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

Juliana Du, Barbara Pascal, Patrice Abry

▶ **To cite this version:**

Juliana Du, Barbara Pascal, Patrice Abry. Compared performance of Covid19 reproduction number estimators based on realistic synthetic data. Colloque Francophone de Traitement du Signal et des Images (GRETSI), GRETSI, Aug 2023, Grenoble, France. hal-04032614v4

HAL Id: hal-04032614
https://hal.science/hal-04032614v4

Submitted on 17 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

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]ENSL, CNRS, Laboratoire de physique, F-69342 Lyon, France

[2]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Résumé – Estimer précisément le coefficient de reproduction est un enjeu majeur pour surveiller et contrôler une épidémie. Afin d'évaluer et de comparer plusieurs estimateurs proposés durant la pandémie de Covid19, une procédure de génération de données épidémiologiques synthétiques réalistes est élaborée, reflétant à la fois la propagation épidémique, suivant le modèle épidémiologique proposé par Cori et col., et un processus de collecte de nombre de cas dégradé (erreur de report, données manquantes). Des simulations de Monte Carlo sur un large jeu de données synthétiques aux dynamiques variées fournissent une évaluation quantitative des performances d'estimation du coefficient de reproduction.

Abstract – Accurate estimation of the reproduction number is a major challenge for the surveillance and control of epidemics. To evaluate and compare several estimators proposed during the Covid19 pandemic, a procedure for generating realistic synthetic epidemiological data is designed, reflecting both the epidemic spread, following the epidemiological model proposed by Cori et al., and the degraded collection of infection counts (erroneous figures, missing data). Monte Carlo simulations over a large set of synthetic data with different dynamics yield quantitative assessment of reproduction number estimation performance.

1 Introduction

Contexte. Durant la pandémie de Covid19 des efforts conséquents ont été dévolus à la construction d'outils précis pour le suivi de la situation sanitaire au jour le jour. L'indicateur de l'intensité épidémique 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, 7]. Plusieurs estimateurs de R_t sont documentés [9, 5, 1, 8], dont certains ont été développés pour répondre à la crise engendrée par la pandémie de Covid19. Néanmoins, faute de données épidémiologiques annotées, par exemple *synthétiques*, aucun de ces estimateurs n'a reçu de validation systématique de ses performances et aucune comparaison quantitative n'est possible. Or, le contrôle d'une pandémie, aux enjeux sanitaires, sociaux et économiques majeurs, requiert des outils validés scientifiquement.

État-de-l'art. Une revue de la littérature récente montre que peu de travaux se sont intéressés à la génération de données synthétiques pour la Covid19. Les modèles compartimentaux [7, 3], décrivant directement la dynamique épidémique, sont bien adaptés à la génération de données, et ont été mis à profit, par exemple pour la prédiction [2]. Cependant, la qualité des données synthétiques produites dépend de la précision dans l'estimation des paramètres du modèle, non seulement coûteuse en ressources de calcul, mais également fortement dégradée par la faible qualité des données relatives à la Covid19 rapportées par les autorités sanitaires (cf. exemple Fig. 1).

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. Le schéma bayésien privilégié par [5] pour estimer R_t à partir des comptes Z_t, approprié pour des données consolidées *a posteriori*, est sévèrement mis en difficulté par les données quotidiennes liées à la Covid19 qui sont de qualité limitée.

Objectifs et contributions. Le but est de comparer les estimateurs du coefficient de reproduction introduits par [1, 8] sur la base de performances d'estimation quantitatives. Une première contribution consiste en la construction d'une stratégie de génération de données synthétiques réalistes, fournissant des comptes d'infection accompagnés de la vérité terrain sur le coefficient de reproduction. Une seconde contribution est la comparaison quantitative systématique des performances d'estimation et du temps de calcul, menée *via* des simulations numériques de Monte Carlo intensives.

2 Estimation de R_t

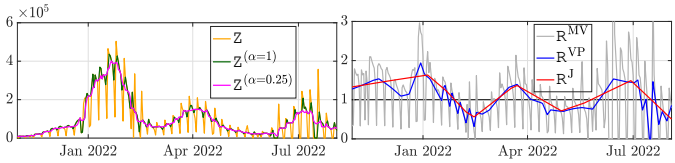
Modèle épidémiologique. Le modèle de Cori [5] fait l'hypothèse que Z_t suit une loi de Poisson dont l'intensité p_t varie dans le temps, de sorte que la log-vraisemblance s'écrit

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

Données relatives à la Covid19. Depuis le début de la pandémie, l'Université Johns Hopkins (JHU) a collecté et mis en ligne le nombre quotidien d'infections par la Covid19 dans plus de 200 pays[1]. À titre d'exemple, la Fig. 1a présente les comptes de nouveaux cas Z_t en France sur une durée de dix mois, qui présentent des fluctuations rapides, parfois très importantes (cf. courbe en orange, Fig. 1a), sans lien avec la dynamique épidémique et reflétant avant tout l'impact des jours chômés sur les stratégies de test et de report des cas. Une

[†]J. Du est financée par le projet 80PRIME-2021 CNRS « *CoMoDécartes* ».

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



(a) Comptes bruts $\mathbf{Z}$ (en orange) et pré-traités par médian glissant $\mathbf{Z}^{(\alpha)}$, avec α grand (resp. petit) en vert (resp. en rose)

(b) Coefficient de reproduction estimé par Max. Vrais. (MV), Vrais. Pén. (VP) et estimation jointe (J).

FIGURE 1 : Infections par la Covid19 en France et coefficient de reproduction entre le 1er novembre 2021 et le 3 août 2022.

estimation précise du coefficient de reproduction sous-jacent nécessite de gérer la mauvaise qualité de ces données réelles.

Maximum de Vraisemblance. Originellement, l'intensité p_t d'une épidémie est le produit du coefficient de reproduction et de la *contagiosité effective* dans la population [5] :

$$\mathsf{p}_t^{(1)} = \mathsf{R}_t \Phi_t^{\mathbf{Z}}, \quad \text{où} \quad \Phi_t^{\mathbf{Z}} = \sum_{s=1}^{\tau_\phi} \phi(s) \mathsf{Z}_{t-s}. \tag{2}$$

La *fonction d'intervalle de série* ϕ encode les caractéristiques de l'épidémie et est modélisée, pour le virus de la Covid19, par une distribution Gamma de moyenne (resp. écart-type) 6,6 (resp. 3,5) jours tronquée à $\tau_\phi = 25$ jours. L'estimateur de Maximum de Vraisemblance (MV) est obtenu par maximisation directe de (1) :

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

et tracé en gris sur la Fig. 1b, dans l'exemple susmentionné des données françaises. La grande variabilité des décomptes d'infections induit un comportement très irrégulier de $\widehat{\mathsf{R}}_t^{\mathrm{MV}}$, non réaliste du point de vue épidémiologique.

Débruitage puis estimation régularisée. Les fluctuations erratiques des comptes reportés proviennent majoritairement 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 pseudo-périodicités, et seulement dans une moindre mesure du caractère aléatoire de la propagation de la pandémie.

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] applique un médian glissant aux données brutes de nouvelles infections. 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

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

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 françaises pré-traitées *via* la procédure de médian glissant, notées $\mathbf{Z}^{(\alpha)}$, sont tracées en Fig. 1a : lorsque 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 le 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 :

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

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, c.-à-d., 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 diminue (cf. courbe en bleu en Fig. 1b). Pour palier cette limitation et gérer efficacement les comptes erronés et les données manquantes dans les données récentes, un modèle étendu de l'intensité de l'épidémie a été proposé dans [8] :

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

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

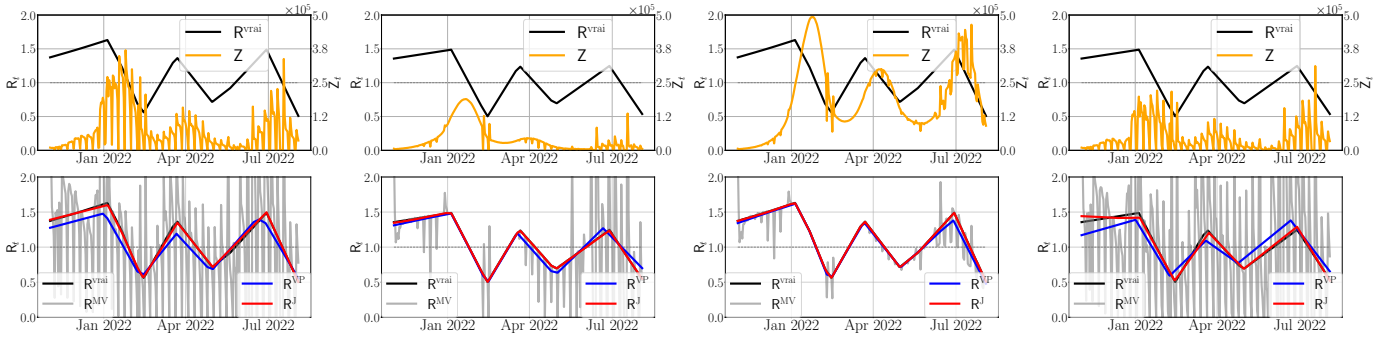
$$\widehat{\mathbf{R}}^{\mathrm{J}}, \widehat{\mathbf{O}}^{\mathrm{J}} = \operatorname*{argmin}_{\mathbf{R} \in \mathbb{R}_+^T, \mathbf{O} \in \mathbb{R}^T} \sum_{t=1}^{T} \mathrm{d}_{\mathsf{KL}}(\mathsf{Z}_t | \mathsf{p}_t^{(2)}) + \lambda_{\mathsf{R}} \|\mathbf{D}_2\|_1 + \lambda_{\mathsf{O}} \|\mathbf{O}\|_1.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

Le terme $\|\mathbf{O}\|_1$ promeut la parcimonie de la variable d'erreur O_t, dont le niveau est contrôlé par un paramètre $\lambda_{\mathsf{O}} > 0$. L'estimée de R_t *via* la procédure jointe apparaît significativement plus robuste à la qualité limitée des données (voir courbe en rouge, Fig. 1b) ce qui sera quantifié en Sec. 4.

3 Données synthétiques

L'évaluation et la comparaison quantitatives des stratégies d'estimation de R_t décrites en Sec. 2 requièrent des comptes $\mathbf{Z}$ accompagnés des coefficients de reproduction $\mathbf{R}$ constituant la vérité terrain. L'Alg. 1 est construit pour générer efficacement des nombres de cas reportés synthétiques sous le modèle (6) à partir de vérités terrain $\mathbf{R}$, $\mathbf{O}$. La production de données synthétiques *réalistes* repose sur la capacité à 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. Soit $\mathsf{Z}_1, \ldots, \mathsf{Z}_{t-1}$ les comptes passés, échantillonner 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) (cf. l. 2 de l'Alg. 1), puis de tirer une variable aléatoire sous la loi de Poisson d'intensité $\mathsf{p}_t^{(2)}$ (cf. 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 *via* une normalisation par n_t^ϕ (cf. l. 1 de l'Alg. 1).



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

FIGURE 2 : Haut : données synthétiques $\mathbf{Z}$ (en orange) et vérité terrain $\mathbf{R}^{\text{vrai}}$ (en noir). Bas : estimations de $\mathbf{R}$ par maximum de vraisemblance (MV, en gris), médian glissant puis vraisemblance pénalisée (VP, en bleu) et procédure jointe (J, en rouge) sur des données synthétiques de vérité terrain R^{vrai} (en noir).

Algorithme 1 : Génération de comptes 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 : $\mathbf{Z} = (\mathsf{Z}_1, \mathsf{Z}_2, \ldots, \mathsf{Z}_T) \in \mathbb{N}^T$

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

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

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

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

fin

Vérités terrains réalistes pour R et O. Obtenir des comptes de nouvelles infections synthétiques *réalistes*, nécessite que $\mathbf{R}^{\text{vrai}}$ et $\mathbf{O}^{\text{vrai}}$ imitent fidèlement à la fois la dynamique de propagation sous-jacente, et la dégradation des données par le processus de collecte. Or, ni le coefficient de reproduction R_t évoluant au cours du temps, ni la variable de corruption non stationnaire O_t, qui dépendent fortement de l'épidémie considérée, ne sont accessibles à une modélisation exhaustive. Afin de contourner cet écueil, des R_t, O_t réalistes sont construits à partir d'échantillons de données réelles extraits de la base JHU. Pour un pays et une période de temps donnés, les comptes d'infections reportés sont analysés au moyen de la stratégie jointe (7) pour des paramètres de régularisation fixés ; les estimées obtenues $\widehat{\mathbf{R}}^{\text{J}}$ et $\widehat{\mathbf{O}}^{\text{J}}$ fournissent alors des vérités terrains réalistes. En considérant différentes configurations, c.-à-d. différents réglages des paramètres de régularisation $(\lambda_{\mathsf{R}}, \lambda_{\mathsf{O}})$, cette méthode permet de construire des vérités terrains réalistes, ensuite combinées pour générer, *via* l'Alg. 1, des données synthétiques diversifiées (cf. Fig. 2a à 2d, haut).

4 Performances d'estimation

Optimisation convexe non lisse. Les estimées par vraisemblance pénalisée (5) et par la procédure jointe (7) sont calculées par les schémas primaux-duaux de Chambolle-Pock [4] développés dans [1, 8], recourant à des opérateurs proximaux pour la minimisation de fonctionnelles non-lisses. Les pas de descente sont choisis de la même manière que dans [1, 8]. Cependant, pour renforcer la robustesse de l'estimation vis-à-vis de dynamiques épidémiques très différentes et de différents niveaux de régularisation, l'algorithme est stoppé lorsque le maximum, sur les T composantes, des incréments relatifs du coefficient de reproduction entre deux itérations successives $\mathbf{R}^{[k]}$ et $\mathbf{R}^{[k+1]}$, lissé sur 500 itérations, est inférieur à 10^{-6}, c.-à-d. quand

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

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

Données synthétiques. Les performances des estimateurs décrits en Sec. 2 sont évaluées sur des données synthétiques produites *via* la procédure exposée en Sec. 3 : les comptes de nouvelles infections reportés en France entre le 1er novembre 2021 et le 3 août 2022 (cf. Fig. 1a) sont analysés au moyen de (7), avec deux réglages de paramètres de régularisation très différents[2], fournissant une collection de $\mathbf{R}^{\text{vrai}}$, $\mathbf{O}^{\text{vrai}}$. Ces vérités terrains sont ensuite combinées en quatre configurations :

C. I : beaucoup de changements de pente de R_t, beaucoup de valeurs non nulles de O_t (cf. Fig. 2a haut) ;
C. II : peu de changements de pentes de R_t, peu de valeurs non nulles de O_t (cf. Fig. 2b haut) ;
C. III : beaucoup de changements de pentes de R_t, peu de valeurs non nulles de O_t (cf. Fig. 2c haut) ;
C. IV : peu de changements de pentes de R_t, beaucoup de valeurs non nulles de O_t (cf. Fig. 2d haut).

Ces vérités terrains réalistes et diversifiées, avec des dynamiques de R_t et O_t similaires pour C. I et II et « croisées » pour C. III et IV, sont ensuite mises à profit pour générer des comptes de nouvelles infections synthétiques $\mathbf{Z}$ au moyen de l'Alg. 1. Pour chaque couple $(\mathbf{R}^{\text{vrai}}, \mathbf{O}^{\text{vrai}})$, c.-à-d. pour chaque C. I à IV, 20 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

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

quantifie l'adéquation entre l'estimée $\widehat{\mathbf{R}}$ et la vérité terrain $\mathbf{R}^{\text{vrai}}$. La qualité des estimées fournie par les méthodes en deux vs. une étape(s) dépend fortement du réglage des hyperparamètres, c.-à-d. du choix du seuil α dans (4), 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 paramètres espacés logarithmiquement.

[2] $(\lambda_{\mathsf{R}}, \lambda_{\mathsf{O}}) = (3{,}7 \cdot 10^5 ;\ 0{,}03)$ et $(\lambda_{\mathsf{R}}, \lambda_{\mathsf{O}}) = (5{,}3 \cdot 10^6 ;\ 0{,}75)$

	Max. Vrais. (MV)	Vrais. Pén. (VP)	Estim. Jointe (J)
SNR (dB)			
C. I	$0{,}90 \pm 0{,}03$	$20{,}65 \pm 0{,}06$	$\mathbf{35{,}68 \pm 0{,}39}$
C. II	$-2{,}31 \pm 0{,}06$	$26{,}55 \pm 0{,}15$	$\mathbf{48{,}56 \pm 0{,}90}$
C. III	$19{,}14 \pm 0{,}08$	$29{,}79 \pm 0{,}09$	$\mathbf{56{,}34 \pm 0{,}77}$
C. IV	$-2{,}32 \pm 0{,}01$	$18{,}79 \pm 0{,}02$	$\mathbf{31{,}06 \pm 0{,}20}$
Indice de Jaccard sur le laplacien (%)			
C. I	$0{,}79 \pm 0{,}00$	$57{,}63 \pm 1{,}33$	$\mathbf{84{,}28 \pm 1{,}48}$
C. II	$0{,}57 \pm 0{,}00$	$47{,}71 \pm 1{,}10$	$\mathbf{96{,}46 \pm 0{,}77}$
C. III	$2{,}28 \pm 0{,}01$	$73{,}76 \pm 0{,}73$	$\mathbf{98{,}03 \pm 0{,}90}$
C. IV	$0{,}67 \pm 0{,}00$	$13{,}58 \pm 0{,}65$	$\mathbf{76{,}79 \pm 0{,}57}$

TABLE 1 : SNR et indice de Jaccard pour trois estimateurs et quatre configurations, moyennées sur $N = 20$ réalisations.

Pour chaque estimateur et chaque configuration, le SNR maximal moyen et son intervalle de confiance gaussien à 95% sont reportés dans la Tab. 1, l. 3 à 6. Puis, pour la meilleure estimation au sens du SNR, l'indice de Jaccard[3] [6] entre le laplacien discret binarisé de $\widehat{\mathbf{R}}$ et celui de $\mathbf{R}^{\text{vrai}}$, avec une fenêtre $\mathbf{g}$ gaussienne d'écart-type 1 jour, reporté dans la Tab. 1, l. 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, les Fig. 2a à 2d, en bas, présentent 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. Quelle que soit la configuration C. I à IV, l'estimée $\mathbf{R}^{\mathsf{J}}$ y est systématiquement la plus fidèle à $\mathbf{R}^{\text{vrai}}$. L'étude quantitative du SNR sur 20 réalisations de chaque configuration, l. 3 à 6 du Tab. 1, montre la qualité nettement supérieure des estimées 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 la vraisemblance pénalisée, col. 3 du Tab. 1.

Robustesse de la procédure d'estimation. La diversité des données synthétiques C. I à IV permet de sonder la robustesse des estimateurs de R_t vis-à-vis de différentes combinaisons de dynamiques épidémiques et de niveaux de corruption des données. Pour toutes les configurations, l'indice de Jaccard entre le laplacien de l'estimée jointe et celui de la vérité terrain, Tab. 1 col. 8 à 11, varie entre 76% et 97%, attestant à la fois d'une détection efficace et d'une localisation correcte des changements de dynamique épidémique, plaidant pour l'utilisation de la procédure jointe [8] pour surveiller une épidémie, préférentiellement aux estimateurs de [5, 1].

Temps de calcul. Les meilleures performances d'estimation de la procédure jointe se font au prix d'un coût de calcul plus élevé (cf. Fig. 3). Pour des paramètres de régularisation équivalents pour les deux méthodes, le critère de convergence (8) atteint la précision requise (10^{-6}) en 20 fois plus d'itérations pour la procédure jointe (J, en rouge) que pour la méthode par vraisemblance pénalisée (VP, en bleu), et ce de manière robuste, comme l'attestent les barres d'erreur très étroites calculées sur $N = 20$ réalisations (cf. encart de droite, Fig. 3).

[3] Pour $\mathbf{X},\mathbf{Y} \in \{0,1\}^T$, $\mathsf{Jac}^{\mathbf{g}}(\mathbf{X},\mathbf{Y}) := \dfrac{\sum_{t=1}^{T} \sqrt{\mathsf{X}_t^{\mathbf{g}} \times \mathsf{Y}_t^{\mathbf{g}}}}{\sum_{t=1}^{T} \mathsf{X}_t^{\mathbf{g}} + \mathsf{Y}_t^{\mathbf{g}} - \sqrt{\mathsf{X}_t^{\mathbf{g}} \times \mathsf{Y}_t^{\mathbf{g}}}}$
où $\mathbf{X}^{\mathbf{g}}$ est la convolution de $\mathbf{X}$ avec la fenêtre de lissage $\mathbf{g}$.

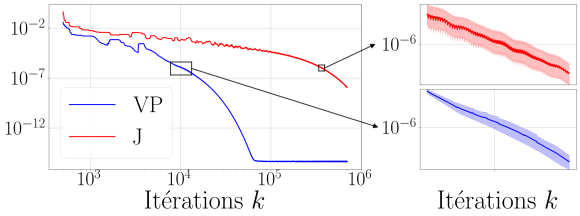


FIGURE 3 : Incréments relatifs (8) lors de l'estimation de R_t via VP (bleu) ou J (rouge) sur $N = 20$ réalisations de C. I.

5 Conclusion

Une procédure de génération de comptes d'infections synthétiques *réalistes*, reproduisant à la fois la dynamique épidémique et les erreurs dans le report des cas, a été proposée et mise à profit pour construire des données synthétiques diversifiées sur lesquelles plusieurs estimateurs du coefficient de reproduction ont été évalués et comparés *via* des simulations de Monte Carlo intensives, démontrant la supériorité d'une approche *jointe*, corrigeant les comptes et estimant le coefficient de reproduction simultanément. La génération de données synthétiques constitue une étape cruciale pour évaluer systématiquement les estimateurs d'indicateurs épidémiologiques. En outre, cette étude ouvre la voie à l'élaboration de méthodes automatiques et pilotées par les données pour le réglage des paramètres de régularisation de la procédure jointe.

Références

[1] P. ABRY, N. PUSTELNIK, S. ROUX, P. JENSEN, P. FLANDRIN, R. GRIBONVAL, C.-G. LUCAS, É. GUICHARD, P. BORGNAT et N. GARNIER : Spatial and temporal regularization to estimate COVID-19 reproduction number $\mathsf{R}(t)$: Promoting piecewise smoothness via convex optimization. *PLOS One*, 15(8):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] 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.

[7] 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.

[8] 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.

[9] 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.