Journal of Water & Health

Journal of Water and Health Vol 00 No 0, 1 doi: 10.2166/wh.2024.043

Watching the guards: A data-driven method to trigger warnings in national wastewater surveillance networks

Ll. Bosch[a,b], J. Pueyo-Ros[a,b], M. Comas-Cufí[c], J. Saldaña[c], J. Ripoll[c], E. Calle[d], P. Fonseca[e], J. Garcia[e], C. M. Borrego[a,f] and Lluis Corominas[a,b,*]

[a] Catalan Institute for Water Research (ICRA-CERCA), Emili Grahit 101, Girona 17003, Catalonia, Spain
[b] University of Girona, Plaça Sant Domènec 3, Girona 17004, Catalonia, Spain
[c] Department of Computer Science, Applied Mathematics, and Statistics, University of Girona, Girona 17003, Catalonia, Spain
[d] Institute of Informatics and Applications, University of Girona, Maria Aurèlia Capmany 61 (Building PIV), Girona 17003, Catalonia, Spain
[e] Polytechnic University of Catalonia – Barcelona Tech., Jordi Girona 31, Barcelona 08034, Catalonia, Spain
[f] Group of Molecular Microbial Ecology, Institute of Aquatic Ecology, University of Girona, Girona 17003, Catalonia, Spain
*Corresponding author. E-mail: lcorominas@icra.cat

JP, 0000-0002-1236-5651; MC, 0000-0001-9759-0622; JS, 0000-0001-6174-8029; JR, 0000-0002-1186-0175; EC, 0000-0003-2361-602X; CMB, 0000-0002-2708-3753; LC, 0000-0002-5050-2389

ABSTRACT

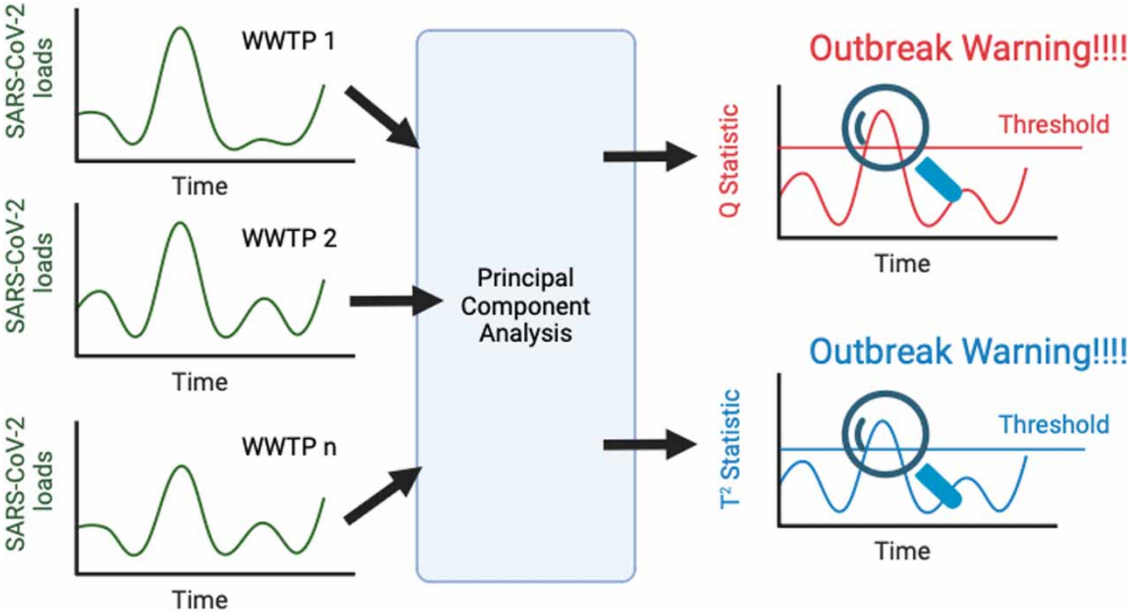
Surveillance networks have been established in many countries worldwide to monitor SARS-CoV-2 in sewage and to estimate the communal prevalence of COVID-19 cases. Despite their popularity, gaining a rapid understanding of how infectious diseases spread across the territory covered by the network is difficult because of the many factors involved. To improve the detection of warning signals within the territory, we propose to apply a principal component analysis (PCA) to screen time-series data generated from wastewater treatment plants (WWTPs) under surveillance. Our analysis allows us to identify single WWTPs deviating from the normal behavior as well as deviations of a cluster of WWTPs (indicative of an intermunicipal outbreak). Our approach is illustrated through the analysis of the dataset generated by the Catalan Surveillance Network of SARS-CoV-2 in Sewage (SARSAIGUA). Using 10 principal components, we captured 78.6% of the variance in the original dataset of 51 variables (WWTPs). Our analysis identified exceedance of the Q-statistic threshold as evidence of anomalous performance of a single WWTP, and exceedance of the T^2-statistic as a sign of an intermunicipal outbreak. Our approach provides a comprehensive picture of the spread of the COVID-19 pandemic, enabling decision-makers to make informed decisions to better manage future pandemics.

Key words: COVID-19 pandemic, fault detection, PCA, wastewater-based surveillance

HIGHLIGHTS

- PCA is applied to SARS-CoV-2 measurements in sewage from multiple WWTPs.
- Q- and T^2-statistics are useful to trigger warnings of COVID-19 outbreaks at city and national levels.
- Warnings across waves are triggered by different subsets of WWTPs.
- The contribution of WWTPs to triggering warnings is independent of their size.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Wastewater-based epidemiology (WBE) has emerged as a powerful approach for extracting valuable health-related information from wastewater, enabling comprehensive surveillance of population health (Singer *et al.* 2023). While its initial focus was on tracking poliovirus, WBE techniques have rapidly expanded to encompass a broad spectrum of chemical and biological targets. These include monitoring illicit drug use (González-Mariño *et al.* 2020), assessing pharmaceutical consumption patterns (Escolà Casas *et al.* 2021), and analyzing various health and lifestyle biomarkers (Daughton 2018; Shao *et al.* 2023). Moreover, the global outbreak of SARS-CoV-2 has significantly heightened the importance of wastewater as a valuable source of information (Bivins *et al.* 2020). This has led to the establishment of surveillance networks at regional and country scales, with over 4,107 monitoring sites across 72 countries worldwide dedicated to monitor SARS-CoV-2 in sewage (Naughton *et al.* 2023). These networks play a crucial role in enhancing our understanding of chemical and biological target dynamics and facilitating targeted interventions. These networks usually involve monitoring multiple wastewater treatment plants (WWTPs), ranging from 8 (e.g., in Slovenia) to 352 (i.e., in The Netherlands) (Table 1). Once established, the data generated are analyzed and transformed into useful information to support decision-making of health authorities. Results are commonly presented by plotting normalized loads of the targeted SARS-CoV-2 genetic marker over time per WWTP. Also, such plots usually include a smoothing operation (e.g., median over 7 days or the moving average of 7 days). In some cases, loads are summed up to have a general overview of the network at the country level (Guerrero-Latorre *et al.* 2022). Trends in SARS-CoV-2 loads are then estimated by applying the percent change recommended by the CDC, which entails utilizing the regression slope of at least the three most recent measurements and categorizing it based on the statistical significance of the slope (https://www.cdc.gov/nwss/reporting.html). Alternatively, other trend estimates, such as the relative strength index and the Mann-Kendall trend test, are also considered (Chan *et al.* 2023). Some surveillance networks transform the results into qualitative indicators (of concentrations or loads and/or of trends) shown on a map (e.g., the French network). In some cases, such as the Swiss and Austrian surveillance networks, the effective reproductive number of SARS-CoV-2, the time-varying analog of the basic reproductive number along one or several waves of the disease, is estimated from wastewater data at each monitored WWTP (Huisman *et al.* 2022).

Beyond the implementations at the national level, several studies have proposed mathematical methods: (i) to convert raw wastewater data to prevalence estimates (Ahmed *et al.* 2020; Morvan *et al.* 2022), (ii) to smooth the time-series data based on an autoregressive model (Courbariaux *et al.* 2022), (iii) to reduce and correct the noise associated with the quantification of SARS-CoV-2 gene targets (Cluzel *et al.* 2022), (iv) to model single outbreaks for short-term forecasting (Joseph-Duran *et al.*

Table 1 | Data analysis made publicly available by health authorities of a selection of national wastewater surveillance networks for SARS-CoV-2

Name	# WWTPs	Cov. (%)	SF	Data analysis
Austria	48	58	2	Figures: Gene copies/inhab./day per site; Sum of loads of all WWTPs; Effective reproductive number per WWTP
Belgium	43	45	2	Figures: Gene copies/day and per 100,000 inhab. (rolling average); Increasing trend indicators
Canada	38	NA	3	Figures: 7-day rolling average of gene copies/day for each site. Trend expressed as: Statistically significant increase or decrease, possible increase, no change
Catalonia	56	80	1*	Figures: 7-day rolling average of gene copies/day for each site and aggregated at the national level. Trend expressed as: Increasing, stable, decreasing
Finland	16	49	1*	Figures: Gene copies/1,000 inhab./day; the trend and its uncertainty are estimated using a statistical model
France	200	33	NA	Maps: qualitative assessment, trend over last 30 days, trend over last 7 days
Germany	135	NA	NA	Figures: Gene copies/L per site and aggregated; Maps and Heatmaps of trends
Hungary	21	NA	1	Maps: Concentrations and trend (qualitative, 4 categories)
Ireland	68	80	1	Figures: Qualitative (Positive, positive DNQ, weak positive, negative)
Luxembourg	13	NA	1	Figures: Gene copies/day/100,000 equivalent inhab.
Netherlands	352	98	2–3	Figures and Map: average number of gene copies per 100,000 inhab. per municipality
Slovenia	8	25	1	Figures: N2 gene copies/PMMoV; estimated cases from wastewater after linear regression
Sweden	13	25	1	Figures: Gene copies/PMMOV
Switzerland	14	27	3–6	Figures: Gene copies/100,000 inhab./day; Effective reproductive number per WWTP
England	270	60	3	National and regional means of gene copies/L
US (wastewaterSCAN)	200	60	2–3	Figures and maps: Gene copies/PMMoV per site, and national levels categorized into bottom, middle, and upper thirds
US (Biobot)	700 +	30	1	Figures: Nationwide, region, county, expressed as gene copies per mL; case estimates from effective concentrations

Cov., population coverage (in %); SF, sampling frequency (# per week); NA, not available.
* A subset of WWTPs is sampled at less frequency than once per week.

2022), and (v) to model the transmission of SARS-CoV-2 (SEIR model) in combination with the fate of SARS-CoV-2 in sewage after fecal shedding (Nourbakhsh *et al.* 2022; Mattei *et al.* 2023) conferring long-term forecasting prediction capabilities and a way to simulate non-pharmaceutical interventions. Yet, all these approaches are applied to data obtained from a single WWTP. When it comes to aggregating information from multiple WWTPs within the network, it becomes more complex to gain a rapid understanding of how the pandemic is evolving across the entire territory covered by the network. As such, there is a need for reliable and scalable mathematical methods that can provide a more comprehensive picture of the pandemic's spread at the national level. Two surveillance networks employ mathematical methods to integrate data from multiple WWTPs. The WastewaterSCAN network (wastewaterscan.org) represents the 5-sample trimmed average of SARS-CoV-2 concentrations normalized by the concentration of the Pepper Mild Mottle Virus, used as an indicator of human fecal shedding. Aggregated trend lines present population-weighted averages across groups of sites. National levels, depicted in charts, offer a relative interpretation of current wastewater levels over the last 365 days. These levels are categorized into bottom, middle, and upper thirds, based on a retrospective analysis of data from the past 365 days across all sites. The US Biobot (https://biobot.io) approach includes weekly averaging within reporting counties, weighted by population, using a 3-sample rolling average for greater emphasis on recent measurements. This data is subsequently averaged at both the national and regional levels.

In this work, we applied a statistical process control technique to analyze the time-series data generated from multiple WWTPs at once. The usefulness of the approach is illustrated by the application to the data obtained from the Catalan Surveillance Network of SARS-CoV-2 in Sewage (Corominas *et al.* 2021; Guerrero-Latorre *et al.* 2022) (SARSAIGUA). Our

analysis is based on principal component analysis (PCA), which allows us to feed the gene target loads into an algorithm that can identify deviations in single WWTP from normal behavior (a COVID-19 early warning at a city scale) as well as generic deviations from normal behavior for a cluster of WWTPs (occurrence of an outbreak at the national level).

2. METHODS

2.1. The Catalan Surveillance Network of SARS-CoV-2

Catalonia is a region with more than 7.7 million inhabitants in north-eastern Spain. The Catalan Water Agency (ACA) and the Public Health Agency of Catalonia (ASPCAT) promoted and funded the deployment of SARSAIGUA (Guerrero-Latorre *et al.* 2022). This network started in July 2020 monitoring 56 WWTPs evenly distributed across Catalonia and serving 80% of the total population. The sample collection and analysis started on the 6th of July 2020, approximately 4 months after the detection of the first clinical COVID-19 case in Catalonia (25 February 2020). Out of the 532 WWTPs in Catalonia, 56 were included in the surveillance network; these 56 represent a high population coverage (80%) and are evenly distributed across the territory (at least 1 WWTP per county). The sampling frequency was set to one sample per week in 36 of the selected 56 WWTPs and fortnightly for the remaining 18, thus resulting in the collection and analysis of 45 samples per week. Notably, some WWTPs are only surveyed during the summer season to better monitor municipalities receiving high tourism (e.g., Castell-Platja d'Aro, Vilaseca-Salou). Refrigerated flow-based composite samples are collected for most WWTPs at the entrance. Details on the WWTPs and the sampling can be found in Supplementary Table S1. The 45 weekly samples are distributed to the three reference laboratories with wide expertise in molecular diagnosis and environmental virology. Each laboratory receives 15 samples per week that are analyzed for SARS-CoV-2 genome abundance using optimized protocols. Quantification of SARS-CoV-2 genomes is accomplished using RT-qPCR targeting a common genetic marker (N1) and two complementary targets, N2 and IP4 (CDC 2020; Pasteur 2020). Manual data quality control on the results generated from the laboratories is executed once a week. This manual quality control involves evaluating: (i) RT-qPCR standard curve parameters; (ii) recovery values from process control; and (iii) correlation between RT-qPCR target genes. Whenever large deviations from previous accumulated data in parameters like recovery percentage of a process control are identified, the laboratories are requested to repeat the sample.

2.1.1. WWTP data

Data used for this study is obtained from an API-created *ad-hoc* (https://apicovid.icradev.cat/n1). When accessed, this API reads the SQL database where the results of the SARS-CoV-2 gene targets are uploaded weekly by the laboratories responsible for the quantification, and outputs the data retrieved in CSV format. The data includes information about the concentration of N1 gene copies at each WWTP and influent flows (in m^3/day). For the analysis, we only included data corresponding to the 51 WWTPs sampled weekly and biweekly, while those WWTPs that were sampled on a seasonal basis (5) were excluded from the analysis. The data used as input for the statistical method applied in this study was the normalized N1 gene load for each WWTP, that is the concentration of N1 (in gene copies (GC)/L) multiplied by the daily flow of the WWTP (L/day) and divided by the number of inhabitants assisted by the corresponding WWTP. Data interpolation was applied to the biweekly sampled WWTPs (N1 concentration and flows) to obtain a weekly value, and missing values (due to sampling or analytical issues) on the entire dataset were interpolated as well. We are aware that this interpolation is not feasible in a real-world surveillance network where only weekly WWTPs can be used. However, we decided to interpolate fortnightly sampled WWTP to increase the size of dataset and provide more robust evidence of our method. In total, 138 observations collected from the 51 WWTPs were used for the period between July 2020 and August 2023.

2.1.2. Reported clinical cases

Daily reported clinical cases were aggregated for municipalities in the catchment of each WWTP using the official API from the Information Systems of the Department of Health and the Catalan Health Service (Departament de Salut n.d.) (https://analisi.transparenciacatalunya.cat/resource/jj6z-iyrp.json).

2.1.3. Simulated COVID-19 active cases using a compartmental epidemic model

The outcomes from the SCVEIR ID RHUD model described in Fonseca Casas *et al.* (2023) and continuously updated and calibrated in the SDL-PAND dashboard (http://pand.sdlps.com) were used in our work to cover the gap of reported clinical cases after April 2022. The model is a cellular automation (CA) model, hence the 'CA' in the name of the model, CA-SCVEIR

ID RHUD, where each one of the different cells of the automata implements a complete SCVEIR ID RHUD model. The different levels of the compartmental model are: (i) S: susceptible, (ii) C: confined, (iii) V: vaccinated, (iv) E: exposed, (v) I: infective, with IR: infective real (real cases) and ID: infective detected, where the evolution is based on the IR and ID are only for validation purposes, (vi) R: recovered, (vii) H: hospitalized, (viii) U: critical hospitalized, and (ix) D: Dead. The presented data encompasses all simulated active cases throughout Catalonia, extending beyond the confines of the 51 WWTPs specifically included in the PCA. Nevertheless, utilizing this broader dataset for comparative analysis remains valuable, given that these 51 WWTPs represent 80% of the Catalan population.

2.2. PCA and statistical process control

PCA is a multivariate statistical method for exploratory data analysis. It can be used to reduce the number of variables in a dataset while retaining as much information as possible (Jolliffe 2002). PCA can be used for a variety of applications, such as identifying patterns in data, reducing the dimensionality of data for visualization, and making predictions using machine learning algorithms. Here, we propose to use PCA as a basis for an automated warning system procedure using the Q-statistic (a.k.a. squared prediction error) (Jackson & Mudholkar 1979) and the Hotelling's T^2 (Johnson & Wichern 2002).

The application of PCA followed comprehensive manual data quality control procedures for both flow and N1 measurements. Thus, the primary objective of the PCA analysis was not to identify issues arising from measurement errors, but rather to elucidate the correlations among N1 loads across WWTPs. For the PCA, each of the 51 WWTPs was treated as a variable (matrix column) and an observation (matrix row) is a week in time between July 2020 and August 2022. An input matrix without gaps is required when running PCA; hence, linear interpolation was applied to the biweekly sampled WWTPs (N1 concentration and flows) and to the missing values due to incidences in sampling (e.g., missing flow data). Scaling was applied to each column, to ensure each WWTP is given equal weight in the monitoring process. This involves subtracting each variable by its sample mean (to capture the variation from this mean) and then, divide by its standard deviation. This gives equal importance to all columns and removes variabilities caused by different sampling and analytical methods (e.g., analytical differences between laboratories). After performing PCA, the Q-statistic was calculated to identify structural changes in the process. An increase in Q suggests that the correlation structure captured by the PCA model is not maintained for that observation (1 observation is the set of N1 loads from the 51 WWTPs). In our context, this indicates that one or more WWTPs deviate from the average behavior. Additionally, T^2 was computed to measure the distance of each observation from the center of the PCA model, represented by the mean. An increase in T^2 signifies that the observation maintains the model structure but with values that are further from the mean, indicating a general deviation of N1 loads from 'normal' values and signaling an outbreak scenario. We set up the threshold values for Q and T^2 using the first set of observations ($n = 51$) (training dataset) and using an arbitrary level of significance (α) of 0.05 which gave us acceptable results to illustrate the potential of the PCA. We did not follow a rigorous approach for setting up the thresholds which would require data labeling, training, and validation. Finally, a contribution analysis is performed to determine which variable (WWTP) or variables in the original dataset space are responsible for the detected warning, specifically identifying instances where thresholds are exceeded. We implemented adaptive PCA, which is an extension of traditional PCA that allows for the continuous updating of the principal components as new data becomes available. We used an incremental window to update the PCA, that is, on day i, we used the observations from 1 to i. To improve the stability of the method, in each new PCA, we rejected all observations where Q or T^2 values were higher than the respective thresholds of 0.05.

2.3. PCA and thresholds implementation

The PCA, the calculations of Q and T^2 and the thresholds (see detailed equations in the Supplementary Material) were implemented in JavaScript programming language, which is native to the most-used web browsers at the time of writing (e.g., Google Chrome, Mozilla Firefox, Microsoft Edge, etc.) and it is independent of the operating system and device type. We chose JavaScript rather than other languages keener to statistical methods (such as R or Python) to facilitate the integration of the warning system in the SARSAIGUA platform (https://sarsaigua.icra.cat), already developed using JavaScript. However, the proposed warning system method has only been applied retrospectively for the purpose of this research but not into the SARSAIGUA programme since it was discontinued in December 2023. A matrix library was created to compute basic matrix operations (e.g., sum, multiplication, transposition, determinant, inverse, etc.). Then, the SVD algorithm and PCA procedures, along with the computation of the Hotelling's T^2 and Q statistics and contribution analysis were implemented on top of this first layer. This matrix library is freely available at https://github.com/icra/matrix-library-js. To

validate the JavaScript implementation, numeric tests were conducted using R language (https://www.r-project.org/) since it is a well-known and established numeric platform with PCA functions built in. A visual web interface was built on top of the matrix library, coded in HTML and CSS, and using VueJS (https://vuejs.org/) as the front-end JavaScript framework.

3. RESULTS

3.1. Raw time series

The dynamics of the COVID-19 pandemic in Catalonia is displayed in Figure 1, covering six waves between July 2020 and August 2022. The first wave started in March 2020 and occurred before the implementation of the SARSAIGUA programme. The progression of pandemic waves was explained both by reported cases and N1 loads measured from wastewater; and showed a Spearman correlation coefficient of 0.69 between both variables using data from July 2020 to March 2023 (6th wave). In the 7th wave (April 2022), the Catalan Health authorities discontinued the reporting of COVID-19 cases and launched the sentinel system SIVIC (https://sivic.salut.gencat.cat) based on the screening of 10 symptomatic patients per week in 33 carefully selected primary care health centers. The red line in Figure 1 shows the N1 loads from all 51 WWTPs included in this study; clinical reported COVID-19 cases correspond to the cases detected and reported in the catchment communities serving these 51 WWTPs. Data have been standardized in order to assign equal importance to all WWTPs no matter their size. Otherwise, WWTPs serving large populations (i.e., El Prat de Llobregat with 1M inhabitants or Besòs 1.4M inhabitants) would bias the entire analysis.

3.2. PCA results (scores and loadings)

The PCA was conducted on a dataset comprising 51 WWTPs and 138 observations, resulting in a matrix of dimensions 138 × 51. Employing 10 PCs enabled us to capture 78.6% of the total variance within the original dataset, while the first two components retained 49.3% of the variance. Specifically, PC1 accounted for 30.9% of the variance, and PC2 explained 18.4%. The analysis revealed distinct patterns, as certain WWTPs tended to cluster together, indicating strong correlations among them. For instance, during the 6th wave (January 2022), WWTPs such as DMAT, DIGU, and DGVC displayed significant correlations. Similarly, during the 2nd wave (November 2020) and the 8th wave (November 2022), a cluster of WWTPs,

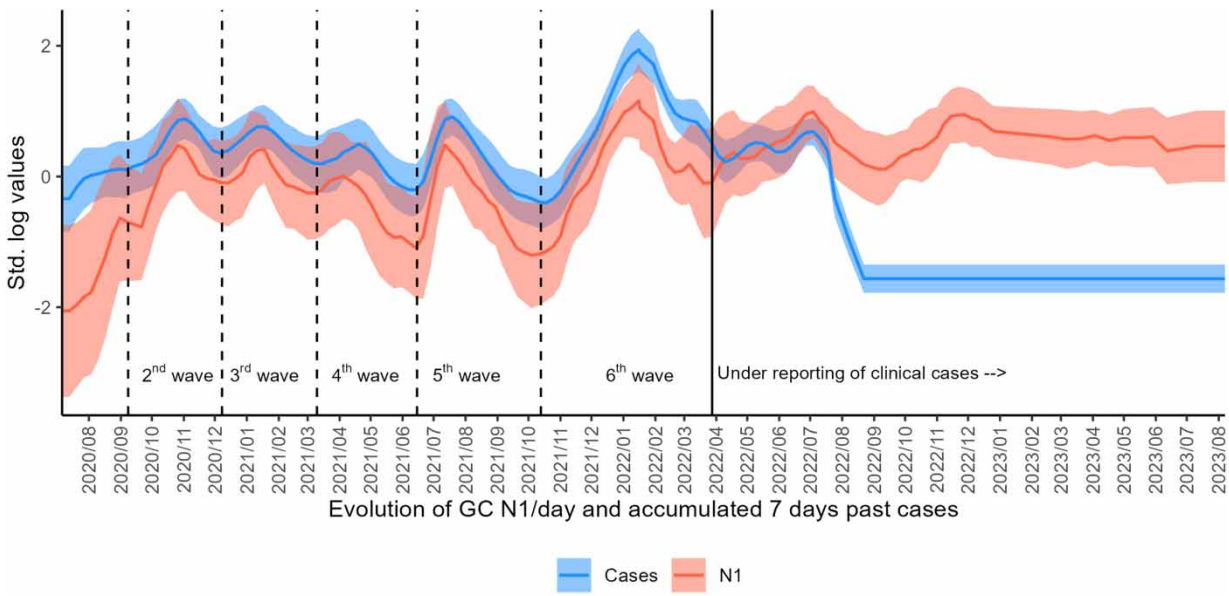


Figure 1 | Comparison of reported COVID-19 cases (blue) and N1 gene loads (red) in wastewater collected across the Catalan territory since the deployment of the SARSAIGUA program in July 2020. Values were standardized to allow comparison among WWTP and smoothed using a moving average of size 3. The line represents the average value while the shade areas comprise the values within 1 standard deviation. Dashed lines distinguish the different pandemic waves as indicated. 1st to 4th waves were caused by the Alpha variant, the 5th wave was caused by the Delta variant, and the Omicron variant was responsible for the 6th and upcoming waves. Vaccination started during the 3rd wave (27 December 2020).

including DLLL, DBER, DFIG, and DABR, exhibited notable correlations. It is worth noting that Figure 2 is a simplification of the PCA outcome (we captured 10 PCs and hence the PCA model works in a 10-dimensional space) comprising only the two first components.

3.3. Triggering warnings for outbreaks

3.3.1. Threshold exceedances

Figure 3 (left) illustrates a warning generated by the exceedance of the Q-statistic. It depicts the abrupt increase in the N1 load at WWTP DSRS on the 16th of May 2022 as compared with the other WWTPs. Figure 3 (right) provides an example of the exceedance of the T^2-statistic on the 28th of June 2021. It is evident that there is a general increase in N1 loads across the WWTPs analyzed. WWTPs highlighted in color contribute the most to the warning. In terms of PCA interpretation, all WWTPs show deviations from the center of the 11-dimensional space but exhibit strong correlations with each other.

Figure 4 illustrates that exceedances of Q- and T^2-statistics were observed in all COVID-19 waves examined in our study, except for the 4th wave, which experienced a minor surge in infections. No threshold exceedances during interwave periods indicate the absence of false alarms. During the 2nd, 3rd, 5th, and 6th outbreaks, the T^2-statistic increases alongside diagnosed cases, with the threshold being exceeded early in the outbreak stages. Subsequently, the T^2-statistic declines as the outbreaks reach their peaks (at the peak or 1 week later). However, during the 7th and 8th outbreaks, the T^2-statistic surpasses the threshold less frequently. Instead, the Q-statistic triggers more warnings, indicating a divergence in the behavior of these outbreaks compared with previous ones. This shift between the 6th and the 7th outbreaks coincides with changes in government policy regarding non-pharmacological interventions. Specifically, the government repealed mandates for mask-wearing in public health spaces (effective April 2022), discontinued mandatory case reporting, and relaxed confinement measures, including the cessation of contact tracing and quarantine obligations. Notably, during this period, outbreaks occurred asynchronously both temporally and spatially.

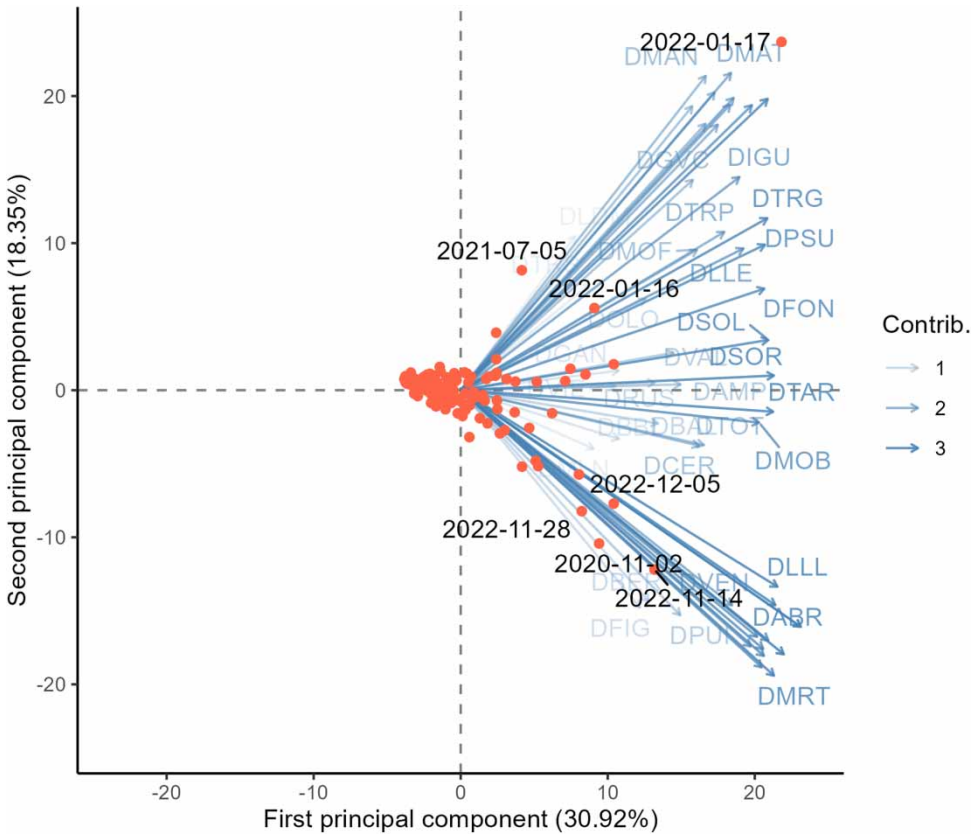


Figure 2 | Biplot of first two dimensions of the PCA. The labels are shown for distinguished observations.

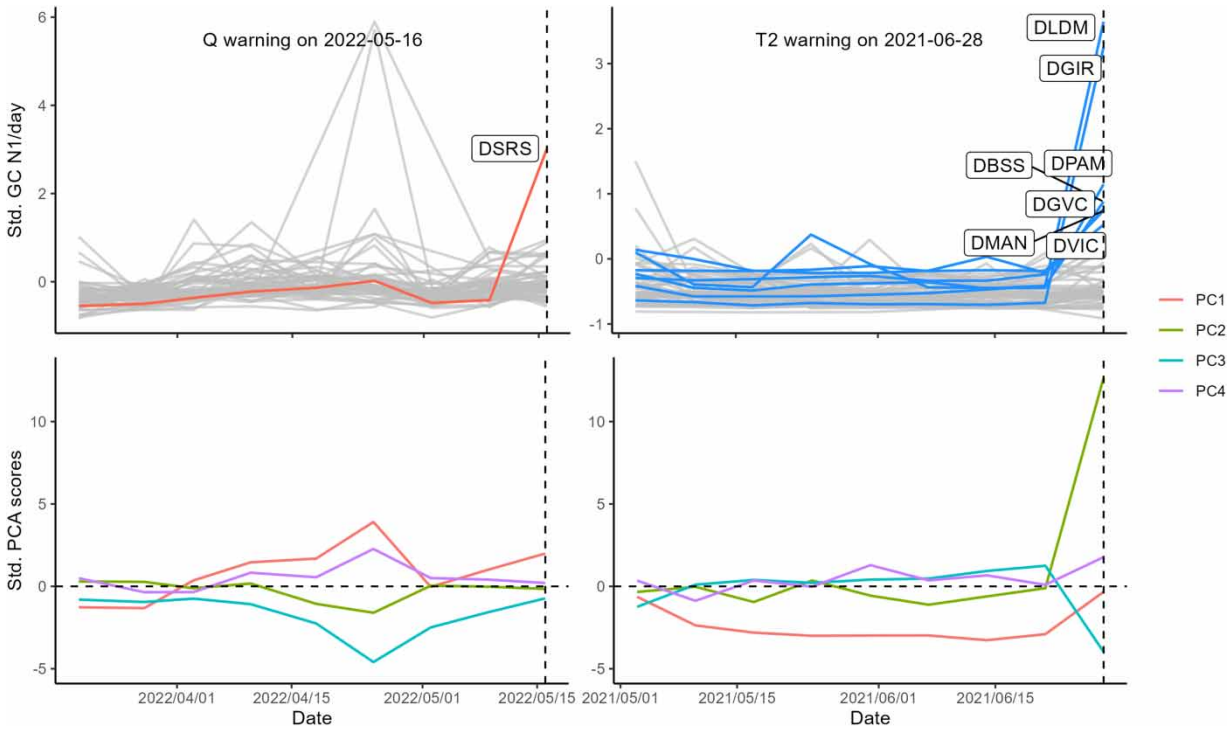


Figure 3 | Illustration of standardized N1 loads in all WWTP for different warnings on Q and T^2 and the respective PCA scores for those events. The highlighted WWTPs are the ones with the highest contribution in Q-statistic or T^2 values, respectively.

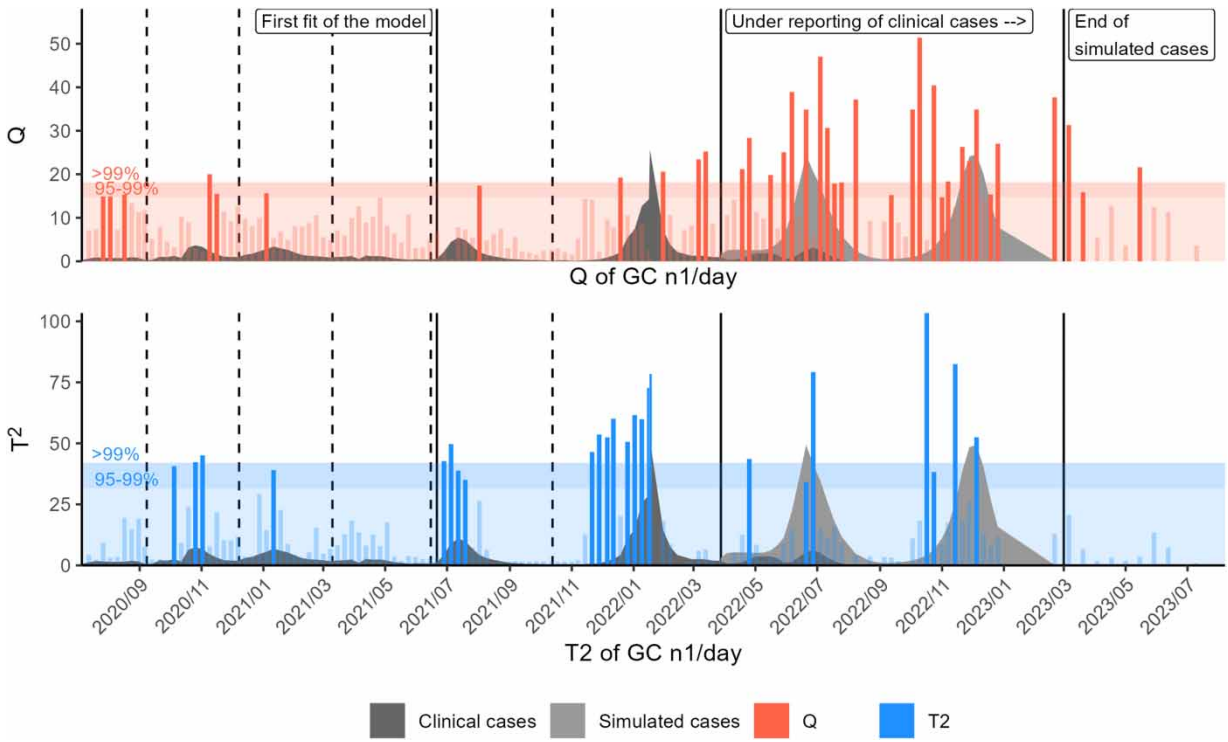


Figure 4 | Control chart based on Q- and T^2-statistics applied to WWTP data. Dark gray areas represent the clinical cases, and the light gray areas represent the simulated active cases obtained from the epidemiological model. The outcomes of the epidemiological model contribute to interpreting warnings during the period of underreporting cases (post-April 2022).

3.3.2. Contributions of the different WWTPs to the Q and T^2 warnings

Figure 5 illustrates the average ranking score of each WWTP concerning Q and T^2 warnings. The contribution of each WWTP in the warning generation can be obtained whenever a warning is triggered (see Materials and Methods). For each warning, the WWTPs are sorted in descending order of contribution. Subsequently, we calculated the average of these orders across all warnings. A lower score indicates that, on average, a specific WWTP contributes more to triggering Q (or T^2) warnings. Concerning T^2, the fact that some WWTPs contribute minimally to warning generation implies a high correlation with principal components, aligning with the general pattern of N1 concentrations (WWTPs converging toward the center of the 11-dimensional space). For the Q-statistic, WWTPs with low-ranking scores are more likely to detect early a rise in cases at the city level, exhibiting higher sensitivity in detecting changes in N1 loads or having experienced a surge in cases earlier than other communities (i.e., cities). The rankings for all WWTPs fall within a narrow range of 19–32. It is important to highlight that the rankings for Q- and T^2-statistics differ, indicating that no specific subset of WWTPs consistently contributes significantly more than others to the warnings, and this contribution is independent of WWTP size. This suggests that different subsets of WWTPs are involved in triggering warnings, rendering all WWTPs equally informative.

4. DISCUSSION

The method proposed in this work allows triggering warnings for COVID-19 outbreaks using data of SARS-CoV-2 N1 gene loads measured in influent wastewater of several WWTPs monitored weekly by SARSAIGUA surveillance network. Despite SARSAIGUA prioritized maximizing spatial coverage over relying on high-frequency measurements, our method allows the effective triggering of outbreak alerts using 1 measurement per week at high spatial resolution. In this regard, it is worth mentioning that the distance from any place to the nearest WWTP included in the network is less than 13 km (SD: 6.6, max: 38) representing 141k inhabitants/WWTP and 569 km^2/WWTP. By prioritizing the spatial coverage over the high-frequency data

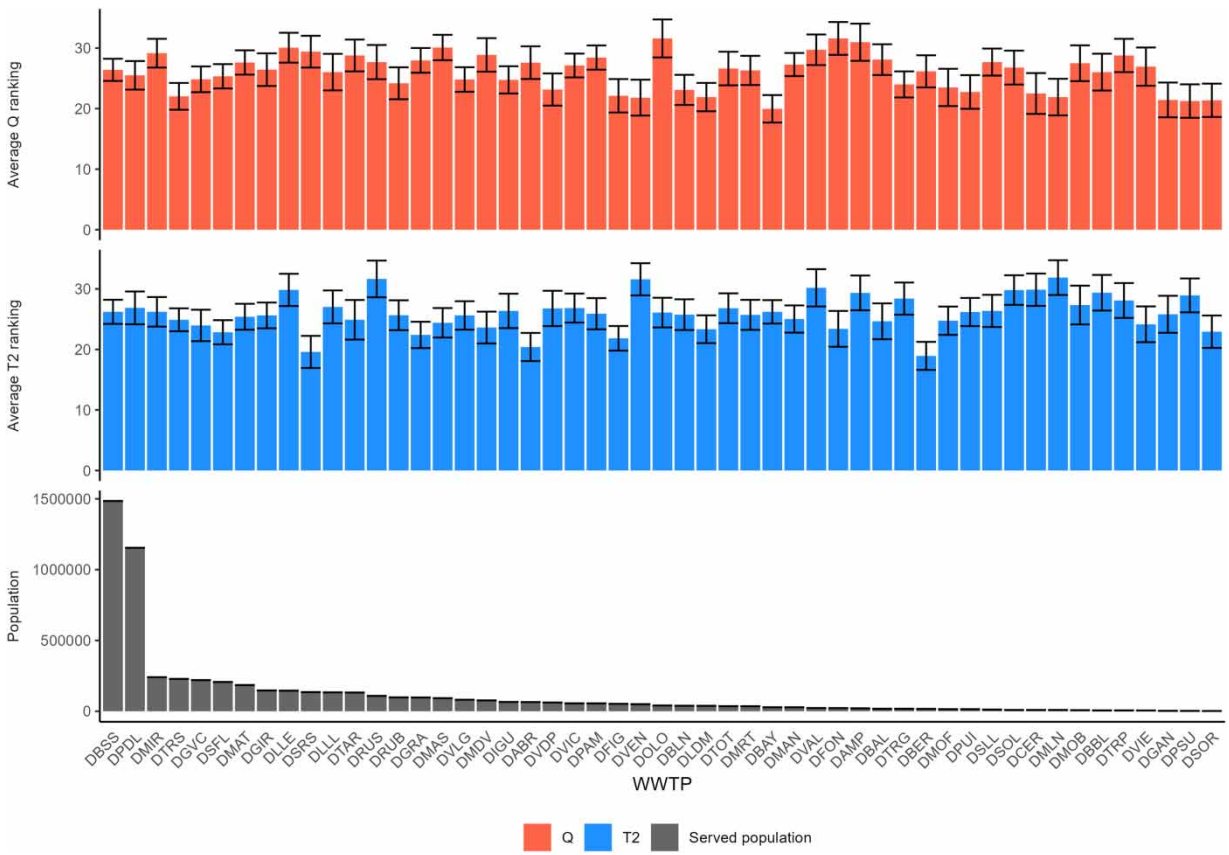


Figure 5 | Average ranking score of the Q and T^2 warnings obtained for each WWTP sorted by assisted inhabitants in descending order. The overlap in Q and T^2 for most WWTPs shows that there is not a subgroup of WWTPs consistently responsible for alarm triggering.

acquisition, it is thus possible to enhance the robustness of the warning system. In practice, the national wastewater surveillance networks have adopted different strategies for data collection. Most networks typically collect between 3 and 7 samples per week (see Table 1) to identify trends in SARS-CoV-2 concentration in wastewater and estimate the communal prevalence of COVID-19 in the surveilled territory. Such high frequencies are required for properly estimating trends (Chan *et al.* 2023) and for the estimation of the effective reproductive number of the virus (Re) (Huisman *et al.* 2022). Conversely, networks such as SARSAIGUA (Guerrero-Latorre *et al.* 2022) opted for a broader coverage by monitoring a larger number of WWTPs at a lower sampling frequency. With such low temporal resolution, it is not possible to get a reliable trend analysis per each WWTP, but as shown here it is possible to trigger warnings at a national scale. Discerning which of these approaches is better requires future investigations conducted on a comprehensive dataset encompassing a large number of WWTPs sampled at high frequency, enabling the downscaling of the temporal and the spatial resolutions. This debate has gained prominence, especially in the current scenario, where budget constraints for wastewater surveillance networks prompt considerations about optimizing either the sampling locations or the sampling frequency. To make informed decisions under these budgetary limitations, it is imperative to determine the most effective approach for ensuring timely warnings and accurate detection of anomalies within the wastewater treatment infrastructure.

The activation of warnings across the progressive pandemic waves does not consistently involve the same set of WWTPs. Interestingly, the larger WWTPs assisting populations exceeding 1 million inhabitants did not significantly contribute to the triggering of warnings but largely dictate the overall (or 'average') behavior. In contrast, WWTPs serving populations smaller than 1 million bore the responsibility for initiating warnings. In our analytical framework, all WWTPs hold equal weight due to the statistical standardization of N1 loads of each WWTP. Consequently, an outbreak involving only a few individuals in a small population (i.e., small WWTP) may exert a substantial influence on the outcome of warning initiation. This underscores the heightened impact of even minor outbreaks within these smaller WWTPs on the warning-triggering results. Our investigations in Catalonia provide substantial support for the notion that the propagation of COVID-19 across waves originates differentially at a spatial scale. Assuming rigorous sampling protocols were consistently implemented in all WWTPs across the territory, particularly employing flow-based composite sampling, we propose that our approach effectively identifies the originating sources of each outbreak using wastewater data.

Our approach has several limitations. While data-driven methodologies like PCA offer operational simplicity by requiring few parameters, their effectiveness hinges on the careful selection of these parameters. In our PCA method, two parameters – the number of principal components retained in the model and the level of statistical significance (α) – need to be determined. The selection of the latter entails data labeling, dataset separation into training and validation sets, and refinement of the Q and T^2 thresholds. Although in our application, these thresholds were kept fixed after training, monitoring non-stationary processes may necessitate adjusting the parameters to maintain the desired false detection rate (Schmitt *et al.* 2016). Moreover, as the behavior of the virus may evolve with the emergence of new variants, there is a need to readjust the thresholds after each outbreak. Secondly, the first PCA can only be applied when the number of observations (elapsed weeks) matches the number of WWTPs. Consequently, if wastewater monitoring begins after the outbreak onset, there will be a delay before the method becomes useful as an 'early warning system'. This issue becomes more pronounced for larger networks. For example, a network of 100 WWTPs, collecting and analyzing 1 sample per week, would require approximately 2 years of training data. In such cases, either the number of WWTPs included in the PCA model should be reduced or the sampling frequency should be increased. The lead time of warnings generated by our method relative to other approaches warrants further investigation. Another limitation is the utilization of interpolation, particularly in the test dataset. While linear interpolation is acceptable for training datasets, its application in the test dataset can lead to data leakage and overly optimistic performance estimates. Another option to address data gaps is to use the last available value and maintain it until a new value is received. However, this strategy may lead to false alarms. For example, when an outbreak has reached the peak and starts descending, the majority of WWTPs may experience a decrease in values, while those interpolated with the last available value remain unchanged. Consequently, this discrepancy may trigger a Q alarm, erroneously signaling an anomaly in the data. Addressing this interpolation issue and refining the adaptation of Q and T^2 thresholds would require further research.

The wastewaterSCAN national surveillance network (USA) facilitates integrated data visualization from multiple WWTPs, featuring two key aspects. Firstly, it enables the plotting of population-weighted average trend lines for selected site groups, accessible by choosing predefined groups (e.g., state or county) from the dropdown menu. Secondly, the charts incorporate national-level indicators, categorizing data into bottom, middle, and upper percentiles based on the past 365 days across all wastewaterSCAN sites. The 33rd and 66th percentiles delineate these categories, offering a relative context for interpreting

the last year's wastewater levels. While effective for contextualizing SARS-CoV-2 RNA trend lines, the approach lacks warning capabilities. This is where the PCA approach proposed in this paper stands out. Finally, the PCA approach was coded in JavaScript and is open source so that it can be easily integrated in the existing web dashboards of the different national surveillance networks. The developed system allows detecting such abnormal behavior and tracing back the origin of the warning to a single WWTP. Notwithstanding this automation, a manual check is recommended for each warning generated to detect potential errors in the input data that may conduct to false alarms that mislead health authorities. Such a situation occurred once in our dataset, where an unexpected increase in 4.5 standard deviations was observed. All analytical parameters were within the quality control ranges and the only explanation was the discharge of sewage from cruise ships to the target WWTP. Yet, the system allowed identifying such anomaly and allowed the quality check of the data corresponding to that observation.

5. CONCLUSIONS

In this study, we have presented a method for generating network-level warnings in sewage surveillance networks for SARS-CoV-2. Our approach maximizes the value contributed by each individual WWTP to the whole network, highlighting that 'the whole is greater than the sum of its parts'. By utilizing a statistical process control technique, our method can identify deviations from normal behavior at the level of individual WWTPs, providing early warnings for potential epidemic outbreaks providing that a wastewater monitoring programme for the pathogen is already in place. Furthermore, our analysis shows that network-level warnings are not generated by a specific set of WWTPs on a recurrent basis, but rather all WWTPs contribute with relevant information at different times, and this contribution is independent of WWTP size. Our method for generating network-level warnings can provide a comprehensive picture of the spread of a pandemic across a territory, which can inform decision-making processes and intervention strategies by health authorities. By leveraging the value of individual WWTPs in the whole network, our method can maximize the potential for early detection and intervention, ultimately leading to more effective control of the pandemic. Last but not least, despite being applied for SARS-CoV-2 datasets, our method can be applied to other biological or chemical targets detected in urban sewage.

ACKNOWLEDGEMENTS

The study is within the frame of the virWASTE project, which has received funding from the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) under the call 'Pandèmies 2020' (Ref. 2020 PANDE 00044). ICRA authors acknowledge the Economy and Knowledge Department of the Catalan Government through Consolidated Research Groups 2021-SGR-01283 ICRA-TECH and 2021 SGR 01282 ICRA-ENV, and the funding from the CERCA program (Generalitat de Catalunya). The authors also acknowledge the Catalan Surveillance Network of SARS-CoV-2 in Sewage (SARSAIGUA) for the provision of data. We thank Miquel Calvo from UB for his contributions during the design phase of the study.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J. W., Choi, P. M., Kitajima, M., Simpson, S. L., Li, J., Tscharke, B., Verhagen, R., Smith, W. J. M., Zaugg, J., Dierens, L., Hugenholtz, P., Thomas, K. V. & Mueller, J. F. 2020 First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Sci. Total Environ.* **728**. https://doi.org/10.1016/j.scitotenv.2020.138764.

Bivins, A., North, D., Ahmad, A., Ahmed, W., Alm, E., Been, F., Bhattacharya, P., Bijlsma, L., Boehm, A. B., Brown, J., Buttiglieri, G., Calabro, V., Carducci, A., Castiglioni, S., Cetecioglu Gurol, Z., Chakraborty, S., Costa, F., Curcio, S., De Los Reyes, F. L., Delgado Vela, J., Farkas, K., Fernandez-Casi, X., Gerba, C., Gerrity, D., Girones, R., Gonzalez, R., Haramoto, E., Harris, A., Holden, P. A., Islam, M. T., Jones, D. L., Kasprzyk-Hordern, B., Kitajima, M., Kotlarz, N., Kumar, M., Kuroda, K., La Rosa, G., Malpei, F., Mautus, M., McLellan, S. L., Medema, G., Meschke, J. S., Mueller, J., Newton, R. J., Nilsson, D., Noble, R. T., Van Nuijs, A., Peccia, J., Perkins, T. A., Pickering, A. J., Rose, J., Sanchez, G., Smith, A., Stadler, L., Stauber, C., Thomas, K., Van Der Voorn, T., Wigginton, K., Zhu, K. & Bibby, K. 2020

Wastewater-based epidemiology: Global collaborative to maximize contributions in the fight against COVID-19. *Environ. Sci. Technol.* **54**. https://doi.org/10.1021/acs.est.0c02388.

CDC 2020 *CDC 2019-Novel Coronavirus (2019-nCoV) Real-Time RT-PCR Diagnostic Panel for Emergency Use Only Instructions for Use*. Atlanta.

Chan, E. M. G., Kennedy, L. C., Wolfe, M. K. & Boehm, A. B. 2023 Identifying trends in SARS-CoV-2 RNA in wastewater to infer changing COVID-19 incidence: Effect of sampling frequency. *PLoS Water* **2**, e0000088. https://doi.org/10.1371/journal.pwat.0000088.

Cluzel, N., Courbariaux, M., Wang, S., Moulin, L., Wurtzer, S., Bertrand, I., Laurent, K., Monfort, P., Gantzer, C., Guyader, S. L., Boni, M., Mouchel, J. M., Maréchal, V., Nuel, G. & Maday, Y. 2022 A nationwide indicator to smooth and normalize heterogeneous SARS-CoV-2 RNA data in wastewater. *Environ. Int.* **158**. https://doi.org/10.1016/j.envint.2021.106998.

Corominas, L., Collado, N., Guerrero-Latorre, L., Abasolo-Zabalo, N., Anfruns-Estrada, E., Anzaldi-Varas, G., Bofill-Mas, S., Bosch, A., Bosch-Lladó, L., Caimari-Palou, A. & Borrego, C. *et al.* 2021 Catalan Surveillance Network of SARS-CoV-2 in Sewage (1.58) [Data Set]. Zenodo [WWW Document].

Courbariaux, M., Cluzel, N., Wang, S., Maréchal, V., Moulin, L., Wurtzer, S., Mouchel, J. M., Maday, Y. & Nuel, G. 2022 A flexible smoother adapted to censored data with outliers and its application to SARS-CoV-2 monitoring in wastewater. *Front. Appl. Math. Stat.* **8**. https://doi.org/10.3389/fams.2022.836349.

Daughton, C. G. 2018 Monitoring wastewater for assessing community health: Sewage Chemical-Information Mining (SCIM). *Sci. Total Environ.* **619–620**, 748–764. https://doi.org/10.1016/j.scitotenv.2017.11.102.

Departament de Salut n.d. Registre de casos de COVID-19 a Catalunya per àrea bàsica de salut (ABS) i sexe [WWW Document]. Dades obertes de Catalunya. Available from: https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-a-Catalunya-per-rea-/xuwf-dxjd (accessed 4 May 2022).

Escolà Casas, M., Schröter, N. S., Zammit, I., Castaño-Trias, M., Rodriguez-Mozaz, S., Gago-Ferrero, P. & Corominas, L. 2021 Showcasing the potential of wastewater-based epidemiology to track pharmaceuticals consumption in cities: Comparison against prescription data collected at fine spatial resolution. *Environ. Int.* **150**, 106404. https://doi.org/10.1016/j.envint.2021.106404.

Fonseca Casas, P., Garcia Subirana, J. & Garcia Carrasco, V. 2023 Modeling SARS-CoV-2 true infections in Catalonia through a digital twin. *Adv. Theor. Simul.* **6**. https://doi.org/10.1002/adts.202200917.

González-Mariño, I., Baz-Lomba, J. A., Alygizakis, N. A., Andrés-Costa, M. J., Bade, R., Bannwarth, A., Barron, L. P., Been, F., Benaglia, L., Berset, J. D., Bijlsma, L., Bodík, I., Brenner, A., Brock, A. L., Burgard, D. A., Castrignanò, E., Celma, A., Christophoridis, C. E., Covaci, A., Delémont, O., Devoogt, P., Devault, D. A., Dias, M. J., Emke, E., Esseiva, P., Fatta-Kassinos, D., Fedorova, G., Fytianos, K., Gerber, C., Grabic, R., Gracia-Lor, E., Grüner, S., Gunnar, T., Hapeshi, E., Heath, E., Helm, B., Hernández, F., Kankaanpaa, A., Karolak, S., Kasprzyk-Hordern, B., Krizman-Matasic, I., Lai, F. Y., Lechowicz, W., Lopes, A., de Alda, M. L., López-García, E., Löve, A. S. C., Mastroianni, N., McEneff, G. L., Montes, R., Munro, K., Nefau, T., Oberacher, H., O'brien, J.W., Oertel, R., Olafsdottir, K., Picó, Y., Plósz, B.G., Polesel, F., Postigo, C., Quintana, J.B., Ramin, P., Reid, M.J., Rice, J., Rodil, R., Salgueiro-Gonzàlez, N., Schubert, S., Senta, I., Simões, S.M., Sremacki, M.M., Styszko, K., Terzic, S., Thomaidis, N.S., Thomas, K. V., Tscharke, B.J., Udrisard, R., van Nuijs, A.L.N., Yargeau, V., Zuccato, E., Castiglioni, S. & Ort, C. 2020 Spatio-temporal assessment of illicit drug use at large scale: Evidence from 7 years of international wastewater monitoring. *Addiction* **115**, 109–120. https://doi.org/10.1111/add.14767.

Guerrero-Latorre, L., Collado, N., Abasolo, N., Anzaldi, G., Bofill-Mas, S., Bosch, A., Bosch, L., Busquets, S., Caimari, A., Canela, N., Carcereny, A., Chacón, C., Ciruela, P., Corbella, I., Domingo, X., Escoté, X., Espiñeira, Y., Forés, E., Gandullo-Sarró, I., Garcia-Pedemonte, D., Girones, R., Guix, S., Hundesa, A., Itarte, M., Mariné-Casadó, R., Martínez, A., Martínez-Puchol, S., Mas-Capdevila, A., Mejías-Molina, C., Rafa, M. M. i., Munné, A., Pintó, R. M., Pueyo-Ros, J., Robusté-Cartró, J., Rusiñol, M., Sanfeliu, R., Teichenné, J., Torrell, H., Corominas, L. & Borrego, C. M. 2022 The Catalan Surveillance Network of SARS-CoV-2 in sewage: Design, implementation, and performance. *Sci. Rep.* **12**. https://doi.org/10.1038/s41598-022-20957-3.

Huisman, J. S., Scire, J., Caduff, L., Fernandez-Cassi, X., Ganesanandamoorthy, P., Kull, A., Scheidegger, A., Stachler, E., Boehm, A. B., Hughes, B., Knudson, A., Topol, A., Wigginton, K. R., Wolfe, M. K., Kohn, T., Ort, C., Stadler, T. & Julian, T. R. 2022 Wastewater-based estimation of the effective reproductive number of SARS-CoV-2. *Environ. Health Perspect.* **130**. https://doi.org/10.1289/EHP10050.

Jackson, J. & Mudholkar, G. 1979 Control procedures for residuals associated with principal component analysis. *Technometrics* **21**, 341–349.

Johnson, R. & Wichern, D. 2002 *Applied Multivariate Statistical Analysis*, 5th edn. Prentice-Hall Inc, Upper Saddle River, NJ, USA.

Jolliffe, I. T. 2002 *Principal Component Analysis, Springer Series in Statistics*. Springer-Verlag, New York. https://doi.org/10.1007/B98835.

Joseph-Duran, B., Serra-Compte, A., Sàrrias, M., Gonzalez, S., López, D., Prats, C., Català, M., Alvarez-Lacalle, E., Alonso, S. & Arnaldos, M. 2022 Assessing wastewater-based epidemiology for the prediction of SARS-CoV-2 incidence in Catalonia. *Sci. Rep.* **12**. https://doi.org/10.1038/s41598-022-18518-9.

Mattei, M., Pintó, R. M., Guix, S., Bosch, A. & Arenas, A. 2023 Analysis of SARS-CoV-2 in wastewater for prevalence estimation and investigating clinical diagnostic test biases. *Water Res.* **242**. https://doi.org/10.1016/j.watres.2023.120223.

Morvan, M., Jacomo, A. L., Souque, C., Wade, M. J., Hoffmann, T., Pouwels, K., Lilley, C., Singer, A. C., Porter, J., Evens, N. P., Walker, D. I., Bunce, J. T., Engeli, A., Grimsley, J., O'Reilly, K. M. & Danon, L. 2022 An analysis of 45 large-scale wastewater sites in England to estimate SARS-CoV-2 community prevalence. *Nat. Commun.* **13**. https://doi.org/10.1038/s41467-022-31753-y.

Naughton, C. C., Roman, F. A., Alvarado, A. G. F., Tariqi, A. Q., Deeming, M. A., Kadonsky, K. F., Bibby, K., Bivins, A., Medema, G., Ahmed, W., Katsivelis, P., Allan, V., Sinclair, R. & Rose, J. B. 2023 Show us the data: Global COVID-19 wastewater monitoring efforts, equity, and gaps. *FEMS Microbes* **4**. https://doi.org/10.1093/femsmc/xtad003.

Nourbakhsh, S., Fazil, A., Li, M., Mangat, C. S., Peterson, S. W., Daigle, J., Langner, S., Shurgold, J., D'Aoust, P., Delatolla, R., Mercier, E., Pang, X., Lee, B. E., Stuart, R., Wijayasri, S. & Champredon, D. 2022 A wastewater-based epidemic model for SARS-CoV-2 with application to three Canadian cities. *Epidemics* **39**. https://doi.org/10.1016/j.epidem.2022.100560.

Pasteur, I. 2020 *Protocol: Real-Time RT-PCR Assays for the Detection of SARS-CoV-2*.

Schmitt, E., Rato, T., De ketelaere, B., Reis, M. & Hubert, M. 2016 Parameter selection guidelines for adaptive PCA-based control charts. *J. Chemom.* **30**, 163–176. https://doi.org/10.1002/CEM.2783.

Shao, X. T., Zhao, Y. T., Jiang, B., Li, Y. Y., Lin, J. G. & Wang, D. G. 2023 Evaluation of three chronic diseases by selected biomarkers in wastewater. *ACS ES&T Water* **3**, 943–953. https://doi.org/10.1021/acsestwater.2c00452.

Singer, A. C., Thompson, J. R., Filho, C. R. M., Street, R., Li, X., Castiglioni, S. & Thomas, K. V. 2023 A world of wastewater-based epidemiology. *Nat. Water* **1**, 408–415. https://doi.org/10.1038/s44221-023-00083-8.

First received 28 January 2024; accepted in revised form 3 June 2024. Available online 12 June 2024