Compared performance of Covid19 reproduction number estimators based on realistic synthetic data

Juliana Du, Barbara Pascal, Patrice Abry

Performances comparées d'estimateurs du coefficient de reproduction de la Covid19 à l'aide de données synthétiques réalistes

Juliana DU[1] Barbara PASCAL[2] Patrice ABRY[1]

[1]Laboratoire de Physique de l'ENS de Lyon (UMR CNRS 5672), 46 allée d'Italie F-69364 Lyon, France

[2]Laboratoire des Sciences du Numérique de Nantes (UMR 6004), 2 Chemin de la Houssinière F-44000 Nantes Cedex 3, France

Résumé – La pandémie de Covid19 a vu une explosion du nombre de travaux en traitement du signal pour l'épidémiologie. Plusieurs méthodes d'analyse de données épidémiologiques ont notamment été développées pour estimer le *coefficient de reproduction*, quantifiant la dynamique de propagation du virus, et ont relevé plusieurs défis soulevés par le traitement quotidien des données relatives à la Covid19 : les données sont de qualité limitée, une assise épidémiologique solide doit sous-tendre l'analyse, et les algorithmes qui en découlent doivent être suffisamment efficaces pour effectuer un suivi quotidien. S'appuyant sur le modèle épidémiologique de Cori, des stratégies variatonnelles ont été proposées, fournissant des estimateurs du coefficient de reproduction à la fois précis et calculables rapidement. Néanmoins, à l'heure où nous écrivons, aucune analyse de performance systématique n'a été conduite, la raison principale étant le manque de données annotées. Une évaluation quantitative est pourtant cruciale, tout d'abord, pour valider la capacité des estimateurs du coefficient de reproduction à gérer des données sévèrement corrompues, et par suite pour envisager d'étendre leur utilisation à d'autres épidémies. Dans ce but, une procédure de génération de données synthétiques est élaborée, rendant compte à la fois du modèle épidémiologique de Cori et de la qualité limitée des données réelles afin de fournir des séries temporelles de nouvelles infections réalistes, accompagnées du coefficient de reproduction variable dans le temps constituant la vérité terrain. Cette procédure est mise à profit pour comparer différents estimateurs au moyen de simulations numériques de Monte Carlo intensives, explorant différentes dynamiques épidémiques et différents niveaux de corruption des données.

Abstract – Covid19 pandemic has seen a boom in signal processing research for epidemiology. In particular, numerous epidemiological data processing strategies to infer the *reproduction number*, quantifying the virus spread dynamics, have been designed to address the main challenges raised by daily analysis of Covid19 data, namely their limited quality, the requirement of epidemiologically grounded methods, and the need for efficient algorithms for daily monitoring of the pandemic. Elaborating on Cori's epidemiological model, variational strategies have been proposed, yielding consistent and fast reproduction number estimators. Though, at the time of writing, no systematic accuracy analysis have been carried out, the main reason being the absence of annotated data. Quantitative assessment is crucial, first, to validate the ability of reproduction number estimators to manage highly corrupted data, and second, to envision enlarging their use to other epidemics. To this purpose, a synthetic data generation procedure is designed, capturing both Cori's epidemiological model and the limited quality of real-world data, to yield realistic synthetic infection counts accompanied with ground truth time-varying reproduction numbers. This procedure is then leveraged to compare the performance of different estimators through intensive Monte Carlo numerical experiments, exploring different epidemiological dynamics and levels of data corruption.

1 Introduction

Contexte. Depuis le début de la pandémie de Covid19 beaucoup d'efforts ont été dévolus à la construction d'outils de surveillance efficaces. L'indicateur de l'intensité de l'épidémie le plus populaire est le *coefficient de reproduction* R_t, défini comme le nombre moyen de personnes contaminées par un individu infectieux, quantifiant la transmissibilité du virus au jour t [5, 8]. Parmi les estimateurs populaires coefficient de reproduction d'une épidémie on peut citer [6, 12, 5, 10], certains ayant été développés très récemment suite à l'émergence de la Covid19. Néanmoins, faute de données annotées, par exemple *synthétiques*, aucun de ces estimateurs n'a reçu de validation systématique de ses performances et il est donc impossible de déterminer quel estimateur s'avère le plus précis.

État-de-l'art. Une revue de la littérature récente montrent en effet que peu de travaux se sont intéressés à la génération de données synthétiques pour la Covid19. Les modèles compartimentaux [8, 3], décrivant directement la dynamique épidémique, sont bien adaptés à la génération de données, et ont été mis à profit, e.g., pour la prédiction [11, 2]. Cependant, la qualité des données synthétiques produites dépend de la précision dans l'estimation des paramètres du modèle compartimental, qui est non seulement coûteux en temps, mais également fortement dégradée par la faible qualité des données relatives à la Covid19 reportées par les autorités sanitaires.

Alternativement, dans le modèle statistique de Cori [5], le nombre de nouvelles infections Z_t au jour t dépend du coefficient de reproduction R_t et des nombres d'infections passés $\{\mathsf{Z}_1,\mathsf{Z}_2,\ldots,\mathsf{Z}_{t-1}\}$, pondérés par la fonction d'intervalle de série ϕ modélisant le délai aléatoire entre le début des symptômes dans l'infection primaire et l'infection secondaire.Pour l'estimation du coefficient de reproduction d'une épidémie à partir des nombres de cas, [5] privilégie un schéma bayésien, qui s'est avéré approprié lorsqu'appliqué à des données consolidées *a posteriori*, mais qui est sévèrement mis en difficulté par la qualité limitée des données liées à la Covid19.

Objectifs et contributions. Le but premier de cet article est de comparer les procédures en deux vs. une étape(s) introduites par [1, 10] pour l'estimation du coefficient de reproduction

de la Covid19 sur la base d'une évaluation quantitative de leurs performances. Cette évaluation systématique nécessite des données de nouvelles infections accompagnées de leur vérité terrain, motivant la première contribution : la construction d'une stratégie de génération de données synthétiques suffisamment rigoureuse pour permettre une évaluation objective et pertinente. Une seconde contribution est de réalisée une comparaison détaillée, quantitative et systématique, des performance de différents estimateur *via* des simulations numériques de Monte Carlo intensives.

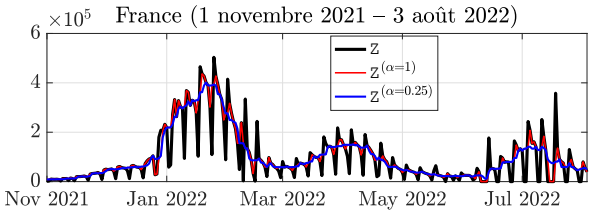


FIGURE 1 – Comptes bruts de nouvelles infections (en noir) $\mathbf{Z}$, comptes pré-traités par médian glissant $\mathbf{Z}^{(\alpha)}$, avec α grand (resp. petit) en rouge (resp. en bleu).

2 Estimation de R_t

Modèle épidémiologique. Selon le modèle original proposé par [5], le nombre de nouvelles infections Z_t au jour t suit une loi de Poisson dont l'intensité p_t, variable dans le temps, dépend du coefficient de reproduction R_t et d'une moyenne pondérée des précédentes infections $\Phi_t^{\mathbf{Z}} = \sum_{s=1}^{\tau_\phi} \phi(s)\mathsf{Z}_{t-s}$, de sorte que la log-vraisemblance du modèle s'écrit

$$\ln \mathbb{P}(\mathsf{Z}_t|\mathsf{Z}_1,\dots,\mathsf{Z}_{t-1};\mathsf{p}_t) = \mathsf{Z}_t \ln(\mathsf{p}_t) - \mathsf{p}_t - \ln(\mathsf{Z}_t!). \quad (1)$$

La *fonction d'intervalle de série* ϕ encode les caractéristiques de la pandémie de Covid19 et est modélisée par une distribution Gamma de moyenne (resp. écart-type) 6,6 (resp. 3,5 jours) tronquée à $\tau_\phi = 25$ jours, reflétant la contagiosité maximale entre 3 et 10 jours après l'apparition des symptômes.

Données relatives à la Covid19. Depuis le tout début de la pandémie, en janvier 2020, l'Université Johns Hopkins (JHU) a collecté et mis en ligne les nombre d'infections quotidiennes par la Covid19 dans plus de 200 pays[1]. À titre d'exemple, la Fig. 1 présente le décompte de nouveaux cas en France sur une durée de dix mois, qui présente des fluctuations rapides, parfois très importantes (voir la courbe en noir sur la Fig. 1), qui sont sans lien avec la dynamique épidémique mais reflètent l'impact des jours chômés sur le report des cas. Une estimation précise du coefficient de reproduction sous-jacent nécessite de gérer la mauvaise qualité de ces données réelles.

Estimateur de Maximum de Vraisemblance (MV). Selon le modèle épidémiologique de Cori [5], l'intensité p_t, variable au cours du temps, d'une épidémie est le produit du coefficient de reproduction et de la contagiosité effective dans la population :

$$\mathsf{p}_t^{(1)} = \mathsf{R}_t \Phi_t^{\mathbf{Z}}. \quad (2)$$

Une maximisation directe de (1) fournit l'estimateur de Maximum de Vraisemblance (MV) :

$$\widehat{\mathsf{R}}_t^{\mathrm{MV}} = \mathsf{Z}_t / \Phi_t^{\mathbf{Z}}, \quad (3)$$

qui est tracé en noir sur la Fig. 2 dans l'exemple susmentionné des données françaises. À cause de la grande variabilité des décomptes d'infections (voir Fig. 1), le maximum de vraisemblance présente un comportement très irrégulier, absolument non réaliste.

Débruitage puis estimation régularisée. Les fluctuations erratiques des nouvelles infections reportées proviennent, en petite partie, du caractère aléatoire de la propagation de la pandémie, mais surtout, du processus de collecte fortement impacté par les jours non-travaillés entraînant des erreurs de report, des données aberrantes, des décomptes nuls le week-end et des effets de pseudo-périodicité.

i) Pré-traitement des comptes erronés. Une stratégie naturelle est d'appliquer aux données corrompues une étape de débruitage en amont de l'estimation de R_t. Pour retirer les décomptes anormalement bas/élevés tout en préservant autant que possible la dynamique sous-jacente, [1] ont proposé d'appliquer un médian glissant aux données brutes de nouvelles infections. En pratique, pour chaque jour t, le médian de Z_s sur une fenêtre W_t de deux semaines centrée en t est comparé au compte observé Z_t : pour un seuil fixé $\alpha > 0$, Z_t est remplacé par le médian local si

$$|\mathsf{Z}_t - \mathrm{med}\{\mathsf{Z}_s, s \in \mathsf{W}_t\}| \geq \alpha \times \mathrm{mad}\{\mathsf{Z}_s, s \in \mathsf{W}_t\} \quad (4)$$

où med (resp. mad) désigne le médian (resp. la déviation au médian en valeur absolue) d'une série temporelle ; sinon Z_t est laissé inchangé. Les données française pré-traitées *via* la procédure de médian glissant, notées $\mathbf{Z}^{(\alpha)}$, sont tracées en Fig. 1 : lors α est grand seules les valeurs extrêmement anormales sont corrigées, tandis qu'un petit α induit un débruitage plus fort, au prix d'une dégradation de la dynamique sous-jacente, (voir discussion en Sec. 4).

ii) Estimation régularisée. Une fois les données corrigées de leurs valeurs aberrantes par médian glissant, [1] estime R_t *via* la minimisation d'une Vraisemblance Pénalisée (VP) favorisant un comportement linéaire par morceaux qui permet de saisir la tendance globale de la dynamique épidémique, et résout le problème d'optimisation :

$$\widehat{\mathbf{R}}^{\mathrm{VP}} = \operatorname*{argmin}_{\mathbf{R} \in \mathbb{R}_+^T} \sum_{t=1}^{T} \mathrm{d}_{\mathsf{KL}}(\mathsf{Z}_t|\mathsf{p}_t^{(1)}) + \mu_{\mathsf{R}} \|\mathbf{D}_2\mathbf{R}\|_1 \quad (5)$$

où la divergence de Kullback-Leibler, définie comme

$$\mathrm{d}_{\mathrm{KL}}(\mathsf{Z}|\mathsf{p}) = \begin{cases} \mathsf{Z}\ln(\mathsf{Z}/\mathsf{p}) + \mathsf{p} - \mathsf{Z} & \text{si } \mathsf{Z} > 0 \text{ et } \mathsf{p} > 0 \\ \mathsf{p} & \text{si } \mathsf{Z} = 0 \text{ et } \mathsf{p} \geq 0 \\ \infty & \text{sinon,} \end{cases}$$

coïncide avec l'opposé de la log-vraisemblance de Poisson. La parcimonie du laplacien discret de l'estimée, i.e., le caractère linéaire par morceaux de $\widehat{\mathbf{R}}^{\mathrm{VP}}$, est favorisé par $\|\mathbf{D}_2\mathbf{R}\|_1 = \sum_{t=3}^{T} |\mathsf{R}_t - 2\mathsf{R}_{t-1} + \mathsf{R}_{t-2}|$. Le paramètre $\mu_{\mathsf{R}} > 0$ contrôle le niveau de régularisation (voir Sec. 4).

Modèle de Cori étendu et procédure jointe. Bien que fournissant des estimées régularisées satisfaisantes sur certaines données peu corrompues, cette méthode en deux étapes est plus difficile à mettre en œuvre lorsque la qualité des données

[1] https://coronavirus.jhu.edu/

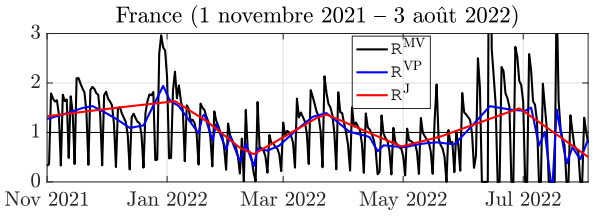


FIGURE 2 – Estimation du coefficient de reproduction à partir des données françaises de la Fig.1 par maximum de vraisemblance (MV), médian glissant suivi de la vraisemblance pénalisée (VP) et estimation jointe (J).

diminue. Pour palier cette limitation et gérer des comptes erronés et des données manquantes efficacement, [10] ont proposé un modèle étendu de l'intensité de l'épidémie :

$$\mathsf{p}_t^{(2)} = \mathsf{R}_t \Phi_t^{\mathbf{Z}} + \mathsf{O}_t, \tag{6}$$

où la variable O_t modélise l'écart entre comptes reportés et nombre réel de nouvelles infection et est estimée conjointement avec R_t *via* un algorithme joint original résolvant

$$\widehat{\mathbf{R}}^{\mathrm{J}}, \widehat{\mathbf{O}}^{\mathrm{J}} = \underset{\mathbf{R}\in\mathbb{R}_+^T, \mathbf{O}\in\mathbb{R}^T}{\operatorname{argmin}} \sum_{t=1}^{T} \mathrm{d}_{\mathsf{KL}}(\mathsf{Z}_t | \mathsf{p}_t^{(2)}) + \lambda_{\mathsf{R}} \|\mathbf{D}_2\|_1 + \lambda_{\mathsf{O}} \|\mathbf{O}\|_1. \tag{7}$$

Au terme $\|\mathbf{D}_2\|_1$ favorisant la linéarité par morceaux de R_t est adjoint un terme $\|\mathbf{O}\|_1$ promouvant la parcimonie de la variable d'erreur O_t, dont le niveau est contrôlé par un paramètre $\lambda_{\mathsf{O}} > 0$.

3 Données synthétiques

L'évaluation et la comparaison quantitatives des stratégies d'estimation du coefficient de reproduction décrites Sec. 2 requiert des décomptes de nouvelles infections $\mathbf{Z}$ accompagnés de coefficients de reproduction $\mathbf{R}$ constituant la vérité terrain. L'Algorithme 1 est construit soigneusement pour générer des nombres de cas reportés synthétiques, présentant une dégradation modélisée par la variable d'erreur O_t, de façon à être aussi réalistes que possible du point de vue épidémiologique. Un second défi à relever pour la production de données synthétiques réalistes est de fabriquer des vérités terrains appropriées pour R_t et O_t, pour cela une procédure s'appuyant sur les données réelles disponibles est proposée.

Échantillonnage itératif sous le modèle de Cori étendu. Étant donné les décomptes passés $\mathsf{Z}_1, \ldots, \mathsf{Z}_{t-1}$, échantillonner un nouveau compte d'infections Z_t sous le modèle de Cori étendu (1) revient à calculer l'intensité de Poisson instantanée $\mathsf{p}_t^{(2)}$ à partir des nombres de cas passés selon (6) (voir l. 2 de l'Alg. 1), puis d'échantillonner une variable aléatoire sous la loi de Poisson d'intensité $\mathsf{p}_t^{(2)}$ (voir l. 3 de l'Alg. 1). La contagiosité effective $\Phi_t^{\mathbf{Z}}$ intervenant dans le calcul de $\mathsf{p}_t^{(2)}$ dépend des τ_ϕ précédents comptes. Pour $t < \tau_\phi$ les effets de bords dans le calcul de $\Phi_t^{\mathbf{Z}}$ sont gérés en imposant que la somme des poids appliqués vaille un par une normalisation par n_t^ϕ (voir l. 1 de l'Alg. 1).

Vérités terrains réalistes pour R et O. Pour obtenir des décomptes de nouvelles infections synthétiques réalistes, la

Algorithme 1 : Génération de séries temporelles de nouvelles infections quotidiennes synthétiques.

Données : $\mathbf{R}^{\text{vrai}} \in \mathbb{R}_+^T$, $\mathbf{O}^{\text{vrai}} \in \mathbb{R}^T$,
$\mathsf{Z}_0 \in \mathbb{N}$ *# nombre d'infections initial*

Résultat : $\mathbf{Z} = (\mathsf{Z}_1, \mathsf{Z}_2, \ldots, \mathsf{Z}_T) \in \mathbb{N}^T$

1 **pour** $t = 1, 2, \ldots, T$ **faire**

2 $\quad n_t^\phi = \sum_{s=1}^{\max(t,\tau_\phi)} \phi(s)$ *# effets de bords*

3 $\quad \mathsf{p}_t = \mathsf{R}_t^{\text{vrai}} \times \sum_{s=1}^{\max(t,\tau_\phi)} \frac{\phi(s)}{n_t^\phi} \mathsf{Z}_{t-s} + \mathsf{O}_t^{\text{vrai}}$ *# intensité*

4 $\quad \mathsf{Z}_t \sim \mathrm{Poiss}\left(\max(\mathsf{p}_t, 0)\right)$ *# nombre d'infections*

5 **fin**

construction de $\mathbf{R}^{\text{vrai}}$ et $\mathbf{O}^{\text{vrai}}$ imitant fidèlement à la fois la dynamique de propagation sous-jacente et la dégradation des données par le processus de collecte est crucial. Or, ni le coefficient de reproduction R_t dépendant du temps, ni la variable de corruption non stationnaire O_t ne sont accessibles à une modélisation exhaustive, notamment à cause de leur dépendance radicale vis-à-vis de l'épidémie considérée. Ainsi la fabrication directe de $\mathbf{R}^{\text{vrai}}$ et $\mathbf{O}^{\text{vrai}}$ est exclue. Afin de contourner cet écueil, des coefficients de reproduction et erreurs réalistes sont extraits à partir du traitement d'échantillons de données réelles mis à dispositions par l'Université Johns Hopkins. Pour un pays et une période de temps donnés, les décomptes d'infections reportés sont analysés au moyen de la stratégie jointe (en une étape) (7) pour des paramètres de régularisation fixés ; les estimées obtenues $\widehat{\mathbf{R}}^{\mathrm{J}}$ et $\widehat{\mathbf{O}}^{\mathrm{J}}$ sont ensuite considérées comme des vérités terrains crédibles. En considérant différentes configurations, c.-à-d. différents réglages des paramètres de régularisation $(\lambda_{\mathsf{R}}, \lambda_{\mathsf{O}})$, cette méthode permet de construire un jeu de vérités terrains réalistes qui peuvent ensuite être combinées pour générer des données synthétiques à la fois réalistes et diversifiées à partir de l'Alg. 1 (voir Fig. 3).

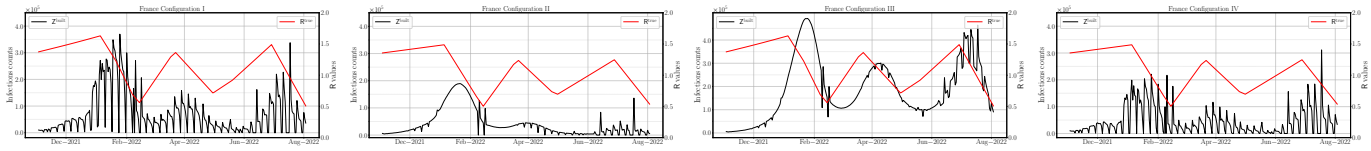
4 Comparaison des performances

Optimisation convexe non lisse. Les estimées par vraisemblance pénalisée et par la procédure jointe, correspondant aux problèmes d'optimisation (5) et (7), sont calculées par les schémas primaux-duaux de Chambolle-Pock [4] développés dans [1, 10], recourant à des *opérateurs proximaux* [9] pour la minimisation de fonctionnelles *non-lisses*. Les pas de descente sont choisis de la même manière que dans [1] et [10], cependant, pour renforcer la robustesse de l'estimation vis-à-vis des différentes configurations et différents niveaux de régularization, la convergence est mesurée au moyen des incréments du coefficient de reproduction entre deux itérations successives $\mathbf{R}^{[k]}$ et $\mathbf{R}^{[k+1]}$. L'algorithme est stoppé lorsque le maximum des incréments relatifs sur les T composantes, lissé sur une fenêtre de $5 \cdot 10^2$ itérations, est inférieur à 10^{-6}, i.e., quand

$$\max_{k-500<k'\leq k} \max_{t=1,\ldots,T} \frac{\left|\mathsf{R}_t^{[k'+1]} - \mathsf{R}_t^{[k']}\right|}{\mathsf{R}_t^{[k']}} \leq 10^{-6} \tag{8}$$

ou bien après $7 \cdot 10^5$ itérations.

Données synthétiques. Les performances d'estimation des méthodes à deux vs. une étape(s) sont évaluées sur des données



(a) C. I (b) C. II (c) C. III (d) C. IV

FIGURE 3 – Données synthétiques $\mathbf{Z}$ (en noir) et vérité terrain $\mathbf{R}^{\text{vrai}}$ (en rouge).

synthétiques produites *via* la procédure décrite à la Sec. 3. Les comptes de nouvelles infections reportés en France entre le 1er novembre 2021 et le 3 août 2022 sont analysés au moyen de (7), avec deux réglages de paramètres de régularisation très différents, fournissant une collection de $\mathbf{R}^{\text{vrai}}$, $\mathbf{O}^{\text{vrai}}$. Ces vérités terrains sont ensuite combinées entre elles pour fournir quatre configurations (C.) :

C. I R_t possède beaucoup de ruptures de pente, O_t a beaucoup de composantes non nulles (voir Fig. 3a) ;

C. II peu de changements de pentes de R_t, peu de valeurs non nulles de O_t (voir Fig. 3b) ;

C. III beaucoup de changements de pentes de R_t, peu de valeurs non nulles de O_t (voir Fig. 3c) ;

C. IV peu de changements de pentes de R_t, beaucoup de valeurs non nulles de O_t (voir Fig. 3d).

Ces vérités terrains réalistes et diversifiées, avec des dynamiques de R_t, O_t similaires pour C. I et II et « croisées » pour C. III et IV, sont ensuite mises à profit pour générer des comptes $\mathbf{Z}$ synthétiques au moyen de l'Alg. 1. Pour chaque couple $(\mathbf{R}^{\text{vrai}}, \mathbf{O}^{\text{vrai}})$, c.-à-d. chaque configuration, 5 réalisations de $\mathbf{Z}$ sont générées.

Évaluation des performances. Pour mesurer la qualité des estimées du coefficient de reproduction, deux critères sont utilisés. Tout d'abord, le rapport signal à bruit (SNR), indicateur standard mesuré en décibels (dB), défini comme

$$\mathsf{SNR} := 10 \log_{10} \left(\frac{\|\mathbf{R}^{\text{vrai}}\|_2^2}{\|\widehat{\mathbf{R}} - \mathbf{R}^{\text{vrai}}\|_2^2} \right) \tag{9}$$

quantifie l'adéquation entre l'estimée $\widehat{\mathbf{R}}$ et la vérité terrain R^{vrai}. La qualité des estimées fournie par les méthodes en deux vs. une étape(s) dépend fortement du réglage des hyperparamètres, c'est-à-dire du choix du seuil α dans (4) et du paramètre μ_{R} dans (5), et de λ_{R}, λ_{O} dans (7). Pour chaque estimateur, chaque configuration et chaque réalisation, le SNR est maximisé sur une grille de 20×20 hyperparamètres espacés logarithmiquement. Pour chaque estimateur et chaque configuration, le SNR maximal moyen et son intervalle de confiance gaussien à 95% sont reportés dans la Tab. 1, lignes 3 à 6. Dans un second temps, pour la meilleure estimation obtenue au sens du SNR, l'indice de Jaccard [7] entre le laplacien discret de $\widehat{\mathbf{R}}$ et celui de $\mathbf{R}^{\text{vrai}}$, reporté dans la Tab. 1, lignes 8 à 11, permet d'évaluer si les points de rupture dans le comportement linéaire de R_t, correspondant aux changements brutaux de dynamique épidémique, sont détectés correctement.

Comparaison entre les estimateurs. Pour chaque configuration C. I à IV, la Fig. 4 présente les estimées de $\mathbf{R}$ obtenues par maximum de vraisemblance (en gris), médian glissant puis vraisemblance pénalisée (en bleu), et procédure jointe (en rouge) sur une réalisation de données synthétiques : l'estimation R^{J} y est systématiquement la plus fidèle à $\mathbf{R}^{\text{vrai}}$. L'étude

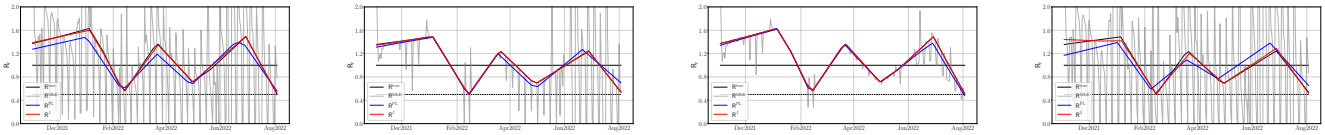
	Max. Vrais. (MV)	Vrais. Pén. (VP)	Estim. Jointe (J)
SNR (dB)			
C. I	$0{,}92 \pm 0{,}04$	$20{,}70 \pm 0{,}08$	$35{,}69 \pm 0{,}63$
C. II	$-2{,}25 \pm 0{,}08$	$26{,}68 \pm 0{,}18$	$47{,}84 \pm 2{,}24$
C. III	$19{,}11 \pm 0{,}14$	$29{,}64 \pm 0{,}15$	$56{,}22 \pm 1{,}36$
C. IV	$-2{,}33 \pm 0{,}01$	$18{,}78 \pm 0{,}01$	$31{,}22 \pm 0{,}44$
Indice de Jaccard (%)			
C. I	$0{,}79 \pm 0{,}01$*	$57{,}23 \pm 0{,}56$	$85{,}04 \pm 2{,}25$
C. II	$0{,}57 \pm 0{,}01$	$47{,}60 \pm 1{,}57$	$96{,}40 \pm 0{,}85$
C. III	$2{,}26 \pm 0{,}04$	$73{,}20 \pm 2{,}16$	$97{,}61 \pm 1{,}25$
C. IV	$0{,}67 \pm 0{,}01$*	$13{,}63 \pm 0{,}44$	$76{,}69 \pm 1{,}18$

TABLE 1 – SNR et indice de Jaccard pour trois estimateurs et quatre configurations, moyennées sur $N = 5$ réalisations.
* intervalle de confiance plus petit que 0,01.

quantitative du SNR sur 5 réalisations de chaque configuration, l. 3 à 6 du Tab. 1 montre la supériorité indéniable des méthodes régularisées, col. 3 et 4, par rapport au maximum de vraisemblance, col. 2. En outre, la procédure jointe, col. 4 du Tab. 1, atteint des SNR significativement plus élevés, de plus de 10 dB, que l'estimation par médian glissant suivi de la vraisemblance pénalisée, col. 3 du Tab. 1. Ces observations se vérifient de manière robuste pour les quatre configurations de données synthétiques C. I à IV, montrant que la procédure jointe surpasse systématiquement le maximum de vraisemblance, ainsi que la méthode de médian glissant puis vraisemblance pénalisée.

Robustesse de la procédure d'estimation. La diversité des données synthétiques C. I à IV fournit un cadre adéquat pour sonder la robustesse d'un estimateur de R_t vis-à-vis de combinaisons de dynamiques épidémiques et de niveaux de corruption des données très différents. Pour toutes les configurations C. I à IV, l'indice de Jaccard entre le Laplacien de l'estimée jointe et celui de la vérité terrain, col. 8 à 11 du Tab. 1, varie entre 76% et 97% attestant à la fois d'une détection efficace des ruptures dans la dynamique épidémique et de leur bonne localisation. Ces performances robustes plaident pour l'utilisation de la procédure d'estimation jointe dans le cadre de la surveillance d'une épidémie, à privilégier par rapport aux méthodes précédemment développées [5, 1].

Temps de calcul. La procédure jointe permet d'obtenir les estimations de $\mathbf{R}$ les plus fidèles aux vérités terrains $\mathbf{R}^{\text{vrai}}$ (cf. Tab. 1), la contrepartie est un coût de calcul significativement plus élevé comme le montre la Fig. 5. L'évolution du critère de convergence (8) pour la méthode par vraisemblance pénalisée (en bleu) et pour la procédure jointe (en rouge) montre qu'atteindre la précision requise de 10^{-6} nécessite 20 fois plus d'itérations pour la procédure jointe que pour la vraisemblance pénalisée.



(a) C. I (b) C. II (c) C. III (d) C. IV

FIGURE 4 – Estimations de R par maximum de vraisemblance (en gris), médian glissant puis vraisemblance pénalisée (en bleu) et procédure jointe (en rouge) sur des données synthétiques de vérité terrain R^{vrai} (en noir).

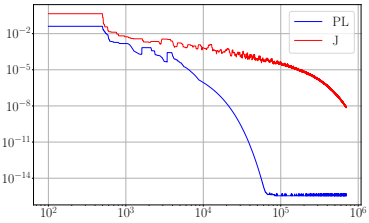


FIGURE 5 – Incréments relatifs (8) en fonction des itérations lors de l'estimation de R_t à partir de données synthétiques (**C I**) via VP (en bleu) ou J (en rouge).

5 Conclusion

Une procédure de génération de décomptes d'infections a été proposée, fournissant des données synthétiques réalistes, reproduisant les principales caractéristiques des données liée à la Covid19, c'est-à-dire à la fois la dynamique épidémique et les irrégularités et les comptes manquants dus aux aléas dans la collecte des nouveaux cas. À partir de l'algorithme proposé 1, un jeu de donné diversifié de décomptes synthétiques accompagnés de la vérité terrain sur le coefficient de reproduction a été construit et mis à profit pour évaluer et comparer quantitativement, *via* des simulations de Monte Carlo intensives, deux estimateurs récents du coefficient de reproduction. La supériorité de l'approche jointe, réalisant en une seule étape la correction des comptes erronés et l'estimation d'un coefficient de reproduction régularisé, a été démontrée clairement, aussi bien en terme de SNR que d'indice de Jaccard sur la localisation des ruptures de pente de $\widehat{R}_t$.

L'algorithme de génération de décomptes d'infections synthétiques constitue une étape cruciale vers une évaluation systématique des estimateurs d'indicateurs épidémiologiques, permettant de valider l'utilisation élargie des outils développés récemment pour la surveillance épidémique. En outre, cette étude ouvre la voie à l'élaboration de méthodes automatiques et pilotées par les données pour le réglage des hyperparamètres λ_R,λ_O de la méthode jointe en une seule étape, fournissant ainsi un estimateur du coefficient de reproduction sans paramètre et auto-adaptable à différentes épidémies.

Références

[1] P. ABRY, N. PUSTELNIK, S. ROUX, P. JENSEN, P. FLANDRIN, R. GRIBONVAL, C.-G. LUCAS, E. GUICHARD, P. BORGNAT et N. GARNIER : Spatial and temporal regularization to estimate covid-19 reproduction number r (t) : Promoting piecewise smoothness via convex optimization. *PLOS One, vol. 15, no. 8, p. e0237901*, 2020.

[2] N. BANNUR, V. SHAH, A. RAVAL et J. WHITE : Synthetic Data Generation for Improved covid-19 Epidemic Forecasting. *medRxiv*, pages 2020–12, 2020.

[3] F. BRAUER, C. CASTILLO-CHAVEZ et Z. FENG : *Mathematical models in epidemiology*. Springer, New York, 2019.

[4] A. CHAMBOLLE et T. POCK : A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vis.*, 40(1):120–145, 2011.

[5] A. CORI, N. M. FERGUSON, C. FRASER et S. CAUCHEMEZ : A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epidemiol.*, 178(9):1505–1512, 2013.

[6] O. DIEKMANN, J. A. P. HEESTERBEEK et J. A. J. METZ : On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. *J. Math. Biol.*, 28:365–382, 1990.

[7] P. JACCARD : Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines. *Bull. Soc. Vaudoise Sci. Nat.*, 37:241–272, 1901.

[8] Q.-H. LIU, M. AJELLI, A. ALETA, S. MERLER, Y. MORENO et A. VESPIGNANI : Measurability of the epidemic reproduction number in data-driven contact networks. *PNAS*, 115(50):12680–12685, 2018.

[9] N. PARIKH et S. BOYD : Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[10] B. PASCAL, P. ABRY, N. PUSTELNIK, S. ROUX, R. GRIBONVAL et P. FLANDRIN : Nonsmooth convex optimization to estimate the Covid-19 reproduction number space-time evolution with robustness against low quality data. *IEEE Trans. Signal Process.*, 70:2859–2868, 2022.

[11] IHME COVID-19 Forecasting TEAM et Hay S. I. : COVID-19 scenarios for the United States. *medRxiv*, 2020.

[12] J. WALLINGA et P. TEUNIS : Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *Am. J. of Epidemiol.*, 160(6):509–516, 09 2004.