Détection de rupture dans des données climatiques et dans des données de santé

Change detection in some climatic data and in some health data

Joseph Ngatchou-Wandji[a*] et Marwa Ltaifa[b]

[a*] EHESP Sorbonne Paris Cité et Institut Élie Cartan de Lorraine, 54506 Vandoeuvre-lès-Nancy cedex, France.
E-mail : joseph.ngatchou-wandji@univ-lorraine.fr
[b] Institut Élie Cartan de Lorraine, 54506 Vandoeuvre-lès-Nancy cedex, France et LAMMDA-ESST Hammam-Sousse, University of Sousse, 4011 Hammam-Sousse, Tunisia.
E-mail : ltaifa.marwa@gmail.com

RÉSUMÉ. Nous appliquons de nouvelles méthodes de détection de ruptures à une série de données sur les anomalies climatiques annuelles, et à trois séries de données journalières relatives à la première vague du Covid-19 en France en 2020. Dans chacune de ces séries, plusieurs ruptures sont détectées et leurs localisations sont estimées.
ABSTRACT. We apply some new change point detection methods to a series of data relative to annual climatic anomalies, and to three series of data relative to the first wave of the Covid-19 in France in 2020. For each of these series, many change-points are detected and their locations are estimated.
MOTS-CLÉS. Séries chronologiques, Test du rapport de vraisemblance, Ruptures, Données climatiques, Données du Covid-19 en France.
KEYWORDS. Time series, Likelihood ratio test, Breaks, Climatic data, Data from Covid-19 in France.

Introduction

Les changements *discrets* dans un phénomène temporel peuvent se manifester par la présence de faibles *ruptures* dans les données recueillies. Afin de comprendre ce phénomène, le contrôler ou le prévoir, il peut s'avérer très important de détecter au préalable ces ruptures avant toute inférence statistique.

Une rupture dans une série de données chronologiques peut être vue comme un changement brusque dans la dynamique de ces données, ou plus généralement comme un changement dans leur distribution. Les ruptures de faibles amplitude peuvent être confondues à des effets de bruits aléatoires, et, de ce fait, peuvent échapper aux méthodes de détection classiques. Ainsi, un dysfonctionnement dans un système risque de n'être détecté qu'à un stade critique. Ngatchou-Wandji et Ltaifa (2021) proposent des méthodes pour détecter ce type de ruptures et des stratégies pour estimer leurs localisations.

Depuis Page (1955), la littérature sur les ruptures s'est beaucoup enrichie, évoluant du contexte des données indépendantes et identiquement distribuées (iid) à celui des données dépendantes comme par exemple, les séries chronologiques. Les monographies de Basseville et Nikiforov (1993 et Csorgö et Horváth (1997) présentent les notions de base de la théorie des ruptures. Aue et Horvath (2013) et Truong et el. (2020) font la revue des méthodes récentes et des techniques d'étude des ruptures en séries chronologiques. Cependant, la plupart de ces méthodes de détection de ruptures ainsi que les stratégies d'estimation de leurs localisations sont construites pour détecter les ruptures assez significatives. Ils peuvent ainsi s'avérer inaptes à la détection des ruptures faibles, celles-là même susceptibles d'être annonciatrices du dysfonctionnement du phénomène d'intérêt étudié.

Dans cet article, nous appliquons les méthodes développées dans Ngatchou-Wandji et Ltaifa (2021) à la détection des ruptures dans la moyenne des données sur les anomalies de température issues de la

base de données de la *National Oceanic and Atmospheric Administration* (NOAA), et aux moyennes de trois séries de données relatives à la première vague du Covid-19 ayant sévi en France au cours du deuxième trimestre de l'année 2020. Ces dernières données ont été obtenues à partir du site web de la Worldometers.

Pour les donnés climatiques, il s'agit pour nous de nous vérifier que les méthodes que nous appliquons détectent bien les différentes anomalies climatiques, et qu'elles fournissent des estimations raisonnables de leurs périodes approximatives de survenue. L'analyse des données sur le Covid-19 est dans l'air du temps. Tout comme pour les données climatiques, l'un des buts est de leur ajuster un modèle statistique, ce qui peut à la fois permettre de comprendre l'évolution de cette maladie et évaluer son impact sur la santé, la finance et l'économie française. Pour ces données du Covid-19, après avoir vérifié le bien-fondé des méthodes de Ngatchou-Wandji et Ltaifa (2021), nous voulons déceler les moments des débuts des changements importants survenus dans l'évolution de cette pandémie au cours de la première vague en 2020 en France.

1. Les méthodes

Les méthodes étuditées dans Ngatchou-Wandji et Ltaifa (2021) reposent essentiellement sur la puissance théorique d'un test du rapport de vraisemblance (voir par exemple Biesmans (2016)) pour discriminer entre modèles conditionnellement hétéroscédastiques non-linéaires. Plus précisément, soit $k, n \in \mathbb{N}$ et k un entier très petit devant n. Supposons les observations $X_1, \ldots, X_n$ issues du modèle

$$X_t = T_\rho(Z_{t-1}) + \sum_{j=1}^{k+1} \gamma_j \mathbb{1}\left(t \in [t_k, t_{k+1})\right) + V_\theta(Z_{t-1})\varepsilon_t, \quad t \in \mathbb{Z}, \qquad [1]$$

où T_ρ et V_θ sont des fonctions paramétriques à valeurs réelles, définies sur $\mathbb{R}^p$ avec $\inf_{x\in\mathbb{R}^p} V_\theta(x) > 0$, $\mathbb{1}\left(t \in A\right)$ est la fonction indicatrice qui vaut 1 si $t \in A$ et 0 sinon, $\gamma_j \in \mathbb{R}$, $j = 1, \ldots, k+1$, $1 = t_0 < t_1 < \ldots < t_k < t_{k+1} = n$ sont de potentielles localisations des ruptures, $(X_t)_{t\in\mathbb{Z}}$ est ergodique et stationnaire par morceaux sur les $[t_j, t_{j+1})$, $(\varepsilon_t)_{t\in\mathbb{Z}}$ est un bruit blanc de densité f, et pour tout $t \in \mathbb{Z}$, $Z_t = (X_t, \ldots, X_{t-p+1})^\top$ et ε_t est indépendent des Z_s, $s < t$.

L'équation stochastique (1) décrit la grande classe des modèles dits CHARN d'ordre p dont la moyenne dépend du temps. L'acronyme CHARN introduit dans Härdle et al. (1998) signifie en anglais Conditional Heteroskedastic AutoRegressive Nonlinear. Cette classe de modèles non-linéaires contient plusieurs modèles usuels tels que les modèles AR(p), ARCH(p), TARCH(p), EXPAR(p) dont les propriétés statistiques (estimation, tests) et probabilistes (stationnarité, ergodicité) sont largement étudiées dans la littérature (voir Guégan (1994), Tong (1993)). Un tel modèle d'ordre ∞ est considéré dans Bardet et Kengne (2014).

On note $\gamma = (\gamma_1, \ldots, \gamma_k, \gamma_{k+1})^\top$. Dans Ngatchou-Wandji et Ltaifa (2021) un test du rapport de vraisemblance est construit pour tester

$$H_0 : \gamma = \gamma_0 \text{ contre } H_\beta^{(n)} : \gamma = \gamma_0 + \frac{\beta}{\sqrt{n}} = \gamma_n, \quad n > 1,$$

pour un $\gamma_0 \in \mathbb{R}^{k+1}$ et $\beta \in \mathbb{R}^{k+1}$ dependant des t_j.

Les deux hypothèses ci-dessus se rapprochent lorsque la taille de l'échantillon grandit. On montre qu'elles sont contigues au sens de Le Cam (voir Droesebeke et Fine (1996)). Cette propriété permet l'étude de la puissance du test contruit, et l'obtention d'une expression explicite de sa puissance. En effet, soit $\psi_0 = (\rho_0^\top, \theta_0^\top)^\top$ le vrai paramètre de nuisance, sous des hypothèses techniques que nous ne rappelons pas, on montre que pour tout $\beta \in \mathbb{R}^{k+1}$, le test du rapport de vraisemblance construit est asymptotiquement optimal, de puissance asymptotique $\mathcal{P}_{k,t^k} = 1 - \Phi\left(u_\alpha - \varpi(\psi_0, \gamma_0, \beta)\right)$, où $\alpha \in (0,1)$ est le niveau de significativité du test, u_α le quantile d'ordre $(1-\alpha)$ de le loi normale centrée réduite de fonction de répartition Φ, et ϖ est une fonction à valeurs réelles définie sur $\mathbb{R}^{p\times p\times(k+1)\times(k+1)}$ dont l'expression explicite est donnée dans Ngatchou-Wandji et Ltaifa (2021). Il est à rappeler qu'un test optimal est un test dont la puissance est supérieure à celle de tout autre test de même niveau de significativité.

Le problème de test ci-dessus est très général. Il peut être appliqué à l'étude des ruptures pour des γ_0 et β particuliers, à savoir, aux γ_0 pour lesquels la première composante est la moyenne théorique μ_1 des données d'indices dans $[t_0, t_1)$, et pour des β dont la première composante est nulle. Une décision sur ce problème de test peut être prise sur la base d'un estimateur $\widehat{\mathcal{P}}_{k,t^k}$ de la puissance. Un tel estimateur peut être obtenu en substituant aux paramètres, leurs estimateurs dans l'expression de $\mathcal{P}_{k,t^k}$. On peut considérer comme estimateurs de $\psi_0 = (\rho_0^\top, \theta_0^\top)^\top$ et γ_0 leurs estimateurs du maximum de vraisemblance respectifs $\widehat{\psi}_n = (\rho_n^\top, \theta_n^\top)^\top$ et $\widehat{\gamma}_{0,n}$. Le vecteur $\beta = (0, \beta_2, \ldots, \beta_{k+1})$ peut être estimé par $\widehat{\beta}_n = (0, \widehat{\beta}_2, \ldots, \widehat{\beta}_{k+1})$ où pour $j = 2, \ldots k+1$, $\widehat{\beta}_j = \sqrt{n}(\overline{X}_{j,n} - \widehat{\widetilde{\gamma}}_{0j,n})$, avec $\widehat{\widetilde{\gamma}}_{0j,n}$ un estimateur de la $j$eme composante de γ_0, et $\overline{X}_{j,n}$ est la moyenne empirique des observations dont les indices de temps sont dans $[\tau_j, \tau_{j+1})$. Un possible $\widehat{\widetilde{\gamma}}_{0j,n}$ est la moyenne empirique des observations dont les indices sont dans $[t_{j-1}, t_j)$. Avec ces quantités, on peut donc prendre $\widehat{\mathcal{P}}_{k,t^k} = 1 - \Phi\left(u_\alpha - \varpi(\widehat{\psi}_n, \widehat{\gamma}_{0,n}, \widehat{\beta}_n)\right)$ comme estimateur de $\mathcal{P}_{k,t^k}$.

La première étape des méthodes décrites dans Ngatchou-Wandji et Ltaifa (2021) consiste à déterminer, à partir du chronogramme, les m premières données $X_1, X_2, \ldots, X_m$ qui sont à peu près stationnaires, le nombre maximum K de ruptures potentielles et la distance minimale $h << n$ entre elles.
Pour détecter la présence de rupture, prendre $k = 1$ et appliquer le test à tous les t_1 tels que $m \le t_1 \le n - h$. On convient que $\mathcal{P}_{0,t^0} = \alpha$. Soit $\zeta \in (0, .1)$.

i- Si $|\widehat{\mathcal{P}}_{1,t^1} - \mathcal{P}_{0,t^0}| \le \zeta$ pour tous ces t_1, alors aucune rupture n'est détectée dans la série.

ii- Si $|\widehat{\mathcal{P}}_{1,t^1} - \mathcal{P}_{0,t^0}| > \zeta$ pour un t_1, alors, il existe au moins une rupture dans la série.

Pour estimer les localisations des ruptures, pour $k = 1, \ldots, K$, on suppose que $m < \tau_1^0 < \ldots < \tau_k^0 \le n - h$, $\tau_j^0 - \tau_{j-1}^0 \ge h$, $j = 2, \ldots, k$, sont de potentielles localisations des ruptures obtenues du chronogramme. Soit C_j un ensemble arbitraire d'indices autour des τ_j^0, $j = 1, \ldots, k$. On considère $S_k = C_1 \times C_2 \times \ldots \times C_k$. Pour tout k-uplet $\tau^k = (\tau_1, \ldots, \tau_k) \in S_k$, on applique le problème de test ci-dessus avec $t_j = \tau_j$, $j = 0, \ldots, k+1$ et on calcule $\widehat{\mathcal{P}}_{k,t^k}$.

• À l'étape $k+1$:

i- Si $|\widehat{\mathcal{P}}_{k+1,t^{k+1}} - \widehat{\mathcal{P}}_{k,t^k}| \le \zeta$ et $|\widehat{\mathcal{P}}_{k,t^k} - \widehat{\mathcal{P}}_{k-1,t^{k-1}}| > \zeta$, alors un estimateur $(\widehat{k}, \widehat{t^k})$ du couple formé par le nombre des ruptures et le vecteur de leurs localisations peut s'obtenir de la manière suivante :

$$(\widehat{k}, \widehat{t^k}) = \arg\max_{t^k \in S_k} \widehat{\mathcal{P}}_{k,t^k}.$$

ii- Si $|\widehat{\mathcal{P}}_{k+1,t^{k+1}} - \widehat{\mathcal{P}}_{k,t^k}| > \zeta$, répéter l'item i avec k remplacé par $k+1$.

2. Application aux données réelles

Nous appliquons ici, les méthodes décrites dans le paragraphe précédent à quelques données réelles. Le premier jeu de données est relatif aux anomalies climatiques. Comme dit plus haut, ces données sont issues de la base de données de la NOAA. Les autres jeux de données sont relatifs à la première vague du COVID-19 en France. Ces données concernent le taux journalier de décès, le nombre journalier de décès et le nombre journalier de cas. Elles proviennent des bases de données de la Worldometers.

2.1. *Modélisation*

Le chronogramme de chacune des séries brutes (Y_t) considerées dans cette étude semble présenter une tendance et aucune saisonnalité (voir Figures 1 (a), 2 (a), 3 (a) et 4 (a)). Les résultats établis dans Ngatchou-Wandji et Ltaifa (2021) ne peuvent s'appliquer directement à ces séries. Nous les décomposons alors en la somme de deux composantes comme suit :

$$Y_t = Z_t + X_t, \qquad [1]$$

où (X_t) est une série stationnaire par morceaux de moyenne (μ_t) et de variance (σ_t), et (Z_t) représente la tendance inconnue supposée continue, et estimée ici par la moyenne-mobile d'ordre 5 donnée par

$$\widehat{Z}_t = \frac{1}{5}\sum_{j=-2}^{2} Y_{t+j}. \qquad [2]$$

Cette estimation est produite dans cet article en utilisant la routine "ma" du logiciel R.
Étant donné que Z_t est supposée continue, on peut espérer que toutes les éventuelles ruptures dans (Y_t) se trouvent dans (X_t) qu'on peut estimer par $(Y_t - \widehat{Z}_t)$. Les tests de Box-Ljung et Box-Pierce (voir Brockwell et Davis (2002)) appliqués à chacune de ces séries résiduelles ont rejeté l'hypothèse qu'elles sont iid. De même, le test de Shapiro-Wilk appliqué aux séries résiduelles et à divers sous-ensembles de ces séries, ont conduit à l'acceptation de l'hypothèse qu'elles sont gaussiennes. Toutes ces investigations nous ont suggéré l'hypothèse d'hétéroscédasticité de toutes les séries résiduelles étudiées dans le présent article, et le postulat d'un modèle de la forme suivante :

$$X_t = \mu + \frac{\beta_j}{\sqrt{n}} + \sigma_j \varepsilon_t,\ t \in [t_{j-1}, t_j),\ j = 1, \ldots, k+1, \qquad [3]$$

où les $t_1, t_2, \ldots, t_k$ sont k potentielles localisations de ruptures dans les différentes séries $Y_1 - \widehat{Z}_1, Y_2 - \widehat{Z}_2, \ldots, Y_n - \widehat{Z}_n$, $\mu + (\beta_j/\sqrt{n})$ et σ_j sont respectivement la moyenne et la variance de X_t sur l'intervalle $[t_{j-1}, t_j)$, $\beta_1 = 0$, pour tout $j = 2, \ldots, k+1$, $\beta_j \in \mathbb{R}$, et (ε_t) est un bruit blanc gaussien centré réduit. Ce modèle (3) est une sous-classe de (1) pour $T(x) = 0$, $V(x) = \sigma_j$ sur chaque intervalle $[t_{j-1}, t_j)$, et pour le problème de test, $\gamma_0 = (\mu, \mu, \ldots, \mu)^\top$, $\gamma_n = (\mu, \mu + (\beta_2/\sqrt{n}), \mu + (\beta_3/\sqrt{n}), \ldots, \mu + (\beta_{k+1}/\sqrt{n}))^\top$ et $\beta = (0, \beta_2, \beta_3, \ldots, \beta_{k+1})^\top$.

La méthode de détection de rupture décrite dans Ngatchou-Wandji et Ltaifa (2021), rappelée plus haut, appliquée à chacune des séries résiduelles, pour $\zeta = 0.01$ a conduit à la conclusion que toutes contiennent au moins un point de rupture. Nous avons ensuite utilisé la deuxième stratégie d'estimation des localisations des ruptures décrite dans Ngatchou-Wandji et Ltaifa (2021), elle aussi, rappelée plus haut. Dans les paragraphes suivants, nous détaillons et commentons les résultats obtenus pour chacune d'elles.

2.2. *Données annuelles sur les anomalies climatiques*

Nous rappelons tout d'abord qu'une anomalie climatique est un épisode pouvant durer de quelques mois à plusieurs siècles pendant lequel des variables climatiques et météorologiques sont éloignées d'un état climatique moyen. Dans le contexte de cet article, c'est l'écart entre la température mondiale moyenne annuelle et la température moyenne calculée sur la période de 1901 à 2000. Nous avons considéré les températures relevées par les stations au sol et par celles dans les océans.

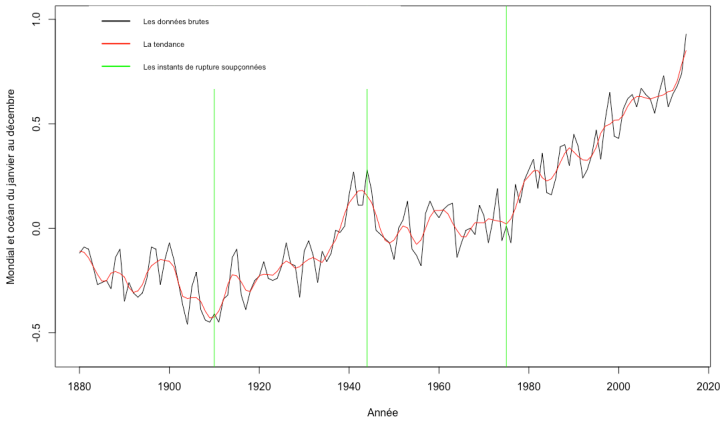
La Figure 1 (a) présente le chronogramme des données brutes. Il semble faire apparaître 3 points potentiels de ruptures correspondants aux années 1910, 1944 et 1975. Il s'agit pour nous de chercher des points de ruptures dans les voisinages de ces années candidates. Nous commençons par estimer la tendance par la méthode de moyenne mobile. Cette estimation est représentée en rouge dans la Figure 1 (a). La série résiduelle est représentée dans la Figure 1 (b). De vue, elle ressemble à un bruit blanc. Les méthodes classiques auraient du mal à détecter des ruptures dans sa moyenne. Nous lui ajustons un modèle de la forme (3) avec $k = 3$, $t_0 = 1880$, $t_1 = 1910$, $t_2 = 1944$, $t_3 = 1975$ et et $t_4 = 2015$. Nous obtenons par la méthode de Ngatchou-Wandji et Ltaifa (2021) les estimations suivantes : $\widehat{t}_1 = 1915$, $\widehat{t}_2 = 1944$ et $\widehat{t}_3 = 1975$ représentées dans la Figure 1 (c). Ce sont les points vus comme potentiels points de rupture à l'exception de $\widehat{t}_1$ qui est cependant proche de $t_1 = 1910$. Ces points semblent coller à la réalité. En effet, dans le premier intervalle la tendance globale est à la baisse. Elle est à la hausse dans le second intervalle, puis constante dans le suivant et de nouveau à la hausse dans le dernier.

2.3. *Données journalières sur le COVID-19 en France*

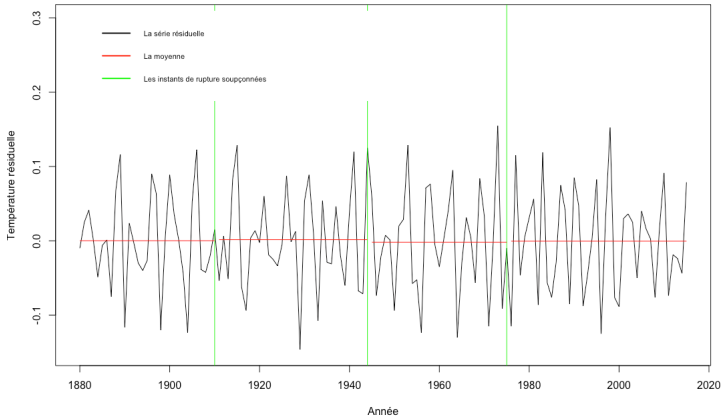
Dans cette sous-section, nous considérons trois séries de données journalières sur la première vague du COVID-19 en France. Il s'agit des séries du taux de mortalité quotidien, du nombre de décès quotidiens et de celle du nombre de cas quotidiens. Nous désignons par $1, 2, \ldots, n$, les numéros d'ordre des données : 1 correspond au premier jour, 2 au deuxième jour, $\ldots$, n au dernier jour. Dans la première, la deuxième et la troisième séries, 1 correspond respectivement aux dates des 15/02/2020, 27/02/2020 et 25/02/2020 et n correspond à la date du10/07/2020.

Avant de passer aux applications, nous donnons tout d'abord quelques éléments de la chronologie de cette première vague du Covid-19 qui peuvent aider à l'interprétation des ruptures dans les séries étudiées.

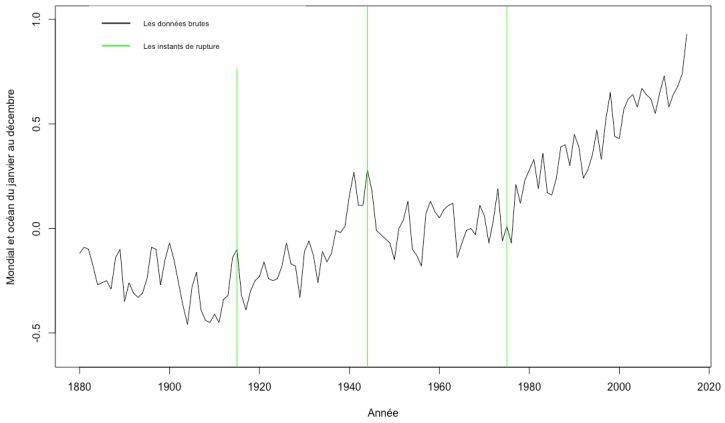
— Le 14/02/2020 : Enregistrement du premier décès en France, celui d'un touriste chinois de 80 ans.
— Le 25/02/2020 : Enregistrement du premier décès français et cinq nouveaux cas.
— Le 15/03/2020 : Fermeture des lieux publics.
— Le 17/03/2020 : Mise en place du confinement.
— Le 11/05/2020 : Début de la phase 1 du déconfinement.
— Le 02/06/2020 : Début de la phase 2 du déconfinement.
— Le 22/06/2020 : Début de la phase 3 du déconfinement.



(a) Série brute



(b) Série résiduelle



(c) Années de rupture estimées

Figure 1. – Anomalies climatiques annuelles mondiales de 1880 à 2015

2.3.1. Taux de mortalité

Nous commençons par la série du taux de mortalité sur la période du 15/02/2020 au 10/07/2020. La Figure 2 (a) présente le chronogramme des données brutes. Ce graphique semble montrer 5 potentiels points de rupture correspondants aux jours $t_1 = 11, t_2 = 31, t_3 = 47, t_4 = 58$ et $t_5 = 110$, qui correspondent respectivement aux dates des 25/02/2020, 16/03/2020, 01/04/2020, 12/04/2020 et 03/06/2020. Nous pouvons remarquer très facilement que ces potentiels candidats apparaissent sur le chronogramme de la série résiduelle représentée dans la Figure 2 (b). Nous ajustons à cette série un modèle de la forme (3) pour $k = 5$, $t_0 = 1$, $t_1 = 11, t_2 = 31, t_3 = 47, t_4 = 58$ et $t_5 = 110$ et $t_6 = n$. Nous obtenons par la méthode de Ngatchou-Wandji et Ltaifa (2021) les estimations suivantes : $\widehat{t_1} = 11$, $\widehat{t_2} = 32$, $\widehat{t_3} = 46$, $\widehat{t_4} = 60$, $\widehat{t_5} = 110$. Elles sont représentées dans la Figure 2 (c). Pour interpréter ces estimations, nous dirons qu'autour du 25/02/2020, le taux de mortalité qui est à la baisse, va brusquement augmenter pour atteindre son pic le 15/03/2020 et chuter drastiquement le lendemain puis rester presque constant jusqu'au 30/03/2020, avant de remonter lentement jusqu'au 11/04/2020 (ce qui justifie peut-être la décision de prolonger le confinement), puis redescendre lentement jusqu'au 03/06/2020 (lendemain de la phase 2 du déconfinement) avant de se stabiliser jusqu'à la phase 3 du déconfinement et même longtemps après.

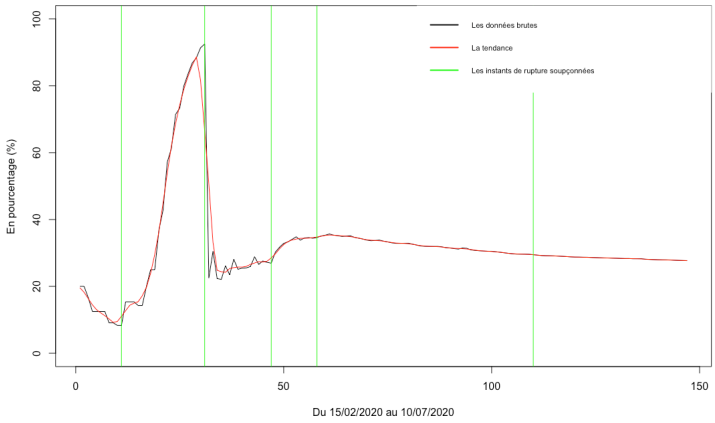
2.3.2. Nombre de décès quotidiens

Nous cherchons dans cette partie les points de rupture dans la série des nombres de décès quotidiens du COVID-19 en France sur la période du 27/02/2020 au 10/07/2020. La Figure 3 (a) exhibe le chronogramme des données brutes tracé en vert. Les potentiels points de rupture sont $t_1 = 31, t_2 = 38, t_3 = 55, t_4 = 86$ et $t_5 = 116$ qui correspondent respectivement aux dates suivantes : 28/03/2020, 04/04/2020, 21/04/2020, 22/05/2020 et 21/06/2020. Comme dans les cas précédents, nous ajustons à la série résiduelle représentée sur la Figure 3 (b), un modèle du type (3) avec les t_j ci-dessus. Nous obtenons les estimations suivantes : $\widehat{t_1} = 35, \widehat{t_2} = 40, \widehat{t_3} = 56, \widehat{t_4} = 88$ et $\widehat{t_5} = 117$. Ces dates, représentées dans la Figure 3 (c), correspondent respectivement aux dates 02/03/2020, 06/04/2020, 22/04/2020, 21/05/2020 et 25/06/2020, différentes mais assez proches des dates potentielles de rupture.

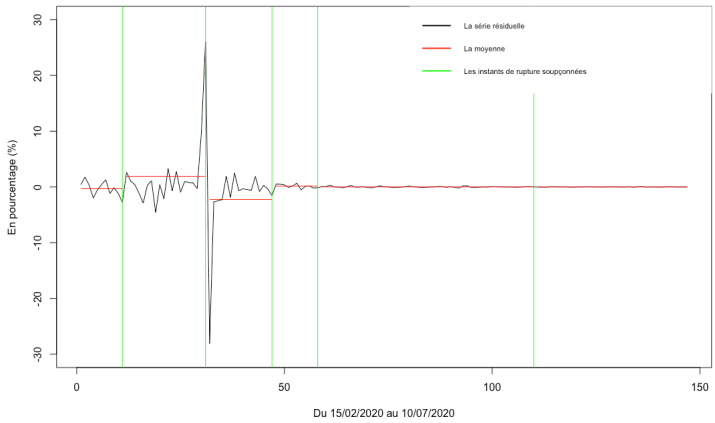
Une interprétation possible des dates estimées est la suivante : Au tour du 02/03/2020, le nombre de décès augmente drastiquement, atteint son pic et redescend autour du 06/04/2020, puis oscille significativement jusqu'aux environs du 22/04/2020, et un peu moins significativement entre les premiȁere et deuxième phases du déconfinement, jusqu'au 21/05/2020, date à partir laquelle il se stabilise avant de se réduire considérablement à partir du 25/06/2020, peu après la troisième phase du déconfinement.

2.3.3. Nombre de nouveaux cas quotidiens

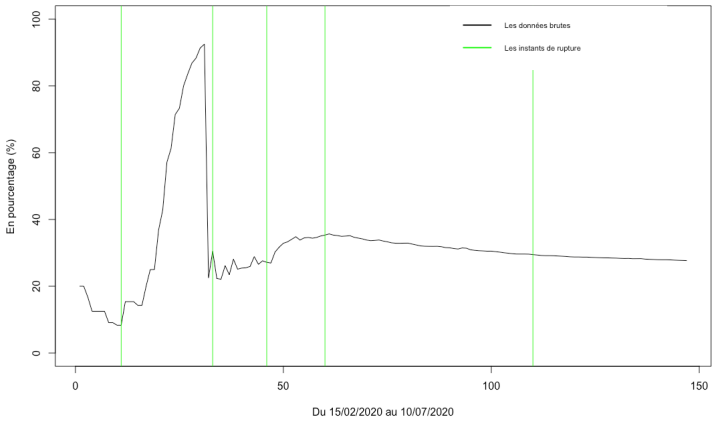
Nous considérons enfin les nouveaux cas quotidiens de COVID-19 en France sur la période du 25/02/2020 au 10/07/2020. Les données présentées sur la Figure 4 (a) montrent 6 potentiels points de ruptures re-



(a) Série brute

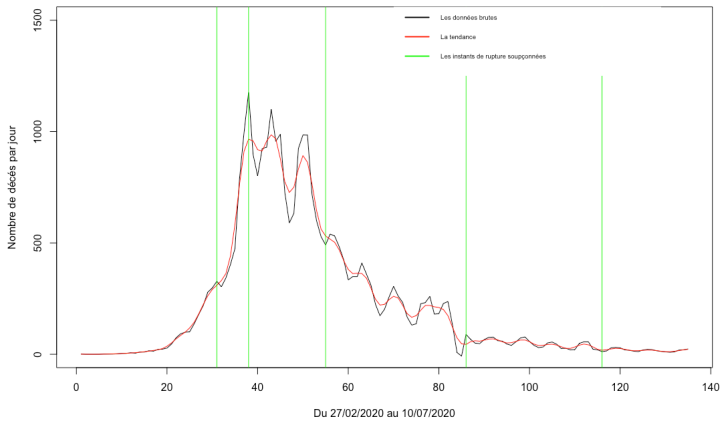


(b) Série résiduelle

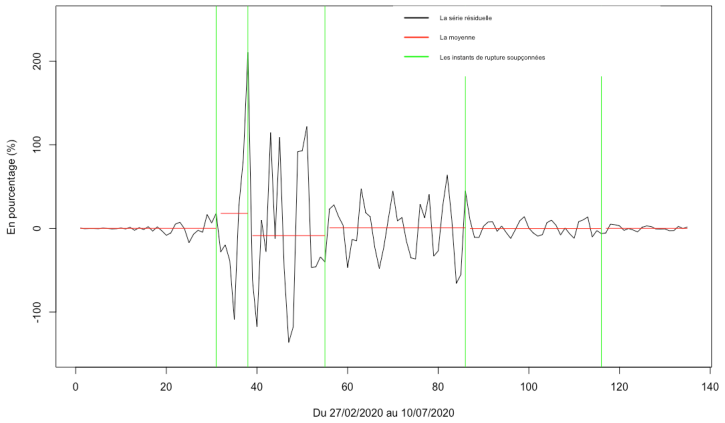


(c) Dates de rupture estimées

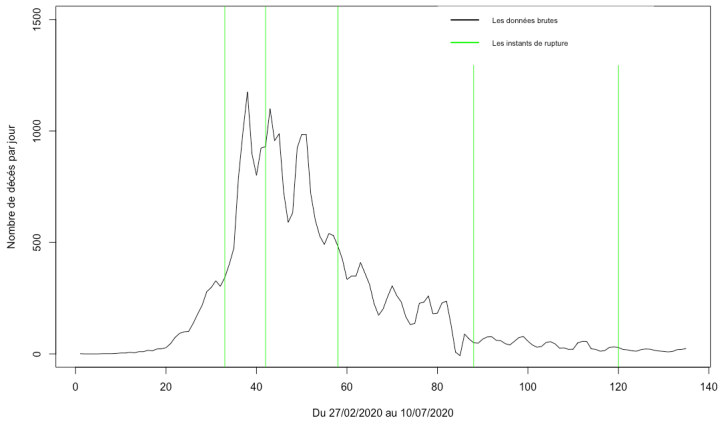
Figure 2. – Taux de mortalité du COVID-19 en France du 15/02/2020 au 10/07/2020



(a) Série brute



(b) Série résiduelle



(c) Dates de rupture estimées

Figure 3. – Nombre de décès quotidiens de COVID-19 en France du 27/02/2020 au 10/07/2020

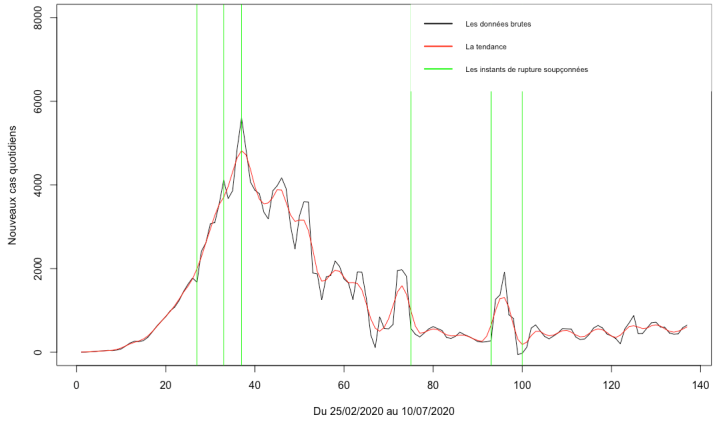
présentés en vert. Ce sont les instants $t_1 = 27, t_2 = 33, t_3 = 37, t_4 = 75, t_5 = 93$ et $t_6 = 100$ qui correspondent respectivement aux dates des 23/03/2020, 28/03/2020, 01/04/2020, 09/05/2020, 27/05/2020 et 03/06/2020. Nous ajustons aux données résiduelles un modèle du type (3) avec $k = 6$, $\widehat{t}_1 = 26, \widehat{t}_2 = 34, \widehat{t}_3 = 40, \widehat{t}_4 = 80, \widehat{t}_5 = 93$ et $\widehat{t}_6 = 110$ à compter à partir du 25/02/2020. Ces dates, représentées dans la Figure 4 (c), correspondent respectivement aux 22/03/2020, 29/03/2020, 04/04/2020, 4/05/2020, 27/05/2020 et 13/06/2020. Les dates estimées sont différentes des potentiels points de rupture mais elles en sont proches.

On peut comprendre ces estimations de la manière suivante : Autour du 22/03/2020, le nombre de nouveaux cas augmente rapidement, fléchi, puis reprend de la vitesse avant de connaître un nouveau fléchissement vers le 29/03/2020. Après avoir atteint son pic, il décroît de manière importante jusqu'au 04/04/2020, puis oscille avec une amplitude relativement grande jusqu'au 4/05/2020 et avec une petite amplitude jusqu'au 27/05/2020, quinze jours après la première phase du déconfinement, et de manière assez régulière à partir de 13/06/2020 entre la deuxième et la troisième phase du déconfinement.

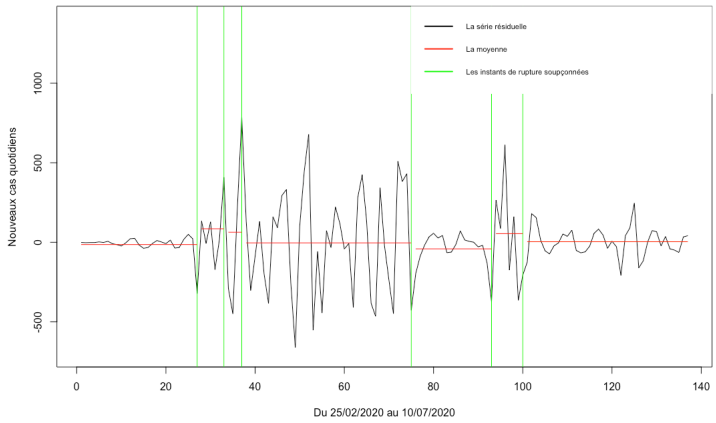
Remarque - Les séries résiduelles des trois séries de données sur le covid se caractérisent par le fait que pour les potentiels points de rupture considérés, les segmentations qui en résultent ne font pas toujours apparaître clairement les écarts entre les moyennes des observations sur certains intervalles consécutifs. Ceci peut faire penser qu'il n'y a pas de rupture où il y en a une qui est faible. Les méthodes de détection classiques peuvent s'avérer inefficaces dans la détection de telles ruptures.

3. Conclusion

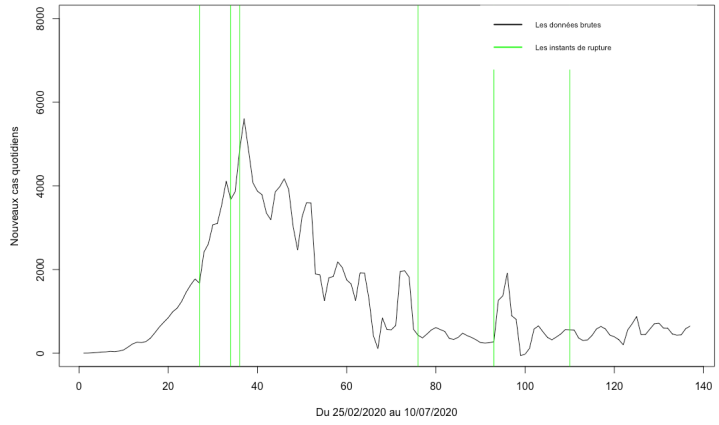
Le travail effectué dans cet article montre l'intérêt des méthodes dédiées à la détection des ruptures faibles, comme par exemple, celles développées dans Ngatchou-Wandji et Ltaifa (2021). En effet, certaines ruptures détectées ont une amplitude si petite qu'elles n'auraient pas été détectées par les méthodes classiques. Il est évident que ceci peut sérieusement compromettre la compréhension du phénomème étudié, son contrôle ou sa prévision. Cela peut aussi impacter l'étude d'autres phénomènes liés.



(a) Série brute



(b) Série résiduelle



(c) Dates de rupture estimées

Figure 4. – Nouveaux cas quotidiens de COVID-19 en France du 25/02/2020 au 10/07/2020

Bibliographie

AUE A., HORVATH L., *Structural breaks in times series.* J. Time Ser. Anal., 34 :1–16, 2013.

BARDET J.-M. KENGNE W., *Monitoring procedure for parameter change in causal time series.* J. Mult. Anal., 125 :204–221, 2014.

BASSEVILLE M. , NIKIFOROV I., *Detection of abrupt changes : Theory and Applications.* Prentice Hall, Inc, 1993.

BBIESMANS, F., *Probabilités et statistique inférentielle.* Ellipses, Coll. Référence Sciences, Paris, 600 pages, 2016.

BROCKWELL P., DAVIS R., *Introduction to time series and forecasting.* Springer, 2002.

CSORGÖ M., HORVÁTH L., *Limit Theorems in Change-Point Analysis.* John Wiley & Sons, New York ; Chichester, 1997.

DREOSEBEKE J.-J., FINE J., *Inférence non paramétrique. Les statistiques de rangs.* Éditions de l'Université de Bruxelles, 1996.

GUÉGAN D., *Séries chronologiques non-linéaires à temps discret.* Economica, 1994.

HÄRDLE W., TSYBAKOV A., YANG L. *Nonparametric vector autoregression.* J. Statist. Plann. Inference, 68 :221–245, 1998.

NGATCHOU-WANDJI J., LTAIFA M., *On detecting weak changes in the mean of CHARN models.* ArXiv : 2101.08597. 2021.

PAGE E., *A test for a change in a parameter occurring at an unknown point.* Biometrika, 42(3/4) :523–527, 1955.

TONG H., *Non-linear Time Series : A Dynamical System Approach.* Oxford Science Publications. 1993.

TRUONG C., OUDRE L., VAYATIS N., *Selective review of offline change point detection methods.* ArXiv :1801.00718v3. 2020.