Synthetic Spatiotemporal Covid19 Infection Counts to Assess Graph-Regularized Estimation of Multivariate Reproduction Numbers

Juliana Du, Barbara Pascal, Patrice Abry

Synthetic Spatiotemporal Covid19 Infection Counts to Assess Graph-Regularized Estimation of Multivariate Reproduction Numbers

Juliana Du
CNRS, ENS de Lyon
Laboratoire de physique
F-69007 Lyon, France
juliana.du@ens-lyon.fr

Barbara Pascal
Nantes Université, École Centrale Nantes
CNRS, LS2N, UMR 6004
F-44000 Nantes, France
barbara.pascal@cnrs.fr

Patrice Abry
CNRS, ENS de Lyon
Laboratoire de physique
F-69007 Lyon, France
patrice.abry@ens-lyon.fr

***Abstract*—The heavy impacts of Covid19 pandemic triggered significant research efforts to monitor the virus transmission. Several strategies were devised to estimate the reproduction number, quantifying the pandemic intensity, jointly along time and across territories while being robust to the limited quality of the reported Covid19 infection counts. However, because the true evolution of the pandemic intensity is unknown (lack of ground truth) estimation performance assessments and comparisons are impaired. The first contribution of this work is thus to design an original graph-based regularization strategy for the construction of spatially correlated synthetic ground truth reproduction number time series, further enabling the synthesis of realistic spatiotemporal infection counts. A second contribution consists in using such synthetic counts to compare the performance of several state-of-the-art reproduction number estimators, showing the superiority of multivariate estimation strategies compared to univariate procedures.**

I. Introduction

Context. The Covid19 pandemic resulted in a major sanitary crisis and still constitutes a significant threat. Devising efficient and balanced public health measures during epidemic outbreak requires real-time monitoring of the pathogen transmissibility, robust to limited quality data, triggering massive research efforts worldwide in computational epidemiology [1]–[8]. Recent works [5], [9] extended state-of-the-art epidemiological tools to monitor the pathogen transmissibility dynamics simultaneously on different territories, providing useful insight on the *spatial* dynamics of the pandemic[1].

Related work. During pandemic outbreaks, the intensity of the virus propagation is quantified through the instantaneous *reproduction number* R_t, defined as the expected number of infections generated by an individual infected on day t [2], [3], [10]. Standard reproduction number estimators [3] show severe accuracy drop when processing low quality new infection counts time series, such as the Covid19 incidence data, collected by national health authorities of 200+ countries and made available by Johns Hopkins University[2]. Indeed, Covid19 infection counts are corrupted by missing or outlier samples and pseudo-seasonalities. The imperious need for accurate, robust, real-time monitoring tools thus motivated the design of several regularized estimation procedures [5], [7]–[9], [11], [12], capable of estimating the reproduction number either independently per territory (e.g., for a country) or jointly for related territories (e.g., for counties or states of a same country).

Work supported by ANR-23-CE48-0009 "*OptiMoCSI*" . J. Du's PhD is funded by 80PRIME-2021 CNRS project "*CoMoDécartes*".

[1]https://www.covidatlas.eu/World/

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

However the assessment and comparison of their performance are impaired by the lack of knowledge of the true evolution of the pandemic intensity. A natural way to circumvent this issue consists in designing Covid19 synthetic reproduction number time series and resulting synthetic infection counts. This, however, raises two major difficulties. First, given a ground truth reproduction number, generating realistic infection counts requires to account for errors in the reporting process [9]. Second, designing synthetic reproduction number time series consistent with Covid19 pandemic is still an open problem in epidemiology. Agent-based models [13]–[16] simulate contacts and resulting infections between individuals; thus they require to specify the precise characteristics of the population and of the contagion process by fine calibration on consolidated data, which restricts their use to very restricted territory and time period [13], [14], [17]. Although compartmental models describe the epidemic at a coarser level [6], [18], their ability to produce realistic infection counts also depends decisively on fine calibration of their parameters. In contrast, the recent innovative procedure introduced in [19] relies only on the epidemiological model of [3] and on the corrupted counts model proposed in [9] to produce realistic synthetic infection counts accompanied with their ground truth reproduction number. Most attempts remained limited to a single territory, while pandemic monitoring naturally calls for multiple territories (or multivariate) synthesis and estimation, an issue addressed in the present work.

Goals, contributions and outline. The double aim of the present work is: i) to propose a methodology for the design of realistic infection counts on a set of connected territories, characterized by their multivariate reproduction number ground truth, and ii) to leverage these annotated synthetic counts to compare the accuracy of several state-of-the-art, univariate or multivariate, reproduction number estimators. Section II recalls the definition and illustrates the compared reproduction number estimators. Section III is devoted to the first contribution of this work: the generation of spatiotemporal synthetic infection counts, with a careful design of ground truth multivariate reproduction numbers encapsulating correlations between the epidemic temporal dynamics in connected territories. The second contribution of this work is reported in Section IV, where the designed annotated synthetic incidence data are leveraged to perform quantitative comparisons of the state-of-the-art reproduction number estimators performance, under several inter-territory connectivity structures and levels of correlation. Monte Carlo based simulations demonstrate the significant superiority of the multivariate strategy when epidemic dynamics are correlated amongst territories.

II. Reproduction number estimation procedures

A. Daily new infection count models and univariate estimators

Epidemiological model. The seminal model for viral epidemics proposed in [3] states that, conditionally to past daily new infection counts $\mathsf{Z}_1, \mathsf{Z}_2, \ldots, \mathsf{Z}_{t-1}$, the new infection count at day t, Z_t, follows a Poisson distribution, driven by a time-dependent intensity, p_t, yielding the following univariate log-likelihood:

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

The intensity $\mathsf{p}_t = \mathsf{R}_t \Phi^{\mathsf{Z}}_t$ is the product of the reproduction number R_t and a weighted sum of past counts $\Phi^{\mathsf{Z}}_t = \sum_{s=1}^{\tau_\phi} \phi(s) \mathsf{Z}_{t-s}$. The *serial interval function* ϕ, accounting for the random delay between primary and secondary infections, is well-modeled, for Covid19 pandemic, by a Gamma distribution with mean (resp. standard deviation) of 6.6 (resp. 3.5) days [20], [21], vanishing after $\tau_\phi = 26$ days.

Maximum likelihood estimator. Independent maximization of (1) for each t yields the Maximum Likelihood Estimator:

$$\widehat{\mathsf{R}}^{\mathsf{MLE}}_t = \mathsf{Z}_t / \Phi^{\mathsf{Z}}_t. \quad (2)$$

By design, $\widehat{\mathsf{R}}^{\mathsf{MLE}}_t$ (gray curve in Fig. 1, bottom plot), irrelevantly follows the high irregularities of the reported counts Z_t (black curve, top plot), and thus highly deviates from the smooth and slow temporal evolution expected from an epidemic intensity index [5], [9].

Bayesian estimator. To improve estimation accuracy, it was proposed in [3] to enhance temporal consistency by estimating the reproduction number R_t *as if* it were constant during the past τ days. Under a Gamma prior on R_t with shape and scale parameters $a = 1$ and $b = 5$ [3], the a posteriori mean estimator expresses as:

$$\widehat{\mathsf{R}}^{\Gamma}_{\tau,t} = \frac{\langle \mathsf{Z} \rangle_{\tau,t} + a}{\langle \Phi^{\mathsf{Z}} \rangle_{\tau,t} + 1/b}, \text{ where } \langle \mathsf{Z} \rangle_{\tau,t} = \sum_{s=0}^{\tau-1} \mathsf{Z}_{t-s}. \quad (3)$$

$\widehat{\mathsf{R}}^{\Gamma}_{\tau,t}$ for $\tau = 15$ days estimated from Covid19 infection counts for France is displayed in green in Fig. 1 (bottom). While showing a smoother temporal behavior compared to $\mathsf{R}^{\mathsf{MLE}}_t$ (in blue), $\widehat{\mathsf{R}}^{\Gamma}_{\tau,t}$ still suffers from under-reported counts on week-ends.

Time-regularized variational estimation. Along another line, it has been proposed in [5], [9] to ensure smooth and slow time evolution of the estimated R_t by augmenting the log-likelihood (1) with a regularization term favoring piecewise linearity in time, yielding the univariate variational estimator:

$$\widehat{\mathbf{R}}^{\mathsf{U}} = \underset{\mathbf{R} \in \mathbb{R}^T_+}{\operatorname{argmin}} \, \mathrm{D}_{\mathsf{KL}}(\mathbf{Z}|\mathbf{p}) + \lambda^{\mathsf{U}}_{\mathsf{L}} \|\mathsf{L}\mathbf{R}\|_1, \quad \mathsf{p}_t = \mathsf{R}_t \Phi^{\mathsf{Z}}_t \quad (4)$$

with $\mathrm{D}_{\mathsf{KL}}(\mathbf{Z}|\mathbf{p}) = -\sum_{t=1}^{T} \ln \mathbb{P}(\mathsf{Z}_t|\mathsf{Z}_1, \ldots, \mathsf{Z}_{t-1}; \mathsf{p}_t)$ the opposite of the log-likelihood of the epidemiological model (1), L the discrete second-order differential operator, $\|\cdot\|_1$ the ℓ_1-norm favoring sparsity and $\lambda^{\mathsf{U}}_{\mathsf{L}} > 0$ an hyperparameter controlling the level of enforced time regularity. The estimate $\widehat{\mathsf{R}}^{\mathsf{U}}_t$ (blue curve in Fig. 1, bottom plot) shows a regular and slowly-varying time evolution, far more realistic from an epidemiological viewpoint than the fast fluctuations observed on $\widehat{\mathsf{R}}^{\mathsf{MLE}}_t$, and to a lesser extent on $\widehat{\mathsf{R}}^{\Gamma}_t$.

B. Spatio-temporal counts and multivariate estimation

Space-time pandemic intensity evolution. While univariate estimation on a per country basis is relevant in epidemiology, notably during lockdown periods, the variation in space of the pandemic intensity, for instance across territories (*states*, *counties* of a same country) ruled by the same sanitary policy is also of interest. Yet, the joint space and time evolution of the pandemic intensity has mostly been considered in very constrained situations [16]. A highly flexible extension of the univariate model [3] was proposed in [5], [9], considering that infection counts in different territories, referred to as *counties* in the following, are *independent* random variables, each following the univariate model [3], hence assuming a multivariate intensity, $\mathsf{p}_{c,t} = \mathsf{R}_{c,t} \Phi^{\mathsf{Z}}_{c,t}$, $\Phi^{\mathsf{Z}}_{c,t} = \sum_{s=1}^{\tau_\phi} \phi(s) \mathsf{Z}_{c,t-s}$, with c labeling the multiple counties. Importantly, the serial interval function is considered unchanged across counties.

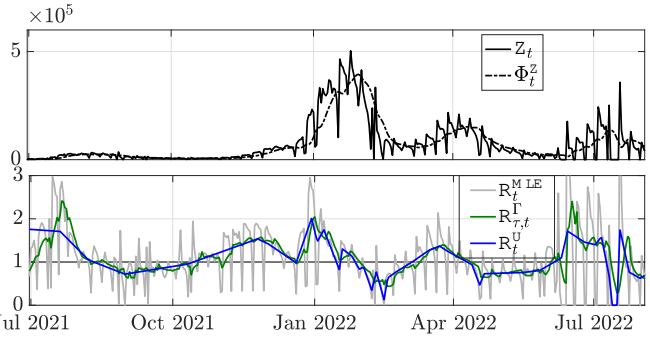

Fig. 1: **Covid19 pandemic in France** from Nov. 2021 to Aug. 2022. (top) Infection counts collected by National Health Authorities and made available by Johns Hopkins University.[2] (bottom) Univariate estimates of the reproduction number.

Multivariate estimators. To estimate the multivariate $\mathbf{R} \in \mathbb{R}^{C \times T}_+$ during T days jointly in C counties from multivariate infection counts $\mathbf{Z} \in \mathbb{N}^{C \times T}$, the regularizing functional of (4) has been augmented with a *spatial* regularization term, enforcing consistency between the estimated reproduction numbers in *connected* counties [5], [9], yielding the multivariate variational estimator :

$$\widehat{\mathbf{R}}^{\mathsf{M}} = \underset{\mathbf{R} \in \mathbb{R}^{C \times T}_+}{\operatorname{argmin}} \, \mathrm{D}_{\mathsf{KL}}(\mathbf{Z}|\mathbf{p}) + \lambda^{\mathsf{M}}_{\mathsf{L}} \|\mathsf{L}\mathbf{R}^\top\|_1 + \lambda^{\mathsf{M}}_{\mathsf{G}} \|\mathsf{G}\mathbf{R}\|_1, \quad (5)$$

where L computes the time second-order derivative independently on each time series, G encodes the connectivity between counties, and $\lambda^{\mathsf{M}}_{\mathsf{L}}, \lambda^{\mathsf{M}}_{\mathsf{G}} > 0$ are hyperparameters controlling the level of regularity in time and space enforced. In [5], [9], the counties are the French metropolitan departments, and the graph-based regularization reads

$$\|\mathsf{G}\mathbf{R}\|_1 = \sum_{t=1}^{T} \sum_{c \sim c'} |\mathsf{R}_{c,t} - \mathsf{R}_{c',t}| \quad (6)$$

where $c \sim c'$ if counties c and c' share a terrestrial border. Other connectivity patterns can be encoded designing different operators G.

III. Synthesis of multivariate infection counts

Principle. Quantitative assessment and comparison of univariate and multivariate reproduction number estimators crucially relies on the access to an evaluation dataset consisting of (multivariate) infection counts accompanied with their (multivariate) reproduction number ground truth. Recently, an efficient procedure to generate realistic synthetic Covid19 counts has been proposed [19].[3] First, an univariate reproduction number consistent with the Covid19 pandemic dynamics is constructed (Fig. 2, bottom); then it is used as ground truth to generate synthetic univariate infection counts under Model (1) (Fig. 2, top plot) through an original strategy specifically devised to reproduce the low quality of real-world Covid19 counts (Fig. 1, top plot).

The present work proposes to generate multivariate synthetic infection counts by, first, designing synthetic multivariate reproduction

[3] https://github.com/juliana-du/Covid-R-estim *(in english)*

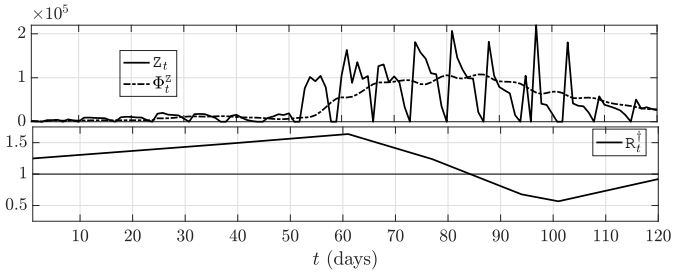

Fig. 2: **Synthetic univariate Covid19 data from [19].** (top) Realistic synthetic infection counts under Model (1). (bottom) Ground truth piecewise linear underlying reproduction number.

numbers encapsulating correlations in the pandemic temporal dynamics between connected territories, and then leveraging the synthesis procedure from [19] to synthesize resulting realistic multivariate infection counts. The major bottleneck lies in the construction of synthetic *multivariate* reproduction numbers properly implementing correlations between connected territories while remaining consistent with a Covid19 pandemic-like temporal dynamic.

Synthetic multivariate reproduction numbers. To generate synthetic Covid19 data along T consecutive days, in C territories, abstractly representing *connected counties*, the present work proceeds in two steps. First, C synthetic reproduction number time series, consistent with Covid19 pandemic temporal dynamics, are independently constructed following [19] and concatenated into a $C \times T$ matrix $\mathbf{R}^\dagger$. Second, correlations between connected counties are implemented by regularizing these raw synthetic multivariate reproduction numbers *across counties* through

$$\mathbf{R}^\star(\mathbf{R}^\dagger;\delta) = \underset{\mathbf{R}\in\mathbb{R}^{C\times T}}{\operatorname{argmin}} \|\mathbf{R}^\dagger - \mathbf{R}\|_2^2 + \delta\|\mathsf{G}\mathbf{R}\|_2^2 \tag{7}$$

where G encodes the pairwise differences between connected counties, and $\delta > 0$ controls the level of regularization. In the limit $\delta \to 0$, the raw reproduction numbers remain unchanged though (7), while as $\delta \to \infty$, the components of the resulting multivariate reproduction numbers tends to be all equal. Interpreting the counties as vertices and connections between them as edges, the minimization (7) amounts to perform a Tikhonov graph-smoothing procedure [22], widely used in graph signal processing [23]. Problem (7) has an explicit solution [22]

$$\mathbf{R}^\star(\mathbf{R}^\dagger;\delta) = (\mathrm{Id} + 2\delta\mathsf{G})^{-1}\mathbf{R}^\dagger. \tag{8}$$

Further, G acting separately on each time t, (8) reduces to the resolution of T linear systems and can thus be solved very efficiently.

Evaluation dataset construction. Considering $T = 120$ days and $C = 5$ counties, assumed to follow the `Line` graph connectivity structure (cf. Fig. 3, left), consistent raw synthetic multivariate reproduction numbers are first built using [19],[3] and displayed in Fig. 5 (leftmost plot). Then, the proposed inter-county regularization (7) is applied with four logarithmically spaced smoothing parameters from $\delta_{\mathrm{I}} = 0.01$ (low inter-county correlation, Fig. 5, second plot) to $\delta_{\mathrm{IV}} = 1.49$ (high inter-county correlation, Fig. 5, rightmost plot).

The level of correlation enforced is monitored through the magnitude of the inter-county regularization term in (7), $\|\mathsf{G}\mathbf{R}^\star\|_2^2$, displayed for a wide range of δ in Fig. 4. The regularization term decreases smoothly from its maximal value, corresponding to independent raw reproduction numbers, to zero for multivariate reproduction numbers with all components equal. The four regularization parameters δ_{I}, δ_{II}, δ_{III}, δ_{IV} are chosen to span regularly increasing levels of correlation.

Fig. 5 shows that, first, for the four inter-county correlation levels, the temporal dynamics of the synthetic multivariate reproduction

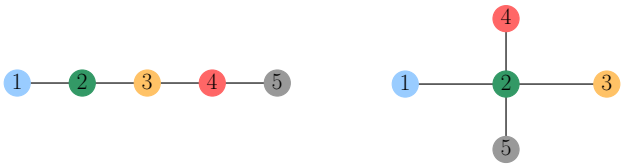

Fig. 3: **Graph-encoding of connectivity structure.** (left) `Line` graph and (right) `Hub` graph.

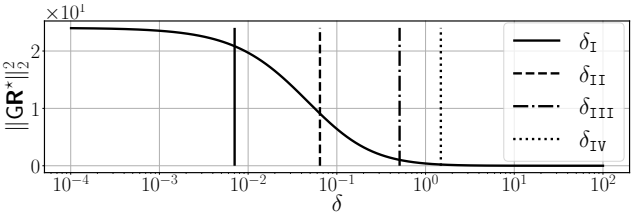

Fig. 4: **Inter-county correlation level.** The vertical lines indicates the correlation levels illustrated in Fig. 5 and explored in Sec. IV.

numbers are consistent with Covid19 pandemic: for each county, the reproduction number is varying slowly in time, in a piecewise linear manner reflecting the sudden changes in the virus spread observed during the outbreak of the pandemic, e.g., at the beginning of an epidemic wave. Second, the reproduction numbers in connected counties tend to synchronize as the enforced correlation level increases (Fig. 5, from left to right). The procedure provides similar results for counties following the `Hub` graph connectivity pattern (cf. Fig. 3, right plot).

The designed synthetic multivariate reproduction numbers can then be used as ground truth to generate multiple realizations of realistic synthetic infection counts via the procedure proposed in [19], constituting a rich evaluation dataset leveraged in Sec. IV to assess and compare univariate and multivariate reproduction number estimators.

IV. Compared estimation performance

A. Experiment set up

Performance assessment. The accuracy of the univariate and multivariate reproduction number estimators presented in Sec. II is measured through the *averaged normalized Mean Square Error*:

$$\mathsf{MSE} := \frac{1}{C}\sum_{c=1}^{C} \frac{\sum_{t=1}^{T}\left(\widehat{\mathsf{R}}_{c,t} - \mathsf{R}^\star_{c,t}\right)^2}{\sum_{t=1}^{T}\left(\mathsf{R}^\star_{c,t}\right)^2} \tag{9}$$

where $\widehat{\mathsf{R}}_{c,t}$ (resp. $\mathsf{R}^\star_{c,t}$) denotes the estimated (resp. ground truth) reproduction number in county c at day t. It consists of the average over the C considered counties of the univariate normalized quadratic errors along the T consecutive days.

The performance of the four estimators are assessed and compared over the evaluation datasets of Sec. III, considering both a *sparse* connectivity structure, encoded in the `Line` graph (cf. Fig. 3, left), and a *dense* connectivity structure, encoded in the `Hub` graph (cf. Fig. 3, right). For each connectivity structure, the four levels of inter-county correlation described in Sec. III are considered, together with the case of *independent* counties corresponding to $\delta = 0$.

Finally, for each reproduction number ground truth, characterized by a connectivity structure and an inter-county correlation level, performance are computed over $N = 15$ independent realizations of synthetic infection counts and mean MSE accompanied with 95% Gaussian confidence intervals are reported.

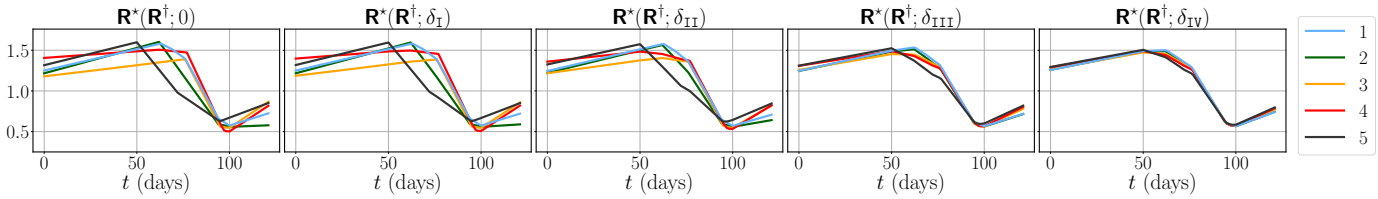

Fig. 5: **Synthetic multivariate reproduction numbers with different inter-county correlation levels.** Five abstract counties are considered, along 120 days; the enforced connectivity structure is encoded in the `Line` graph (Fig. 3, left). According to (7), the level of correlation is controlled by δ, ranging from $\delta = 0$ (no correlation) to $\delta_{\mathtt{IV}} = 1.49$ (high correlation). Intermediate correlation levels are indicated in Fig. 4.

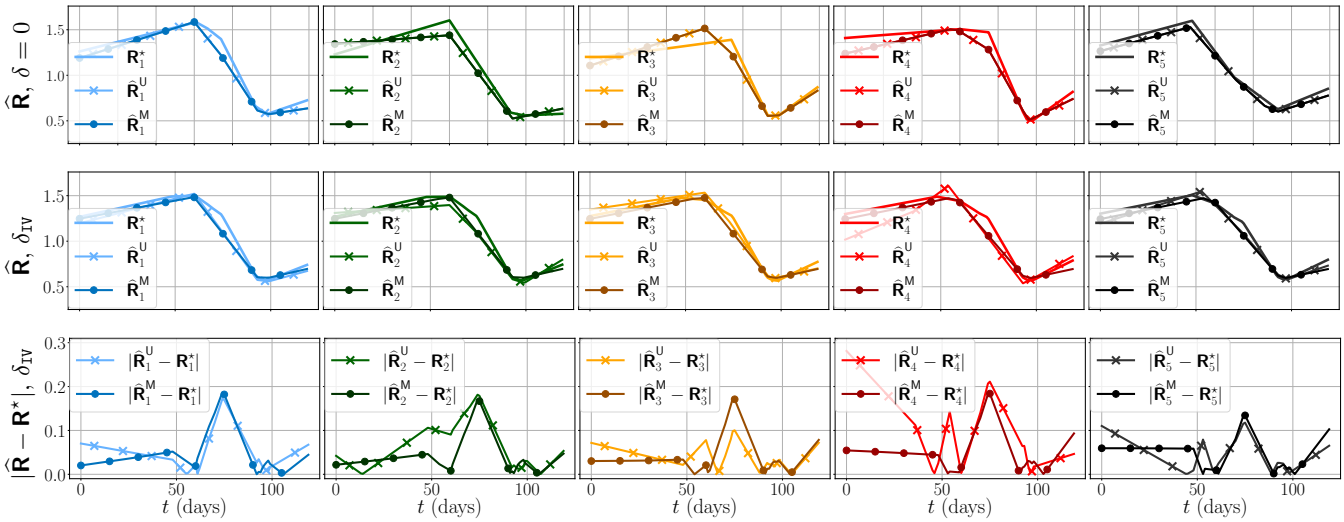

Fig. 6: **Comparison of univariate and multivariate variational estimators.** Illustration for the `Line` connectivity structure, for independent (first row) and highly correlated (second and third rows) counties. (first and second rows) Univariate (o-mark, faded color) and multivariate (x-mark, dark color) reproduction number estimates compared to ground truth (plain line, faded color). (third row) Absolute error w.r.t. ground truth of the univariate (x-mark, faded color) and multivariate (o-mark, dark color) estimates for high correlation level.

Optimization. The minimization problems defining the regularized univariate (4) and multivariate (5) estimators are solved using the Chambolle-Pock primal-dual scheme [24], which handles *nonsmooth* convex objective functions by resorting to *proximity operators* [25]. Following [5], descent steps are chosen so as to saturate the convergence condition. Further, convergence is monitored using the robust criterion proposed in [19], consisting in smoothed normalized increments of the reproduction number iterates; the algorithm is stopped whenever these increments fall below the precision $\epsilon = 10^{-7}$, or when the total number of iterations reaches $K_{\max} = 7 \cdot 10^5$. The authors have made publicly available a PYTHON implementation of both the univariate and multivariate reproduction number estimators, including the robust iterate increment-based stopping criterion.[4]

Estimator hyperparameter selection. The accuracy of the estimators described in Sec. II crucially relies on the fine-tuning of their *hyperparameters*. For each estimator, each connectivity structure and each inter-county correlation level, the reported MSE corresponds to the *optimal* hyperparameter choice, i.e., the hyperparameter setting reaching the smallest MSE. In practice, the integer parameter τ involved in the Bayesian estimator (3) is optimized over 20 regularly spaced integers ranging from 1 to 50. The regularization parameter(s) $\lambda^{\mathsf{U}}_{\mathsf{L}} > 0$ (resp. $\lambda^{\mathsf{M}}_{\mathsf{L}}, \lambda^{\mathsf{M}}_{\mathsf{G}} > 0$) of the univariate (resp. multivariate) estimator (4) (resp. (5)) are selected minimizing the MSE over a logarithmic grid of 20 (resp. 20×20) parameters.

[4]https://github.com/juliana-du/Covid-R-estim

B. Results

Table I shows that, whatever the connectivity structure and inter-county correlation level, the regularized univariate and multivariate estimates (third and fourth columns) significantly outperform the MLE and Bayesian estimators (first and second columns) reaching MSEs smaller by one order of magnitude. This supports the empirical observations drawn from real Covid19 data in Sec. II (cf. Fig. 1) and demonstrates that regularization is key to accurately estimating epidemiological indicators from low quality Covid19 incidence data.

While for low inter-county correlation levels, the univariate and multivariate estimators achieve equivalent performance (cf. Tab. I, third, fourth, tenth, eleventh rows), for medium to high correlation levels, the multivariate estimate yields more accurate reproduction number estimates (cf. Tab. I, fifth, sixth, twelfth, thirteenth rows). The superiority of the multivariate estimator is illustrated in Fig. 6 for the `Line` graph. On the first row, for *independent* counties, the univariate (x-mark, fade color) and multivariate (o-mark, dark color) estimates perfectly coincide and approach equally closely the ground truth (plain line, faded color). In contrast, for highly correlated counties, in Fig. 6 second and third rows, the multivariate estimate (o-mark, dark color) better fits the ground truth (plain line, faded color) than the univariate estimate (o-mark, faded color). Tab. I and Fig. 6 demonstrate the significant advantage of the multivariate estimator, leveraging the known connectivity structure, w.r.t. unvariate strategies as soon as the inter-county correlation reaches a medium level.

	$\widehat{\mathbf{R}}^{\mathsf{MLE}}$	$\widehat{\mathbf{R}}^{\Gamma}$	$\widehat{\mathbf{R}}^{\mathsf{U}}$	$\widehat{\mathbf{R}}^{\mathsf{M}}$
`Line` connectivity structure				
$\delta = 0$	19.44 ± 0.49	2.66 ± 0.03	$\mathbf{0.40 \pm 0.01}$	$\mathbf{0.39 \pm 0.01}$
$\delta_{\mathtt{I}}$	19.34 ± 0.51	2.63 ± 0.04	$\mathbf{0.41 \pm 0.02}$	$\mathbf{0.40 \pm 0.02}$
$\delta_{\mathtt{II}}$	21.29 ± 1.08	2.62 ± 0.05	0.47 ± 0.02	$\mathbf{0.38 \pm 0.01}$
$\delta_{\mathtt{III}}$	25.59 ± 0.74	2.70 ± 0.05	0.48 ± 0.01	$\mathbf{0.29 \pm 0.01}$
$\delta_{\mathtt{IV}}$	27.43 ± 0.96	2.74 ± 0.05	0.47 ± 0.01	$\mathbf{0.24 \pm 0.01}$
`Hub` connectivity structure				
$\delta = 0$	19.50 ± 0.52	2.67 ± 0.03	$\mathbf{0.41 \pm 0.02}$	$\mathbf{0.38 \pm 0.01}$
$\delta_{\mathtt{I}}$	19.52 ± 0.39	2.67 ± 0.03	$\mathbf{0.40 \pm 0.02}$	$\mathbf{0.37 \pm 0.02}$
$\delta_{\mathtt{II}}$	19.80 ± 0.66	2.61 ± 0.03	0.42 ± 0.02	$\mathbf{0.37 \pm 0.02}$
$\delta_{\mathtt{III}}$	22.94 ± 0.84	2.61 ± 0.04	0.48 ± 0.01	$\mathbf{0.34 \pm 0.01}$
$\delta_{\mathtt{IV}}$	24.55 ± 0.61	2.64 ± 0.03	0.48 ± 0.01	$\mathbf{0.29 \pm 0.00}^*$

TABLE I: **Compared** MSE **estimation performance.** For the sake of readability the reported figures correspond to $10^2 \times$ MSE. Two connectivity structures (cf. Fig. 3), `Line` (top rows) and `Hub` (bottom rows), are explored, and, for each, five inter-county correlation levels are considered. Performance are averaged over $N = 15$ realizations of synthetic infection counts and accompanied with 95% Gaussian confidence intervals. **confidence interval of order* 10^{-3}.

V. Conclusions and perspectives

A procedure has been designed permitting the joint synthesis of realistic multivariate Covid19 data on connected territories, consisting in multivariate synthetic infection counts accompanied with their multivariate reproduction ground truth. Correlations in the epidemiological temporal dynamics between connected territories are enforced through a Tikhonov graph-smoothing regularization, allowing significant versatility in both the choice of a connectivity structure and of epidemiological dynamics. The obtained evaluation datasets are then leveraged to compare quantitatively several univariate and multivariate state-of-the-art reproduction number estimators via intensive Monte Carlo simulations, demonstrating the superiority in terms of MSE of the multivariate graph-regularized estimator, jointly assessing time and space pandemics dynamics.

The realistic epidemiological evaluation datasets will further permit the assessment of advanced estimators, explicitly accounting for the low quality of Covid19 data [9]. It also paves the way toward the design of a procedure simultaneously estimating the multivariate reproduction number and the inter-territory connectivity structure [26].

A new release of the companion toolbox of [19] is publicly available,[5] including the generation of *multivariate* synthetic Covid19 data and the graph-regularized reproduction number estimator.

References

[1] A. Flahault, "COVID-19 cacophony: is there any orchestra conductor?" *The Lancet*, vol. 395, no. 10229, p. 1037, 2020.

[2] O. Diekmann, J. A. P. Heesterbeek, and 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.*, vol. 28, pp. 365–382, 1990.

[3] A. Cori, N. Ferguson, C. Fraser, and S. Cauchemez, "A new framework and software to estimate time-varying reproduction numbers during epidemics," *Am. J. Epidemiol.*, vol. 178, no. 9, pp. 1505–1512, 2013.

[4] R. N. Thompson, J. E. Stockwin, R. D. Van Gaalen, J. A. Polonsky, Z. N. Kamvar, P. A. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, J. Lessler, S. Cauchemez, and A. Cori, "Improved inference of time-varying reproduction numbers during infectious disease outbreaks," *Epidemics*, vol. 29, p. 100356, 2019.

[5] P. Abry, N. Pustelnik, S. Roux, P. Jensen, P. Flandrin, R. Gribonval, C.-G. Lucas, É. Guichard, P. Borgnat, and N. Garnier, "Spatial and temporal regularization to estimate COVID-19 reproduction number $\mathsf{R}(t)$: Promoting piecewise smoothness via convex optimization," *PLoS One*, vol. 15, no. 8, p. e0237901, 2020.

[5] https://github.com/juliana-du/Covid-R-estim

[6] IHME COVID-19 Forecasting Team and S. I. Hay, "COVID-19 scenarios for the United States," *medRxiv*, 2020.

[7] R. K. Nash, P. Nouvellet, and A. Cori, "Real-time estimation of the epidemic reproduction number: Scoping review of the applications and challenges," *PLoS Digit. Health*, vol. 1, no. 6, p. e0000052, 2022.

[8] R. K. Nash, S. Bhatt, A. Cori, and P. Nouvellet, "Estimating the epidemic reproduction number from temporally aggregated incidence data: A statistical modelling approach and software tool," *PLoS Comput. Biol.*, vol. 19, no. 8, p. e1011439, 2023.

[9] B. Pascal, P. Abry, N. Pustelnik, S. Roux, R. Gribonval, and 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.*, vol. 70, pp. 2859–2868, 2022.

[10] J. Wallinga and P. Teunis, "Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures," *Am. J. of Epidemiol.*, vol. 160, no. 6, pp. 509–516, 09 2004.

[11] S. Bhatia, J. Wardle, R. K. Nash, P. Nouvellet, and A. Cori, "Extending EpiEstim to estimate the transmission advantage of pathogen variants in real-time: SARS-CoV-2 as a case-study," *Epidemics*, p. 100692, 2023.

[12] G. Fort, B. Pascal, P. Abry, and N. Pustelnik, "Covid19 Reproduction Number: Credibility Intervals by Blockwise Proximal Monte Carlo Samplers," *IEEE Trans. Signal Process.*, vol. 71, pp. 888–900, 2023.

[13] S. L. Chang, N. Harding, C. Zachreson, O. M. Cliff, and M. Prokopenko, "Modelling transmission and control of the COVID-19 pandemic in Australia," *Nature communications*, vol. 11, no. 1, p. 5710, 2020.

[14] N. Popper, M. Zechmeister, D. Brunmeir, C. Rippinger, N. Weibrecht, C. Urach, M. Bicher, G. Schneckenreither, and A. Rauber, "Synthetic reproduction and augmentation of COVID-19 case reporting data by agent-based simulation," *medRxiv*, pp. 2020–11, 2020.

[15] P. C. Silva, P. V. Batista, H. S. Lima, M. A. Alves, F. G. Guimarães, and R. C. Silva, "COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions," *Chaos Solit. Fractals*, vol. 139, p. 110088, 2020.

[16] I. Mahmood, H. Arabnejad, D. Suleimenova, I. Sassoon, A. Marshan, A. Serrano-Rico, P. Louvieris, A. Anagnostou, S. JE Taylor, D. Bell, and D. Groen, "FACS: a geospatial agent-based simulator for analysing COVID-19 spread and public health measures on local regions," *J. Simul.*, vol. 16, no. 4, pp. 355–373, 2022.

[17] G. Rykovanov, S. Lebedev, O. Zatsepin, G. Kaminskii, E. Karamov, A. Romanyukha, A. Feigin, and B. Chetverushkin, "Agent-based simulation of the COVID-19 epidemic in Russia," *Her. Russ. Acad. Sci.*, vol. 92, no. 4, pp. 479–487, 2022.

[18] N. Bannur, V. Shah, A. Raval, and J. White, "Synthetic Data Generation for Improved covid-19 Epidemic Forecasting," *medRxiv*, pp. 2020–12, 2020.

[19] J. Du, B. Pascal, and P. Abry, "Compared performance of Covid19 reproduction number estimators based on realistic synthetic data," in *GRETSI'23 XXIXème Colloque Francophone de Traitement du Signal et des Images*, Grenoble, France, Aug. 28 - Sept. 1 2023.

[20] D. Cereda, M. Tirani, F. Rovida, V. Demicheli, M. Ajelli, P. Poletti, F. Trentini, G. Guzzetta, V. Marziano, A. Barone *et al.*, "The early phase of the COVID-19 outbreak in Lombardy, Italy," *Preprint arXiv:2003.09320*, 2020.

[21] F. Riccardo, M. Ajelli, X. D. Andrianou, A. Bella, M. Del Manso, M. Fabiani, S. Bellino, S. Boros, A. M. Urdiales, V. Marziano *et al.*, "Epidemiological characteristics of COVID-19 cases and estimates of the reproductive numbers 1 month into the epidemic, Italy, 28 January to 31 March 2020." *Euro Surveillance*, 2020.

[22] A. Elmoataz, O. Lezoray, and S. Bougleux, "Nonlocal Discrete Regularization on Weighted Graphs: A Framework for Image and Manifold Processing," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1047–1060, 2008.

[23] Y. Pilavcı, P.-O. Amblard, S. Barthelmé, and N. Tremblay, "Graph Tikhonov Regularization and Interpolation Via Random Spanning Forests," *IEEE Trans. Signal Inf. Process.*, vol. 7, p. 359–374, 2021.

[24] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.*, vol. 40, no. 1, pp. 120–145, 2011.

[25] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[26] X. Dong, D. Thanou, P. Frossard, and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Sig. Proc.*, vol. 64, no. 23, pp. 6160–6173, 2016.