Mamann Aaron

ENSAE 2ème année
Stage d'application
2020-2021

Détection de ruptures multiples dans des signaux épidémiologiques

CENTRE BORELLI,
Laboratoire de l'ENS Paris-Saclay

Maître de stage: Mathilde MOUGEOT
1er Juin - 18 Septembre 2020

Contents

1 Introduction

Ce stage répond au besoin croissant de pouvoir prévoir le plus tôt possible l'apparition d'épidémies. Ce besoin s'est notamment fait remarqué avec la crise du COVID-19. Plus on peut prévoir tôt l'apparition d'une épidémie (par exemple d'une deuxième vague dans le cadre du coronavirus), plus les moyens envisagés par les gouvernements pour contrer l'épidémie seront efficaces.

Pour détecter les épidémies, j'ai donc fait appel à des algorithmes de détection de ruptures. Par conséquent, ce stage a consisté, dans un premier temps, à comprendre les travaux faits dans la théorie de détection de ruptures sur des signaux puis à les appliquer sur des signaux épidémiologiques (univariés et multivariés) dans le but de détecter des épidémies.

Il est important de comprendre ce que veut dire concrêtement, dans le cadre de mon travail, de détecter les ruptures dans un signal épidémiologique. Dans le cas d'un signal épidémiologique univarié qu'on peut alors représenter par une courbe décrivant un indice épidémiologique (comme le nombre d'infectés pour une certaine maladie) en fonction du temps, il s'agit alors de segmenter la courbe en différentes périodes. Dans le cadre des signaux épidémiologiques, les segments (i.e. périodes) identifiés dans la courbe sont soit des périodes d'épidémie soit des périodes non-épidémiques. Un point de la courbe marquant le passage d'une période à une autre est appélé point de rupture ("change point" en anglais). Ainsi, les points de rupture peuvent marquer soit le passage d'une période non-épidémique à une période d'épidémie, soit le passage d'une période d'épidémie à une période non-épidémique.

Le choix de travailler sur les épidémies de grippe a été pris par le laboratoire. Ce choix peut s'expliquer par le fait que les données sur le coronavirus étaient insuffisantes (au moment où j'ai fait mon stage). Cela peut aussi s'expliquer par le fait que la grippe ressemble sur beaucoup de critères au coronavirus notamment sur les modes de transmissions, les symptômes contractés... Ainsi, en travaillant sur des signaux concernant la grippe, le laboratoire vise par la suite d'utiliser les résultats trouvés pour déctecter les épidémies de coronavirus.

2 Précisions nécessaires

Insistons ici sur un des objectifs rappelé en introduction. Dans le cadre de maladies infectieuses, nous avons un besoin de développer des méthodes de détection d'épidémie dans un seul but : prévoir le plus tôt possible les épidémies de manière à ce que les mesures prises par les Etats (pour contrer l'épidémie) soient prises le plus tôt possible. Plus on prend ces mesures de manière anticipée, plus ces mesures seront efficaces. Nous pouvons actuellement illustrer cette idée par la crise du coronavirus. L'Italie qui fut le premier Etat considérablement contaminé en Europe (par le coronavirus) a donc pris des mesures de manière tardive et a donc connu beaucoup plus de mal à stopper l'expansion de la maladie dans son pays que l'Allemagne. En effet, l'Allemagne a été touché après l'Italie et la France et donc a eu le temps de se préparer et de prendre des mesures plus tôt qui se sont avérés plus efficaces dans le contrôle de l'épidémie.

Il est également important de préciser que pour segmenter un signal en plusieurs périodes épidémique et non-épidémique et trouver les points de ruptures (ce vocabulaire a été expliqué dans l'introduction) nous allons travailler seulement sur le signal en lui-même. Il est crucial de le préciser car souvent on utilise des variables explicatives pour travailler sur nos variables d'intérêts alors qu'ici nous allons uniquement travailler sur un signal (univarié ou multivarié) sans faire appel à des variables extérieures. Par exemple, si la variable qui nous intéresse est le taux de nouveaux cas de grippe (évoluant en fonction du temps) alors nous allons travailler que sur ce signal pour détecter d'éventuelles épidémies grippales.

Enfin, il est crucial de préciser que l'analyse du signal est effectuée ici a posteriori. On ne prédit pas le signal, on l'analyse seulement a posteriori, et ce dans le but de distinguer une éventuelle période d'épidémie. On parle de détection de ruptures dite hors ligne.

3 Cadre mathématique

Posons le cadre mathématique qui a été utilisé tout au long de mon stage. Considérons un processus aléatoire multivarié non stationnaire $y = (y_1, ..., y_T)$ où $y_t \in \mathrm{R}^d$ (avec $d \geq 1$) correspond au temps t du processus. La détection de ruptures revient à segmenter le signal y, autrement dit à décomposer le signal y en morceaux comme cela est illustré dans la Figure 1 dans laquelle 5 régimes (i.e. morceaux) différents sont identifiés. Ces morceaux sont délimités par $t_1 < t_2 < \ < t_K$ qui sont donc inconnus et qu'il faut donc déterminer. Ainsi, par exemple, la première période identifiée est le segment $[t_1, t_2]$, la seconde est $[t_2, t_3]$ jusqu'à la dernière période qui est $[t_{K-1}, t_K]$. Le nombre K de ruptures est soit connu soit inconnu.

Présentons ici quelques notations que nous utiliserons par la suite. Le sous-signal du signal y (où $y = \{y_t\}_{t=1}^{T}$) est $\{y_t\}_{t=i}^{j} (1 \leq i < j \leq T)$ et est donc noté $y_{i..j}$. Chaque segmentation est un ensemble de point de rupture qui est noté de la manière suivante : $\Upsilon = \{t_1, ..., t_K\}$. Notons le cardinal d'$\Upsilon$ avec l'écriture suivante $|\Upsilon|$. Remarquons alors que $\Upsilon \subset \{1, ..., T\}$, que $t_0 := 0$ et que $t_{K+1} := T$.

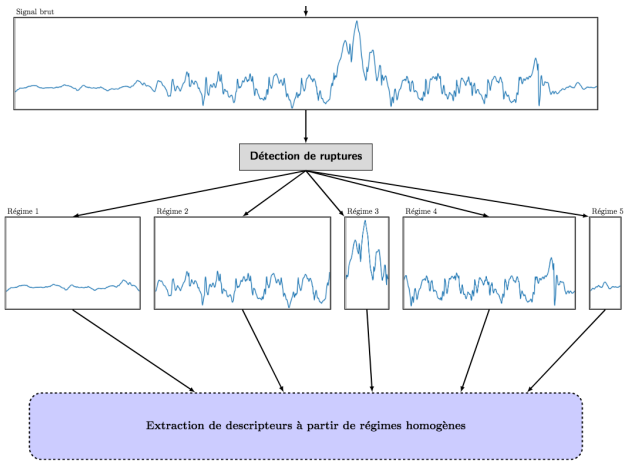

Figure 1: Illustration de la détection de ruptures sur un signal univarié

Il est maintenant logique de se poser la question : sous quels critères détermine t-on la segmentation (i.e. sous quels critères détermine t-on les $(t_1, t_2...)$? Cette segmentation est faite de manière à avoir des morceaux dans lequel le signal est homogène comme on peut le voir dans la Figure 1. Nous allons préciser, dans le paragraphe à la fin de cette page, ce qu'on entend ici par le mot homogène.

Nous essayons plusieurs segmentations avant de trouver la segmentation optimale. Imaginons qu'on teste une segmentation $(t_1, t_2...)$. Nous essayons d'approximer le signal sur chaque segment $([t_1, t_2], [t_2, t_3]...)$ par un modèle (les modèles choisis sont différents d'un segment à l'autre). Par exemple, on peut approximer le sous-signal $y_{t_1..t_2}$ (i.e. le signal sur le segment $[t_1, t_2]$) par une régression linéaire et le sous-signal $y_{t_2..t_3}$ par une autre régression linéaire. Introduisons alors c$(\cdot)$ une fonction dite fonction de coût qui indique l'erreur d'approximation d'un sous-signal par un modèle. On aurait donc avec l'exemple précédent que $c(y_{t_1..t_2})$ est l'erreur d'approximation du sous-signal $y_{t_1..t_2}$ par une régression linéaire. On note alors $V(\Upsilon, y)$ (ou $V(\Upsilon)$ si on sait qu'on travaille sur le signal y) la somme des erreurs d'approximation de tous les segments. On a donc :

$$V(\Upsilon, y) = \sum_{i=0}^{K} c(y_{t_i..t_{i+1}}) \tag{1}$$

On cherche alors à minimiser cette somme. Pour comprendre avec quelle démarche on minimise $V(\Upsilon, y)$ reprenons l'exemple du paragraphe précédent. On a que $V(\Upsilon, y)$ est la somme des erreurs d'approximation par les régressions linéaires effectués séparément sur chaque segment. Il est important que ces régressions linéaires sont faites séparément car les droites qui approximent chaque sous-signal ne sont pas les mêmes. Par conséquent, pour chaque segmentation testée $(t_1, t_2...)$ on associe cette somme d'erreurs d'approximation $V(\Upsilon, y)$. Ainsi, on choisit la segmentation qui a le $V(\Upsilon, y)$ minimal. On entend donc par le mot homogène, les morceaux finalement trouvés par cette démarche. Cette segmentation qui a le $V(\Upsilon, y)$ minimal, on l'appelle la "meilleure segmentation" ou la "segmentation optimale" et on la note $\widehat{\Upsilon}$. On cherche donc:

$$\min_{|\Upsilon|=K} V(\Upsilon) \tag{2}$$

Comme dans notre cas, le nombre K de ruptures est inconnu, il peut-être déterminé indirectement à l'aide d'une pénalisation (qu'on va expliquer dans le prochain paragraphe) :

$$\min_{\Upsilon} V(\Upsilon) + PEN(\Upsilon) \tag{3}$$

La fonction PEN est une fonction de pénalisation. Notons que pour une segmentation Υ plus le nombre de segments (i.e. périodes) est grand plus la quantité $PEN(\Upsilon)$ sera grande (positivement). Or, le critère a minimisé est maintenant $V(\Upsilon) + PEN(\Upsilon)$. Ainsi, on parvient à pénaliser les segmentations où le nombre de morceaux (i.e. segments) est trop important.

Il est crucial de pénaliser le nombre de segments : si le nombre de segments est excessivement élevé cela veut dire que la segmentation divise le signal en morceaux très petits. Or, on peut très bien visualiser le fait que les segmentations comportant un nombre de segments (i.e. périodes) excessivement élevé auront souvent une somme d'erreur d'approximation $V(\Upsilon)$ plus petite, ce qui ferait que la segmentation optimale $\widehat{\Upsilon}$ comporterait souvent un nombre de segments (i.e. périodes) excessivement élevé, ce qui nous intéresse pas. En effet, pour notre cas, cela reviendrait à identifier trop de périodes épidémiques et non-épidémiques (cf. Introduction) et donc notamment à identifier des périodes épidémiques pour la moindre petite évolution sur le signal.

Il est important de savoir ajuster la pénalisation. La pénalité peut-être plus ou moins sévère : par exemple si la pénalité s'écrit de la manière suivante : $PEN(\Upsilon) = \lambda|\Upsilon|$ (où λ est positif), on remarque que plus λ est grand plus le nombre de segments est pénalisé. Ainsi, on peut durcir ou adoucir la pénalité à notre guise, ce qui est utile car dans notre cas même si le nombre K de ruptures est inconnu nous avons une petite idée de la fourchette dans laquelle il se trouve. Ainsi, si on trouve que le nombre K de ruptures est trop faible (comparé à la fourchette qu'on a en tête) il faut adoucir la pénalité (et la durcir si le le nombre K de ruptures est trop important).

Enfin, remarquons qu'on fait face à une difficulté lorsque l'on veut trouver la segmentation optimale $\widehat{\Upsilon}$ via le problème d'optimisation décrit par

l'équation 3. En effet, on minimise la quantité $V(\Upsilon)+PEN(\Upsilon)$ sur l'ensemble $\{\Upsilon \text{ s.c. } 1 \leq |\Upsilon| < T\}$. Or, cet ensemble comporte $\sum_{K=1}^{T-1} \binom{T-1}{K-1}$ éléments, ce qui correspond à toutes les segmentations possibles de tous les cardinaux possibles. Il existe donc des manière de ne pas tester toutes les segmentations possibles. Autrement dit, il existe des méthodes pour trouver la segmentation optimale ou du moins une bonne segmentation sans tester toutes les segmentations, ce qui permettrait à l'algorithme d'être plus rapidement exécuté sans trop perdre en qualité. On discutera du choix de la méthode de recherche des segmentations (permettant de ne pas tester toutes les segmentations) par la suite, au moment de l'application de la détection de ruptures sur notre cas (sur les signaux concernant la grippe). On discutera également à ce moment là des modèles choisis pour fitter chaque sous-signal.

4 Application de la théorie de la détection de ruptures pour la détection d'épidémies grippales

4.1 Détection de rutpures sur un signal épidémiologique univarié

Dans mon travail, pour mettre en pratique les détections de ruptures, j'ai utilisé la librarie python "ruptures" développée par Charles Truong durant son doctorat au laboratoire de l'ENS Paris-Saclay.

4.1.1 Discussion sur les modèles pour fitter les sous-signaux

Le signal qui nous intéresse ici est le taux de nouveaux cas grippales par semaine pour 100 000 habitants au niveau de la France. Nous allons donc, rappelons-le, détecter sur ce signal des points de ruptures et ainsi segmenter le signal en périodes de deux types : période épidémique et période non-épidémique. Pour mettre en oeuvre cette détection de ruptures, il faut d'abord choisir les modèles avec lesquelles on approxime chaque sous-signal. En effet, pour avoir la segmentation optimale $\widehat{\Upsilon}$, nous essayons plusieurs segmentations. Sur chaque segmentation essayée, nous calculons la somme des erreurs d'approximation des modèles sur chaque segment et nous retenons la

segmentation $\widehat{\Upsilon}$ qui minimise cette somme d'erreurs. On doit donc choisir ces modèles.

Nous allons donc d'abord essayer d'approximer les sous-signaux par des processus autorégressifs. Pour tout sous-segment $[t_i, t_{i+1}]$ (avec $i \in [0, K]$ où K est le nombre de ruptures) on a :

$$\forall i \in [0, K], \forall t \in [t_i, t_{i+1}], \widehat{y_t} = c + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} \quad (4)$$

où $(\alpha_1, \alpha_2, ..., \alpha_p)$ sont les paramètres du modèle et "c" est une constante.

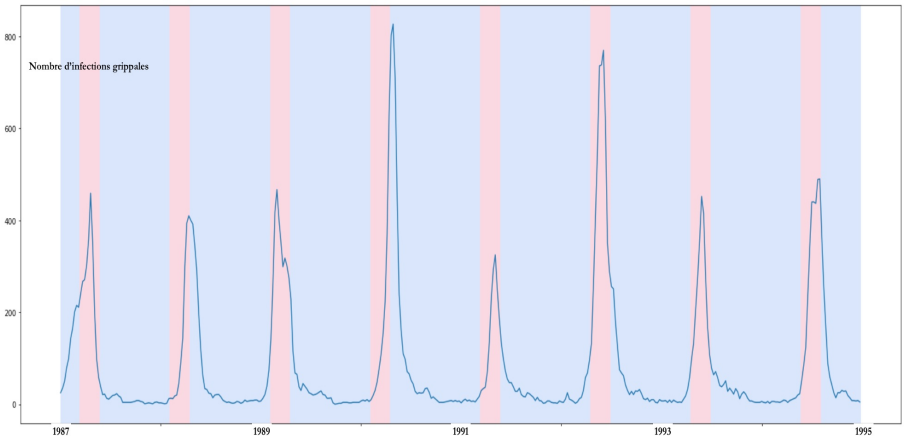

Figure 2: Détection de ruptures à l'aide de modèles autorégressifs sur le signal indiquant le taux de nouveaux cas grippales par semaine pour 100 000 habitants en France

Remarque sur mon code Python : J'ai réalisé la Figure 2 grâce à la librairie Python "ruptures" qui permet de simuler des algorithmes de détection de ruptures. J'ai pu préciser dans l'algorithme que le signal allait être fitté sur chaque segment par des modèles autorégressifs d'ordre 5.

On remarque directement quelques problèmes dans les résutats trouvés dans le Figure 2. Avant tout, précisons qu'on a ici que les segments bleus

sont les périodes épidémiques et les segments rouges sont les périodes non-épidémiques. On remarque que la période épidémique détectée en 1988 s'arrête au pic des infections grippales ce qui n'est pas normal. On remarque également que la période épidémique détectée en 1990 s'arrête elle aussi juste avant le pic de nouveaux cas grippales, ce qui n'est également pas normal. En effet, cela voudrait dire (que ce soit en 1988 ou en 1990) que l'épidémie serait terminé au moment où elle est la plus forte.

Cela nous fait réaliser une chose sur laquelle il faut absolument insister. Ici, le but est d'identifier les périodes épidémiques. Or, ce qui caractérise une épidémie c'est plusieurs critères dont notamment le taux de nouveaux cas grippales anormalement élevé par rapport à la moyenne. Or, les modèles autorégressifs ne prennent vraiment pas cette composante en compte et se concentre sur la dynamique des infections d'un jour à l'autre (on peut le voir sur l'équation 4). Expliquons cette dernière remarque : notons ici que les périodes (épidémiques et non-épidémiques) reconnus sont considérés comme homogènes dans la mesure où sur la segmentation optimale, les valeurs du signal d'un même segment ont pu être bien approximée par le même modèle autorégressif. Intuitivement, cela veut dire que sur un segment le signal suit à peu près un modèle autorégressif et qu'au niveau du point de rupture le signal suit un autre modèle autorégressif i.e d'un segment à l'autre il y a désormais une nouvelle relation entre le taux de nouveaux cas à un jour t et le taux de nouveaux cas aux jours précédents. C'est dans ce sens qu'on peut dire que les modèles autorégressifs se concentrent surtout sur la dynamique des infections d'un jour à l'autre et pas assez sur l'écart anormalement important entre le taux d'infections à une semaine t et sa moyenne.

Suite aux remarques du paragraphe précédent, il est maintenant logique de proposer comme modèle pour fitter les sous-signaux, le modèle suivant : pour tout sous-segment $[t_i, t_{i+1}]$ (avec $i \in [0, K]$ où K est le nombre de ruptures) on a :

$$\forall i \in [0, K], \forall t \in [t_i, t_{i+1}], \widehat{y_t} = médiane(y_{t_i..t_{i+1}}) \tag{5}$$

Nous avons choisi la médiane plutôt que la moyenne car la moyenne est trop sensible aux valeurs extrêmes relativement présentes dans notre base de données. Ainsi, en prenant ce modèle, on va segmenter des parties du

signal selon leur médiane, ce qui est une bonne chose car une des principales caractérisques d'une épidémie est le niveau particulièrement élevé du taux d'infections (i.e. de nouveaux cas).

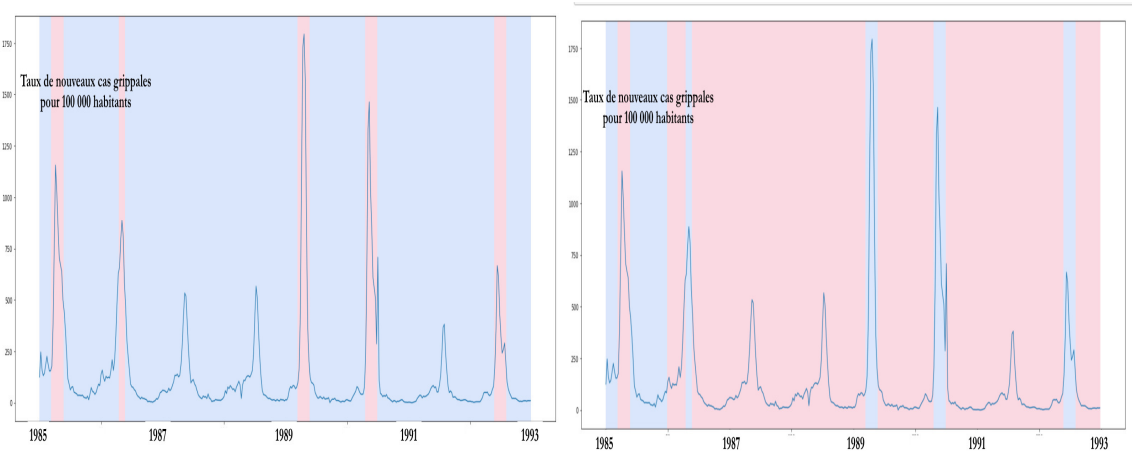

Figure 3: Détection de ruptures à l'aide de modèles décrits dans l'équation 5 sur le signal indiquant le taux de nouveaux cas grippales par semaine pour 100 000 habitants en France. Dans le graphique à droite on a adoucit la pénalisation du nombre de ruptures

Remarque sur mon code Python : J'ai réalisé la Figure 3 grâce à la librairie Python "ruptures" qui permet de simuler des algorithmes de détection de ruptures. Dans cet algorithme, j'ai pu préciser que le signal allait être fitté sur chaque segment par la médiane des valeurs prises par le signal sur le segment en question. J'ai également précisé dans l'algorithme que l'erreur d'approximation du signal par les modèles allait être calculé par la norme L1 (nous en parlerons dans le sous-chapitre suivant).

Sur la Figure 3, il est d'abord important de préciser que les périodes épidémiques ne sont pas forcément colorées en rouge et que les périodes non-épidémiques ne sont pas forcément colorées en bleu. Autrement dit, ici les couleurs n'ont donc pas d'importance.

Sur le graphique de droite on a adoucit la pénalisation du nombre de

ruptures par rapport au graphique de gauche. En effet, la pénalité peut-être plus ou moins sévère et on peut la durcir ou l'adoucir à notre guise : par exemple si la pénalité s'écrit de la manière suivante : $PEN(\Upsilon) = \lambda|\Upsilon|$ (où λ est positif, et Υ est une segmentation), on remarque que plus λ est grand plus le nombre de morceaux est pénalisé.

On remarque alors que dans la segmentation du graphique de droite il y a plus de périodes (i.e. de ruptures) que dans le graphique de gauche. En effet, plus on adoucit la pénalisation du nombre de ruptures plus on aura tendance à choisir comme meilleure segmentation des segmentations comportant davantage de ruptures (cf notre analyse sur les pénalisations à la section 3 de ce rapport à la page 6).

Par conséquent, sur la Figure 3 de gauche, comme j'ai vu que les deux petits pics de l'année 1987 et 1988 n'avait pas été reconnu en tant que période épidémique, je me suis alors dit qu'il fallait adoucir la pénalisation du nombre de ruptures, pour avoir une segmentation comportant plus de périodes et pour que ces deux petits pics de l'année 1987 et 1988 soit reconnus. Or, comme on peut le voir sur 3 de droite, étant le résultat après un adoucissement de la pénalisation, l'algorithme n'a toujours pas reconnu comme des périodes épidémiques les petits pics de 1987 et 1988. A la place, la nouvelle période identifiée se trouve juste avant le pic l'année 1986 : cette période supplémentaire peut être interprétée comme un entre-deux entre la période épidémique et la période non épidémique.

Or, on est sûr que ces deux petits pics de l'année 1987 et 1988 sont deux épidémies grippales en France : en effet, si ces deux petits pics de 1987 et 1988 n'étaient pas deux épidémies cela aurait voulu dire qu'il n'y a pas eu d'épidémie grippale ni en 1987 ni en 1988, or il y en a eu en 1987 et 1988.

On remarque que, si on adoucit la pénalisation du nombre de ruptures de plus en plus, l'algorithme va finalement reconnaitre ces deux petits pics de 1987 et 1988 mais il va aussi reconnaitre des périodes étant des entre-deux entre période épidémique et période non épidémique, ce qui n'est pas ce que nous voulons. Donc, on est face à un problème en chosissant ce type de modèle (on parle du modèle décrit dans l'expression 5).

Ce qui est d'autant embêtant avec ce modèle, c'est que les deux petits

pics de 1987 et 1988 ne sont pas les seuls petits pics à ne pas être reconnus (comme période épidémique). Ainsi, le problème est le suivant : le modèle décrit dans l'équation 5 ne permet pas de reconnaitre les petits pics qui sont pourtant des épidémies.

4.1.2 Discussion sur les fonctions de coût dans le calcul des erreurs d'approximation

Rappelons-le, nous essayons d'approximer le signal sur chaque segment ($[t_1, t_2]$, $[t_2, t_3]$...) par un modèle. Nous avons donc sur chaque segment une erreur d'approximation qu'on va calculer à l'aide d'une fonction de coût c($\cdot$). On a par exemple que $c(y_{t_1..t_2})$ est l'erreur d'approximation du sous-signal $y_{t_1..t_2}$ par un certain modèle. On doit alors préciser que les résultats (sur la Figure 3) de la détection de ruptures à l'aide du modèle décrit par l'équation 5 ont été trouvé avec la fonction de coût suivante :

$$\forall i \in [0, K], c(y_{t_i..t_{i+1}}) = \sum_{t=t_i}^{t_{i+1}} ||y_t - médiane(y_{t_i..t_{i+1}})||_1 \tag{6}$$

On calculait donc les erreurs d'approximation à l'aide de la norme 1. On a choisit la norme 1 par rapport à la norme 2, car la norme 2 grossit trop les grosses erreurs d'approximation.

Nous allons alors opérer un changement à la fois sur la fonction de coût et sur les modèles utilisés pour fitter le signal sur chaque segment. Nous allons voir qu'en faisant cela l'algorithme va reconnaitre les petits pics épidémiques (qui étaient notre problème évoqué dans le sous-chapitre précédent).

Définissons cette fonction de coût et ce modèle. Soit un noyau semi-definie positif k($\cdot$,$\cdot$) : IR $\times$ IR $\mapsto$ IR et une fonction Φ : IR $\mapsto \mathcal{H}$ (où $\mathcal{H}$ est un espace de Hilbert) tel que : $\forall (x, x') \in \text{IR}^2$, $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$. Ici, on choisit comme noyau k($\cdot$,$\cdot$) une fonction de base radiale :

$\forall (x, x') \in \text{IR}^2$, $k(x, x') = exp(-\gamma ||x - x'||^2)$ où $||\cdot||$ est la norme euclidienne et $\gamma > 0$ est ce qu'on appelle en anglais "bandwidth parameter".

Voici donc la fonction de coût qu'on propose pour calculer l'erreur d'approximation du signal par modèle sur chaque segment :

$$\forall i \in [0, K], c(y_{t_i..t_{i+1}}) = \sum_{t=t_i}^{t_{i+1}} ||\Phi(y_t) - \overline{\Phi(y_t)_{t_i..t_{i+1}}}||^2_{\mathcal{H}} \tag{7}$$

où $\overline{\Phi(y_t)_{t_i..t_{i+1}}}$ est la moyenne empirique de $\{\Phi(y_t)\}_{t\in[t_i,t_{i+1}]}$

Remarque sur mon code Python : J'ai réalisé la Figure 4 (ci-dessous) grâce à la librairie Python "ruptures" qui permet de simuler des algorithmes de détection de ruptures. J'ai pu préciser dans l'algorithme le modèle et la fonction de coût décrits dans l'expression 7.

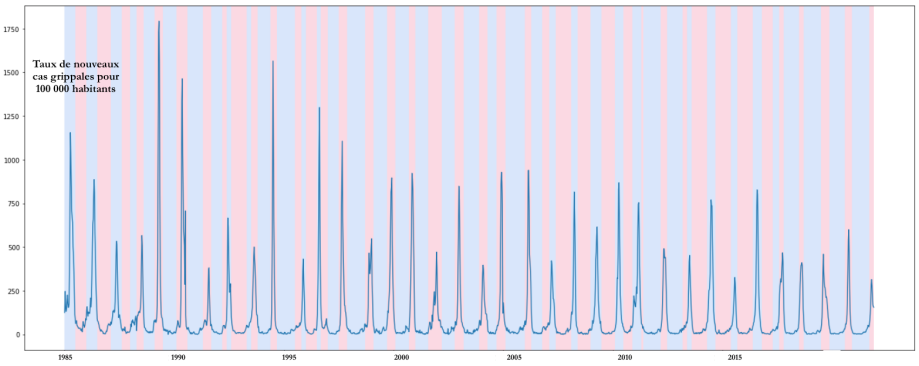

Figure 4: Détection de ruptures, sur le signal indiquant le taux de nouveaux cas grippales par semaine pour 100 000 habitants en France, à l'aide du modèle et de la fonction de coût décrits dans l'expression 7

Encore une fois, on insiste sur le fait que les couleurs rouges et bleues n'ont pas d'importance.

Sur la Figure 4, j'ai fait exprès d'inclure le signal entier (i.e. sur la période 1984-2020) car je voulais montrer que le problème qu'on avait précédemment est résolu : toutes les petits pics épidémiques sont reconnus. Nous pouvons voir aussi que le problème qu'on avait quand on avait fait appel aux modèles autorégressifs n'est également pas présent. En effet, la détection de ruptures faite à l'aide des modèles autorégressifs désignait des périodes épidémiques

qui se finissait quand le taux de nouveaux cas était à sa plus haute valeur. On voit donc sur la Figure 4 qu'on n'a pas ce problème.

Ainsi, on peut conclure que pour la détection de ruptures dans le signal indiquant le taux de nouveaux cas de grippe en France, le modèle et la fonction de coût décrits dans l'équation 7 n'a pas les limites des modèles et des fonctions de coût précédement évoqués.

4.2 Détection de rutpures sur un signal épidémiologique multivarié

Nous allons voir maintenant les avantages d'utiliser le signal multivarié suivant : le taux de nouveaux cas de grippe pour 100 000 habitants dans les 22 régions françaises.

4.2.1 Motivations

L'avantage de travailler sur le taux de nouveaux cas des 22 régions françaises et pas seulement sur le taux de nouveaux cas **agrégé** en France est qu'on capture les dynamiques d'infections entre les différentes régions. Or, ces dynamiques d'infections entre les différentes régions sont des facteurs clés de l'épidémie.

Nous allons illustrer notre propos : si par exemple, 3 régions en France sont durement touchées par l'épidémie et les autres sont épargnées c'est pire que si toutes les régions étaient légèrement touchés. En effet, s'il y a 10 000 personnes nouveaux cas dans une région l'augmentation des nouveaux cas au niveau de la France sera bien plus importante que s'il y avait 1000 nouveaux dans 10 régions différentes. Or, on ne voit pas ces différences sur le taux de nouveaux cas **agrégé** en France : en effet, on somme les nouveaux cas de toutes les régions donc qu'il y ait 10 000 nouveaux cas dans une région ou qu'il y ait 1000 nouveaux cas dans 10 régions la valeur de la somme sera la même.

Ainsi, le signal multivarié indiquant le taux de nouveaux cas au niveau des 22 régions françaises permet de prendre en compte la répartition de la maladie au sein du territoire, ce qui est un facteur clé pour la détection d'épidémie. Nous verrons par la suite qu'on a différentes manières d'évaluer la répartition de la maladie au sein du territoire à travers des indices comme l'entropie par

exemple ou en écrivant sur Python un code permettant d'afficher un graphe représentant la carte de France avec le taux de nouveaux cas sur chaque région ce qui nous permettrait de voir d'un coup d'oeil cette répartition.

4.2.2 Choix du modèle et de fonction de coût

Rappelons-le on travaille désormais sur le signal multivarié indiquant le taux de nouveaux cas de grippe des 22 régions françaises.

On doit maintenant choisir les modèles qui nous permettront d'approximer ce signal (multivarié) sur chaque segment identifié. On doit aussi choisir la fonction de coût qui nous permettra de calculer l'erreur d'approximation du modèle sur chaque segment. On prend alors les mêmes modèles et la même fonction de coût qu'on avait retenu à la fin pour le signal univarié. Définissons les maintenant dans le cas multivarié :

Soit un noyau semi-definie positif $\mathrm{k}(\cdot,\cdot) : \mathrm{IR}^{22} \times \mathrm{IR}^{22} \mapsto \mathrm{IR}$ et une fonction $\Phi : \mathrm{IR}^{22} \mapsto \mathcal{H}$ (où $\mathcal{H}$ est un espace de Hilbert) tel que :
$\forall (x, x') \in \mathrm{IR}^{22} \times \mathrm{IR}^{22}, k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$. Ici, on choisit comme noyau $\mathrm{k}(\cdot,\cdot)$ une fonction de base radiale :
$\forall (x, x') \in \mathrm{IR}^{22} \times \mathrm{IR}^{22}$, $k(x, x') = exp(-\gamma ||x - x'||^2)$ où $||\cdot||$ est la norme euclidienne et $\gamma > 0$ est ce qu'on appelle en anglais "bandwidth parameter".

Voici donc la fonction de coût qu'on propose pour calculer l'erreur d'approximation du signal par le modèle sur chaque segment :

$$\forall i \in [0, K], c(y_{t_i..t_{i+1}}) = \sum_{t=t_i}^{t_{i+1}} ||\Phi(y_t) - \overline{\Phi(y_t)_{t_i..t_{i+1}}}||_{\mathcal{H}}^2 \tag{8}$$

où $\overline{\Phi(y_t)_{t_i..t_{i+1}}}$ est la moyenne empirique de $\{\Phi(y_t)\}_{t \in [t_i, t_{i+1}]}$

4.2.3 Résultats

Ainsi, en prenant le modèle (expliqué dans le sous-chapitre précédent) et la fonction de coût expliquée dans l'expression 8, on a :

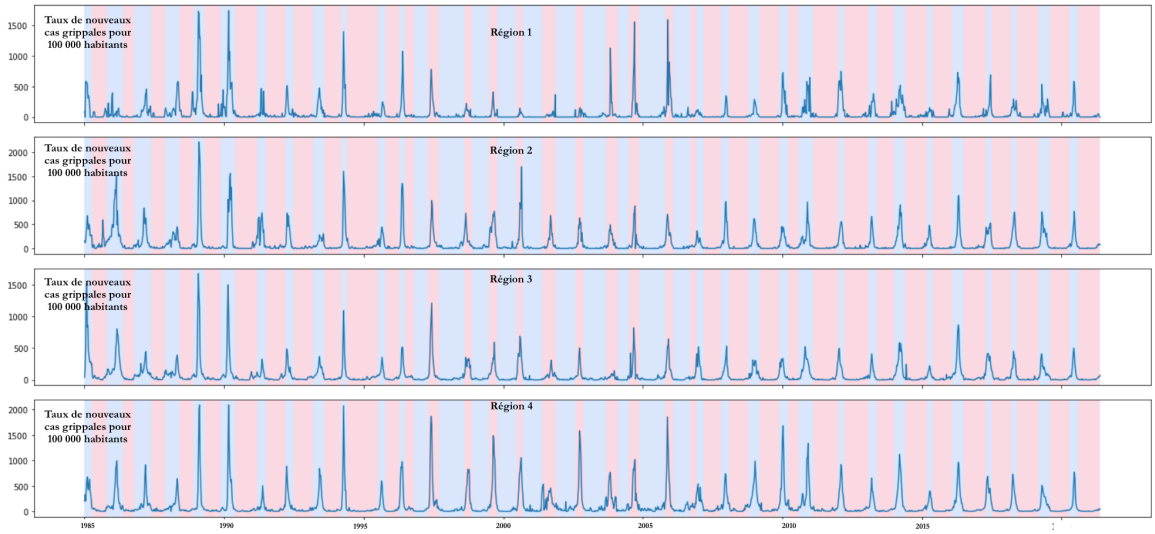

Figure 5: Détection de ruptures, sur le signal mutivarié indiquant le taux de nouveaux cas grippales par semaine pour 100 000 habitants dans les 22 régions françaises, à l'aide du modèle et de la fonction de coût décrits dans l'équation 8

Remarque sur mon code Python : Pour faire la Figure 5, j'ai du faire appel à la librairie Python "ruptures" qui permet également de travailler sur les signaux multivariés et permet donc de détecter dans ces signaux multivariés les ruptures. J'ai pu préciser dans l'algorithme le modèle et la fonction de coût décrits dans l'expression 8.

Sur la Figure 5, chacun des 4 graphiques correspond au taux de nouveaux cas de grippe pour une région. Il y avait donc 22 graphiques à la base pour les 22 régions françaises mais je ne les ai pas mis pour ne pas que cela prenne trop de place dans le rapport. On va maintenant commenter les performances de cette algorithme qui a été fait à l'aide du modèle et de la fonction de coût décrits dans l'équation 8. On remarque d'abord que les petits pics sont reconnus, ce qui est une bonne chose. En effet, lorsqu'on avait dans le cas du signal univarié comme modèle (pour fitter le signal sur les sous-segments) la médiane, l'algorithme ne reconnaissait pas les petits pics qui étaient pourtant des périodes épidémiques. Ici, ce problème est donc

résolu. On remarque également qu'on a plus le problème auquel on avait été face quand on avait choisi comme modèles les modèles autorégressifs. En effet, on avait vu qu'à l'aide des modèles autorégressifs, il y avait un nombre non-négligeable de périodes épidémiques qui s'arrêtaient au pic du taux de nouveaux cas, ce qui n'est pas normal. En effet, cela voudrait dire que l'épidémie se serait terminé au moment où elle est la plus forte. Ainsi, ce problème est ici également résolu. Ainsi, on en conclut que nous allons retenir comme modèles et comme fonction de coût ceux décrits dans l'équation 8.

5 Comparaison de mes résultats et méthodes avec ceux déja établis dans des articles épidémiologiques

Il est maintenant temps de comparer mes résultats et méthodes utilisés avec ceux déja établis notamment sur le sujet du taux d'incidence grippales (i.e. taux de nouveaux cas de grippe). Pour cela, je suis allé chercher sur le site "Sentinelles". Ce site dévoile de nombreux travaux en épidémiologie sur plusieurs maladies. **C'est d'ailleurs de ce site que j'ai pu avoir les bases de données sur lesquelles j'ai travaillé tout au long de mon stage.** Ce site propose parmi ces travaux, des méthodes de détection d'épidémie notamment sur le cas de la grippe et se concentre également sur le taux de nouveaux cas de grippe pour 100 000 habitants en France. La qualité des travaux sur ce site (**ainsi que la qualité des bases de données**) est garantie dans la mesure où ce site est une collaboration entre l'Inserm, Médecine Sorbonne Université et Santé Publique France. Il est maintenant temps d'étudier en profondeur la méthode qu'ils utilisent pour détecter les épidémies grippales.

5.1 Explication de la méthode de détection d'épidémies grippales mise en place dans un article du site "Sentinelles"

Ici, nous nous intéressons à un article publié sur le site "Sentinelles". Ce papier est réalisé par l'UPMC et l'Inserm. Il est important d'abord de re-

marquer que cet article travaille sur le signal univarié qu'on avait étudié (le taux de nouveaux cas de grippe en France pour 100 000 habitants). Nous allons maintenant expliquer la méthode exposée sur ce papier.

Il est d'abord très important de préciser que la méthode utilisée dans cet article ne suit pas la même logique que les méthodes que j'ai utilisé et qui sont exposées dans toute la section 4. En effet, dans les méthodes que j'ai utilisé on testait plusieurs segmentations différentes, et pour chaque segmentation on calculait la somme sur tous les segments des erreurs d'approximations faites par les modèles sur le signal. Ici, la méthode utilisée dans l'article publié sur le site "Sentinelles" fonctionne autrement.

La méthode consiste à approximer le signal entier (qui est le même signal que celui étudié la section 4.1 "Détection de ruptures sur un signal épidémiologique univarié c'est à dire celui représentant le taux de nouveaux cas de grippe pour 100 000 habitants au niveau de la France). L'approximation du signal se fait par la fonction suivante :

$$t \longmapsto \alpha_0 + \alpha_1 t + \sum_{k=1}^{n} [\alpha_{2,k} cos(\frac{2k\pi t}{52}) + \alpha_{3,k} sin(\frac{2k\pi t}{52})] \tag{9}$$

(la fonction est évaluée en t qui représente la semaine où l'on mesure le taux de nouveaux cas et donc t varie entre 1 et 52. De plus, n est égal à 2 pour l'étude sur le taux de nouveaux cas de grippe)

On parle de régression périodique dite de "Serfling" pour désigner l'approximation du signal univarié par la fonction décrite dans l'expression 9. Quand on voit la permière partie de la fonction on reconnait une fonction affine puis on voit une somme de fonctions de cosinus et de sinus. On voit également des poids devant le cosinus et le sinus et on remarque qu'on somme des cosinus et des sinus ayant différentes périodes. Il est assez intuitif de voir qu'il est bien d'avoir des cosinus et des sinus ayant différentes périodes pour approximer un signal car cela permet d'approximer de manière plus personalisée différents endroits.

Suite à l'approximation du taux de nouveaux cas, l'article approxime un nouveau signal (disponible dans la base de données) qui est la borne supérieure de l'intervalle de prédiction à 90% du taux de nouveux cas de

grippe (pour 100 000 habitants). On approxime cette nouvelle variable par la même fonction c'est à dire la fonction décrite dans l'expression 9. Quand le signal est supérieur à cette dernière approximation, on peut considérer qu'on détecte une nouvelle épidémie. Cela va être plus claire avec le graphique suivant :

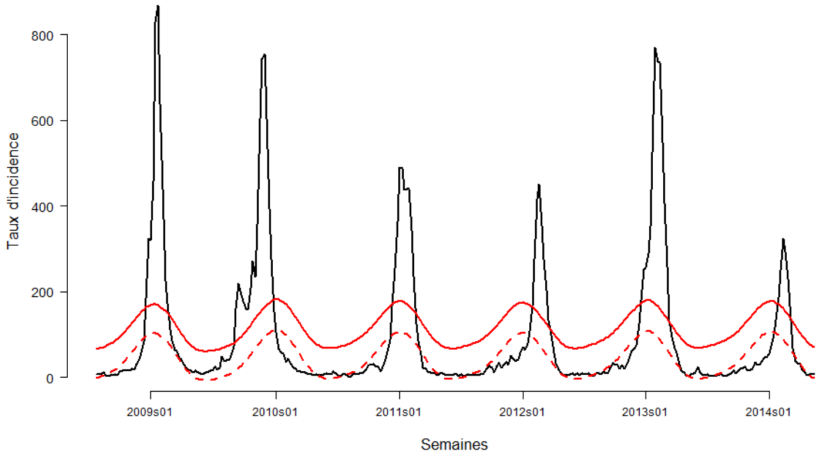

Figure 6: Détection d'épidémie grippales venant de l'article (publié sur "Sentinelles"). Taux d'incidence des syndromes grippaux pour 100 000 habitants (noir). Valeurs estimées par le modèle de régression périodique (rouge pointillé), borne supérieure de l'intervalle de prédiction à 90% des valeurs attendues (rouge plein).

Autrement dit, quand le signal est supérieur à la ligne rouge pleine dans le graphique ci-dessous alors on détecte une nouvelle épidémie. Le ligne rouge pleine désigne l'approximation faite par la fonction décrite dans l'expression 9 sur la nouvelle variable étant la borne supérieure de l'intervalle de prédiction à 90% du taux de nouveux cas de grippe (pour 100 000 habitants en France).

Rappelons ici que le taux d'incidence (cf Figure 6) signifie le taux de nouveaux cas sur une période donnée (ici la période donnée est une semaine).

5.2 Comparaison de mes résultats avec ceux de l'article

Nous allons comparer mes résultats avec ceux tirés de la méthode décrite dans l'article. Je vais notamment comparer les résultats de l'article avec mes résultats que j'ai eu en faisant la détection de ruptures sur le signal multivarié à l'aide du modèle et de la fonction de coût décrite dans l'équation 8 (cf section 4.2.2). Nous allons notamment comparer les périodes épidémiques trouvées dans les deux méthodes. Pour cela, nous allons nous aider de la Figure suivante :

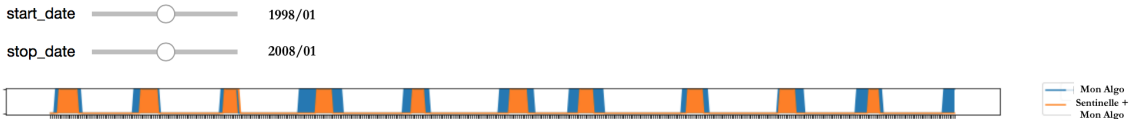

Figure 7: Comparaison, sur la période 1998-2008, de mes résultats tirés de la méthode expliquée dans la section 4.2.2 avec ceux de l'article publié sur le site "Sentinelles". Les zones bleues sont les périodes épidémiques reconnues par la méthode expliquée dans la section 4.2.2. et les zones oranges sont les périodes épidémiques reconnues à la fois par la méthode de l'article "Sentinelles" et par la méthode expliquée dans la section 4.2.2

On remarque, de manière frappante, que l'algorithme sur lequel j'ai travaillé (voir section 4.2.2) détecte les épidémies avant que ne les détecte l'algorithme de l'article "Sentinelles".

Cependant, le fait de détecter une épidémie plus tôt est une bonne chose que dans certains cas. Autrement dit, si on détecte une épidémie trop tôt avant qu'elle commence, la période dans laquelle l'Etat va prendre des mesures pour stopper l'épidémie risque d'être trop longue. Or, plus la période dans laquelle l'Etat va prendre des mesures est longue plus cela va lui coûter de l'argent et plus cela peut également faire des problèmes à l'économie. C'est ce que nous sommes entrain de voir actuellement avec la crise du coronavirus. La période dans laquelle l'Etat prend des mesures (comme le couvre-feu à partir de 21h, ou la fermeture des bars et restaurants à partir d'une certaine heure) coûte très cher à l'économie. Par conséquent, le fait de détecter une épidémie plus tôt est bénéfique que si cela est bien justifié.

Ainsi, suite à la remarque du paragraphe précédent, on va tenter de voir s'il est bénéfique que l'algorithme sur lequel j'ai travaillé (voir section 4.2.2) détecte les épidémies avant que ne les détecte l'algorithme de l'article "Sentinelles".

5.2.1 Une épidémie est d'autant plus dangereuse qu'elle part d'endroits différents

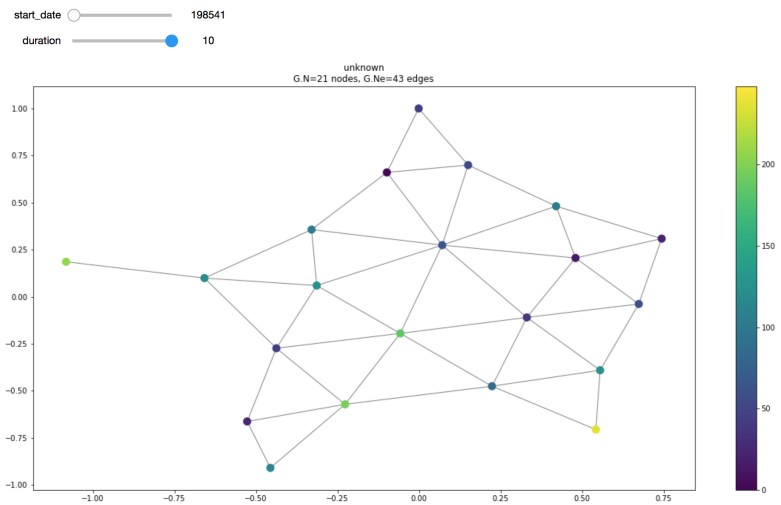

Figure 8: Graphe de la France, dans lequel chaque noeud est une des 22 régions françaises, et dans lequel chaque noeud a une couleur correspondant à la moyenne du taux de nouveaux cas de grippe (par semaine) entre la 41ème semaine et 51ème semaine de l'année 1985

La Figure 8 montre dans un cas particulier qu'il était bénéfique que l'algorithme sur lequel j'ai travaillé (voir section 4.2.2) détecte l'épidémie avant que ne le détecte l'algorithme de l'article "Sentinelles". En effet, mon algorithme a désigné comme période épidémique la période allant de la 41ème semaine de l'année 1985 à la 19ème semaine de l'année 1986 alors que l'algorithme de Sentinelles a désigné comme période épidémique la période

allant de la 2ème semaine de l'année 1986 à la 12ème semaine de l'année 1986.

Ainsi, la période allant de la 41ème semaine de l'année 1985 à la 1ère semaine de l'année 1986 est considérée comme une période épidémique pour mon algorithme et elle n'est pas considérée comme une période épidémique pour l'algorithme de Sentinelles.

Nous pouvons voir sur la Figure 8 que l'épidémie de grippe avait effectivement commencé au moment où mon algorithme a commencé à détecter une épidémie. En effet, la Figure 8 porte sur la période (allant de la 41ème semaine de l'année 1985 à la 51ème semaine de l'année 1985) considérée comme une période épidémique pour mon algorithme mais pas pour l'algorithme de Sentinelles. Or, on voit sur cette période que l'épidémie a commencé : au Nord-Ouest, au Sud-Est et au Sud-Ouest on peut voir qu'il y a des régions exposées à un grand nombre de nouveaux cas. Or, le fait que ces régions, ayant beaucoup de nouveaux cas, sont dispersés à différents endroits de la France fait que l'épidémie est d'autant plus compliquée à maîtriser.

Cependant, nous avons montré que dans un cas particulier qu'il était bénéfique que l'algorithme sur lequel j'ai travaillé (voir section 4.2.2) détecte l'épidémie avant que ne les détecte l'algorithme de l'article "Sentinelles". Néanmoins, on peut voir ce phénomène sur beaucoup d'exemples mais cela supposerait pour chaque exemple de mettre un nouveau graphe de la France dans le rapport, ce qui prendrait trop de place.

Nous avons vu avec la Figure 8 qu'il était bénéfique de détecter de manière très anticipée une épidémie notamment quand des régions relativement dispersées comportent un taux de nouveaux cas élevé car plus les clusters d'infections sont dispersés plus cela est compliqué à maitriser. Nous allons maintenant voir, avec un autre critère, quand il est bénéfique de détecter de manière très anticipée une épidémie, et ce critère s'appelle l'entropie.

5.2.2 Entropie et Hétérogénéité

Définissons d'abord ce que nous voulons dire par entropie.

$$H(t) = -\sum_{i=1}^{22} P_i(t) \times log(P_i(t)) \tag{10}$$

avec $P_i(t) = \frac{inc_i(t)}{\sum_{k=1}^{22} inc_k(t)}$ où $inc_i(t)$ est le taux d'incidence (i.e. le taux de nouveaux cas) pour 100 000 habitants à la semaine t dans une région i, où $P_i(t)$ est la probabilité d'apparition de nouveaux cas dans une région i à la semaine t, et H(t) est l'entropie à la semaine t.

L'entropie définie dans l'expression 10 va nous permettre de voir s'il est bénéfique que l'algorithme sur lequel j'ai travaillé (voir section 4.2.2) détecte l'épidémie avant que ne les détecte l'algorithme de l'article "Sentinelles".

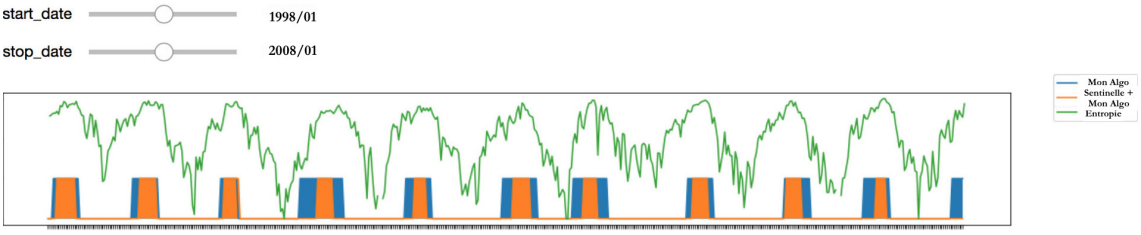

Figure 9: Courbe (en vert) de l'entropie évoluant à travers les semaines. Les zones bleues sont les périodes épidémiques reconnues par la méthode expliquée dans la section 4.2.2. et les zones oranges sont les périodes épidémiques reconnues à la fois par la méthode de l'article "Sentinelles" et par la méthode expliquée dans la section 4.2.2

Avant de commenter la Figure 9, il est important de préciser qu'une augmentation de l'entropie dans notre cas de figure signifie une augmentation de l'hétérogénéité des $(P_i(t))_{i\in[1,22]}$ qui désignent le taux de nouveaux cas dans chaque région normalisée (voir la formule des P_i dans l'équation 10). Par conséquent, quand l'entropie augmente cela veut dire que le taux de nouveaux cas diffère davantage entre chaque région.

Sur la Figure 9, on voit d'abord en général que l'entropie augmente avec les épidémies (que ce soit pour les périodes épidémiques reconnu par mon

algorithme ou par l'algorithme Sentinelles). Ceci est intuitif car on a dit dans le paragraphe précédent que quand l'entropie augmente cela veut dire que le taux de nouveaux cas différe davantage entre chaque région. En effet, il y a plus de probabilité d'avoir une épidémie quand les taux de nouveaux cas diffèrent beaucoup entre chaque région puisque les régions infectées vont contaminer les régions non-infectées ce qui peut probablement créer une épidémie.

Comme on l'a dit précédemment on remarque que sur chaque épidémie, l'algorithme sur lequel j'ai travaillé (voir section 4.2.2) détecte l'épidémie avant que ne les détecte l'algorithme de l'article "Sentinelles". Or, à ces moments détectés que par mon algorithme, on voit en majorité l'entropie monter fortement, ce qui veut dire que les périodes détectées par mon algorithme sont des périodes où les taux de nouveaux cas diffèrent beaucoup entre chaque région ce qui veut dire les régions infectées vont contaminer les régions non-infectées ce qui est le début caractéristique des épidémies. Ainsi, l'entropie montre qu'il est bénéfique mon algorithme détecte les épidémies avant que ne les détecte l'algorithme de l'article "Sentinelles".

5.2.3 L'avance de mon algorithme sur la détection d'épidémie peut être utile lors des grosses épidémies

Il est bénéfique de détecter de manière très anticipée une épidémie notamment quand on attend une grosse épidémie. On a vu précédemment qu'il était couteux (sur le plan économique) de détecter de manière très anticipée une épidémie mais cela vaut le coup lorsqu'il s'agit d'une grosse épidémie.

Pour cela, nous allons voir si l'algorithme de Sentinelles a eu beaucoup de retard relativement à mon algorithme sur des grosses épidémies. Mais d'abord comment définir ce qu'est une grosse épidémie ?

Le premier critère pour définir une grosse épidémie est le niveau de son pic c'est à dire le niveau le plus haut de nouveaux cas de grippe durant l'épidémie. Ici on parle du pic du signal univarié (indiquant le taux de nouveaux cas de grippe sur 100 000 habitants agrégé au niveau de la France). On introduit ainsi la fonction :

$$\forall i \in [0, \frac{K-1}{2}], pic(y_{t_{2i}..t_{2i+1}}) = \frac{max((y_t)_{t\in[t_{2i},t_{2i+1}]})}{max(max((y_t)_{t\in[t_{2i},t_{2i+1}]})_{i\in[0,\frac{K-1}{2}]})} \quad (11)$$

Remarquons d'abord qu'on n'a pas pris tous les segments pour cette fonction (décrite dans l'expression 11) dans la mesure où seulement un segment sur deux est une période épidémique (l'autre segment est une période non-épidémique). Remarquons ensuite que le numérateur correspond au niveau le plus haut du taux nouveaux cas de grippe dans l'épidémie, qu'on appellera plus brièvement le pic de l'épidémie. Le dénominateur, quant à lui, correspond au pic le plus haut de toutes les épidémies confondues (entre 1984 et 2020). Le dénominateur est juste là pour normaliser et avoir une idée de la hauteur du pic de l'épidémie i. On dira donc que la fonction pic de l'épidémie i représente la hauteur du pic de l'épidémie i relativement au plus grand pic épidémique observé.

Il s'agit maintenant de calculer le retard de l'algorithme "Sentinelles" sur mon algorithme pour chaque épidémie. En notant $(t_i)_{i\in[0,K]}$ les bornes temporelles des épidémies reconnus par mon algorithme et en notant $(t'_i)_{i\in[0,K]}$ les bornes temporelles des épidémies reconnus par l'algorithme Sentinelles, on a :

$$\forall i \in [0, \frac{K-1}{2}], lag(\textit{épidémie}_i) = (t_{2i} - t'_{2i})^2 \quad (12)$$

Comme on peut le voir sur l'expression 12, la fonction lag est l'écart de temps (exprimé en semaines) mis au carré entre la date du début de l'épidémie i donnée par mon algorithme et la date du début de l'épidémie i donnée par l'algorithme Sentinelles.

On va maintenant pour chaque épidémie entre 1984 et 2019 voir dans la Figure 10 (ci-dessous) à la fois l'écart de détection entre les deux algorithmes et leur pic normalisé (décrit dans l'expression 11).

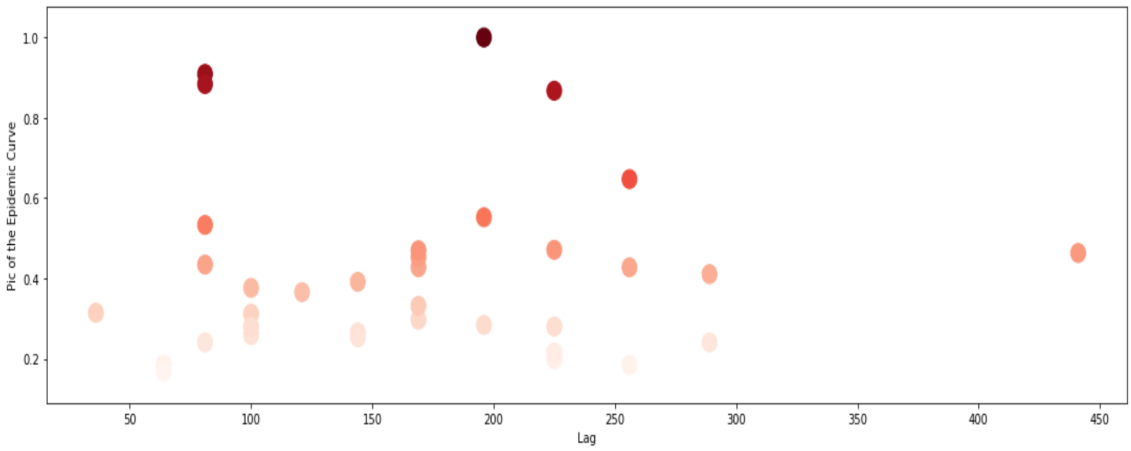

Figure 10: Chaque point représente une épidémie. En ordonnée, pour une épidémie i, on recense la hauteur de son pic normalisé (par la hauteur du plus grand pic de toutes les épidémies), cf expression 11. En abscisse, on recense le temps de retard au carré (pour détecter l'épidémie i) de l'algorithme Sentinelles sur mon algorithme,ce temps étant exprimé en semaines, cf expression 12

On voit sur la Figure 10 qu'il y a un nombre non-négligeable d'épidémies qui ont été détecté avec un temps de retard important (par l'algorithme Sentinelles par rapport à mon algorithme) et qui ont un pic élevé. Cela veut dire, au pire des cas, que mon algorithme a détecté de manière très anticipé (par rapport à l'algorithme Sentinelles) des grosses épidémies, ce qui est légitime.

On a également une autre manière de définir les grosses épidémies. Ici, on ne va pas mesurer le pic des épidémies pour définir leur grosseur mais on va sommer les taux de nouveaux cas durant toute l'épidémie, ce qui donne la fonction suivante :

$$\forall i \in [0, \frac{K-1}{2}], area(épidémie_i) = \frac{\sum_{t=t_{2i}}^{t_{2i+1}} y_t}{max((\sum_{t=t_{2i}}^{t_{2i+1}} y_t)_{i \in [0, \frac{K-1}{2}]})} \tag{13}$$

La fonction area définit, pour une épidémie i, la somme normalisée des taux de nouveaux cas (de grippe pour 100 000 habitants en France) durant toute l'épidémie. On appelera la version non normalisée de cette somme, l'aire sous la courbe de l'épidémie i. On a donc que la fonction area définit l'aire sous la courbe de l'épidémie i divisée par la plus grande aire sous la courbe parmi toutes les épidémies confondues, cette division étant là pour normaliser la quantité et aussi pour avoir une idée de la grandeur de l'aire sous la courbe de chaque épidémie.

On va maintenant, pour chaque épidémie entre 1984 et 2019, voir dans la Figure 11 (ci-dessous) à la fois l'écart de détection entre les deux algorithmes (mon algorithme et celui de Sentinelles) et leur aire sous la courbe normalisée (décrite dans l'expression 13).

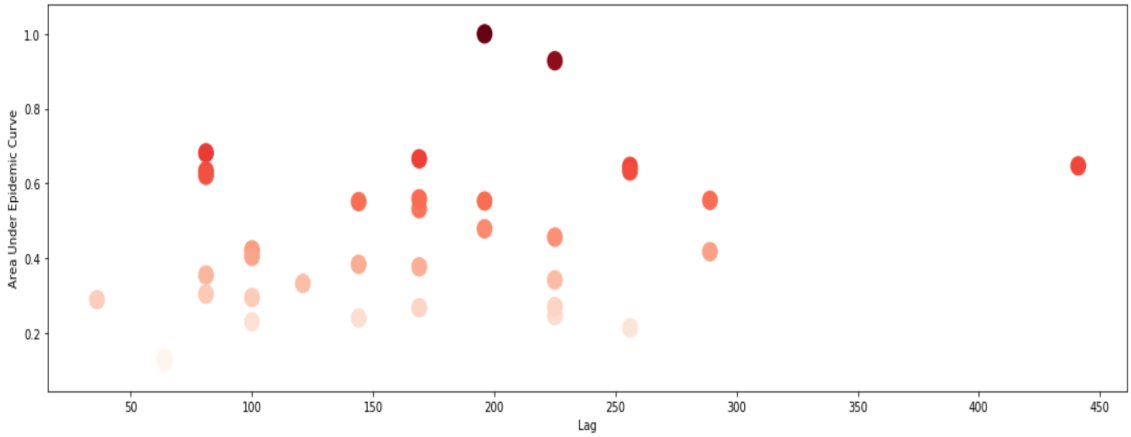

Figure 11: Chaque point représente une épidémie. En ordonnée, pour une épidémie i, on recense son aire sous la courbe normalisée (par la plus grande aire sous la courbe parmi toutes les épidémies confondues), cf expression 13. En abscisse, on recense le temps de retard au carré (pour détecter l'épidémie i) de l'algorithme Sentinelles sur mon algorithme, ce temps étant exprimé en semaines, cf expression 12

On voit sur la Figure 11 qu'il y a un nombre non-négligeable d'épidémies qui ont été détecté avec un temps de retard important (par l'algorithme Sentinelles par rapport à mon algorithme) et qui ont une aire sous leur courbe

importante. Cela veut dire, au pire des cas, que mon algorithme a détecté de manière très anticipé (par rapport à l'algorithme Sentinelles) des grosses épidémies, ce qui est légitime.

Ainsi, on peut conclure qu'on a pu démontrer avec plusieurs critères (la répartition spatiale des nouveaux cas, l'entropie, la grosseur des épidémies mesurée par leur pic et leur aire sous la courbe) qu'il a été bénéfique dans de nombreux cas que l'algorithme sur lequel j'ai travaillé (dans la section 4.2.2) ait permis de détecter de manière anticipée les épidémies par rapport à l'algorithme de l'article de Sentinelles.

6 Conclusion

Dans ce stage, j'ai pu apprendre de nombreuses notions sur la théorie de détection de ruptures dans des signaux. Ces notions étaient souvent en lien avec ce qu'on nous a enseigné à l'ENSAE notamment le cours de "Séries temporelles", le cours de "Python pour les data scientists" (vu que tous mes résultats se sont faits par Python en implémentant des algorithmes de détection de ruptures), le cours d'"Optimisation différentiable", le cours de "Séminaire statistique" où j'ai pu découvrir les applications statistiques en épidémiologie.

J'ai donc testé plusieurs algorithmes de détection de ruptures sur des signaux concernant la grippe dans le but de détecter des épidémies de grippe. J'ai pu tester différents modèles pour approcher sur chaque segment les signaux et ai également testé différentes fonctions de coût pour calculer l'erreur d'approximation des signaux par les modèles. A chaque test, que ce soit pour les modèles ou les fonctions de coût, j'ai pu critiquer les résultats que j'ai eu. Chaque critique m'a fait penser à un nouveau modèle et à une nouvelle fonction de coût permettant de contrer la critique. Le modèle et la fonction de coût finalement retenus sont ceux exposés dans l'expression 7 pour le signal univarié, et l'expression 8 pour le signal multivarié. Rappelons que le signal univarié indique le taux de nouveaux cas de grippe pour 100 000 habitants agrégé au niveau de la France et que le signal multivarié indique le taux de nouveaux cas de grippe pour 100 000 habitants au niveau des 22 régions françaises.

L'avantage d'avoir travaillé sur le signal multivarié est qu'on parvient à capturer avec ce signal multivarié les dynamiques d'infection entre les différentes régions de France, ce qui est essentiel pour détecter une épidémie.

J'ai ensuite comparé mes résultats et méthodes faites sur le signal multivarié (exposés dans la section 4.2.2) avec ceux de l'article publié dans "Sentinelles". On a démontré avec plusieurs critères (la répartition des nouveaux cas, l'entropie, la grosseur des épidémies mesurée par leur pic et leur aire sous la courbe) qu'il a été bénéfique dans de nombreux cas que l'algorithme sur lequel j'ai travaillé ait permis de détecter de manière anticipée les épidémies par rapport à l'algorithme de l'article publié dans "Sentinelles".

References

[1] C. Truong, L. Oudre, N. Vayatis. Selective review of offline change point detection methods. Signal Processing, 167:107299, 2020

[2] Collaboration Inserm/UPMC, Détection des épidémies de grippe et de gastro-entérite par le réseau Sentinelles, Article, Sentinelles

[3] Serfling RE: Methods for current statistical analysis of excess pneumonia-influenza deaths. Public health reports 1963, 78(6):494-506.

[4] R. Killick, P. Fearnhead, and I. Eckley. Optimal detection of change points with a linear computational cost. Journal of the American Statistical Association, 107(500):1590–1598, 2012.

[5] J. Bai. Estimating multiple breaks one at a time.Econometric Theory, 13(3):315–352, 1997.

NOTES DE SYNTHESE

Note de Synthèse *: Détection de ruptures multiples dans des signaux épidémiologiques*

Ce stage, réalisé au laboratoire de l'ENS Paris-Saclay (Centre Borelli), répond au besoin croissant de pouvoir prévoir le plus tôt possible l'apparition d'épidémies. Ce besoin s'est notamment fait remarquer avec la crise du COVID-19. Plus tôt on peut prévoir l'apparition d'une épidémie (par exemple l'apparition d'une deuxième vague dans le cadre du coronavirus), plus les moyens envisagés par les gouvernements pour contrer l'épidémie seront efficaces. Le but de ma démarche a été donc d'appliquer des algorithmes de détection de ruptures pour détecter les épidémies.

Mais d'abord qu'est ce qu'un algorithme de détection de ruptures ?

Comme on peut le voir sur la Figure 1, un algorithme de détection de ruptures doit reconnaitre différentes périodes dans un signal et donc diviser ce signal en plusieurs segments, où chaque segment représente une période. Une rupture est le point de passage d'une période à une autre, autrement dit une rupture marque la fin d'un segment et le début d'un autre.

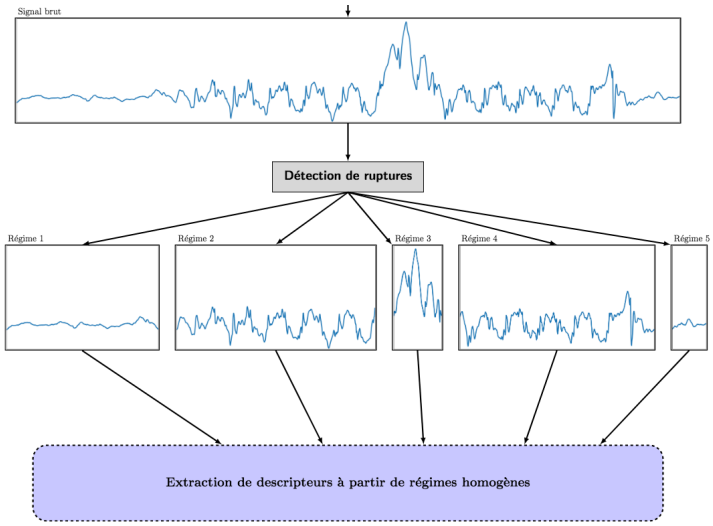

Figure 1 : *Fonctionnement d'un algorithme de détection de ruptures*

Ici, on ne cherche pas à segmenter n'importe quels signaux : on cherche à segmenter des signaux épidémiologiques. Un signal épidémiologique est un signal décrivant une mesure de la diffusion d'une épidémie au cours du temps. Par exemple, un signal épidémiologique peut être le nombre d'infections dans une population ou le nombre de personnes en réanimation…

Ainsi, chaque segment (qui représente une portion de temps) représente soit une période épidémique soit une période non-épidémique.

Mais de quelle manière notre algorithme de détection de ruptures va t-il segmenter un signal (en périodes d'épidémie et périodes non épidémiques) ?

L'algorithme teste plusieurs segmentations différentes (i.e l'algorithme teste différentes découpes du signal). Sur chaque segmentation testée, notée Υ, l'algorithme utilise des modèles (modèles

autorégressifs, médiane…) pour approximer le signal sur chaque segment. On a donc par segment une erreur d'approximation du signal par le modèle. On somme les erreurs d'approximation de tous les segments. Ainsi, la segmentation (i.e. la découpe) du signal qui aura la plus petite somme d'erreurs d'approximation sera gardée et considérée comme la segmentation optimale notée $\widehat{\Upsilon}$. On notera la somme d'erreurs d'approximation de tous les segments $V(\Upsilon)$ (où Υ est une segmentation i.e. un ensemble de segments qu'on peut écrire $:\Upsilon = \{[t_1, t_2], [t_2, t_3], \ldots, [t_{K-1}, t_K]\}$ où les t_i sont les bornes des segments i.e. les points de ruptures). On cherche donc la segmentation optimale $\widehat{\Upsilon} = argmin_{\Upsilon} V(\Upsilon)$ qui nous permettra donc de détecter de la manière la plus juste les périodes d'épidémie (et aussi les périodes non épidémiques).

Dans notre cas, on cherche à segmenter deux signaux épidémiologiques précis. Le premier signal à segmenter est un signal univarié : le taux de nouveaux cas de grippe (chaque semaine entre 1984 à 2020) sur 100 000 habitants en France. Le deuxième signal à segmenter est un signal multivarié : le vecteur où chaque coordonnée représente le taux de nouveaux cas de grippe (chaque semaine entre 1984 à 2020) sur 100 000 habitants dans une des 22 régions françaises (ce vecteur a donc 22 coordonnées).

On a donc, avec le **librairie "Ruptures" sur Python**, testé plusieurs découpes (i.e. segmentations) des signaux décrits dans le paragraphe précédent. Cependant, pour chaque découpe on doit faire des choix : le premier choix concerne le modèle avec lequel on veut approximer le signal sur chaque segment et le deuxième choix concerne la manière avec laquelle on calcule l'erreur d'approximation i.e. la fonction de coût.

Pour approcher le signal sur chaque segment, on a d'abord testé les modèles autorégressifs. Cependant, on voit que certaines épidémies identifiées (en choisissant comme modèles les modèles autorégressifs) s'arrêtent au moment où le taux de nouveaux cas est le plus fort, ce qui n'est pas normal.

Pour approcher le signal sur chaque segment, on a également pensé à la médiane des valeurs prises par le signal sur chaque segment, ce qui pourrait bien fonctionner car ce qui caractérise une épidémie c'est notamment le taux de nouveaux cas anormalement élevé par rapport à la moyenne. On a choisi la médiane à la place de la moyenne car elle est plus robuste aux valeurs extrêmes. Cependant, on voit (en choisissant comme modèles les médianes des valeurs prises par le signal sur chaque segment) certains pics entiers que l'algorithme ne reconnait pas comme épidémiques alors qu'on sait qu'ils correspondent à des épidémies (cela concerne les petits pics). On a donc choisi (pour notre signal univarié et multivairié) comme type de modèle et comme fonction de coût ceux décrits dans l'expression (1). Ce type de modèle et cette fonction de coût ont été retenu car ils n'ont pas les défauts reprochés aux deux autres types de modèle précédemment testés (modèles autorégressifs et médiane).

$$\forall i \in [0,K], c(y_{t_i..t_{i+1}}) = \sum_{t=t_i}^{t_{i+1}} || \Phi(y_t) - \overline{\Phi(y_t)_{t_i..t_{i+1}}} ||^2_{\mathscr{H}} \qquad (1)$$

où $y_{t_i..t_{i+1}}$ est le signal restreint au segment $[t_i, t_{i+1}]$, $c(.)$ est la fonction de coût, $k(.,.)$ un noyau semi-definie positif tel que $k(y, y') = \langle \Phi(y), \Phi(y') \rangle_{\mathscr{H}}$ ($y, y' \in \mathrm{IR}^{22}$ si on est sur le signal multivarié des 22 régions françaises et $y, y' \in \mathrm{IR}$ si on est sur le signal univarié), $\mathscr{H}$ étant un espace de Hilbert et $\overline{\Phi(y_t)_{t_i..t_{i+1}}}$ étant la moyenne empirique de $\{\Phi(y_t)\}_{t \in [t_i, t_{i+1}]}$.

On va garder les périodes épidémiques identifiées par l'algorithme de détection de ruptures sur le signal multivarié (i.e. le taux de nouveaux cas de grippe sur 100 000 habitants au niveau des 22 régions françaises) plutôt que ceux du signal univarié (i.e. le taux de nouveaux cas de grippe sur 100 000 habitants agrégé au niveau de la France) dans la mesure où le signal multivarié nous permet de capturer les dynamiques d'infections entre les différentes régions. Or, ces dynamiques d'infections entre les différentes régions sont des facteurs clés de l'épidémie.

<u>Notre algorithme de détection d'épidémie est-il meilleur que certains déjà établis ?</u>

Nous avons comparé notre algorithme de détection d'épidémie avec celui développé dans le papier nommé "Détection des épidémies de grippe et de gastro-entérite par le réseau Sentinelles" publié sur le site " Sentinelles" (fruit d'une collaboration Inserm/UPMC).

$$\alpha_0 + \alpha_1 t + \sum_{k=1}^{n} [\alpha_{2,k} cos(\frac{2k\pi t}{52}) + \alpha_{3,k} sin(\frac{2k\pi t}{52})] \qquad (2)$$

(où t représente la semaine de l'année en question et donc t varie entre 1 et 52. De plus, n est égal à 2 pour l'étude sur le taux de nouveaux cas de grippe)

L'algorithme de détection d'épidémies dans ce papier fonctionne de la manière suivante. On approxime par un modèle (fonction décrite dans l'expression (2)) notre signal univarié entier et non pas par segment comme on l'a fait pour notre algorithme. Plus précisément, on approche notre signal univarié par le biais une régression périodique dite de "Serfling". Suite à l'approximation du taux de nouveaux cas, on approxime un nouveau signal (disponible dans la base de données) qui est la borne supérieure de l'intervalle de prédiction à 90 % du taux de nouveaux cas de grippe (pour 100 000 habitants). On approxime cette nouvelle variable par la même fonction décrite dans l'expression (2). Quand notre premier signal univarié (celui représentant le taux de nouveaux cas de grippe) est supérieur à cette dernière approximation, on peut considérer qu'on détecte une nouvelle épidémie.

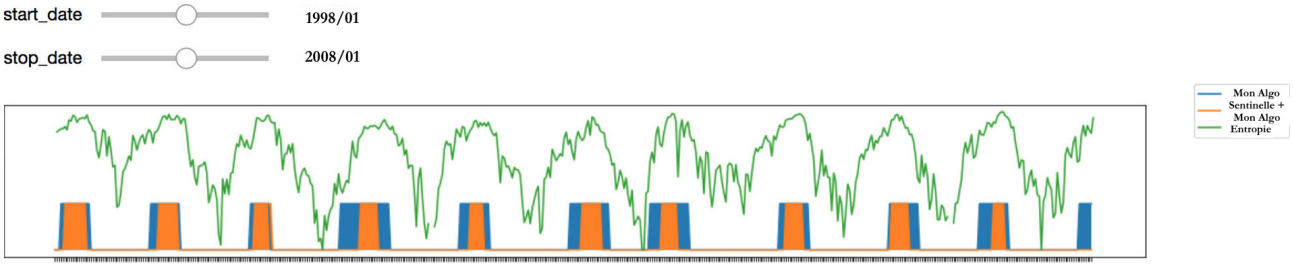

Figure 2 : Les zones bleues sont les périodes épidémiques reconnues (sur la période 1998-2008) par notre algorithme et les zones oranges sont les périodes épidémiques reconnues (sur la période 1998-2008) à la fois par l'algorithme de l'article "Sentinelles" et par notre algorithme. La courbe verte représente l'entropie concernant le taux de nouveaux cas de grippe entre les 22 régions françaises de 1998 à 2008.

La Figure 2 nous permet de comparer les périodes d'épidémie identifiées par les deux algorithmes. **On remarque, de manière frappante, que l'algorithme sur lequel j'ai travaillé détecte les épidémies avant que ne les détecte l'algorithme de l'article "Sentinelles".**

Notre algorithme est davantage sensible à la dispersion des foyers de contamination

Il est important de préciser qu'il est légitime de détecter les épidémies bien à l'avance que dans certains cas. Le premier cas, dans lequel il est utile de détecter très à l'avance une épidémie est lorsque l'épidémie est difficilement contrôlable car elle est dispersée sur différentes régions et il est donc difficile de définir des zones de contrôle. Or, ce qu'on remarque sur la Figure 2, c'est que l'entropie monte fortement au moment où notre algorithme est le seul à détecter l'épidémie (i.e. avant que l'algorithme de Sentinelles ne détecte l'épidémie, cf zones bleues sur la Figure 2) et on voit sur la Figure 2 que ce phénomène se produit pour plusieurs épidémie. Or, une forte augmentation de l'entropie concernant le taux de nouveaux cas entre les 22 régions françaises veut dire que les taux de nouveaux cas sont davantage hétérogènes d'une région à l'autre et donc qu'il y aura des régions avec des taux d'infections plus élevées que d'autres. Ainsi, cela veut dire qu'au moment où notre algorithme est le seul à détecter une épidémie (i.e. avant que l'algorithme de Sentinelles ne détecte l'épidémie), les foyers de cette épidémie se dispersent sur le territoire, ce qui rend l'épidémie difficilement contrôlable. Comme ce phénomène s'est produit pour toutes les épidémies qu'on peut voir sur la Figure 2 (i.e. pour toutes les épidémies entre 1998 et 2008), on peut donc dire qu'il est légitime que notre algorithme détecte ces épidémies bien à l'avance.

On démontre ce qu'on a vu avec l'entropie avec également un graphe de la France (codé aussi sur Python) dans lequel chaque noeud représente une des 22 régions françaises et chaque noeud a une couleur correspondant à un taux spécifique de nouveaux cas dans la région en question (par exemple le jaune représente le taux le plus élevé de nouveaux cas). Ainsi, en un clin d'oeil, nous pouvons voir sur plusieurs semaines les foyers de contamination sur le territoire français. Or, ce qu'on a vu (dans ces graphes de la France) c'est qu'au moment où notre algorithme est le seul à détecter une épidémie (cf zones bleues sur la Figure 2) les foyers de contamination sont dispersés sur le territoire et sont donc difficilement contrôlable.

La détection bien à l'avance de notre algorithme est légitime sur les "grosses épidémies"

On considère qu'une épidémie est "grosse" si la somme des taux de nouveaux cas durant toute l'épidémie est élevée ou si son pic (i.e. le taux de nouveaux cas le plus haut de l'épidémie) est élevé.

On a calculé sur chaque épidémie l'écart en nombre de semaines entre le début de l'épidémie détecté par notre algorithme et le début de l'épidémie détecté par l'algorithme Sentinelles, et ce pour chaque épidémie (on appelle cet écart de détection le lag).

On a donc pu voir que pour plusieurs "grosses épidémies" (dont le pic était élevé et/ou la somme des taux de nouveaux cas durant toute l'épidémie était élevée) le lag était important, ce qui veut dire que le fait que notre algorithme ait beaucoup d'avance sur ces épidémies est légitime car cela concerne des grosses épidémies. On sait qu'il est couteux, d'un point de vue économique, pour l'Etat de prendre des mesures sur une longue période (on peut le voir aujourd'hui avec le couvre-feu à 21h qui menace le secteur des bars, restaurants, cinéma…). Cependant, le coût sanitaire d'une grosse épidémie est tellement important qu'il dépasse l'enjeu économique (on peut aussi le voir avec la première vague de coronavirus qui, à son pic, faisait environ 1000 morts par jour en France).

English version of "Synthèse"*: Multiple change point detection for epidemiological signals*

This internship, at the laboratory of ENS Paris-Saclay (Centre Borelli), aims to predict as soon as possible the apparition of an epidemic. This need has been highlighted by the coronavirus crisis. The sooner we can predict an epidemic (for instance the second wave for coronavirus), the more efficient the actions of the States will be for coping with the epidemic. The goal of my project was to apply this multiple change point detection algorithms to detect epidemics.

First, how does a multiple change point detection algorithm work ?

As we can see in the Figure 1, a multiple change point detection algorithm has to recognise different periods in the signal and then divide this signal into segments, where each segment represents a period. A change point is the transition between two periods.

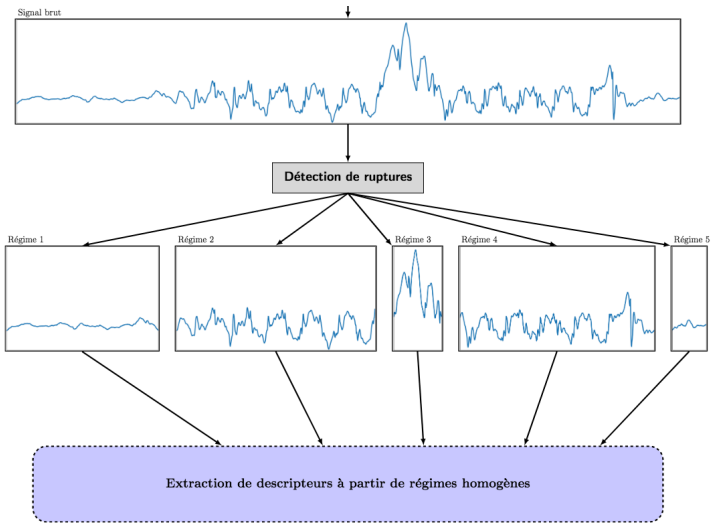

Figure 1 : *Multiple change point detection algorithm*

Here, we focus on a specific signal : an epidemiological signal. An epidemiological signal is a signal describing a measure of the spread of the epidemic evolving through time. For instance, a epidemiological signal can be the number of infections, the number of people in reanimation…

Thus, we will see the segments identified either as epidemic periods or as non-epidemic periods. If one segment is an epidemic period, then the following segment will be a non-epidemic period.

How is a multiple change point detection algorithm going to segment the signal (into epidemic periods and non-epidemic periods) ?

The algorithm tests many different segmentations (i.e. it tests ways to segment the signal). On each tested segmentation, the algorithm uses models (autoregressive models, median…) to approximate the signal on each segment. So, we have an approximation error of the signal on each segment. We

sum the approximation errors of all segments. Thus, the segmentation, denoted by Υ, of the signal which will have the lowest sum of approximation errors will be kept and seen as the best segmentation denoted by $\widehat{\Upsilon}$. The sum of approximation errors of all segments is denoted by $V(\Upsilon)$ (where Υ is a segmentation, so we can write that $\Upsilon = \{[t_0, t_1], [t_1, t_2], \ldots, [t_{K-1}, t_K]\}$ where $(t_i)_{i\in[0,K]}$ are the bounds of the segments i.e. the change points). So, we search for the best segmentation $\widehat{\Upsilon} = argmin_{\Upsilon} V(\Upsilon)$ which will enable us to detect precisely the epidemic periods (and also the non-epidemic periods).

In our case, we want to segment two specific epidemiological signals. The first signal to segment is an univariate signal : the estimated rate of new flu infections per 100,000 inhabitants in France (each week between 1984 and 2020). The second signal to segment is a multivariate signal : the estimated rate of new flu infections per 100,000 inhabitants for the 22 french regions (each week between 1984 and 2020).

So, **with the library "Ruptures" on Python**, we have tested several segmentations for the two signals described in the previous paragraph. However, for each segmentation, we have to make some choices : the first choice is about the model with which we want to approximate the signal on each segment and the second choice is about the way of calculating the approximation errors i.e. the cost function.

When it comes to the models to choose, we have first tested the autoregressive models. However, we can see that some of the epidemic periods identified ends when the rate of new flu infections is the highest, which is a big problem.

When it comes to the models to choose, we have then tested the median of the values taken by the signal on the segment, which could work as what defines an epidemic is namely the rate of new flu infections excessively high compared to the mean. We have chosen the median instead of the mean as the median is more robust for extreme values. However, we can see (by choosing the median of the values taken by the signal on the segment) that some entire small peaks have not been recognized as epidemic periods whereas we know that they were epidemic periods. Therefore, we have chosen (for the univariate and the multivariate signal) as models and cost function those described in the expression (1). This type of models has been chosen as it doesn't have the limits of the two other models previously tested (autoregressive models and median).

$$\forall i \in [0,K], c(y_{t_i..t_{i+1}}) = \sum_{t=t_i}^{t_{i+1}} || \Phi(y_t) - \overline{\Phi(y_t)_{t_i..t_{i+1}}} ||^2_{\mathscr{H}} \qquad (1)$$

where $y_{t_i..t_{i+1}}$ is the signal on the segment $[t_i, t_{i+1}]$, $c(.)$ is the cost function, $k(.,.)$ is positive semi-definite kernel where $k(y, y') = \langle \Phi(y), \Phi(y') \rangle_{\mathscr{H}}$ ($y, y' \in \mathrm{IR}^{22}$ if it is for the multivariate signal of the 22 french regions and $y, y' \in \mathrm{IR}$ if it is for the univariate signal), $\mathscr{H}$ being an Hilbert space and $\overline{\Phi(y_t)_{t_i..t_{i+1}}}$ being the empirical mean of $\{\Phi(y_t)\}_{t\in[t_i,t_{i+1}]}$.

We have selected the epidemic periods identified by the multiple change point detection algorithm on the multivariate signal (i.e. the estimated rate of new flu infections per 100,000 inhabitants for the 22 french regions) instead of those of the univariate signal (i.e. the estimated rate of new flu infections per 100,000 inhabitants in France) because the multivariate signal includes the dynamics of infections between the different regions and we know that the dynamics of infections are key parameters for detecting epidemics.

Is our epidemic detection algorithm better than some epidemic detection algorithms previously made ?

We have compared our epidemic detection algorithm with the one described in the research paper called “Détection des épidémies de grippe et de gastro-entérite par le réseau Sentinelles” published on the website “Sentinelles” (collaboration Inserm/UPMC).

$$\alpha_0 + \alpha_1 t + \sum_{k=1}^{n} [\alpha_{2,k} cos(\frac{2k\pi t}{52}) + \alpha_{3,k} sin(\frac{2k\pi t}{52})] \quad (2)$$

where t is the week and thus t takes its values between 1 and 52 (as there are 52 weeks in a year) and n is equal to 2.

We will now describe how the epidemic detection algorithm (described in this research paper) works. They approximate by a model (function described in the expression (2)) the entire univariate signal (here they don’t work on segments but on the entire signal). This method is called the “Serfling” periodic regression. After that, they approximate a new signal (available in the database) which is the upper bound of the estimated rate incidence 90% confidence interval. They approximate this signal by the same function described in the expression (2). When the former univariate signal (the estimated rate of new flu infections per 100,000 inhabitants in France) is higher than this last approximation, they say that they detect an epidemic.

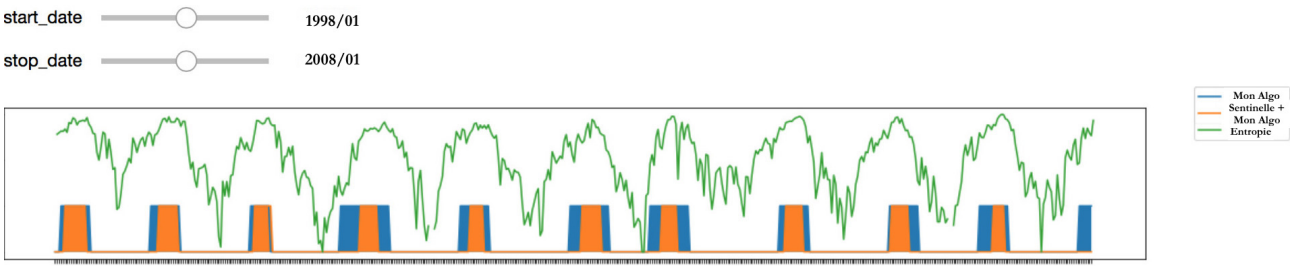

Figure 2 : The blue areas are epidemic periods identified by our multiple change point detection algorithm (between 1998 and 2008) and the orange areas are the epidemic periods identified by both algorithms (our algorithm and the “Sentinelles” algorithm) between 1998 and 2008. The green signal is the entropy for the rates of new flu infections between the 22 french regions between 1998 and 2008.

The Figure 2 enables us to compare the epidemic periods identified by the two algorithms. **We notice directly that our algorithm detects epidemics before the “Sentinelles” algorithm.**

Our algorithm is more sensitive to the dispersal of the contamination hotbeds

It is important to precise that it is legitimate to detect very early epidemics only in some cases. The first case in which it is useful to detect very early an epidemic is when it’s difficult to control the epidemic because the epidemic is already spread in different regions in the territory. We notice on the Figure 2 that the entropy increases strongly when our algorithm is the first to detect an epidemic (cf blue areas on the Figure 2) and we see on the Figure 2 that this phenomenon happens most of the time. An strong increase of the entropy (for the rates of new flu infections between the 22 french regions) means that these rates are more heterogeneous and that there will be regions where the rates of new infections will be higher than others. Thus, it means that when our algorithm is the first to detect an epidemic (cf blue areas on the Figure 2), some contamination hotbeds appear in

different regions so it's more difficult to control the epidemic. As this phenomenon happens most of the time (as we can see on the Figure 2), we can say that it's legitimate that our algorithm detects very early the epidemic.

We can also demonstrate the fact that our algorithm is more sensitive to the dispersal of the contamination hotbeds using a graph of France (coded on Python) in which each node is one of the 22 french regions and the color of each node shows a specific rate of new infections in the region (for instance, the color yellow represents for highest rate of new infections). Therefore, with this graph we can directly see the contamination hotbeds on the territory. So, we have seen on these graphs of France that when our algorithm is the first to detect an epidemic the contamination hotbeds appear in different places on the graph.

The very early epidemic detection by our algorithm is legitimate on "big epidemics"

An epidemic is considered as "big" only if its peak is high or/and the sum of the rates of new infections during the epidemic is high.

We have first calculated for each epidemic the gap between the start of the epidemic identified by our algorithm and the start of the epidemic identified by the "Sentinelles" algorithm. This gap of detection is called the lag.

Therefore, we have seen that for many "big epidemics" (whose peak was high and/or the total rate of new infections during the epidemic was high) the lag was big. So it means that in these specific cases the very early epidemic detection by our algorithm is legitimate because it's for "big epidemics". We know that for the States it has an economic cost to take measures very early (we can see it today with the curfew at 9 pm which threatens the sector of bars, restaurants, cinema…). However, the health cost of a "big epidemic" is more important than the economic cost (we could see it during the peak of the first wave of coronavirus where there were 1000 deaths per day in France).