METHODES STATISTIQUES POUR LA DETECTION D'EVENEMENTS INHABITUELS POUR LA SURVEILLANCE SYNDROMIQUE

Mémoire Santé Publique 2017/2018

Rédigé par :
Johann KUHN – Étudiant au M1 Master Santé Publique

Tuteur : Yann Le Strat – Directeur de l'équipe DATA
(yann.lestrat@santepubliquefrance.fr)

Remerciements

Je tenais à remercier tout particulièrement Yann LE STRAT, directeur de l'équipe DATA à Santé Publique France et mon tuteur, de m'avoir accepté et intégré dans son équipe et de m'avoir accompagné, conseillé et expliqué les points très statistiques durant ce stage, malgré l'emploi du temps chargé.

Je remercie également Isabelle PONTAIS de m'avoir fourni les données rapidement et toute la documentation nécessaire qui m'a aidé à comprendre les différentes sources de données.

Résumé

Contexte : La connaissance de l'état de santé de la population française repose sur le recueil d'informations permettant d'estimer le poids de certaines maladies, leurs dynamiques dans le temps, mais aussi d'identifier les facteurs associés aux pathologies et d'estimer leurs forces d'association. Santé Publique France a mis en place plusieurs types de recueils (systèmes de surveillance, enquêtes), notamment la surveillance syndromique. Celle-ci est née de la canicule de 2003 qui a remis en cause l'efficacité des dispositifs de veille d'alerte sanitaires pour ce type d'événement. À la suite de cet événement, SurSaUD®, un système de surveillance syndromique, a été créé pour être capable de détecter quotidiennement des menaces pour la santé publique à partir de différentes sources de données.

Objectif : Mettre en place des méthodes statistiques de détection temporelle et évaluer leurs performances afin d'aider les épidémiologistes à analyser les données de la surveillance syndromique.

Matériel : Deux sources de données qui composent SurSaUD® ont été utilisées, provenant des réseaux OSCOUR® et SOS Médecins. Des données journalières pour cinq regroupements syndromiques, sur la période de janvier 2010 à décembre 2017, ont été extraites.

Méthodes : Les données journalières ont été agrégées en données hebdomadaires selon trois niveaux géographiques : (1) France métropolitaine, (2) DOM-COM, (3) France entière. Les méthodes statistiques proviennent du package R surveillance et quatre algorithmes ont été retenus : RKI, Bayes, EARS et Farrington.

Résultats : Au cours de la période 2015-2017, les quatre algorithmes ont généré, respectivement, 10, 1, 2 et 1 alarme(s) pour un regroupement syndromique particulier.

Discussion : L'automatisation de ces méthodes est prévue bientôt puis une évaluation des performances et de l'utilité de ces méthodes sera conduite en lien avec les épidémiologistes de l'agence, notamment en régions.

Table des matières

Introduction

La connaissance de l'état de santé de la population française repose sur le recueil d'informations permettant d'estimer le poids de certaines maladies (prévalences), leurs dynamiques dans le temps (incidences), mais aussi d'identifier les facteurs associés aux pathologies et d'estimer leurs forces d'association. Les informations recueillies par une agence comme Santé publique France reposent sur la mise en place de plusieurs types de recueil : (1) les systèmes de surveillance spécifiques qui s'intéressent à une maladie précise, (2) la surveillance syndromique qui ne cible pas une maladie en particulier, (3) les enquêtes épidémiologiques (transversales, cohortes, cas-témoins, etc.) et (4) les bases de données médico-administratives, notamment le système national des données de santé (SNDS).

La surveillance syndromique est née, en France, des conséquences de la canicule de 2003 qui ont conduit à réexaminer les dispositifs de veille et d'alerte sanitaires alors disponibles et quasi-exclusivement basés sur des systèmes spécifiques et par pathologie. L'objectif était de développer une capacité à détecter de nouvelles menaces pour la santé publique. Ces menaces peuvent avoir différentes origines comme par exemple un phénomène environnemental ou bien une pathologie infectieuse émergente.

Dans cette perspective, l'Institut de veille sanitaire (InVS) a développé en 2004 un système de surveillance, baptisé SurSaUD® (surveillance sanitaire des urgences et des décès), centré sur des structures capables de fournir des informations au jour le jour sur l'état de santé de la population.[1] Aujourd'hui ce système inclut quatre sources d'informations : (1) les données des services d'urgences hospitaliers adhérant au réseau OSCOUR® (organisation de la surveillance coordonnée des urgences), (2) les données des associations SOS Médecins, (3) les données de mortalité des services informatisés d'état-civil transmises par l'Insee (institut national de la statistique et des études économiques) et (4) les données de certification électronique des décès (CépiDc).

Les objectifs scientifiques du système sont (1) de détecter un événement sanitaire inattendu, (2) d'estimer l'impact d'un événement environnemental ou sociétal, (3) de surveiller des pathologies en dehors de tout événement et (4) de détecter précocement un événement sanitaire prédéfini, tel qu'une épidémie saisonnière, en mesurer l'impact et les conséquences.

L'objectif de ce travail s'inscrit dans le premier objectif de la surveillance syndromique, c'est-à-dire la détection temporelle d'un événement inhabituel. On définit ici un événement inhabituel comme un intervalle de temps dans lequel le nombre de cas rapportés au système de surveillance est significativement supérieur au nombre attendu de cas. Lorsqu'un événement inhabituel est détecté, une alarme statistique est générée. Cette alarme est un signal statistique qui, avec d'autres signaux,

notamment épidémiologiques, permettent d'aider les épidémiologistes en charge de la surveillance à identifier des menaces potentielles pour la population.

La détection d'un événement inhabituel repose donc sur deux étapes qui font appel, la plupart du temps, à des méthodes statistiques. La première étape consiste à estimer un nombre de cas à partir d'une méthode statistique en se basant sur le nombre de cas déclarés dans le passé. La seconde étape consiste à comparer le nombre attendu au nombre observé en utilisant un test statistique. L'utilisation des méthodes statistiques est d'autant plus incontournable que les analyses doivent se faire en routine et sur un très grand nombre de séries temporelles.

L'objectif final de ce stage est d'appliquer des méthodes statistiques pour la détection temporelle d'événements inhabituels sur les données d'OSCOUR® et de SOS Médecins. Pour parvenir à cet objectif, en dehors des aspects techniques de data-management qui ont pris du temps, une revue de la littérature a été nécessaire afin d'identifier plusieurs méthodes statistiques à retenir parmi la vingtaine de méthodes existantes. Ces méthodes ont été implémentées et des premières analyses ont pu générer des alarmes. La question de recherche la plus importante que l'on s'est posée lors de ce stage est la suivante : **les méthodes implémentées ont-elles des performances suffisantes pour apporter aux épidémiologistes une aide à la décision ?**

La réponse à cette question n'est pas simple car le choix des méthodes s'est fait à partir de la littérature qui est essentiellement basée sur des données de surveillances spécifiques et non syndromique. On sait par ailleurs que les données de surveillance syndromique ont des spécificités : elles sont journalières et présentent des saisonnalités qui peuvent être multiples (saisonnalité dans la semaine, influence des jours fériés, des week-ends, etc.). D'autre part qu'appelle-t-on des performances suffisantes ? Il est d'abord nécessaire de choisir des mesures de performances puis d'identifier des méthodes qui tout en ayant, par exemple, une bonne sensibilité ont également une bonne spécificité afin de ne pas générer un trop grand nombre d'alarmes. En effet, si des alarmes sont générées trop souvent, les épidémiologistes ne les considèrent pas car ils n'ont pas le temps d'analyser chaque alarme.

Ce rapport présente tout d'abord les deux sources de données utilisées pour ce travail, puis les méthodes statistiques choisies et enfin les premiers résultats obtenus. Il se termine par une discussion afin notamment d'identifier les forces et les limites de ce travail et d'envisager les perspectives qui permettront de répondre à la question principale de ce stage.

Matériel

Réseau OSCOUR®

Le réseau OSCOUR® a été mis en place en juillet 2004 à partir de 23 structures d'urgence hospitalières. Il a connu une montée en charge constante. En 2018, 689 structures d'urgence participent au réseau, couvrant 93,2% des passages aux urgences en France, ce qui correspond à l'enregistrement quotidien d'environ 55800 passages (Figure 1). Les données sont enregistrées en routine à partir du dossier médical du patient et du « résumé de passage aux urgences » (RPU). Elles comprennent des variables socio-démographiques (sexe, âge), administratives et médicales (diagnostic principal, diagnostics associés, degré de gravité, mode de transport, etc.). Les diagnostics médicaux sont codés selon la classification internationale des maladies, 10ème version (CIM-10).

La transmission des données se fait automatiquement et quotidiennement à partir des logiciels métier des structures d'urgences par l'intermédiaire de serveurs régionaux hébergés par les Agences Régionales de Santé (ARS) ou d'autres partenaires régionaux, vers Santé publique France qui peut analyser dès 4h du matin les données de la veille.

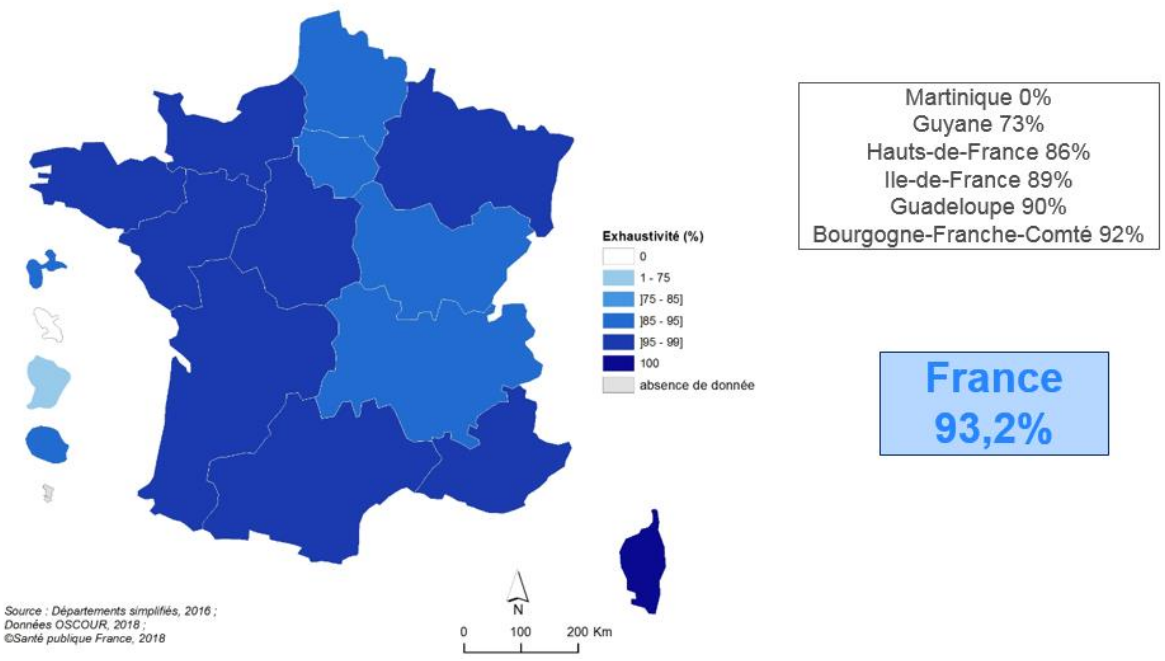

Figure 1 : Couverture nationale du réseau OSCOUR® en 2018

Réseau SOS Médecins

Le réseau SOS Médecins est un réseau d'associations d'urgentistes libéraux qui sont des centres de régulation médicale de statut libéral et qui participent à la permanence des soins ambulatoires en collaboration avec le Samu. Les associations SOS Médecins sont réparties sur l'ensemble du territoire français et couvrent la plupart des grands centres urbains et leur périphérie.

Depuis mai 2006 ces associations participent au système de surveillance SurSaUD®. En 2006, 24 associations participaient, elles sont maintenant, en 2018, 63 associations à participer à la surveillance, sur les 64 existantes. Des permanenciers réceptionnent les appels des patients, enregistrent les variables démographiques ainsi que le ou les motifs d'appel. Ils transmettent les informations à un médecin qui se rend au domicile du patient pour une consultation et pose le diagnostic médical. Les variables récoltées sont similaires à celles collectées par OSCOUR®. Les motifs d'appel et les diagnostics sont codés selon deux thésaurus spécifiques utilisés par chaque association. Au niveau national, le pourcentage de codage est de 94,2%.

Chaque matin, les données de SOS Médecins sont envoyées sur la plateforme nationale SOS Médecins France qui rassemble l'ensemble des données reçues dans une seule base qui est transmise à Santé publique France avant 6h. Le nombre moyen de visites est de 10 518 par jour en 2018.

Construction de regroupements syndromiques

Les diagnostics médicaux et motifs de consultations correspondant à une maladie ou à un symptôme sont codés par les médecins aux urgences avec la CIM-10 qui compte environ 40 000 codes et par SOS Médecins avec leurs propres thésaurus qui comptent environ 1000 codes pour les diagnostics médicaux et 750 codes pours les motifs de recours. Les codes motifs ou diagnostics sont regroupés par Santé publique France en catégories qui ont un sens pour la surveillance sanitaire. Ces catégories sont appelées des syndromes ou « regroupements syndromiques ». Ils sont construits de façon à couvrir une grande partie des diagnostics ou motifs enregistrés.

Quels regroupements étudiés ?

Un groupe de travail, interne à Santé publique France, a proposé une liste d'analyses à réaliser sur les données de la surveillance syndromique. En ce qui concerne la détection d'événements inhabituels, le groupe a choisi la liste suivante de regroupes syndromiques : traumatisme, chute, infections ORL, voies respiratoires hautes, dyspnée, insuffisance respiratoire, bronchite, pneumopathie, bronchiolite,

grippe/syndromes grippaux, asthme, douleurs thoraciques, ischémie myocardique, insuffisance cardiaque, hypertension artérielle, troubles du rythme et de la conduction, gastro-entérite, vomissements, diarrhée, douleurs abdominales, méningite, neurologie autre (épilepsie, céphalées, vertiges), fièvre et éruption cutanée, dermatologie autres, brûlures, colique néphrétique, infection urinaire, troubles anxieux, stress, démence/désorientation, altération de l'état général, malaise, fièvre isolée, déshydratation, conjonctivite, céphalée, hypotension/choc, TIAC, rougeole, oreillon, coqueluche.

Dans le cadre de ce stage, nous avons travaillé sur les regroupements syndromiques présentés dans le Tableau 1. L'identifiant du regroupement syndromique permet d'attribuer un code unique aux différents syndromes et permet d'identifier la source des données (OSCOUR® ou SOS Médecins).

Regroupement syndromique	Code OSCOUR	Code SOS Médecins
Douleurs thoraciques	23	/
Ischémie myocardique	40	DSO1326
Insuffisance cardiaque	15	DSO1323
Hypertension artérielle	33	DSO1315
Troubles du rythme et conduction	53	DSO1228
Toutes causes	TOUS_T_DRP	TOUS_T_DSO

Tableau 1 : Regroupements syndromiques et leurs codes associés pour OSCOUR® et SOS Médecins

Les données utilisées sont issues des bases de données OSCOUR® et SOS Médecins, au cours de la période du 1er janvier 2010 au 31 décembre 2017. Ont été considérés les nombres quotidiens de passages/d'actes aux urgences ou pour SOS Médecins, recueillis sur l'ensemble du territoire français (France métropolitaine, départements d'Outre-mer (DOM) et territoires d'Outre-mer COM)). Les données extraites contiennent les variables suivantes : identifiant de l'établissement, identifiant de la classe d'âge, identifiant du regroupement syndromique, date de passage, nombre total de passages (quel que soit le code du passage ou de la visite).

L'identifiant de la classe d'âge est un codage utilisé pour identifier les classes d'âge que nous avons choisies, présentées dans le Tableau 2.

Identifiant	Classe d'âge	Intervalle d'âge
3	Moins de 15 ans	[0 ; 14]
6	Tous âges	[-1 ; 120]
26	Moins de 2 ans	[0 ; 1]
27	Entre 2 et 14 ans	[2 ; 14]
28	Entre 15 et 74 ans	[15 ; 74]
29	75 ans ou plus	[75 ; 130]
47	65 ans ou plus	[65 ; 130]
126	Entre 15 et 64 ans	[15 ; 64]

Tableau 2 : Classes d'âge sélectionnées et leurs codes respectifs

Le regroupement syndromique « toutes causes » est la somme de tous les passages codés pour un jour, une classe d'âge et un établissement donnés.

Séries temporelles

À titre d'illustration, les Figures 2 et 3 représentent des séries temporelles de l'ischémie myocardique entre 2010 et 2017 en Ile-de-France, pour tous les âges, à partir des services d'urgence (SAU). La Figure 2 représente le nombre de passages pour ischémie myocardique et la Figure 3 représente une proportion (nombre de passages pour une ischémie myocardique rapporté à 10 000 passages codés).

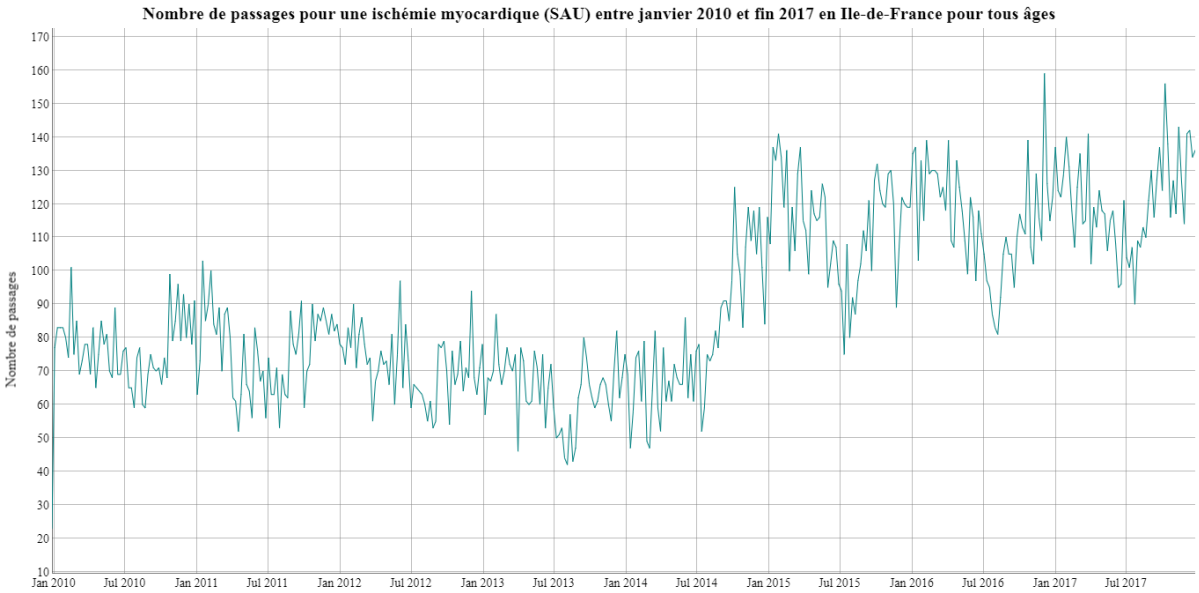

Figure 2 : Nombre de passages pour une ischémie myocardique (SAU) entre janvier 2010 et décembre 2017 en Ile-de-France pour tous âges

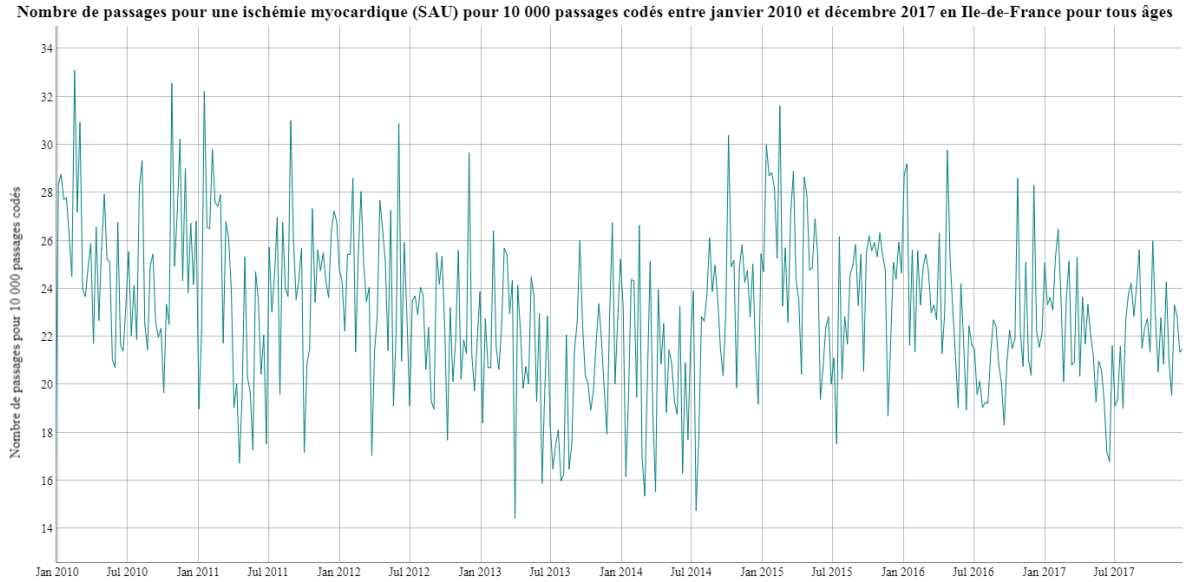

Figure 3 : Nombre de passages pour une ischémie myocardique (SAU) pour 10 000 passages codés entre janvier 2010 et décembre 2017 en Ile-de-France pour tous âges

Méthodes

Toutes les données ont été manipulées et analysées en utilisant le logiciel R version 3.4.4

Agrégation des données

Pour appliquer une détection avec un pas de temps quotidien ou hebdomadaire, sur l'ensemble du territoire français et selon le niveau départemental et régional, l'agrégation s'est faite en plusieurs étapes :

- Étape 1 : Agrégation des données journalières en données hebdomadaires
- Étape 2 : Agrégation des données journalières et hebdomadaires au niveau départemental et régional à partir d'une table de correspondance entre les identifiants des établissements et les codes départementaux et régionaux.
- Étape 3 : Agrégation des données obtenues à l'étape 2 selon trois niveaux : 1/ France métropolitaine, 2/ DOM-COM, 3/ France entière

Au lieu de travailler avec le nombre de passages, nous avons décidé de travailler en proportion. En effet, travailler sur des effectifs est délicat, car le nombre de passages ou de visites dépendent du nombre de participants au système de surveillance. Lorsque le système de surveillance monte en charge, le nombre de participants augmente et par conséquent, le nombre de passages ou visites augmente ce qui ne reflète pas une augmentation épidémiologique comme illustré par les Figures 2 et 3. Des alarmes pourraient ainsi être générées de manière totalement artificielle. Une solution

serait de ne travailler qu'avec des établissements ayant participé depuis 2010 mais cela restreint beaucoup leur nombre. Travailler sur une proportion est une meilleure alternative et représente un avantage de la surveillance syndromique par rapport aux surveillances spécifiques.
Pour une unité de temps et une zone géographique données, la proportion (1 pour 10 000 passages/visites) d'un regroupement syndromique se calcule selon la formule suivante :

$$\frac{\text{Nombre de passages pour un regroupement syndromique}}{\text{Nombre total de passages codés pour tous les regroupements}} \times 10\,000$$

Stockage des données

Afin d'obtenir un système de détection en temps réel, le traitement de données doit être rapide et ce critère n'est pas respecté si le système doit traiter quotidiennement l'ensemble des données antérieures. A titre d'illustration, pour la période du 1er janvier 2010 au 31 décembre 2017 et pour 5 regroupements syndromiques, une vingtaine de millions de lignes ont été extraites.
Pour pallier ce problème, les données du 1er janvier 2010 au 31 décembre 2017 ont été stockées dans des tables dans une base de données SQL hébergée sur PHPMyAdmin et forment l'historique sur lequel s'appuient les algorithmes de détection d'événements inhabituels.
De ce fait, autant de tables ont été créées qu'il y a d'agrégations et il sera possible d'alimenter ces tables avec le flux quotidien de nouvelles données provenant de SurSaUD® et ce, même si le nombre de regroupements syndromiques augmente (environ quarante regroupements syndromiques devront être analysés à court terme).
De plus, chaque table repose sur une même structure : (1) le regroupement syndromique, (2) la classe d'âge, (3) la date, (4) la zone géographique (département ou région), (5) le nombre de visites, (6) le nombre total d'actes codés.

Choix des algorithmes

Des statisticiens du Robert Koch Institute ont développé un package sous R, nommé « surveillance », qui contient un très grand nombre de fonctions permettant de réaliser des analyses statistiques (temporelles ou spatiales) sur des données de surveillance. Parmi elles, une vingtaine d'algorithmes pour la détection d'évènements inhabituels est implémentée pour des séries temporelles hebdomadaires. Le choix des algorithmes à choisir parmi cet ensemble de méthodes n'est pas évident. Nous nous sommes basés sur les résultats présentés dans un article publié récemment.[2] Les auteurs ont évalué les performances des algorithmes en utilisant des séries temporelles simulées présentant différentes caractéristiques (en termes d'effectifs, de tendance, saisonnalité(s), sur-

dispersion), auxquelles ont été ajoutées des épidémies de durées et d'ampleurs différentes [3]. Des algorithmes ont été appliqués sur des milliers de séries temporelles, l'objectif étant de détecter la dernière épidémie simulée.

Les performances des méthodes ont été mesurées en utilisant notamment les critères suivants :

- La sensibilité : nombre de semaines classées épidémiques par l'algorithme parmi le nombre de semaines épidémiques
- La spécificité : nombre de semaines classées non-épidémiques par l'algorithme parmi les semaines non-épidémiques
- Le taux de faux positifs : proportion de semaines classées épidémiques en l'absence d'épidémies
- La probabilité de détection : la probabilité que l'algorithme génère au moins une alarme durant l'épidémie
- La probabilité de détection durant la 1ère semaine de l'épidémie.

Un certain nombre de méthodes sont apparues plus performantes que d'autres. Un prolongement de ce travail est en cours (thèse de Gabriel Bédubourg) pour mesurer les performances de combinaisons de méthodes (combinaisons de deux, trois, quatre ou cinq méthodes) et pour évaluer si un gain de performances est obtenu. Un résultat important de ces analyses (non encore publié) conclut que la combinaison d'algorithmes n'augmente pas de manière significative les performances mais que celles-ci sont moins sujettes aux caractéristiques des séries temporelles (tendance, saisonnalité, etc.). Les résultats obtenus sont plus robustes et permettront à l'épidémiologiste d'avoir davantage confiance dans les alarmes générées par une combinaison de méthodes.

Pour ce stage, une combinaison de quatre algorithmes a été retenue pour la détection d'évènements inhabituels, nommés dans la littérature EARS, RKI, Bayes et Farrington. Nous allons maintenant décrire ces méthodes.

Algorithme EARS

L'algorithme EARS (early aberration report system) a été implémenté pour des systèmes de signalement précoce d'aberrations [4]. Cet algorithme a été créé pour détecter un évènement discret au cours d'une petite période de temps et en ne disposant que de peu de données antérieures voire d'aucune donnée (exemple : les Jeux Olympiques). L'algorithme EARS est un algorithme particulièrement utilisé pour la surveillance syndromique. Il calcule une statistique de test, en déduit

un intervalle de prédiction à un risque α donné. Lorsque le nombre observé de cas dépasse la borne supérieure, l'algorithme génère une alarme.

L'algorithme EARS propose trois versions : C1, C2 et C3. Les niveaux C1 et C2 fonctionnent de la même façon : les deux niveaux calculent une moyenne mobile de l'échantillon ainsi qu'une variance. La différence entre les deux niveaux se trouve dans les données utilisées. Le niveau C1 utilise les données des 7 jours précédant l'observation courante tandis que C2 introduit un délai de 2 jours et prend les 7 jours à partir du 3ème jour précédant l'observation courante. Le niveau C3 utilise la statistique de C2 sur le jour courant et les deux jours antérieurs.

En notant $Y(t)$ le nombre observé de cas au temps t, la statistique de C1, $C_1(t)$, s'écrit :

$$C_1(t) = \frac{Y(t) - \overline{Y_1}(t)}{S_1(t)}$$

où $\overline{Y_1}(t)$ est la moyenne mobile de l'échantillon des observations et $S_1(t)$ l'écart-type de ce même échantillon

$$\overline{Y_1}(t) = \frac{1}{7}\sum_{i=t-1}^{t-7} Y(i) \;\; et \; S_1^2 = \frac{1}{6}\sum_{i=t-1}^{t-7} [Y(i) - \overline{Y_1}(i)]^2$$

Sous l'hypothèse nulle de non épidémie, on suppose que $C_1(t) \sim N(0,1)$. Une alarme est générée si $C_1(t) \geq z_{1-\alpha}$, avec $z_{1-\alpha}$ le $(1-\alpha)^{th}$ quantile d'une loi normale centrée réduite et α le risque consenti. La borne supérieure $U_1(t)$ se définit de la façon suivante : $U_1(t) = \overline{Y_1}(t) + z_{1-\alpha} \times S_1(t)$.

La statistique C2 est similaire mais le calcul de la moyenne et de l'écart-type se fait avec un délai de 2 jours

$$C_2(t) = \frac{Y(t) - \overline{Y_2}(t)}{S_2(t)}$$

avec :

$$\overline{Y_2}(t) = \frac{1}{7}\sum_{i=t-3}^{t-9} Y(i) \;\; et \; S_2^2 = \frac{1}{6}\sum_{i=t-3}^{t-9} [Y(i) - \overline{Y_1}(i)]^2$$

Sous l'hypothèse nulle de non épidémie, on suppose que $C_2(t) \sim N(0,1)$. Une alarme est générée si $C_2(t) \geq z_{1-\alpha}$, avec $z_{1-\alpha}$ le $(1-\alpha)^{th}$ quantile d'une loi normale centrée réduite et α le risque consenti. La borne supérieure $U_2(t)$ se définit de la façon suivante : $U_2(t) = \overline{Y_2}(t) + z_{1-\alpha} \times S_2(t)$.

Le niveau C3 utilise la statistique de C2 mais sur trois jours (du jour courant à deux jours antérieurs) :

$$C_3(t) = \sum_{i=t}^{t-2} \max[0, C_2(i) - 1]$$

Sous l'hypothèse nulle de non épidémie, on suppose que $C_3(t) \sim N(0,1)$. Une alarme est générée si $C_3(t) \geq z_{1-\alpha}$, avec $z_{1-\alpha}$ le $(1-\alpha)^{th}$ quantile d'une loi normale centrée réduite et α le risque consenti. La borne supérieure $U_3(t)$ se définit de la façon suivante :

$$U_3(t) = \bar{Y}_2(t) + S_2(t) \times (z_{1-\alpha} - \sum_{i=t}^{t-2} \max[0, C_2(i) - 1])$$

Algorithme RKI

Pour une semaine d'intérêt, l'algorithme calcule la borne supérieure d'un intervalle de prédiction à partir des cas observés de plusieurs semaines précédentes. Quand le nombre observé de cas de la semaine d'intérêt dépasse cette borne supérieure, une alarme est générée, indiquant une anomalie, c'est-à-dire un nombre observé de cas significativement supérieur au nombre attendu. L'algorithme RKI propose trois versions : RKI1, RKI2 et RKI3. Ces différentes versions appliquent le même mode de fonctionnement et calculent les même statistiques mais à partir des données différentes. Ces données diffèrent par l'utilisation de valeurs différentes pour les paramètres suivants :

- t_0 : la semaine d'intérêt
- $actY$: l'année courante
- b : le nombre d'années qui précèdent l'année courante
- w : la largeur de la moitié de la fenêtre

Ces paramètres déterminent le nombre de semaines pour lesquelles les cas observés seront utilisés pour calculer la borne supérieure du nombre attendu de cas. Soit y_c, le numéro de l'année courante.

Pour RKI1, les paramètres ont pour valeur : $actY = TRUE$ (signifiant que l'année courante est prise en compte), $b = 0$ et $w = 6$. Les nombres de cas observés sont dans l'intervalle de semaines suivant : $[t_0 - (w+1)\ ;\ t_0 - 1]_{y_c}$. Les cas observés sont pris dans les 6 semaines de l'année courante

Pour RKI2, les paramètres ont pour valeur : $actY = TRUE$, $b = 1$ et $w = 6$. Les nombres de cas observés sont pris dans l'intervalle de semaines suivant : $[t_0 - w\ ;\ t_0 + w]_{y_c - b} \cup [t_0 - (w+1); t_0 - 1]_{y_c}$. Les cas observés sont pris dans les 6 semaines de l'année courante et les 7 semaines de l'année précédente

Pour RKI3, les paramètres ont pour valeur : $actY = FALSE$ (l'année courante n'est pas prise en compte), $b = 2$ et $w = 4$. Les nombres de cas observés sont pris dans l'intervalle de semaines suivant : $[t_0 - w\,;t_o + w]_{y_c-b} \cup [t_0 - w\,;t_0 + w]_{y_c-(b-1)} \cup [t_0 - (w+1)\,;t_0 - 1]_{y_c}$. Les cas observés sont pris dans les 9 semaines des deux années précédentes.

L'algorithme va ensuite calculer une moyenne des cas observés considérés pour déterminer la borne supérieure. Si la moyenne dépasse 20, la borne supérieure $upCi$ est égale à $upCi = \mu + 2 \times \sigma$, où μ est la moyenne, σ l'écart type. On remarque que $upCi$ est la borne supérieure de l'intervalle de confiance d'une loi normale $N(\mu, \sigma^2)$: $IC_{1-\alpha} = \mu \pm z_{1-\alpha/2} \times \sigma$, où $\alpha = 5\%$ car $z_{1-\alpha/2} = 1.96 \approx 2$.

Si la moyenne est inférieure à 20, cette dernière sera tronquée à l'unité. La moyenne est alors considérée comme issue d'une loi de Poisson et un autre intervalle de prédiction est calculé. La borne supérieure de cet intervalle est fournie par un jeu interne de données du package surveillance.

Algorithme Bayes

Pour une semaine d'intérêt, l'algorithme calcule la borne supérieure d'un intervalle de prédiction à partir des cas observés de plusieurs semaines précédentes. Quand le nombre observé de cas de la semaine d'intérêt dépasse cette borne supérieure, une alarme est générée, indiquant une anomalie. L'algorithme de Bayes propose trois versions : Bayes1, Bayes2, Bayes3.

Ces différentes versions appliquent le même mode de fonctionnement et calculent les mêmes statistiques mais à partir des données différentes. Ces données diffèrent par l'utilisation de valeurs différentes pour les paramètres suivants :

- t_0 : la semaine d'intérêt
- $actY$: l'année courante
- b : le nombre d'années qui précèdent l'année courante
- w : la largeur de la moitié de la fenêtre
- α : le risque d'erreur consenti

Ces paramètres déterminent le nombre de semaines pour lesquelles les cas observés seront utilisés dans le calcul de la borne supérieure de l'intervalle de prédiction. Soit y_c l'année courante.

Pour Bayes1, les paramètres ont pour valeurs : $actY = TRUE$ (l'année courante est prise en compte), $b = 0$ et $w = 6$. Les nombres observés de cas sont dans l'intervalle de semaines suivant : $[t_0 - (w+1)\,;\, t_0 - 1]_{y_c}$ (l'intervalle de temps pour l'année y_c). Les cas observés sont pris dans les 6 dernières semaines de l'année courante.

Pour Bayes2, les paramètres ont pour valeurs : $actY = TRUE$, $b = 1$ et $w = 6$. Les nombres observés de cas sont dans l'intervalle de semaines suivant : $[t_0 - w\ ;\ t_0 + w]_{y_c - b} \cup [t_0 - (w+1); t_0 - 1]_{y_c}$. Les cas observés sont pris dans les 6 semaines de l'année courante et les 7 semaines de l'année précédente.

Pour Bayes3, les paramètres ont pour valeurs : $actY = FALSE$, $b = 2$ et $w = 4$. Les nombres observés de cas sont dans l'intervalle de semaines suivant : $[t_0 - w\ ; t_o + w]_{y_c - b} \cup [t_0 - w\ ; t_0 + w]_{y_c - (b-1)} \cup [t_0 - (w+1)\ ; t_0 - 1]_{y_c}$. Les cas observés sont pris dans les 9 semaines des deux années précédentes.

L'algorithme calcule ensuite la borne supérieure avec le quantile d'ordre $1 - \alpha$ d'une loi binomiale négative : $NB(r, p)$, avec r la somme des observations, à laquelle on ajoute 0.5 pour être tout le temps strictement positif :

$$r = \frac{1}{2} + \sum obs$$

Et p la probabilité :

$$p = \frac{n}{n+1}$$

où n est le nombre d'observations.

Algorithme de Farrington

Pour une semaine d'intérêt, l'algorithme calcule la borne supérieure d'un intervalle de prédiction à partir des cas observés de plusieurs semaines précédentes. Quand le nombre observé de cas de la semaine d'intérêt dépasse cette borne supérieure, une alarme est générée, indiquant une anomalie.[5]

Cet algorithme utilise les paramètres principaux suivants :

- b : le nombre d'années qui précèdent l'année courante (égal à 3 par défaut)
- w : la largeur de la moitié de la fenêtre (égal à 3 par défaut)
- α : le risque d'erreur consenti (égal à 0,01 par défaut)

L'algorithme de Farrington repose sur un modèle de Poisson en supposant que les nombres de cas observés pour toutes les semaines t_i sont indépendants. De ce fait, le modèle s'écrit :

$$\log(\mu_i) = \alpha + \beta \times t_i$$

où μ_i est le nombre attendu de cas pour la semaine t_i.

Le paramètre de sur-dispersion est un paramètre qui indique l'hétérogénéité des données, en comparant la variance et la moyenne des données. Pour une loi de Poisson, ce paramètre vaut 1.

$$\phi = \frac{V(X)}{E(X)} \text{ et } V(X) = E(X) = \lambda \text{ pour } X \sim P(\lambda)$$

Ici, le paramètre ϕ est estimé par :

$$\hat{\phi} = \max\left\{\frac{1}{n-p}\sum_{i=1}^{n}\omega_i\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, 1\right\}$$

où $n = b(2w+1)$ est le nombre de semaines utilisées pour calculer le nombre de cas attendu et son intervalle de prédiction, p un paramètre indiquant si une tendance est ajustée sur la série temporelle ou non (respectivement égal à 1 ou 2), ω_i un poids pour la semaine t_i (en cas d'épidémie(s) passée(s), le poids des semaines est proche de 0, limitant leur impact dans le calcul du nombre de cas attendu) et $\hat{\mu}_i$ l'estimation du nombre attendu de cas pour la semaine t_i :

$$\hat{\mu}_i = \exp(\hat{\alpha} + \hat{\beta} t_i)$$

La borne supérieure de l'intervalle de prédiction se calcule par :

$$U = \hat{\mu}_0\left\{1 + \frac{2}{3}z_\alpha\left(\frac{\hat{\tau}}{\hat{\mu}_0}\right)^{\frac{1}{2}}\right\}^{\frac{3}{2}}$$

où z_α est le quantile (1-α) d'un loi normale centrée réduite et $\tau = \phi + var(\hat{\mu}_0)/\mu_0$

Résultats

Les résultats présentés se concentrent sur l'analyse des semaines d'intérêts sur les trois dernières années (2015-2017) et plus précisément entre le 29 décembre 2014 et le 25 décembre 2017 (équivalent à 156 semaines soit environ 3 années complètes) pour un regroupement syndromique (l'ischémie myocardique), pour une zone géographique (Ile-de-France), pour la classe « tous âges », à partir des données du réseau OSCOUR® entre janvier 2010 et décembre 2017 avec les quatre algorithmes suivants : RKI3, Bayes2, EARS2, Farrington.

En utilisant les mêmes données et sur la même période d'analyse, les quatre algorithmes génèrent des alarmes différentes (en termes de nombre et de dates). Les algorithmes RKI3, Bayes2, EARS2 et Farrington génèrent respectivement 10, 1, 2 et 1 alarme(s).

Les alarmes de plusieurs de ces algorithmes coïncident pour deux semaines d'intérêt. Pour la semaine du 16 février 2015, les algorithmes RKI3, Bayes 2 et Farrington ont généré une alarme, pour la semaine du 25 avril 2016, ce sont les algorithmes RKI3 et EARS2.

Semaine	RKI3	Bayes2	EARS2	Farrington
12 janvier 2015				
19 janvier 2015				
26 janvier 2015				
16 février 2015				
23 mars 2015				
20 avril 2015				
27 avril 2015				
18 mai 2015				
25 mai 2015				
25 avril 2016				
24 octobre 2016				

Tableau 3 : Semaines avec alarme pour chaque algorithme

Les Figures 4 à 7 montrent la représentation graphique de l'analyse avec les quatre algorithmes suivants : RKI3, EARS2, Farrington et Bayes2. Sur chaque figure, le nombre de passages pour l'ischémie myocardique pour 10 000 passages codés (en vert) et la borne supérieure de prédiction de son estimation (en bleu) sont représentés. Les alarmes, représentées par un carré contenant la lettre « A », sont générées quand le nombre de passages observé est supérieur à la borne supérieure de l'intervalle de prédiction de son estimation.

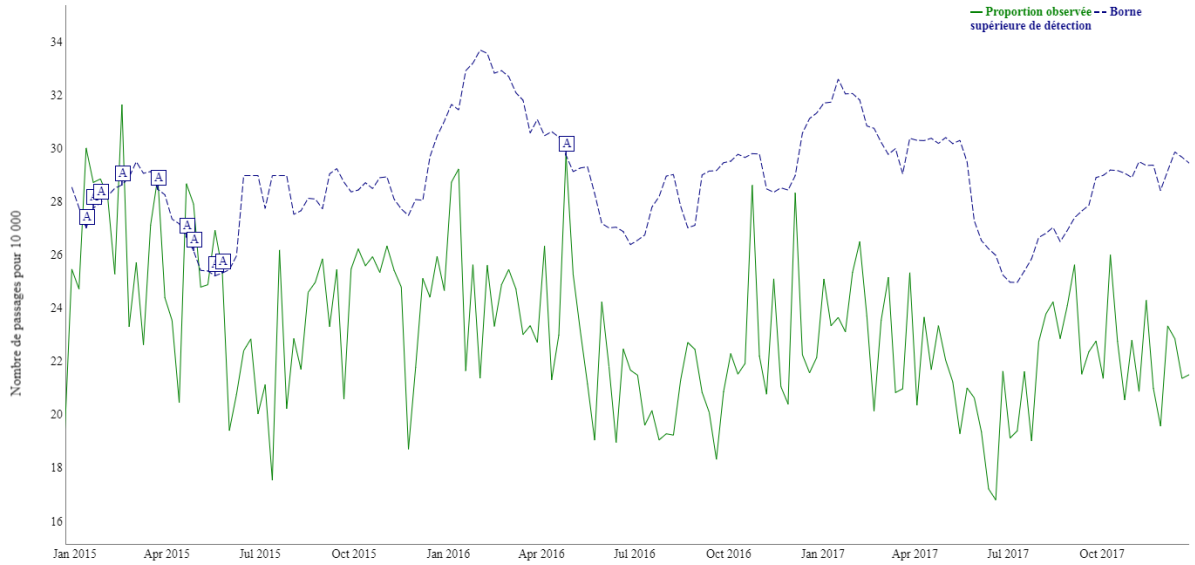

Figure 4 : Nombre de passages pour une ischémie myocardique (SAU) pour 10 000 passages codés entre le 29 décembre 2014 et le 25 décembre 2017 en Ile-de-France pour tous âges avec seuil et alarmes (méthode RKI3)

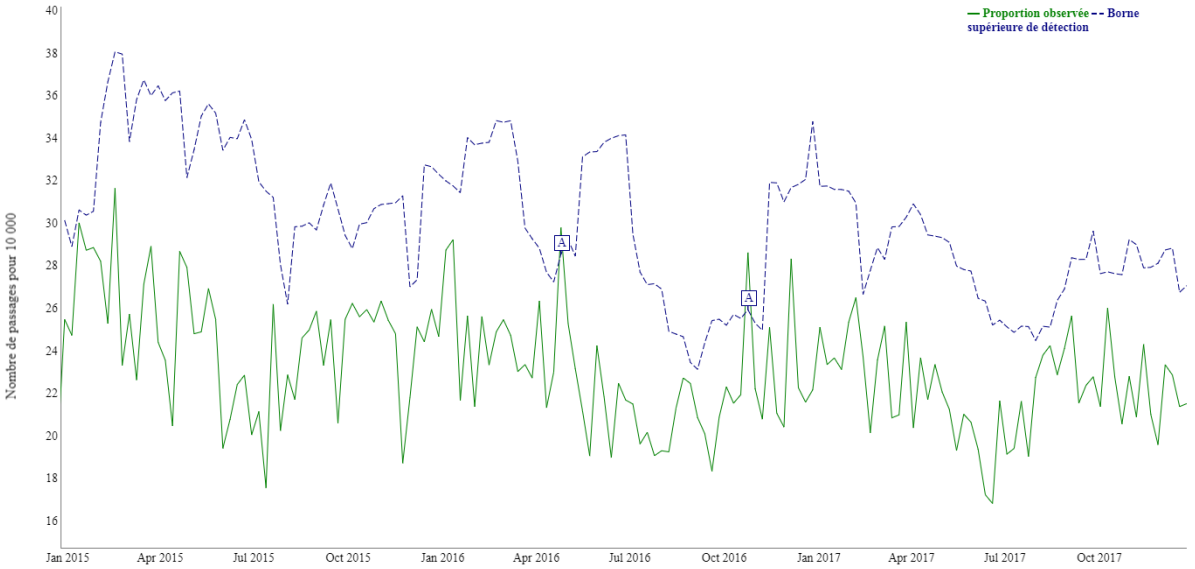

Figure 5 : Nombre de passages pour une ischémie myocardique (SAU) pour 10 000 passages codés entre le 29 décembre 2014 et le 25 décembre 2017 en Ile-de-France pour tous âges avec seuil et alarmes (méthode EARS2)

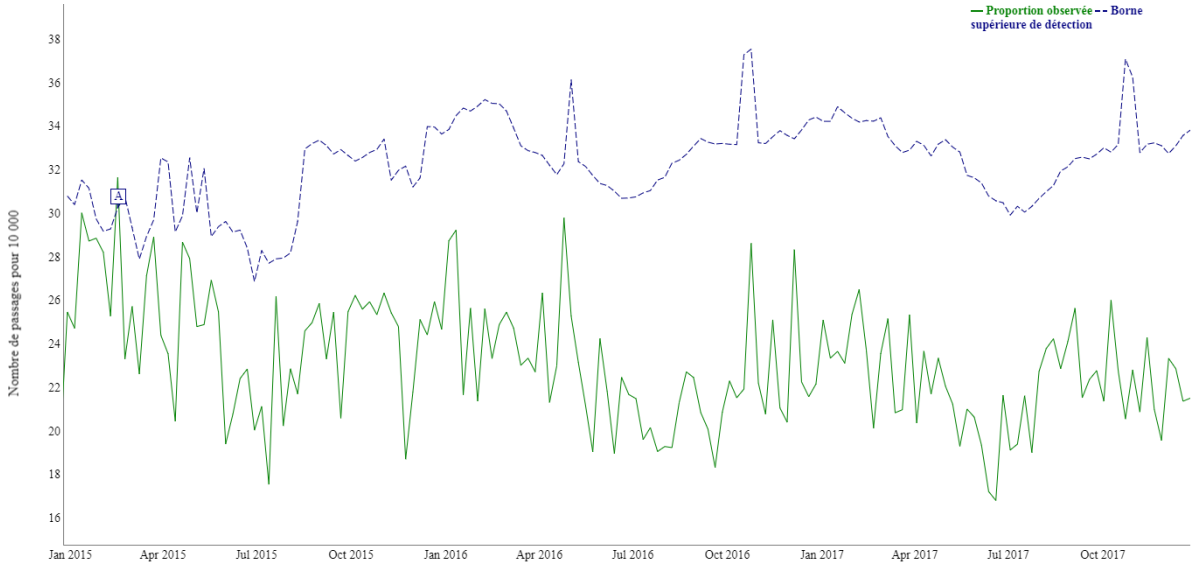

Figure 6 : Nombre de passages pour une ischémie myocardique (SAU) pour 10 000 passages codés entre le 29 décembre 2014 et le 25 décembre 2017 en Ile-de-France pour tous âges avec seuil et alarmes (méthode Farrington)

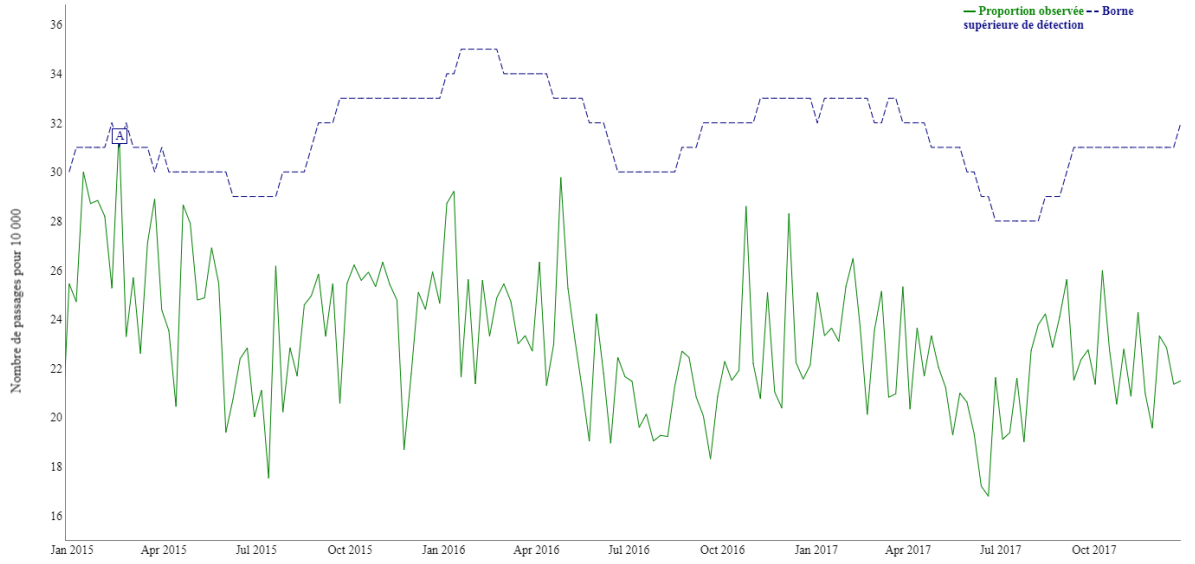

Figure 7 : Nombre de passages pour une ischémie myocardique (SAU) pour 10 000 passages codés entre le 29 décembre 2014 et le 25 décembre 2017 en Ile-de-France pour tous âges avec seuil et alarmes (méthode Bayes2)

Discussion

Les données utilisées pour ce travail sont issues de SurSaUD®. Ce sont des données journalières, recueillies entre 2010 et 2017, rapportant le nombre de passages pour chaque regroupement syndromique étudié et le nombre total de passages codés pour chaque établissement/association participant. Les données journalières ont été agrégées en données hebdomadaires, permettant ainsi l'utilisation des méthodes de détection. Ces données ont été ensuite agrégées selon trois niveaux géographiques : (1) France métropolitaine, (2) DOM-COM, (3) France entière puis stockées dans des tables d'une base de données pour faciliter l'analyse selon ces trois niveaux et pour former l'historique.

La détection d'événements inhabituels, notamment pour la surveillance syndromique, repose sur des méthodes statistiques créées à cet effet. Une revue de littérature a permis d'en retenir quatre parmi la vingtaine évaluée.

Une première analyse sur la période 2015-2017 a été conduite avec les quatre méthodes, pour un regroupement syndromique (l'ischémie myocardique), une classe d'âge (tous âges) et une région (l'Ile-de-France).

Afin de réaliser cette analyse, un travail important de data-management a été réalisé et profitera dans la suite de ce travail à l'analyse d'un plus grand nombre de regroupements syndromiques (une quarantaine prévue) pour chaque région française et pour chaque classe d'âge. Cela fait partie des nombreuses perspectives de ce travail qui devra également étudier l'impact de certaines données manquantes dans les performances des méthodes utilisées.

La question principale auquel ce travail devait répondre portait sur les performances des méthodes implémentées afin d'apporter une aide aux épidémiologistes, notamment en région. L'état actuel du travail ne permet pas de répondre à cette question. Le travail présenté ici n'est que l'introduction d'un travail qui va se poursuivre dans les semaines à venir. Le travail final portera sur l'automatisation du traitement des données (hebdomadaires et journalières) et de leurs analyses avec le développement d'une interface web sous R Shiny. C'est à ce moment que le comportement des méthodes pourra être étudié de manière plus approfondie. Cela ne remplacera cependant pas une évaluation par les épidémiologistes, qui devra s'étaler sur au moins une année, afin de mesurer l'utilité de la mise en place d'une telle application et de proposer si besoin des modifications (présentation des résultats, modifications des paramètres des méthodes, etc.) qui permettront de d'améliorer l'outil.

On peut cependant penser que dans quelques mois, cet outil sera très utilisé par les épidémiologistes de Santé publique France pour l'analyse des données de surveillance syndromique en France, dans un objectif d'alerte.

Bibliographie

1. Caserio-Schönemann C, Bousquet V, Fouillet A, Henry V, pour l'équipe projet SurSaUD®. Le système de surveillance syndromique SurSaUD®. Bull Epidémiol Hebd. 2014;(3-4):38-44

2. Bédubourg G, Le Strat Y. Evaluation and comparison of statistical methods for early temporal detection of outbreaks: A simulation-based study. PloS One. 2017;12(7):e0181227.

3. Noufaily A, Enki DG, Farrington P, Garthwaite P, Andrews N, Charlett A. An improved algorithm for outbreak detection in multiple surveillance systems. Stat Med. 30 mars 2013;32(7):1206-22.

4. Fricker RD, Hegler BL, Dunfee DA. Comparing syndromic surveillance detection methods: EARS' versus a CUSUM-based methodology. Stat Med. 30 juill 2008;27(17):3407-29.

5. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. J R Stat Soc Ser A Stat Soc. 1996;159(3):547-63.