A Flexible Smoother Adapted to Censored Data With Outliers and Its Application to SARS-CoV-2 Monitoring in Wastewater

Marie Courbariaux¹*, Nicolas Cluzel¹, Siyun Wang¹, Vincent Maréchal², Laurent Moulin³, Sébastien Wurtzer³, Obépine Consortium, Jean-Marie Mouchel⁴, Yvon Maday⁵ and Grégory Nuel¹,⁶

¹Maison des Modélisations Ingénieries et Technologies (SUMMIT), Sorbonne Université, Paris, France, ²INSERM, Centre de Recherche Saint-Antoine, Sorbonne Université, Paris, France, ³Eau de Paris, Département de Recherche, Développement et Qualité de l’Eau, Ivry sur Seine, France, ⁴CNRS, EPHE, Sorbonne Université, UMR 7619 Metis, e-LTER Zone Atelier Seine, Paris, France, ⁵Sorbonne Université, CNRS, Université de Paris, Laboratoire Jacques-Louis Lions (LJLL), Paris, France, ⁶Stochastics and Biology Group, Probability and Statistics (LPSM, CNRS 8001), Sorbonne University, Paris, France

A sentinel network, Obépine, has been designed to monitor SARS-CoV-2 viral load in wastewaters arriving at wastewater treatment plants (WWTPs) in France as an indirect macro-epidemiological parameter. The sources of uncertainty in such a monitoring system are numerous, and the concentration measurements it provides are left-censored and contain outliers, which biases the results of usual smoothing methods. Hence, the need for an adapted pre-processing in order to evaluate the real daily amount of viruses arriving at each WWTP. We propose a method based on an auto-regressive model adapted to censored data with outliers. Inference and prediction are produced via a discretized smoother which makes it a very flexible tool. This method is both validated on simulations and real data from Obépine. The resulting smoothed signal shows a good correlation with other epidemiological indicators and is currently used by Obépine to provide an estimate of virus circulation over the watersheds corresponding to about 200 WWTPs.

Keywords: measurement error, smoothing algorithm, outliers, censored data, SARS-CoV-2, autoregressive model

1. INTRODUCTION

A sentinel network, Obépine (Observatoire épidémiologique dans les eaux usées)¹, has been designed to monitor SARS-CoV-2 viral load in wastewaters arriving at about 200² wastewater treatment plants (WWTPs) in France as an indirect macro-epidemiological parameter. This survey was initiated in March 2020 in WWTPs of the Greater Paris, during the first epidemic wave in France [1]. In this system, samples are currently typically taken two times a week, resulting in missing data on most days.

The sources of uncertainty in such monitoring system are numerous [2, 3]. These include, for example:

¹https://www.reseau-obepine.fr/
²In November 2021, representing more that one third of the French population.
• sampling variance (although sampling at the WWTPs is integrated over 24 h),
• variance from the Reverse Transcriptase (RT) and quantitative or digital Polymerase Chain Reaction (qPCR or dPCR) processes used to measure the concentration of SARS-CoV-2 RNA genes in the samples,
• uncertainties on other analytical steps in the labs such as virus concentration, genome extraction, and the presence of inhibitors.

Using the raw viral load measurements right away to monitor the pandemic can, therefore, be misleading, as a large variation in the measured concentration can either be due to a real variation in virus concentration or to a quantification error. Therefore, these data need pre-processing in order to provide an accurate estimate of the real daily amount of virus arriving at each WWTP and to evaluate the uncertainty on this estimate as also underlined by [3], who tested, for this purpose, 13 concurrent smoothing methods on data from 4 Austrian WWTPs.

A solution to estimate the underlying true concentrations (from which the noisy measurements are derived) is to exploit the time dependence of the successive measurements. This temporal dependence is indeed due to the (never observed) underlying process, while the measurement noises from one time step to another are assumed to be independent. A natural algorithm by their conditional expectation with the current smoothing adaptation of Kalman smoothing robust to such outliers (and also to abrupt changes of the underlying state and to switching linear regressions). This is mainly done by modifying the loss function to be minimized. However, these adaptations do not apply to the censoring scheme.

In this study, we here focus on the one-dimensional setting and propose a Smoother adapted to Censored data with OutLIers (SCOU) that answers both the censoring and the outliers detection problems through discretization of the state space of the monitored quantities and more generally permits a very high degree of modeling flexibility.

The proposed model and its implementation are described in section 2. Section 3 provides an illustration and validation of our approach using numerical simulations. Section 4 gives an example of utilization of data from the Obépine network.

2. THE PROPOSED METHOD
2.1. An Auto-Regressive Model
Our method is based on the following state-space model:

\[
\begin{align*}
X_t &= \eta X_{t-1} + \delta + \sigma \xi_{X,t} \\
O_t &\sim B(p) \\
(Y^*_t | O_t = 0) &= X_t + \tau \epsilon_{Y,t} \\
(Y^*_t | O_t = 1) &\sim U([a, b]) \\
Y_t &= \max(Y^*_t, \ell) \\
(\xi_{X,t}, \epsilon_{Y,t})^{i,d} &\sim N(0, I),
\end{align*}
\]

where:

• \( t \) is the time index (ranging from 1 to \( n \)),
• \( X_t \in \mathbb{R} \) is the real quantity at time \( t \) and \( X = (X_t)_{t \in \{1, \ldots, n\}} \) is the vector of real quantities (to be recovered),
• \( Y_t \in \mathbb{R} \) is the measurement at time \( t \). \( Y_t \) is generally only partially observed. We note \( T \subset \{1, \ldots, n\} \) the set of \( t \) at which \( Y_t \) is observed. \( Y = (Y_t)_{t \in T} = Y_T \) is the vector of measurements. \( Y^* \) is an accessory latent variable corresponding to a non-censored version of \( Y \). I is the identity matrix.
• \( \eta \in \mathbb{R}, \delta \in \mathbb{R}, \sigma \in \mathbb{R}^+, p \in [0, 1] \) and \( \tau \in \mathbb{R}^+ \) are parameters (to be estimated).
• \( \ell \) is the threshold below which censorship applies.\(^3\)
• \( O_t \in \{0, 1\} \) is, for any \( t \in T \), the indicator variable of the event “\( Y^*_t \) is an outlier”. \( O = (O_t)_{t \in T} \) stands for the Bernoulli distribution of parameter \( p \) and \( U([a, b]) \) for the Uniform distribution on the interval \([a, b]\). \( a \) and \( b \) have to be chosen, they can for example correspond to quantiles (respectively, very close to 0 and very close to 1) of the empirical marginal distribution of \( Y \).

\(^3\)In practice, \( \ell \) can vary from one day to another, for instance if one works on quantities that correspond to the multiplication of concentrations (with a detection limit) by a fluctuating volume. This can be taken into account within our method with no additional cost.
Markov chains with $D$ states, the $D$ hidden states being the $D$ possible values for each $X$.

The transition matrix, $\pi$, is calculated as follows:

$$
\pi(x,\gamma) = \mathbb{P}(X_{t+1} = \gamma' | X_t = x) = \frac{f_N(\gamma'; \eta x + \delta, \sigma^2)}{\sum_{\gamma'' \in X} f_N(\gamma''; \eta x + \delta, \sigma^2)},
$$

where $f_N(\cdot; \mu, \sigma^2)$ is the Gaussian distribution function with mean $\mu$ and variance $\sigma^2$.

Let $e_i(x) = \mathbb{P}(Y_t = y_i | X_t = x)$. For all $x \in \mathcal{X}$, $e_i(x) = 1$ if $t \notin \mathcal{T}$. If $t \in \mathcal{T}$:

$$
e_i(x) = (1 - p) \left( \mathbb{I}_{[y_t > \ell]} f_N(y_t; x, \tau^2) + \mathbb{I}_{[y_t \leq \ell]} F_N(\ell; x, \tau^2) \right) + p \left( \mathbb{I}_{[y_t > \ell]} \frac{1}{b-a} + \mathbb{I}_{[y_t \leq \ell]} \frac{\ell - a}{b-a} \right),
$$

where $\mathbb{I}_{[A=a]}$ is the indicator variable for the event $[A=a]$ and $F_N(\cdot; \mu, \sigma^2)$ is the Gaussian cumulative distribution function with mean $\mu$ and variance $\sigma^2$.

The $F$ and $B$ quantities, defined in Equations (2) and (3), can then be calculated recursively from $\pi$ and $e$:

- $F_1(x) = \frac{\eta(x)}{\mathcal{R}}$ and, for $t \in \{2, \ldots, n\}$,

$$
F_t(x') = \sum_{x \in \mathcal{X}} F_{t-1}(x) \pi(x, x') e_t(x).
$$

- $B_0(x) = 1$ (by convention) and, for $t \in \{n, \ldots, 2\}$,

$$
B_{t-1}(x) = \sum_{x' \in \mathcal{X}} \pi(x, x') e_t(x') B_t(x').
$$

Numerical tricks like logarithmic re-scaling are moreover used to calculate $e$, $\pi$ and the $F$ and $B$ quantities in order to avoid underflow and overflow problems.

2.2.3. Learning
The parameters $\eta, \delta, \sigma, \tau,$ and $p$ maximize the likelihood, which is given by $\sum x \in \mathcal{X} F_n(x)$, could be obtained with the Expectation-Maximization algorithm. However, it is made challenging by the discretization and renormalization. This is why numerical optimization is preferred. The parameters are more precisely obtained by the numerical optimization of Nelder and Mead (as implemented in the R function optim).

2.2.4. Prediction
For any $t$, the marginal distribution (mean, SD) of $(X_t | Y_T = y_T)$, is given by $P_t(x) B_t(x) = \mathbb{P}(X_t = x, Y_T = y_T) \propto \mathbb{P}(X_t = x | Y_T = y_T)$ (proof left to the reader) that is, for any $x \in \mathcal{X}$, the probability that $X_t$ (once discretized) takes the value $x$ conditionally to $Y_T = y_T$.

2.2.5. Simulations
We similarly show that:

$$
\mathbb{P}(X_{t-1} = x, X_t = x', Y_T = y_T) = F_{t-1}(x) \pi(x, x') e_t(x') B_t(x').
$$

4 Hint: decompose $(Y_T = y_T) = (Y_{[n-\ldots]} = y_T) \cap Y_{[n-\ldots]} = (Y_{[n-\ldots]} = y_T)$, condition on $X_t = x$ and use the conditional independence.
Thus, to simulate from $\mathbb{P}(X_{[1,...,n]}|Y_T = y_T)$, one can process sequentially, simulating $x_t$ with $\mathbb{P}(X_t = x|Y_T = y_T) \propto F_1(x)B_1(x)$, then, for $t \in [2, ..., n]$:

$$\mathbb{P}(X_t = x'|X_{t-1} = x, Y_T=y_T) = \frac{\pi(x,x')\delta_t(x')B_t(x')}{B_{t-1}(x)}.$$ 

2.2.6. Outliers Detection

The probability for each observation $Y_t$ to be an outlier (knowing $Y_T = y_T$ and for the a priori probability of $p$) can be computed as follows:

$$\mathbb{P}(O_t = 1|Y_T = y_T) = \frac{p\mathbb{P}(Y_T = y_T|O_t = 1)}{\mathbb{P}(Y_T = y_T)} = \frac{p\sum_x \sum_{x'} F_{t-1}(x)\pi(x,x')B_t(x') \{1_{[y_T, \infty)}(y_T - a) + \mathbb{I}_{[0, y_T]}(b - y_T)}{\sum_x \sum_{x'} F_{t-1}(x)\pi(x,x')B_t(x')\delta_t(x')}.$$ 

An outlier is detected if $\mathbb{P}(O_t = 1|Y_T = y_T) > h$, where $h$ has to be chosen according to the targeted false positive and false negative rates.

3. NUMERICAL ILLUSTRATIONS ON ARTIFICIAL DATA

The proposed estimation is first evaluated on artificial data. The simulations are designed to study the ability of the algorithm to produce a good estimation of the parameters, to identify correctly the outliers, and to adequately predict the conditional distribution of the underlying process.

3.1. Data Generation and Protocol

Artificial data are simulated according to Model (1) with $n = 150$ time steps, 50% of which are observed. We set $a$ and $b$, respectively, to the quantiles 0.02% and 99.98% of the marginal distribution of $Y'$. For the sake of consistency with the study presented in section 4, parameters are set so as to be realistic with regard to the Obépine data: $\sigma = 0.3$, $\tau = 0.6$, $\eta$ close to 1 and $\delta$ close to 0.

We realize five experiments:

- with no censoring, no outlier, $\delta = 0$, $\eta = 1$,
  - (1) with $\Delta = 0.02$,
  - (2) with $\Delta = 0.1$,
  - (3) with $\Delta = 0.7$,
- (4) with $\Delta = 0.1$, $\ell$ set for each simulated data such that 16% of $Y'$ are censored (medium censoring level), an outlier rate of $p = 7\%$, $\eta = 0.99$ and $\delta = 0.001$,
- (5) same setting as the previous experiment but with 31% of censored data (high censoring level).

Each experiment is replicated 100 times. For each simulation, we compare our approach, SCOU, to:

- a 2-parameter Kalman Smoother implemented in the DLM R package [16]. In the latter method, censoring and outliers are not taken into account and $\delta$ and $\eta$ are not estimated ($\delta$ is set to 0 and $\eta$ is set to 1). Those hypotheses correspond to the settings of the experiments (1), (2), and (3), which aims at checking that the two methods give identical results if $\Delta$ is small enough.
- the moving average smoother.
- the Locally Estimated Scatterplot Smoothing (LOESS) [17] which consists of locally weighted polynomial regressions.

As in the study by [3], the parameters of the two last methods (the number of days of the moving average smoother and the span parameter of the LOESS) are chosen by leave-one-out cross-validation so as to minimize the Root Mean Squared Error (RMSE) for the prediction of the left out observations. When (left-)censoring occurs, considering the observed censored values would result in poor performance of those smoothers in recovering the underlying auto-regressive signal. We compensate for this by taking the estimated mean for the corresponding right-truncated normal distribution in $\ell$ (estimated by assuming that the observations follow a left-censored normal distribution in $\ell$) as the observed value for those smoothers, which results in improved performances in all cases.

![Graph showing Root Mean Square Errors](image-url)
FIGURE 3 | Simulated data with 16% of censored data and $p = 7\%$ of outliers and results from the proposed (smoothing, outlier detection, and prediction) method, the equivalent 2-parameter Kalman Smoother, the 23-days moving average smoother, and the LOESS with 0.24 as span parameter.

FIGURE 4 | Parameters estimates with our method ($\sigma, \tau, \eta, \delta$) and with a 2-parameter Kalman Smoother ($\sigma$-dlm and $\tau$-dlm) for 100 replicates of the simulation experiment (150 time steps, outlier rate of $p = 7\%$, observation rate of 50%), for 16% and 31% of censored data (respectively, white and cyan boxes).
3.2. Results on Artificial Data

3.2.1. Impact of Discretization

The first, second, and third experiments (no outliers, no censoring, \(\delta = 0\) and \(\eta = 1\)) show, as expected, that the 2-parameters Kalman smoother implemented in the DLM R package and our method give identical results for \(\Delta = 0.02\). The two methods show close performances for \(\Delta = 0.1\) and a substantial degradation when \(\Delta = 0.7\), as shown in Figure 2, where the RMSE (Root Mean Square Error) for the prediction of \(X\) is computed for each of the 100 repetitions of each of the three experiments.

In the following, we focus on the two other experiments (medium and high censoring levels, \(p = 7\%, \eta = 0.99, \delta = 0.001\), and \(\Delta = 0.1\)).

3.2.2. Illustration on One Simulation Example

Figure 3 illustrates, on an example, the results of our method on simulated data within the fourth experiment setting (medium censoring level). On this illustration, we can see that the learning process and the use of the smoother permit to finely predict the trajectory of the underlying process, \(X\) (in red), from the observations \(Y_T = y_T\) (represented by dots) while adequately taking into account the time interval during which the censoring applies. On this figure, the more pink-colored the points are, the higher the estimated probability of them being an outlier. For a detection threshold set lower than \(h = 0.95\), two out of the four simulated outliers (pink crosses and circles, two of them being censored) are identified. The 23 days moving average and the LOESS (for span = 0.24) smoothers, whose parameters are selected by leave-one-out cross-validation, give rather good reconstructions but fail in reconstructing the trajectory when censoring or outliers happen.

3.2.3. Parameter Estimation

Figure 4 shows that the parameters are correctly estimated with the proposed method. However, one can notice a little negative bias in \(\eta\) estimates which might be due to the asymmetric impact of the estimation bias, with a strong degradation if the \(\eta\) estimate exceeds one. The parameters learned by our method are indeed close on average to the parameters used for the simulation of the data. The parameters learned with the 2-parameter Kalman smoother (\(\sigma_{dlm}\) and \(\tau_{dlm}\)) present an estimation bias which illustrates the necessity to take into account the censoring of the data and the outliers when they exist.

3.2.4. Outlier Identification

The posteriori probability that an observation \(Y_t\) is an outlier, \(P(O_t = 1|Y_T = y_T)\), is computed as specified Equation (4) with, for each simulation, the estimated parameters. The ROC (Receiver Operating Characteristic) curves for the detection of the simulated outliers (around 525 out of about \(n \times 100 = 7500\) observations of \(Y\)), across all simulations taken altogether, show correct outlier detection performances despite the censoring of some outliers and the possibility for the outliers to take values very close to \(X_s\) (the outliers being simulated from a Uniform distribution). Indeed, the AUC (Area Under the Curve) is \(0.74 \pm 0.01\) for both censoring levels. When the true outlier a priori probability, \(p = 0.07\), is provided, they are, respectively, of \(0.817 \pm 0.014\) and \(0.767 \pm 0.016\) for the medium and high censoring settings. Here, the standard deviations for the
FIGURE 7 | (A) and (B): Application of the LOESS (for a span parameters chosen by leave-one-out) and of the proposed smoother for two WWTPs of the Obépine network: examples of successive predictions for the (never observed) underlying signal, \( X \), with the final parameters, 3 simulations of \( X \) according to the proposed model, 95% prediction intervals for \( X \), and outlier detection results.
computed AUCs are obtained by sampling with discount the 7,500 available observations.

### 3.2.5. Prediction of the Underlying Process Distribution

Finally, we evaluate the ability of our method to predict the distribution of $X$ conditionally on the set of observed $Y$, i.e., the distribution of $(X|Y_T = y_T)$. As illustrated in Figure 5, in both the medium and high censoring level experimental settings, the RMSEs obtained by our method are significantly lower than the ones obtained with the 2-parameter Kalman Smoother, with the LOESS method, and with the moving average even when their parameters (respectively, the span and the number of days) are chosen so as to minimize the RMSE obtained by leave-one-out cross-validation.

As for the variance prediction, the coverage rates of the 95% Prediction Intervals of our method, derived from the predicted distributions of $(X_t|Y_T = y_T)$, are (on average) close to the target of 95%, with a median coverage rate of 93% (compared to 91% with the 2-parameter Kalman Smoother) in the medium censoring level setting (respectively, 93% and 85% in the high censoring level setting) as illustrated in Figure 6.

### 4. APPLICATION TO THE DATA FROM OBEPINE

The developed smoothing method aims to provide an estimate of the actual amount of viral genome arriving at each WWTP and to assess the uncertainty of this estimation.

The concentration measurements provided by Obépine are adjusted beforehand for rainfalls and wastewater sources other than from households, which can, otherwise, distort the conclusions when it is abundant by diluting the water arriving at the WWTPs [18].

The direct application of SCOU on the flow-adjusted measurements shows heteroscedasticity of the residuals [estimated by $y_T - E(X_T|Y_T = y_T)$]: their variance increases with $y$ (even after removal of numerous measurements identified as outliers and of censored measurements). Besides, the underlying process, $X$, follows the dynamic of an epidemic. During the exponential growth, it is, thus, supposed to multiply from one day to another. For both those reasons, a logarithmic transformation of the Obépine data was performed prior to applying the proposed method.

Hence, for this data set, $Y$ is the logarithm of the flow-adjusted measured concentrations, $X$ is the logarithm of the actual virus concentrations (to be estimated), $\ell$ is the logarithm of the limit below which censoring occurs (typically $\log(1000U/L)$ for the quantification of the SARS-CoV-2 E gene, where $U$ stands for RNA Units, but it can fluctuate from one day to another due to flow adjustments).

The application of SCOU to real data from the Obépine network is illustrated for two WWTPs in Figure 7. As shown by the successive predictions, once the parameters are fixed, the predictions are rather stable from one day to another.

The estimates of the smoothing parameters for the 190 Obépine WWTPs with enough available observations (at least 10 measurements) are illustrated in panel (A) of Figure 8. The estimated $\delta$ parameters are above 0 while $\eta$ parameters are very close to 1. Hence, when no new information is provided (no new $Y$ is observed), the smoother predicts, on average, an increase in $X$ values.

The uncertainty of the monitoring system is evaluated by the parameter $\tau$, whose estimates distribution for the 190 Obépine WWTPs is illustrated in panel (B) of Figure 8. The average value for $\tau$ is $0.54\log(U/L)$. This corresponds to an SD of about $0.65x$, where $x$ is the real E-gene RNA concentration (in $U/L$) to which the SD of the measurement error is supposed to be

---

![Figure 8](https://example.com/figure8.png)

**Figure 8** | (A) Parameters estimates for 190 Obépine WWTPs and (B) detailed distribution of the corresponding $\tau$ estimates which give an estimation of the error of the whole measurement system.
FIGURE 9 | Correlation of raw Obépine data and flow-adjusted and smoothed Obépine data with the logarithm of incidence rates of the corresponding cities for 15 cities with enough data available on both sides, for a lag time ranging from 1 to 3 days in both directions and for a period ranging from the beginning of the second wave to the May 21, 2021.

Importantly, the resulting smoothed signal is well correlated with the logarithm of the local COVID-19 incidence rates, and this correlation is most of the time greatly enhanced by the proposed smoothing step as depicted in Figure 9 for the 15 cities with enough data available on both sides. On this figure, the correlations are only computed for dates at which raw data was available and correspond to the best correlation for a time lag ranging from 1 to 3 days in one direction or the other. They are computed for a period ranging from the beginning of the second wave to the May 21, 2021. Those correlations are not expected to be higher since, contrary to incidence rates, the indicator also points to asymptomatic people infected by SARS-CoV-2 and is neither biased by people getting tested outside their city (during holidays for instance) nor by varying testing policies.

In order to produce comparable values from one WWTP to another, Obépine moreover performs scaling of the data after smoothing [18]. This scaling takes into account, e.g., the maximum amount of virus measured during the epidemic wave that occurred in Autumn 2020 in France, the range of volume treated by the WWTP, and the specifics of the laboratory that analyzed the sample by RT-qPCR or RT-dPCR. The final computed indicator (called WWI for WasteWater Indicator) shows good behavior with regard to the corresponding incidence rates [18].

5. DISCUSSION

We developed a method to smooth one dimensional time-series consisting of successive censored measurements with outliers when the associated measurement uncertainty is not known and the measured quantities have an auto-regressive nature. By discretizing the state space of the monitored quantities, the proposed method has the advantage of being easily adaptable to the specificities of the data (such as measurement censoring and the occurrence of outliers). An experiment on artificial data validates the proposed inference and prediction method. Our method has then been successfully applied to
data generated during the Obépine monitoring of SARS-CoV-2 genome concentration in WWTPs [18]. Importantly, the proposed smoothing procedure enhances the correlation of the data with other epidemiological indicators such as the incidence rate of COVID-19. The time lag between the two signals is moreover just a few days [18], making WWI a credible alternative to the evaluation of the incidence rate through massive individual testing. This approach may be especially relevant if massive testing campaigns become less relevant notably with the advancement of the vaccination campaign and the availability of self-tests to the general public. Both of these factors may indeed induce a progressive but significant decline in participation in testing in a few months and a significant dwindling of the population surveyed to monitor the pandemic, potentially making it even more partial than it is now.

The proposed method could be further developed. First, the underlying X process could have a longer time dependency than an AR(1) process. We could, thus, develop an AR(p) version of this method to handle this (with \( p > 1 \)). Besides, the behavior of the marginal X process, and thus its parameters, are expected to change as we move from the propagation of the epidemic stage to a decreasing stage [19]. Joint treatment of the WWTPs time-series could overcome the lack of individual data to face this problem. One could, for example, automatically detect common breakpoints corresponding to a change in the parameters \( \eta \) or \( \delta \). Another possibility would be either to use extrinsic knowledge of the reproduction factor of the epidemic as input data or to add it as a latent variable that slowly evolves from one day to another.

Another way to proceed would be to deduce from other available epidemiological data the shape of the signal to be found in wastewater (and thus an adequate smoothing) based on fine modeling of the whole pathway of SARS-CoV-2 from the human population to wastewater such as the one proposed by [20]. However, such a mechanistic representation includes a large number of unknowns (actual number of infected individuals in the population, rate of RNA degradation in wastewater, etc.) which makes it difficult to exploit for the reconstruction purposes aimed here.

**DATA AVAILABILITY STATEMENT**

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://www.data.gouv.fr/fr/datasets/surveillance-du-sars-cov-2-dans-les-eaux-usees-1.

**THE OBÉPINE CONSORTIUM**

The Obépine consortium’s Scientific Coordination and Steering Committee (CCOS) is composed of Isabelle Bertrand, Mickaël Boni, Christophe Gantzer, Soizick F. Le Guyader, Yvon Maday, Vincent Maréchal, Jean-Marie Mouchel, Laurent Moulin, Rémy Teyssou and Sébastien Wurtzer. The Scientific Interest Grouping (GIS) Obépine is directed by Vincent Maréchal associated with two co-directors: Christophe Gantzer and Laurent Moulin. This group includes the CNRS, Eau de Paris, EPHE, Ifremer, Inserm, IRBA, Sorbonne University, Clermont Auvergne University, Lorraine University, and Université de Paris.

**AUTHOR CONTRIBUTIONS**

YM, J-MM, VM, LM, and SW brought on the scientific problem. MC contributed to the design of the algorithm, performed experiments on artificial and real data, and wrote the manuscript. GN contributed to the design of the algorithm and coordinated the experiments and the writing of the manuscript. NC and SW prepared the data provided by Obépine, contributed to the design of the algorithm, performed experiments on artificial and real data, and discussed the results. NC, SW, J-MM, YM, and VM proofread and discussed the content of the manuscript. YM coordinated the exchanges about the article within the Obépine consortium. All authors contributed to the article and approved the submitted version.

**FUNDING**

This study was carried out within the Obépine project funded by the French Minister of Higher Education, Research and Innovation (Ministre de l’Enseignement Supérieur, de la Recherche et de l’Innovation). Financial support was also obtained from the French National Center for Scientific Research and Sorbonne Université.

**ACKNOWLEDGMENTS**

We would like to thank the WWTPs and the laboratories that contribute day to day to the Obépine network. In particular, the LCPME (Isabelle Bertrand and Christophe Gantzer) produced three quarters of the raw data that was used to produce Figure 9. The experimental reflections of this study were fed by exchanges with the members of the Obépine consortium and notably Karine Laurent.

**REFERENCES**

1. Wurtzer S, Marechal V, Mouchel J, Maday Y, Teyssou R, Richard E, et al. Evaluation of lockdown effect on SARS-CoV-2 dynamics through viral genome quantification in waste water, Greater Paris, France, 5 March to 23 April 2020. *Eurosurveillance*. (2020) 25:2000776. doi: 10.2807/1560-7917.ES.2020.25.50.2000776

2. Li X, Zhang S, Shi J, Luby SP, Jiang G. Uncertainties in estimating SARS-CoV-2 prevalence by wastewater-based epidemiology. *Chem Eng J*. (2021) p. 129039.

3. Arabzadeh R, Gruenbacher DM, Insam H, Kreuzinger N, Markt R, Rauch W. Data filtering methods for SARS-CoV-2 wastewater surveillance. *arXiv preprint arXiv:210411566*. (2021).

4. Kalman RE. A new approach to linear filtering and prediction problems. *J. Basic Eng.* 82:35–45. (1960). doi: 10.1115/1.3662552
5. Kalman RE, Bucy RS. New results in linear filtering and prediction theory. *J. Basic Eng.* 83: 95–108. (1961). doi: 10.1115/1.3658902

6. Leleux D, Claps R, Chen W, Tittel F, Harman T. Applications of Kalman filtering to real-time trace gas concentration measurements. *Appl Phys B.* (2002) 74:85–93. doi: 10.1007/s003400100751

7. Luo R, Cardozo EF, Piovoso MJ, Wu H, Buzon MJ, Martinez-Picado J, et al. Modelling HIV-1 2-LTR dynamics following raltegravir intensification. *J R Soc Interface.* (2013) 10:20130186. doi: 10.1098/rsif.2013.0186

8. Miller C, Allik B, Ilg M, Zurakowski R. Kalman filter-based tracking of multiple similar objects from a moving camera platform. In: 2012 IEEE 51st *IEEE Conference on Decision and Control (CDC).* Maui, HI: IEEE (2012). 5679–84.

9. Wan EA, Van Der Merwe R. The unscented Kalman filter for nonlinear estimation. In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373).* Lake Louise, AB: IEEE. (2000). p. 153–8.

10. Julien SJ, Uhlmann JK, Durrant-Whyte HF. A new approach for filtering nonlinear systems. In: *Proceedings of 1995 American Control Conference-ACC’95.* Vol. 3. Seattle, WA: IEEE. (1995). p. 1628–32.

11. Allik B, Miller C, Piovoso MJ, Zurakowski R. The Tobit Kalman filter: an estimator for censored measurements. *IEEE Trans Control Syst Technol.* (2015) 24:365–71. doi: 10.1109/TCST.2015.2432155

12. Aravkin A, Burke JV, Ljung L, Lozano A, Pillonetto G. Generalized Kalman smoothing: modeling and algorithms. *Automatica.* (2017) 86:63–86. doi: 10.1016/j.automatica.2017.08.011

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Courbariaux, Cluzel, Wang, Marchal, Moulin, Wurtzer, Obépine Consortium, Mouchel, Maday and Nuel. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.