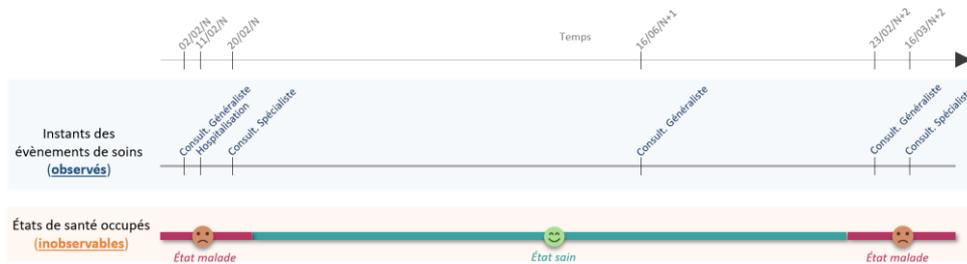




Cas d'usage des MMPP (Markov-modulated Poisson process) en assurance santé : vers une objectivation de l'évolution de l'état de santé des assurés



Un contexte de dérive des prestations qui rend nécessaire une orientation des OCAM vers plus de prévention

Ces dernières années, les acteurs de l'assurance santé complémentaire connaissent une dérive graduelle de leurs prestations. Plusieurs facteurs sont à l'origine de ce phénomène tels que les réformes règlementaires (100% santé, transferts de charge du régime obligatoire vers les OCAM), le vieillissement de la population ou encore le rattrapage des soins post-Covid.

De façon complémentaire, on peut légitimement se demander si en marge de ces effets déjà bien connus et mesurés, ne se profile pas également **une détérioration structurelle de la santé des Français**. C'est une hypothèse qui devient prégnante au regard des problématiques d'accessibilité aux soins ou de changement climatique qui ont émergé.

Pouvoir mesurer cet effet, au moins à l'échelle d'un portefeuille, devient alors crucial pour les organismes d'assurance. Leurs données, notamment celles relatives aux prestations réglées, représentent alors une ressource précieuse, car **elles constituent le meilleur reflet de la santé des assurés qu'il leur est possible d'avoir**. Sous réserve évidemment d'une utilisation respectant le cadre du RGPD, elles peuvent alors former le socle d'une modélisation de l'état de santé, qui permet d'analyser finement et objectivement les tendances d'évolution du risque santé dans un portefeuille. Il serait ainsi possible de définir des stratégies de prévention efficaces permettant de contenir les effets liés à la dégradation de l'état de santé, alternative possible aux indexations tarifaires qui semblent aujourd'hui inéluctables.

La santé vue comme une chaîne de Markov à temps continu qui pilote la chronique de soins

Les données de prestations santé auxquelles ont accès les OCAM recensent une grande partie des soins dont bénéficie un assuré (modulo les prestations prises en charge exclusivement par le régime obligatoire qui sont absentes des SI des OCAM – pharmacie prise en charge à 100%, soins sans dépassement d'honoraire liés à une affection de longue durée).

Par conséquent, et même si elles peuvent parfois manquer de précision dans leur qualification, elles sont l'expression la plus représentative de l'état de santé des assurés du portefeuille. Tout l'enjeu est donc de réussir à construire un modèle permettant de passer de données de prestations santé **observées** à l'objectivation de l'état de santé qui est **inobservable**.

Une idée, inspirée des travaux de [Mews et al., 2023] est alors de considérer l'évolution du nombre d'interactions avec le système de soins de chaque assuré (dont a connaissance l'assureur) comme un **processus de Poisson observable** dont l'intensité est modulée par une **chaîne de Markov à temps continu (CMTC) inobservable à deux états (état « sain » et état « malade »)** : c'est le cadre des **processus de Poisson à modulation markovienne**.

C'est donc la chaîne de Markov qui formalise le concept d'état de santé, dans la mesure où les états de la chaîne dictent de manière sous-jacente l'intensité des événements de soins observés. Le schéma ci-dessus illustre avec la chronique de soins d'un assuré donné.

Il est alors nécessaire d'estimer les paramètres cruciaux de ce modèle, qui sont dans ce cas d'étude :

Paramètre	Les intensités des soins (une par état)	La matrice génératrice Q (matrice carrée d'ordre 2, soit autant que le nombre de états de la CMTC)
Rôle	Elles indiquent le nombre attendu de recours aux soins par unité de temps.	Elle caractérise la dynamique de la chaîne, en définissant notamment les temps de séjours moyens dans chaque état.
Exemple	Un état portant une intensité de 1/365 indique que sur 365 jours de séjour dans cet état, un seul événement de soins sera en moyenne observé, c'est donc un signe d'état sain.	Si $Q = \begin{pmatrix} -1/400 & 1/400 \\ 1/40 & -1/40 \end{pmatrix}$, le temps de séjour moyen en état sain est donné par $-1/Q_{1,1} = 400$ jours et celui en état malade par $-1/Q_{2,2} = 40$ jours.

Par ailleurs, **le modèle est suffisamment souple pour intégrer des variables exogènes**. Il est par exemple possible de considérer que la dynamique de la santé de chaque assuré est pilotée par ses caractéristiques démographiques à travers un effet multiplicatif.

DATA INSIGHTS



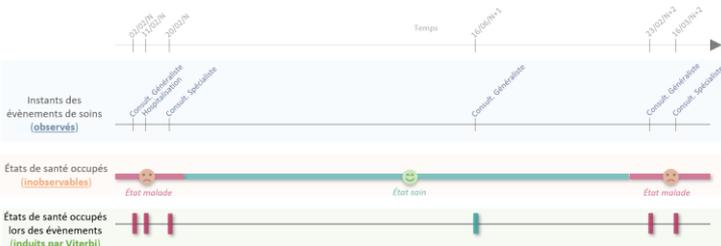
Autrement dit, les éléments du générateur peuvent s'écrire de cette façon :

$$\forall i, j \in \{1,2\}^2, Q_{ij} = \exp(\beta_{ij}^T Z)$$

Z étant un vecteur de variables démographiques (sexe, âge, profil de risque...) et β_{ij} une matrice de coefficients à estimer.

A l'inverse, les intensités de soins par état seront supposées être des paramètres communs à tous les assurés, afin pouvoir garder une comparabilité dans la mesure du niveau de santé.

Avec ces hypothèses, les paramètres peuvent alors être inférés par maximum de vraisemblance à partir des observations de tous les assurés. Cela permet d'appliquer l'intérêt majeur de ce modèle, à savoir **le décodage de la séquence d'états qui a le plus vraisemblablement mené à la chronique de soins observée grâce à l'algorithme de Viterbi**. Pour chaque assuré, il permet en effet de dire de chacun de ses événements de soins observés s'il s'inscrivait plausiblement dans une période de santé saine ou plus dégradée, en s'appuyant notamment sur le voisinage dudit événement dans la chronique.



Dès lors, en interpolant les états décodés, un indicateur dont l'évolution est intéressante à observer est par exemple **la part de temps passé en bonne santé par trimestre**. Cette maille temporelle permet de disposer d'un temps suffisamment long pour dresser des constats pertinents, tout en gardant une finesse pour rester agile dans la prise de décisions. Ces proportions peuvent de plus être moyennées selon des groupes d'individus.

Suivre les tendances d'évolution objective de la santé

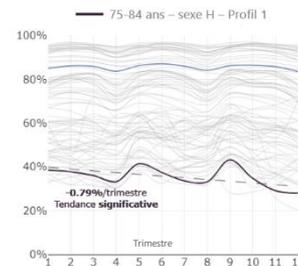
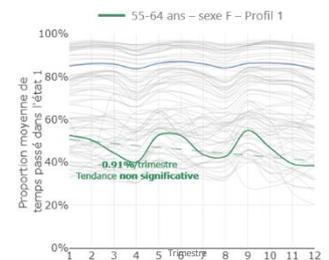
Ainsi, observer l'évolution de cet indicateur sur quelques trimestres dans un portefeuille sert à mener **une étude de tendance afin d'exhiber les typologies d'assurés auprès desquelles des variations sont notables**. Un test de Mann-Kendall permet en effet de déterminer si une tendance monotone significative existe dans cette série chronologique, tendance qui peut être quantifiée de façon robuste par un calcul de pente de Theil-Sen.

Cette publication a été réalisée sous la direction de **Eugénie POYET**, Director Pricing & Analytics Life & Health **Nabil RACHDI**, Head of Data Science avec l'expertise de **Harold ATONZONG**, Consultant Pricing & Analytics Life & Health

Moyenner toutes les proportions trimestrielles pour tous les assurés **revient à obtenir un niveau de santé moyen du portefeuille, qui apparaît stable** avec des fluctuations à la baisse sur les trimestres hivernaux (courbe bleue dans les graphiques ci-après).

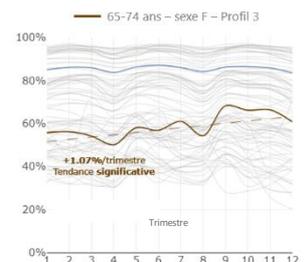
Et en descendant à une maille inférieure, celle des typologies (sexe x tranche d'âge x profil de risque défini a priori), il y a plus de disparité dans les trajectoires d'évolution du risque modélisé (chacune des courbes grises est associée à une persona du portefeuille). Trois d'entre elles ont été mises en exergue pour illustrer le triple intérêt de la méthodologie :

- Identifier les groupes d'assurés **à risque de détérioration saisonnière plus marquée** lors des périodes hivernales par exemple, chez qui des incitations à la vaccination antigrippale pourraient entre autres être menées chaque année.



- Identifier les groupes **à risque de détérioration progressive et significative**, auprès desquels il serait alors pertinent de suggérer et d'étudier des actions de prévention plus spécifiques.

- Identifier les groupes qui voient leur santé ainsi modélisée **s'améliorer de manière manifeste**, ce qui pourrait permettre de voir les effets postérieurs à la mise en place de plans de prévention et de chiffrer leur impact.



Pour conclure, l'exploration de ce nouveau formalisme de modélisation de la santé présente des possibilités d'analyses concrètes, via notamment la flexibilité de la définition des hypothèses. Les possibilités de configuration sont alors nombreuses, et donnent à cette démarche un intérêt certain pour la compréhension fine des risques santé et de leur évolution en utilisant uniquement les données présentées dans les bases de prestations des OCAM.

Bibliographie

Bebbington, M. S. [2007]. Identifying volcanic regimes using Hidden Markov Models. *Geophysical Journal International*, 171(2), 921-942.

Mews, S., Surmann, B., Hasemann, L., and Elkenkamp, S. [2023]. Markov-modulated marked Poisson processes for modeling disease dynamics based on medical claims data. *Statistics in Medicine*.