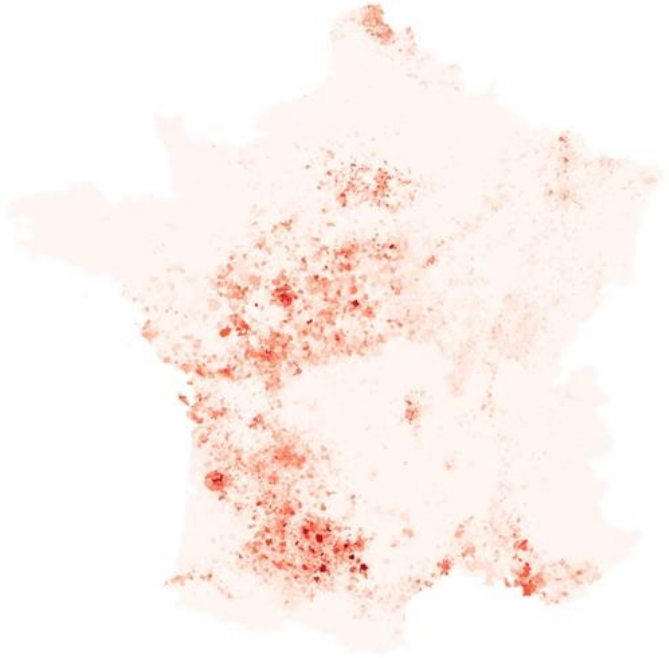
Source des données : https://www.data.gouv.fr/fr/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2020/

[10] https://www.data.gouv.fr/fr/

[11] https://www.data.gouv.fr/fr/pages/onboarding/reutilisateurs/

En assurance habitation, les caractéristiques techniques des sols, les historiques de température et de précipitations, etc. peuvent contribuer à l'amélioration des modèles. Des zoniers peuvent ainsi être établis.

UN ZONIER DU RISQUE SECHERESSE



Source des données : CCR, arrêtés Cat Nat

Les possibilités sont multiples et il ne s'agit donc que d'illustrations.

Qualité De Donnée (QDD) et lutte contre la fraude

La QDD, mais aussi la détection de fraude, peuvent être améliorées par la prise en compte de données externes. Un individu a-t-il bien renseigné ses informations ? La surface déclarée du bien est-elle cohérente avec le cadastre ? Est-ce bien un appartement ou plutôt une maison ? Une fissure n'était-elle pas antérieure à la date supposée du sinistre ? Un orage qui pourrait expliquer les dégâts électriques a-t-il bien eu lieu dans la zone à une date donnée ?

La comparaison des bases de données internes à l'assureur (partagées par le client, et transcrites par le gestionnaire) avec des informations publiques rassemblées (ex : nombre d'étages, piscine et dépendance) peut permettre de compléter des informations manquantes ou de vérifier des données existantes, mais aussi d'aider à la détection de déclaration frauduleuse.

Le traitement automatique d'images (vue Google *Streetview* ou vue aérienne) peut s'avérer également pertinent. Pour autant, le traitement de données non structurées reste aujourd'hui un défi, en témoignent les résultats des travaux de la direction des finances publiques pour repérer les fraudes à la taxe foncière[12].

ILLUSTRATION DE L'API GOOGLE POUR EXTRAIRE UNE VUE STREETVIEW, ET UTILISATION D'OPENCV (SOUS PYTHON)

Source : Google

Souscription et service

L'ajout de données peut permettre d'alléger les questionnaires tarifaires lors d'une nouvelle souscription. Avec une simple adresse, il est désormais possible de répondre automatiquement à un certain nombre de questions qui alimentent les modèles, et éventuellement d'établir des approximations.

L'utilisation des données externes offre la possibilité d'innover en ce qui concerne les services proposés aux assurés. A titre d'illustration, un *dashboard* enrichi de données pertinentes accompagne le futur propriétaire lorsqu'il tape une adresse postale, avec une tarification MRH à l'adresse instantanée mais aussi des informations publiques mises à disposition comme l'état du marché immobilier dans le secteur ou encore le diagnostic de la couverture internet (exemple de service similaire[13]). Cela peut s'accompagner du développement de marketplace (proposant des services de déménagement agréé, artisans, etc.).

Prévention

Enfin, la prévention des risques est également un aspect sur lequel l'Open Data offre des perspectives intéressantes. La plupart des assureurs ont ainsi mis en place des alertes automatiques (conditions météo risquées par exemple). L'idée est à la fois d'apporter un service aux assurés et d'œuvrer à

[12] https://www.larevuedudigital.com/lintelligence-artificielle-pour-detecter-les-piscines-non-declarees-en-question/

[13] https://www.maif.fr/habitation/achat-immobilier/visites#definir-vos-criteres

réduire la sinistralité. Un trafic routier anormalement dense et une météo défavorable pourraient peut-être inciter l'automobiliste qui le peut à prendre les transports en commun.

Perspectives actuarielles

Les cas d'usage sont multiples, de la « simple » utilisation d'une donnée externe structurée pour enrichir un zonier IARD à la mise en œuvre de techniques de deep learning pour détecter des fissures sur des façades de bâtiment ou d'habitations à risques à partir d'images.

Les sujets éthiques sont primordiaux dans l'utilisation des données externes. Le RGPD permet d'assurer la défense des droits à la vie privée, apporte un cadre dans lequel il est possible de développer et invite au questionnement : Les algorithmes qui interprèteront ces données sont-ils éthiques ou reproduiront-ils les biais de notre société ? Tout est bien évidemment une question de nuance, il est certain que l'Intelligence Artificielle a beaucoup à apporter, et que cela passe par la donnée. La tendance actuelle à la transparence, le partage, semble aller dans le bon sens et l'Open Data semble être une condition indissociable des enjeux d'aujourd'hui.

Après l'actuariat et la data science, de nouvelles compétences en *OSINT* et en cartographie deviennent nécessaires : connaissance des sources et de leur qualité, études statistiques, modélisation. Avec une vision transverse du marché, et des expériences dans ces différents domaines, le cabinet Milliman accompagne ses clients sur ces sujets en France et à l'international.

.

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

CONTACTS

Remi BELLINA, IA
remi.bellina@milliman.com

Fanny POUGET, IA
fanny.pouget@milliman.com

Fabrice TAILLIEU, IA
fabrice.taillieu@milliman.com

milliman.com