Modeling the number of people infected with SARS-COV-2 from wastewater viral load in Northwest Spain

Juan A. Vallejo [a,1], Noelia Trigo-Tasende [a,1], Soraya Rumbo-Feal [a,1], Kelly Conde-Pérez [a], Ángel López-Oriona [b], Inés Barbeito [b], Manuel Vaamonde [b], Javier Tarrío-Saavedra [b,*], Rubén Reif [c], Susana Ladra [d], Bruno K. Rodiño-Janeiro [e], Mohammed Nasser-Ali [a], Ángeles Cid [f,g], María Veiga [g], Antón Acevedo [h], Carlos Lamora [i], Germán Bou [a], Ricardo Cao [b,j,*], Margarita Poza [a,f,g,**]

[a] Microbiology Research Group, University Hospital Complex (CHUAC), Institute of Biomedical Research (INIBIC), University of A Coruña (UDC), Servicio de Microbiología, 3ª planta, Edificio Sur, Hospital Universitario, As Xubias, 15006, A Coruña, CIBER de enfermedades infecciosas, Spain
[b] Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña (UDC), Facultade de Informática, Campus de Elviña, 15071 A Coruña, Spain
[c] Centro Singular de Investigación en Química Biolóxica e Materiais Moleculares (CiQUS), Universidade de Santiago de Compostela, 15782-Santiago de Compostela, Spain
[d] Database Laboratory, Research Center for Information and Communication Technologies (CITIC), University of A Coruña (UDC), Campus de Elviña, 15008 A Coruña, Spain
[e] BFlow, Campus Vida, Universidade Santiago de Compostela and Instituto de Investigatión Sanitaria de Santiago de Compostela, Edificio Emprendia, Avenida do Mestre Mateo, 2, 15706 Santiago de Compostela, A Coruña, Spain
[f] Department of Biology, University of A Coruña (UDC), Campus da Zapateira, 15071 A Coruña, Spain
[g] Advanced Scientific Research Center (CICA), University of A Coruña (UDC), As Carballeiras, Campus de Elviña 15071 A Coruña, Spain
[h] General Directorate of Social Health Care, Xunta de Galicia, Edificios Administrativos, San Caetano s/n, 15781 Santiago de Compostela, A Coruña, Spain
[i] Public Wastewater Treatment Plant Company EDAR Bens, S.A., Lugar de Bens, 15010 A Coruña, Spain
[j] Technological Institute for Industrial Mathematics (ITMATI), Universities of A Coruña (UDC), Santiago de Compostela (USC) and Vigo (UVIGO), Campus Vida, Rúa de Constantino Candeira, 15705 Santiago de Compostela, Spain

HIGHLIGHTS

- Spanish seroprevalence study helped to estimate the number of COVID-19 active cases.
- COVIDBENS models estimate the number of people infected from wastewater data.
- A quadratic LOESS model provides the best predictive results.
- A simple linear regression model accurately predicts the number of people infected.
- Information from COVIDBENS helps public health authorities in the decision-making.

GRAPHICAL ABSTRACT
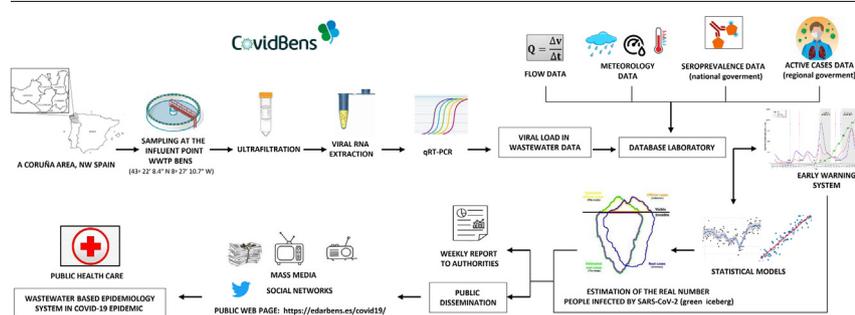
ARTICLE INFO

ABSTRACT

The quantification of the SARS-CoV-2 RNA load in wastewater has emerged as a useful tool to monitor COVID–19 outbreaks in the community. This approach was implemented in the metropolitan area of A Coruña (NW Spain), where wastewater from a treatment plant was analyzed to track the epidemic dynamics in a population of 369,098 inhabitants. Viral load detected in the wastewater and the epidemiological data from A Coruña health system served as main sources for statistical models developing. Regression models described here allowed us to estimate the number

* Corresponding authors.
** Correspondence to: M. Poza, Microbiology Research Group, University Hospital Complex (CHUAC), Institute of Biomedical Research (INIBIC), University of A Coruña (UDC), Servicio de Microbiología, 3ª planta, Edificio Sur, Hospital Universitario, As Xubias, 15006, A Coruña, CIBER de enfermedades infecciosas, Spain.
E-mail addresses: javier.tarrio@udc.es (J. Tarrío-Saavedra), rcao@udc.es (R. Cao), margarita.poza.dominguez@sergas.es (M. Poza).
[1] Authors contributed equally.

of infected people ($R^2 = 0.9$), including symptomatic and asymptomatic individuals. These models have helped to understand the real magnitude of the epidemic in a population at any given time and have been used as an effective early warning tool for predicting outbreaks in A Coruña municipality. The methodology of the present work could be used to develop a similar wastewater-based epidemiological model to track the evolution of the COVID–19 epidemic anywhere in the world where centralized water-based sanitation systems exist.

1. Introduction

During the last decade, wastewater–based epidemiology (WBE) has emerged as a highly relevant discipline with the potential to provide objective information by combining the use of cutting-edge analytical methodologies with the development of ad hoc modeling approaches. WBE has been extensively used in the last years to predict with high accuracy the consumption patterns of numerous substances (EMCDDA, 2020). Several examples from the literature showed different approaches and strategies to tackle the uncertainty associated with WBE studies. For example, Goulding and Hickman assumed three main sources of uncertainty and, using Bayesian statistics, fitted the data to linear regression hierarchical models (Goulding et al., 2020). Other modeling approaches (Croft et al., 2020) considered Monte Carlo simulations to deal with uncertainties such as wastewater inflow variability or stability of the substances in wastewater and their pharmacokinetics. In general, WBE studies showed that despite the wide number of parameters involved in predicting the consumption rate of a specific substance, a correct selection of assumptions combined with a thoughtful modeling process will overcome such uncertainty, leading to accurate results.

Previous studies have evaluated the presence in wastewater of other viruses (Ehlers et al., 2005; Hellmér et al., 2014; Hovi et al., 2012; Lizasoain et al., 2018; Mancini et al., 2019). During the present global COVID–19 pandemic, processes to monitor SARS–CoV–2 in wastewater were first developed in the Netherlands (Medema et al., 2020), followed by other countries (Nemudryi et al., 2020; Wurtzer et al., 2020; Ahmed et al., 2020; La Rosa et al., 2020; Randazzo et al., 2020b; Peccia et al., 2020; Balboa et al., 2021; Weidhaas et al., 2021; Pillay et al., 2021; Kumar et al., 2021; Hart and Halden, 2020).

A range varying from one third to four fifths of patients infected with SARS–CoV–2 are asymptomatic (Bi et al., 2020; Day, 2020; Pollán et al., 2020); a condition that depends on many factors such as the mean age in the population and that promotes the undetected spread of COVID–19. A systematic literature review found that at least an important proportion of COVID–19 infected persons, including symptomatic and asymptomatic people, tested for fecal viral RNA were positive from initial steps of infection (Gupta et al., 2020) and persistently tested positive on rectal swabs even after nasopharyngeal testing was negative (Chen et al., 2020; Xing et al., 2020; Xu et al., 2020; Zhang et al., 2020; Cevik et al., 2021; Miura et al., 2021). For example, it has been reported that excretion of viral RNA in the stool of people infected occurs during a mean prolonged period of 27.9 days after the person has tested negative in their respiratory samples (Y. Wu et al., 2020). Therefore, genetic material of SARS–CoV–2 can be found in wastewater (Lodder and de Roda Husman, 2020), which has made monitoring of viral RNA load in sewage an excellent tool for the epidemiological tracking of the actual pandemic as well as an extremely efficient early warning tool for outbreaks detection (Randazzo et al., 2020a; Ahmed et al., 2020; Medema et al., 2020; Peccia et al., 2020; F. Wu et al., 2020; Wurtzer et al., 2020).

These studies showed the potential of monitoring SARS–CoV–2 levels in wastewater and sewage sludge to track and even pre-empt outbreaks in the community. The works of Lahrich et al. (2021) and Bandala et al. (2021) are two comprehensive monographs on the subject.

Previous studies estimated the number of SARS–CoV-2 infected patients in a community using simple equations and Monte Carlo simulation (Ahmed et al., 2020; Curtis et al., 2020; Hasan et al., 2021). A similar study was made in Barcelona elaborating a simple model to predict the total number of SARS-CoV-2 shedders during the first wave of the pandemic, showing a high proportion of asymptomatic individuals (Chavarria-Miró et al., 2021). In addition, Saththasivam et al. (2021) developed a suitable WBE mathematical model for monitoring the trend in the numbers of infected people in Qatar. The Susceptible-Exposed-Infectious-Removed (SEIR) model based on mass rate of SARS–CoV–2 RNA in wastewater was also used to estimate the number of infected patients in the state of South Carolina (McMahan et al., 2021). Regarding the forecasting of the prevalence of virus shedders within a community, Cao and Francis (2021) have proposed the fitting of VAR models, whereas Li et al. (2021) have used multiple linear regression, artificial neural network, and adaptive neuro fuzzy inference systems.

On March 9th 2020, the city of A Coruña, in the region of Galicia, Northwest of Spain, reported circulation of SARS–CoV–2 for the first time, with data of a COVID–19 outbreak in a civic center that affected 11 people (GCiencia, 2020b, 2020a). At that point, a surveillance phase on approximately 250 people began, and the recommendations from the Health Department on cleaning and mobility restrictions were followed (GCiencia, 2020a). During these initial days of the COVID–19 epidemic in Galicia, most of the cases were in the municipality of A Coruña. Thus, at noon on March 13th, the Xunta de Galicia (Government of the Autonomous Community of Galicia) reported 90 confirmed cases of COVID–19 in Galicia, 43 of them in the A Coruña area (GCiencia, 2020b). The Spanish Government declared a state of alarm on March 14th throughout the country, at which point the Galician community still had few cases. Despite this, in a few days the region went from a monitored and controlled situation to an exponential growth of cases (Rey, 2020), reaching a peak of 1667 active cases. The cases were distributed in an area that covers 37 municipalities in the health area of A Coruña–Cee, as shown by the data provided by SERGAS (Galician Health Service) in https://www.datawrapper.de/_/QrkrZ (SERGAS, 2021).

In this context, the main objective of the present work was to develop parametric and nonparametric statistical models useful to determine the entire SARS–CoV–2 infected population, including symptomatic and asymptomatic people, by tracking the viral load present in the wastewater of the Bens wastewater treatment plant that serves the metropolitan area of A Coruña with near 370,000 residents, without the need of health system data or the number of positive people reported, and obtaining information from population-based seroepidemiological surveys developed in Spain (this represents a contribution with respect to other models). The pursuit of this objective has a public service motivation. At this regard, it is important to note that this research, in the framework of the COVIDBENS project, had high social impact and it was one of the precursors of this type of monitoring in Spain, for the surveillance of SARSCoV–2 in wastewater. COVIDBENS currently provides a public service through weekly reports to municipalities, public health and regional administrations (SERGAS, Xunta de Galicia) for surveillance and early warning tasks.

2. Materials and methods

2.1. Sample collection

The wastewater treatment plant (WWTP) Bens (43° 22′ 8.4″ N 8° 27′ 10.7″ W, A Coruña, Spain) serves the metropolitan area of A Coruña. It

includes the municipalities of A Coruña, Oleiros, Culleredo, Cambre and Arteixo, which correspond to a geographical area of 277.8 km^2 and to a population of 369,098 inhabitants (Fig. 1). The wastewater samples were collected by automatic samplers installed both at the entrance of the WWTP Bens and in a sewer collecting sewage from COVID–19 patients housed on 7 floors of the University Hospital of A Coruña (CHUAC). At the WWTP Bens, 24-h composite samples were collected from April 15th until June 4th, while at CHUAC 24–h composite samples were collected from April 22nd to May 14th 2020. Dataset S1 includes the dates when each sample was collected in the two locations considered: the WWTP Bens and the CHUAC. A total of 33 24–h composite samples were collected from the WWTP and 19 from CHUAC. In addition, samples were collected at 2–h intervals for 24 h on specific days at the WWTP Bens and at CHUAC. A total of 36 2–h intervals samples were collected from the WWTP and 12 from CHUAC. The 24-h composite samples were collected by automatic samplers taking 150 mL of wastewater every 15 min in 24 bottles (1 h per bottle) and, when the 24-h collection ended, the 24 bottles were integrated and a representative sample of 100 mL was collected. For the collection of 2 h intervals, 2 bottles obtained every 2 h were integrated, finally providing 12 bottles per day.

2.2. Sample processing

Samples of 100 mL were processed immediately after collection at 4 °C. Firstly, 100 mL samples were centrifuged for 30 min at 4000 ×g and then filtered through 0.22 μm membranes (Merck Millipore). Regenerated cellulose Amicon Ultra Filters of 30 kDa (Merck Millipore) were used to concentrate and dialyze, keeping the fraction retained, in 500 μL of a buffer containing 50 mM Tris-HCL, 100 mM NaCl and 8 mM MgSO$_4$. Samples were preserved in RNAlater reagent (Sigma-Aldrich) at −80 °C.

2.3. RNA extraction and qRT-PCR assays

RNA was extracted from 100 μL of the concentrated samples using the QIAamp Viral RNA Mini Kit (Qiagen, Germany) according to manufacturer's instructions. Briefly, the sample was lysed under highly denaturing conditions to inactivate RNases and to ensure isolation of intact viral RNA. Then, the sample was loaded in the QIAamp Mini spin column where RNA was retained in the QIAamp membrane. Samples were washed twice using washing buffers. Finally, RNA was eluted in 70 μL of RNase-free water. The quality and quantity of the RNA was checked using a Nanodrop Instrument and an Agilent Bioanalyzer. Samples were kept at −80 °C until use. RT-qPCR assays were done in a CFX 96 System (BioRad, USA) using the qCOVID–19 kit (GENOMICA, Spain) through N gene (coding for nucleocapsid protein N) amplification. Hard–Shell 96–well PCR plates (BioRad, USA) were used and sealed with microseal 'B' PCR Plate Sealing Film (BioRad, USA). Recommendations given by Ahmed et al. (2022) were followed for minimizing errors in RT–PCR detection. RT–qPCR reactions were done following the manufacturer's instructions (GENOMICA, Spain). This kit provided two reaction mixes; Mix A contained DNA polymerase, nucleotides, polymerase buffer and an internal control and Mix B contained primers and probe for N gene. The final reaction contained 5 μL of Mix A, 1 μL of Mix B, 0.2 μL of reverse transcriptase, 5 μL of template and 8.8 μL of water. The internal control allowed discarding the presence of inhibitors. Nuclease–free water was used as negative-control template. The cycling parameters were 50 °C for 20 min for the retrotranscription step, followed a PCR program consisting of a preheating cycle of 95 °C for 2 min, 50 cycles of amplification at 95 °C for 5 s and finally one cycle of 60 °C for 30 s. RT-qPCR assays were done in sextuplicate.

For RNA quantification, a reference pattern was standardized using the Human 2019-nCoV RNA standard from European Virus Archive Glogal (EVAg) (Fig. S1). To build the calibration curve, the decimal logarithm of SARS–CoV–2 RNA copies per μL of control material ranging from 5 to
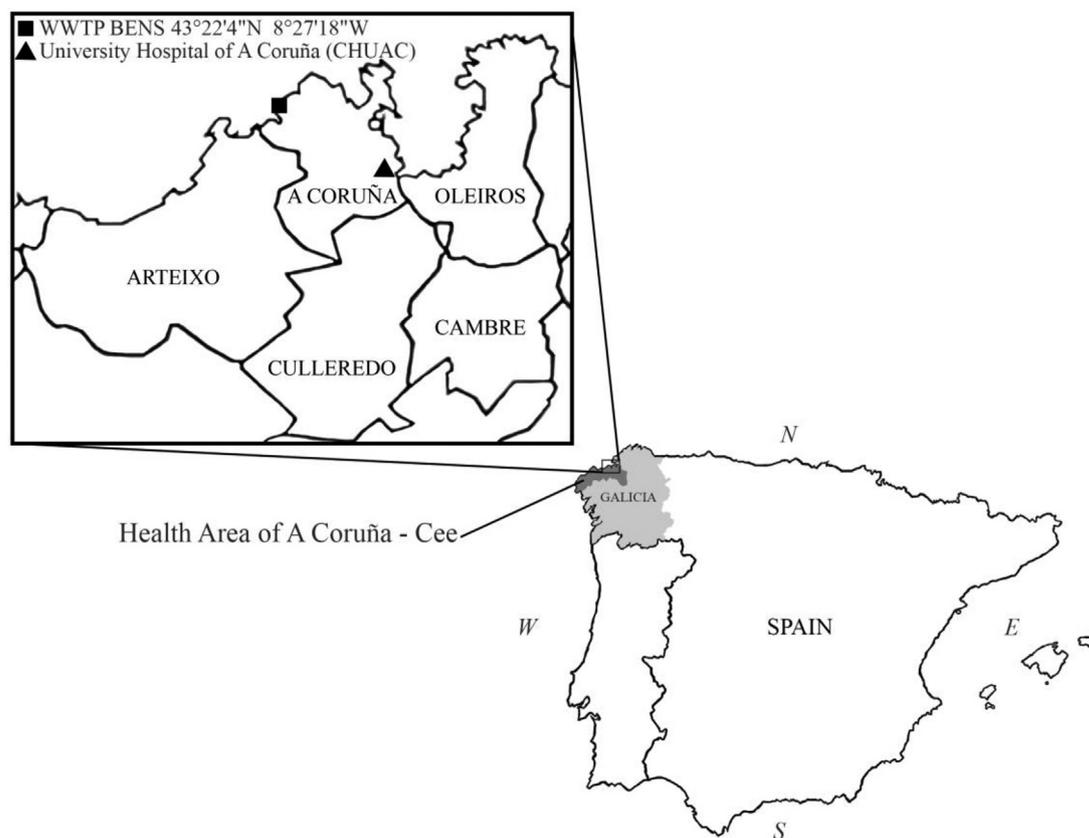
Fig. 1. Map showing Galicia region in Spain, the metropolitan area of A Coruña including Oleiros, Cambre, Culleredo, Arteixo and A Coruña municipalities, and the Health Area of A Coruña-Cee, as well as the specific locations of the University Hospital of A Coruña (CHUAC) and WWTP Bens.

500 were plotted against Cq (quantification cycle) values. Calibration was done amplifying the N gene. Viral load was defined as number of copies of RNA of SARS–CoV–2 per L.

The analytical efficiency of the RT–qPCR was calculated following the criteria recommended by the MIQE guidelines (Bustin et al., 2009). The limit of detection (LOD) of the RT–qPCR was analyzed using serial dilutions of Human 2019-nCoV RNA standard (EVAg). Detection rates are shown in Table S1 (in Supplementary material section), where the LOD was established in 25 copies per reaction and limit of quantification (LOQ) was between 25 and 15 copies per reaction. The slope of the standard curve was -3.56, the R^2 was 0.9984 and the amplification efficiency was above 90%.

2.4. Data collection

In the present work a COVID–19 case was defined as a person with COVID–19 virus infection laboratory-confirmed by RT-PCR of SARS–CoV–2 regardless of clinical signs and symptoms. Active cases mean living persons confirmed with COVID–19 whose symptom onset date is less than or equal to 14 days from the date of the current report.

Data gathered from different sources were used for the present study. Dataset S2 integrates daily observations at the meteorological station of Coruña-Bens, including rainfall, temperature, and humidity, for the period March 1st–May 31st, 2020, obtained from the Galician Meteorology Agency (METEOGALICIA, 2021). Dataset S3 includes cumulative and active number of COVID–19 cases in the metropolitan area of A Coruña and the health area A Coruña–Cee for the period March 1st–May 31st, 2020, reported by the Galician Health Service (SERGAS), the General Directorate of Public Health (Autonomous Government of Galicia) and the University Hospital of A Coruña (CHUAC). Since flow may be an important variable when determining the viral load in the wastewater, an exploratory data analysis for the volume of water pumped at the WWTP Bens during the lockdown period has been performed using flow data. Dataset S4 includes two-minute flow measurements ($m^3 \cdot s^{-1}$) at WWTP Bens for the period January 1st–May 14th. An additional flow study is described in the Supplementary material section.

2.5. Backcasting of COVID–19 active cases

Preliminary statistical methods have been devised to backcast the number of COVID–19 active cases based on reported official cases. Follow-up times (available only until May 7th) for anonymized individual reported COVID–19 cases in Galicia (NW Spain where the WWTP Bens is located, Fig. 1) have been used to count the number of cases by municipality based on patient zip codes. Since the epidemiological discharge date is missing, the number of active cases in the metropolitan area of A Coruña could not be obtained but the cumulative number of cases was computed. On the other hand, the main epidemiological series for COVID–19 were publicly available daily in Galicia at the level of health areas. However, the definition of one of the series changed from cumulative cases to active cases in April 29th. Thus, the epidemiological series for COVID–19 in the health area of A Coruña – Cee (population 551,937) was used to estimate the epidemiological series for COVID–19 for the metropolitan area of A Coruña (population 369,098). To do this, a linear regression model was used to relate the relative cumulative and active cases (cases per million) of COVID–19 for the health area of A Coruña – Cee. Predicting the rate of active cases and considering the population size in the metropolitan area gives the estimated total number of official active cases in the five municipalities. The previous approach is only possible until May 7th, our database update date. To estimate the number of official active cases from May 8th onwards, another linear regression model has been used to relate the number of active cases in the health area of A Coruña – Cee and in the metropolitan area of A Coruña. Since the number of active cases in the health area has been reported until June 5th, the series of estimated official active cases could be backcasted from May 8th until June 5th. Finally, to transform the official number of COVID–19 cases into the real number, the ratio mean of real cases/mean of official cases was estimated using the

official figures of cumulative cases. The Spanish seroepidemiological survey made by the National Center of Epidemiology in Spain, ENECOVID (Pollán et al., 2020), included representative samples within the metropolitan area of A Coruña during the period, April–June 2020. These data have been used to quantify the real proportion of population infected during the period April–June in the area corresponding to WWTP Bens. Accordingly, the number of actual active cases in Galicia was: 56,713 for April 27th–May 11th (prevalence 2.1%) and 59,414 for May 18th–June 1st (prevalence 2.2%). Confronting these numbers with the official numbers in May 11th (10,669) and June 1st (11,308) gives estimated ratios of 5.316 and 5.254 in these two periods, with an average of around 5.29. This conversion factor was used to backcast the series of real active cases based on the estimated daily official COVID–19 cases in the metropolitan area of A Coruña. Some of these series, including the backcasted series of real active cases, are included in the Dataset S3.

2.6. Nonparametric fitting of viral load overtime

In order to properly interpret the results of fitting models applied aiming understand the viral load evolution, a brief description of them is included as follows, namely of the Generalized Additive Models and the Locally Estimated Scatterplot Smoothing (LOESS). In fact, GAM using a basis of cubic regression splines (Hastie and Tibshirani, 1990) and LOESS (Cleveland, 1979) nonparametric regression models have been used to fit the viral load along the day on May 5th, 6th, 11th and 12th, and as a function of time at CHUAC from April 22nd to May 12th and at WWTP Bens from April 16th to June 3rd. Several outliers, showing very small viral load (some of them under the level of detection), have been removed from the data, corresponding to unexpected and intensive pipeline cleaning episodes (8-hour 70 °C water cleaning during Thursday-Friday nights) carried out in April 23rd-24th, April 30th - May 1st and May 7th–8th.

Nonparametric methods do not assume any parametric form (Maity, 2017), allowing us to model complex nonlinear functions (Shokrzadeh et al., 2014). There is a wide range of nonparametric regression models, such as local polynomial regression (Fan and Gijbels, 2018; Breidt and Opsomer, 2000), kernel smoothing (Wand and Jones, 1994), regression splines (Wood, 2017), and LOESS (Cleveland, 1979), among others.

There is other group of models, called semiparametric regression models, that allow the inclusion of linear parametric and flexible nonparametric effects of the predictors on the response, such as the GAM models. In this work, GAM models are applied to define the relation between the viral load and the date, by using penalized regression splines (Wood, 2017) to estimate the smooth effect of date on the viral load. But this type of models can include linear effects in addition to smooth effects. For instance, a GAM model expression consisting in one linear, X_1, and two smooth predictors, T_1 and T_2, can be defined by

$$Y = \beta_0 + \beta_1 X_1 + s_1(T_1) + s_2(T_2) + s_{12}(T_1, T_2) + \varepsilon,$$

whereby β_0 and β_1 are the parameters that define the linear effect of X_1 on the response Y, $s_1(T_1)$ and $s_1(T_1)$ are the smooth effects of T_1 and T_2 variables, $S(T_1, T_2)$ accounts for the smooth effect of their interaction, and ε is the random error, i.e. the difference between Y and its conditional mean provided by the model. In the present case, Y is assumed to be Gaussian distributed, but instead it could be assumed, for instance, Binomial or Poisson distributed, in which case Y should be replaced by $g(\eta)$. In the framework of this section, $Y = ViralLoad$ and $X = Time$, thus:

$$E(Viral\,Load|Time) = \beta_0 + s_{Time}(Time).$$

s accounts for the smooth effects of the predictors (T) and are usually estimated by using penalized regression splines. Splines are piecewise polynomial functions of a specific degree that are continuous in a set of knots $a < t_1 < t_2 < \cdots < t_L < b$ (up to $L - 1$ degree) defined in an interval $[a, b]$.

Thus, the smooth effect of *T* can be expressed by means a splines basis composed of *L + degree* elements,

$$s_j(t) = \sum_{k=1}^{L+degree} \beta_{jk} \phi_{jk}(t) = \beta'_{j} \phi_j,$$

where ϕ_{jk} are the splines functions and *L* accounts for the number of knots. In this work *degree = 3* (cubic regression splines) was chosen.

2.7. Viral load models

In order to find a useful statistical model to estimate the number of real infected cases, including symptomatic and asymptomatic people, well-known regression models, such as simple and multivariate linear models, and more flexible models, such as nonparametric (e.g. local linear polynomial regression) and semiparametric (GAM and LOESS) models, have been formulated.

As in the previous section, in order to be able to correctly interpret the results of the application of the models, a brief introduction from a statistical point of view is provided beforehand.

Parametric models such as linear regression, the simplest one, show a transfer function that defines the relation of dependence between two or more variables involved in a specific process. This function depends on the value of some parameters that can be interpreted from physical or chemical point of views, among others. We can define a simple parametric model by

$$Y = m(X) + \varepsilon, \quad \text{with} \quad E(\varepsilon) = 0,$$

where *Y* is the response variable to be modeled, and *m* is the parametric function, depending on *X*, the independent variable. In the present case, *Y* is the number of real infected people, *X* accounts for the values of the logarithm of viral load in wastewater, while ε are the model random errors. The regression function is assumed to be an element of a set of parametric functions $\mathcal{M} = \{m_\theta / \theta \in \Theta\}$, whereby θ is the vector of parameters that defines the regression model and Θ is a subset of $\mathbb{R}^k$.

On the other hand, the semiparametric models used are flexible ones that allow the introduction of linear and smooth effects of the predictors on the response. All these models have been successfully used to predict the number of COVID–19 active cases based on the measured viral load (number of SARS–CoV–2 RNA copies/L) at WWTP Bens, daily flow in the sewage network as well as other environmental variables, such as rainfall, temperature and humidity. Diagnostic tests (Q-Q plots, residuals versus fitted values plots and Cook's distance) were used for outlier detection, which improved the models fit. The R statistical software was used to perform statistical analyses (R Core Team, 2021). Namely, the mgcv library (Wood, 2017) was applied to fit GAM models and ggplot2 and GGally (Schloerke et al., 2020; Ginestet, 2011) to perform correlation analysis, obtain graphical output and fit LOESS models, respectively. The caret R package was used to fit and evaluate regression models (Kuhn et al., 2008).

Although some RT-qPCR replicates could not be measured when the viral load was scarce, due to the limitation of the detection technique (errors randomly occur when the number of copies/L is under 10,000), 74% of the assays led to three or more measured replications. Conditional mean imputation (Enders, 2010) was used for unmeasured replications. In those cases, the missing values along the replications were imputed with the sample mean of the non-missing values of the replications. When all the replications were missing, they were imputed as the lowest value of all the observed measurements along the whole dataset.

2.8. Public dissemination

With the aim of helping in the awareness of the population and in the management of the pandemic, the data obtained were disseminated at first through private reports to the political and health authorities. Later, data were disseminated through social networks, press, radio, TV and finally a public web page was created in order to inform weekly the society about the COVID–19 pandemic situation in the metropolitan area of A Coruña city assessed by COVIDBENS (COVIDBENS, 2021). The web site can be found in the link https://edarbens.es/covid19/.

3. Results

3.1. Estimated COVID–19 positive cases in the metropolitan area of A Coruña

To model the viral load, the number of COVID–19 positive cases need to be estimated (Fig. 2). Iceberg on Fig. 2 represents the global health of the population of the metropolitan area of A Coruña based in our study, where visible cases are differentiated from the invisible ones separated by a black line. The real (and unknown) iceberg is represented in blue, which includes real cases with or without symptoms, visible and invisible. The iceberg represented in green corresponds to the number of people infected estimated using the statistical models presented here. The green iceberg has been drawn with a slightly different shape and size but very similar to blue, taking into account the 10% margin of error of our models. Visible cases estimated in this study (in yellow) against the real (and unknown) ones (in orange) are also represented in Fig. 2. Finally, the asymptomatic cases known thanks to the national seroprevalence study (Pollán et al., 2020) are represented in purple.

However, due to difficulties in determining the exact number of positive cases, mathematical models had to be developed based on data recovered in Dataset S3. Thus, linear regression models (Fig. 3) were successfully used to estimate COVID–19 positive cases in the A Coruña – Cee health area, where the region of this study is located (Fig. 1). This was based on Intensive Care Unit (ICU) patients before April 29th (Fig. 3A) and, from this date on, on positive cases reported by health authorities in Galicia. A linear regression trend was fitted to predict the proportion of positive cases in the health area of A Coruña – Cee based on the proportion of cumulative cases (Fig. 3B). This linear regression fit was finally used to estimate the positive cases in the metropolitan area of A Coruña served by the WWTP Bens, by means of the proportion of cumulative positive cases in the same area, which was directly obtained from the individual patient data.

3.2. Daily variation of SARS-CoV-2 RNA load in the metropolitan area of A Coruña and hospital

The evolution of the viral load along the day both at the A Coruña metropolitan area and at the University Hospital Complex of A Coruña (CHUAC) was monitored in order to discern which amount of the viral load detected in the WWTP Bens came from the CHUAC and which from the community. An analysis of the viral load of the 24–h and 2-h samples collected at the WWTP Bens and CHUAC, as reflected in sampling schedule presented in Dataset S1, was performed. Dataset S5 includes RT–qPCR results in viral load (number of SARS–CoV–2 RNA copies/L) and Cq values obtained from the 24–h samples. Dataset S6 includes viral load and Cq values obtained from the 2–h samples. Because of the small sample size, nonparametric LOESS models were used in order to prevent the possible overfitting of alternatives such as GAM. Fig. S2 (in Supplementary material) shows the viral load trends at CHUAC (Fig. S2A) and at WWTP Bens (Fig. S2B) depending on the hour of the day, during four different days. Fig. S2A shows the hourly trend at CHUAC, with a maximum around 08:00, whereas the viral load curves at WWTP Bens (Fig. S2B) attained a minimum around 05:00 and a maximum between 14:00 and 15:00. As a consequence of these trends, 24-hour integrated samples are fully justified.

3.3. Lockdown de-escalation in the metropolitan area of A Coruña

As expected, the mean viral load decreased with time when measured at CHUAC (Fig. S3A of Supplementary material, late April–mid May) and at WWTP Bens (Fig. S3B, mid April–early June) following an asymptotic trend (fitted using GAM with cubic regression splines). Time course quantitative detection of SARS–CoV–2 in WWTP Bens correlated with the
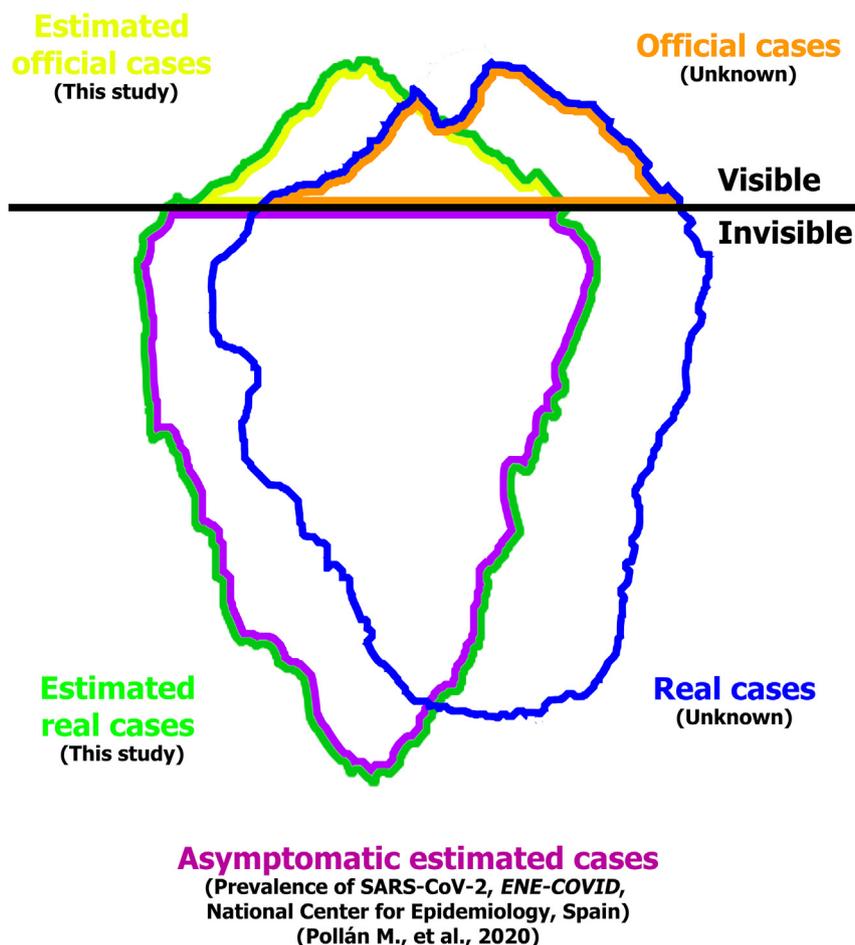
Fig. 2. Iceberg representing the overall health of the population of the metropolitan area of A Coruña infected by SARS–CoV–2, showing the real and official cases estimated in this study. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

estimated number of COVID–19 positive cases, as shown in Fig. 4. The number of copies of viral RNA per liter decreased from around 500,000 to less than 1000, while the estimated cases of patients infected by SARS–CoV–2 decreased approximately 6–fold in the same period, reaching in both cases the lowest levels in the metropolitan area at the beginning of June 2020.

3.4. Wastewater epidemiological models based on viral load for COVID–19 real active cases prediction

Firstly, a correlation analysis (Fig. 5) was done finding that estimated number of COVID–19 cases strongly correlated linearly with the logarithm of daily mean viral load at Bens ($R = 0.923$) and with the mean flow ($R = -0.362$). Nonetheless, a strong inverse linear relationship between estimated COVID–19 positive cases and time was found ($R = -0.99$).

Then, different regression models were tested to predict the real number of COVID–19 active cases based on the viral load and the most relevant atmospheric variables. The best results were obtained using GAM models depending on the viral load and the mean flow (Fig. 6). The effect of the viral load on the COVID–19 estimated active cases showed a logarithmic shape (Fig. 6A), which suggested that the number of COVID–19 real active cases could be modeled linearly as a function of the logarithm of the viral load. On the other hand, the shape of the effect of the mean flow on the estimated number of COVID–19 real active cases appears to be quadratic (Fig. 6B), but its confidence band was wide and contained the horizontal line with height zero, which means that the effect of the mean flow was not significant (p-value $= 0.142$). Therefore, the only significant independent variable was the viral load, with $R^2 = 0.86$.

Since the nonparametric estimation of the viral load effect had a logarithmic shape, a multiple linear model was fitted using the logarithmic transformation of the viral load, daily flow, rainfall, temperature, and humidity. The best option to explain the number of COVID–19 cases is a simple linear model as a function of the as a function of the natural logarithm of viral load. In fact, the R^2 hardly change when adding Flow, Rainfall, Humidity and Temperature variables (see Fig. S4 in Supplementary material that shows the most explanatory models for different number of predictors using the R^2 maximization criterion). In fact, when a multivariate linear model depending on three predictors (viral load, daily flow, and rainfall) was fitted, data showed that the only significant explanatory variable was the viral load (p-value $= 1.32 \times 10^{-8}$). Table 1 shows that the effect of the other two predictors, daily flow (p-value $= 0.186525$) and rainfall (p-value $= 0.099239$), were not clearly significant.

Finally, ignoring the rest of the explanatory variables, only the natural logarithm of the viral load gave a good linear model fit ($R^2 = 0.851$) that was useful to predict the real number of active COVID–19 cases (Fig. 7A). After removing three outliers, the fit improved slightly ($R^2 = 0.894$), as shown in Fig. 7B.

The final fitted linear model became:

$$Y = -7079 + 1059 \times \ln X$$

where Y denotes the real number of active COVID–19 cases, X is the viral load (number of RNA copies per L) and ln stands for the natural logarithm.

For instance, a viral load of $X = 150,000$ copies per liter would lead to an estimated number of $Y = 5543$ active cases.

A)

B)
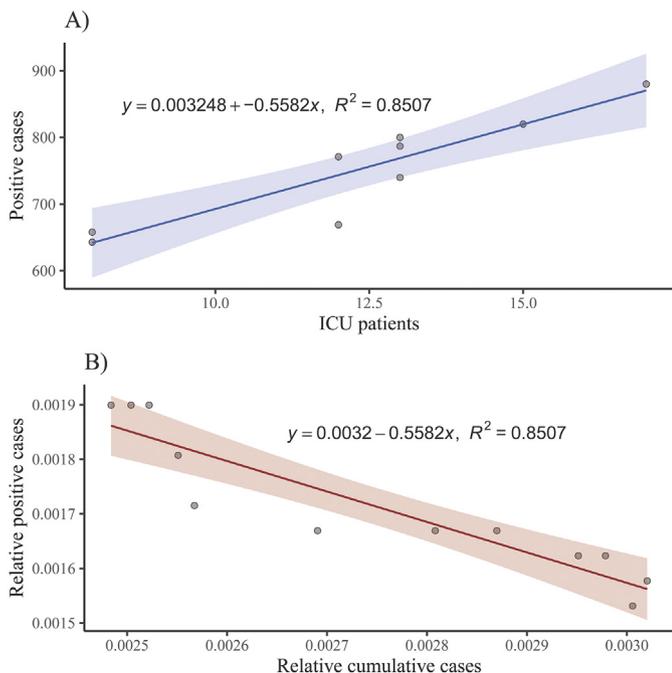
Fig. 3. Estimation of the COVID–19 positive cases using simple linear regression models. A) ICU patients versus COVID–19 positive cases in A Coruña – Cee health area, and its linear fit for the period April 29th on. B) Relative cumulative cases versus relative positive cases in the health area of A Coruña – Cee and their linear fit.

The prediction ability of this fitted linear model, the GAM, and the linear and quadratic LOESS models has been evaluated using a 6-fold cross validation procedure, to prevent overfitting. In all cases, the response variable was the estimated number of real COVID–19 active cases in the metropolitan area (Fig. 2), and the explanatory variable, the natural logarithm of the viral load. Table 2 shows the corresponding prediction R^2 for each one of the four models, along with the root mean squared prediction error (RMSPE). The smaller this error, the better the predictive ability of the model was. All the models provided quite accurate predictions for the real number of COVID–19 active cases using the viral load, with an error of around 10% of the response range. The model with the lowest prediction error, 9.5%, was the quadratic LOESS model. Flexible models, such as LOESS and GAM, slightly improved the predictive performance when compared with the linear model, which has a prediction error of around 11.4%
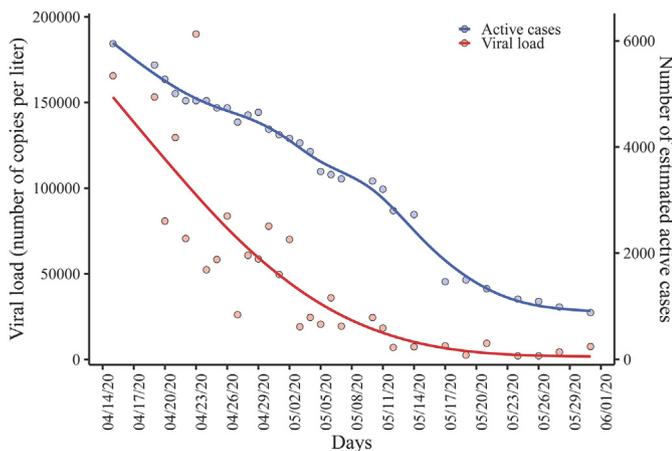
Fig. 4. Viral load detected in the influent of the WWTP Bens (red line) and the estimated number of COVID–19 positive cases (blue line) in the metropolitan area of A Coruña.

of the response range. The quadratic LOESS model was also the one with the largest value for R^2. Therefore, it provided the best predictive results.

A scatterplot of the estimated number of COVID–19 active cases in the metropolitan area versus the natural logarithm of the viral load, along with the quadratic LOESS fitted curve show the linear relationship between these two variables (see Fig. 8A). The actual and predicted values of real number of COVID–19 active cases are very close to the diagonal line, which represents perfect model prediction (see Fig. 8B).

The above mentioned linear model has been used to make estimates of the total number of infected population with COVID–19 in the A Coruña metropolitan area from April 2020 on. In this regard, it is worth mentioning that the correlation coefficient between the logarithm of the viral load and the number of reported active cases 9 days after by the health authorities was 0.87 in the period July–December 2020. Thus, both the number of COVID–19 active cases for each day reported by the health services of Galicia (SERGAS) and the estimates of the number of infected persons (symptomatic and asymptomatic) obtained by the proposed model have been monitored in the framework of the COVIDBENS, so that the estimates can serve as an early warning indicator of possible outbreaks. Fig. 9A shows the trends of reported active cases and estimates of active cases for 2020 (from 17/07/2020), whereas Fig. 9B shows the trends corresponding to 2021 (until 07/04/2020). This figure is split in two panels taking into account that the dominant SARS-CoV-2 variants in A Coruña during 2020 were replaced by SARS-CoV-2 Alpha variant at the beginning of 2021. The evolution of variants can be followed at the link of the website of the Spanish Ministry of Health, Consumer Affairs and Social Welfare (Spanish Health Ministery, 2021), https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/situacionActual.htm.

4. Discussion

In the pandemic context described above, 24–h composite samples from WWTP Bens were continuously analyzed from April 19th until early June, although surveillance has continued until now. The data from wastewater obtained from April 19th to June 1st has confirmed the decrease in COVID–19 incidence. We showed that time course quantitative detection of SARS–CoV–2 in wastewater from WWTP Bens correlated with COVID–19 confirmed cases, which backs up the plausibility of our approach. The previous seroprevalence studies carried out by the Spanish Centre for Epidemiology showed that cases in A Coruña represented about 1.8% of the local population (Pollán et al., 2020). This means that, for a population of about 369,098 inhabitants, the number of people infected with SARS–CoV–2 contributing their sewage into the WWTP Bens would be around 6644, which includes people with symptoms and those who are asymptomatic. Considering that the ratio between people with symptoms (reported by the health service) and the total infected population (including asymptomatic people) was estimated to be 1:5, we estimated that reported cases contributing their wastewater into WWTP Bens would be around 1661, which is close to the maximum number of cases reported in the A Coruña-Cee area (1667 cases on April 28th) therefore these data support our study. Of course, as reflected above, this ratio presents uncertainty and can be discussed. It should be noted that the criteria used by the authorities to report cases varied over time, so this may explain the gap between the graphs reported in the media throughout the epidemic and our Fig. 4, where both a decrease in the viral load and in the estimated COVID–19 cases can be observed from mid–April to early–June. However, the level of the curve at WWTP Bens at the beginning of May was much higher than that corresponding to May 11th. This is due to the effectiveness of the lockdown measures applied in Spain. The May 12th daily curve at CHUAC showed a higher viral load than the one corresponding to May 11th at WWTP Bens, showing that the viral load measured at the hospital tends to be higher than at WWTP Bens due to a lower dilution effect, as expected. One of the key points of this work is the use of statistical regression models for the estimation of real cases of COVID–19 from the information provided by wastewater, in the specific framework of the metropolitan area of A Coruña. In fact, in the present work, nonparametric and even simple
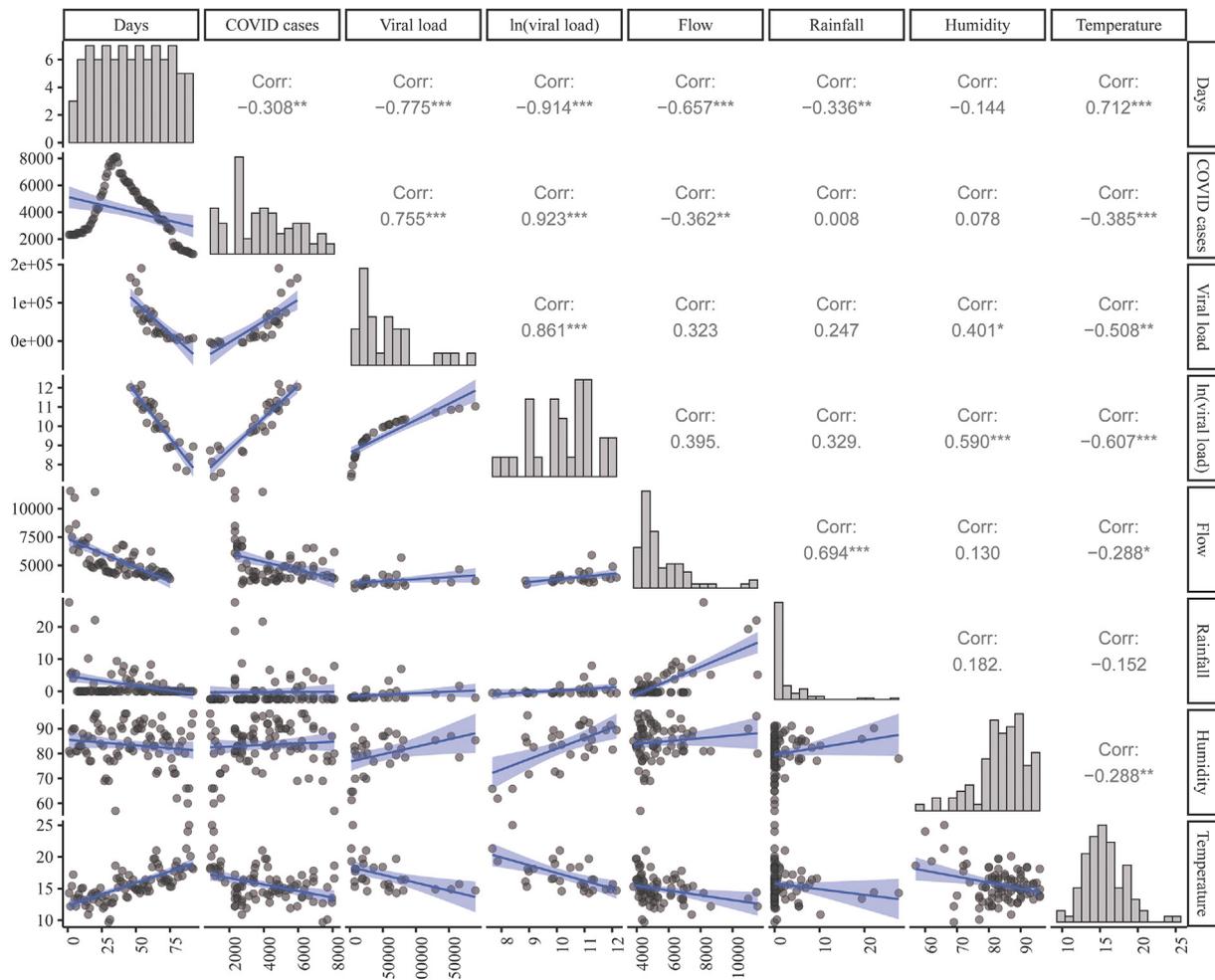
Fig. 5. Correlation analysis between estimated COVID–19 positive cases and different variables. Scatterplot matrix shows fitted linear models and linear correlation coefficients of each pair of variables: time (measured in days from the beginning of reported COVID–19 cases), estimated COVID–19 positive cases (positive cases), daily mean viral load measured in Bens, mean flow of sewage water in Bens, mean rainfall, daily mean humidity and daily mean temperature.

parametric regression models have been shown to be useful tools to estimate the real number of COVID–19 active cases as a function of the viral load. This is a pioneering approach in the context of the SARS-CoV-2 pandemic, specifically in the framework of Spain. To our knowledge, the WBE studies available before the performance of our proposal (Vallejo et al., 2020) tended to be limited to reporting the occurrence of SARS–CoV–2 RNA in WWTPs and sewer networks, in order to establish a direct comparison with declared COVID–19 cases (Randazzo et al., 2020b; Medema et al., 2020; Nemudryi et al., 2020; La Rosa et al., 2020; Randazzo et al., 2020a; Polo et al., 2020). However, there are a number of studies that at similar dates proposed mathematically based models for estimating the number of COVID–19 cases. In this line, Hart and Halden (2020) is one of the precedent works that combines computational analysis and modeling with a theoretical approach in order to identify useful variables and confirm the feasibility and cost-effectiveness of WBE as a prediction tool. We have to mention also the studies of Ahmed et al. (2020) and Curtis et al. (2020) that proposed parametric transfer functions depending on WBE variables to estimate the number of persons infected. In addition, recently, new works focused on providing and comparing regression and time series models for the prediction of active cases have been developed (Cao and Francis, 2021; Li et al., 2021). The present methodology is characterized by focusing on a particular region and studying carefully all the steps, from sample collection to the estimation of the actual number of infected persons. This significantly differentiates our method, and the resulting predictive model, from other valid and useful alternatives. In addition,

it is important to note that we focus on estimating the actual number of infected people rather than the number of cases reported by health authorities, as other methodologies do. For this task, the present approach uses the seroprevalence studies carried out by the Spanish Centre for Epidemiology. Moreover, the statistical regression models included in this methodology range from parametric to non-parametric approaches such as GAM or LOESS, which allow much more flexible data fitting, not subject to a fixed parametric expression relating the response variable and predictors. Other WBE models have been applied to previous outbreaks of other infectious diseases. For example, during a polio outbreak detected in Israel in 2013–2014, a disease transmission model was optimized incorporating environmental data (Brouwer et al., 2018). Given the availability of clinical information on poliovirus, the developed infectious disease model incorporated fully validated parameters such as the transmission and vaccination rates, leading to accurate estimations of incidence. This type of study highlights one of the main challenges we have faced developing our model: the SARS-CoV-2 novelty and the very few epidemiologic information available at the early stage of the pandemic. Considering this, our proposed model has minimized the uncertainty implementing a complete set of hydraulic information from the sewer network of the city of A Coruña. This ad hoc model can be adapted to other scenarios as long as similar hydraulic information can be obtained from the region where it will be implemented. Relevant possible explanatory variables (such as rainfall or the mean flow) did not enter the model. Although this is a bit counterintuitive (dilution should affect the viral load measured), it is important to point out that
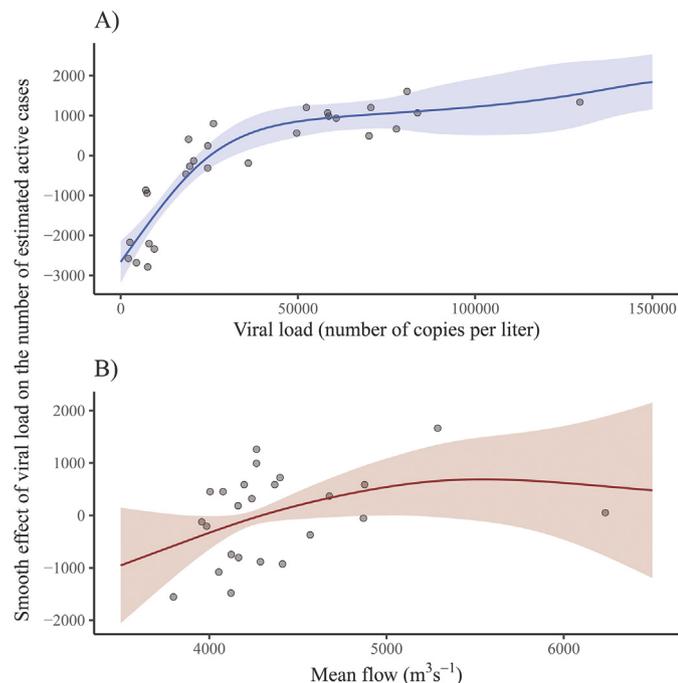
Fig. 6. Smooth effect (solid line) and confidence band (shaded area) for the viral load (A) and for the mean flow (B) on the number of estimated COVID–19 active cases when fitting a GAM.

Table 1
Signification analysis of the multivariate linear model to explain the number of active cases as a function of viral load, and mean temperature.

	Estimate	Standard error	t value	Test *p*-value
Intercept	−5433.37	1805.63	−3.009	0.00562
ln(Viral load)	1008.25	107.32	9.395	5.33e−10
Mean Temperature	−72.66	52.97	−1.372	0.18144

rainfall fluctuated little during the data collection period from mid-April to early–June 2020: its median was 0, its mean was 2.88 L/m^2 and its standard deviation was 6.59 L/m^2. Obviously, flow is an important factor to consider
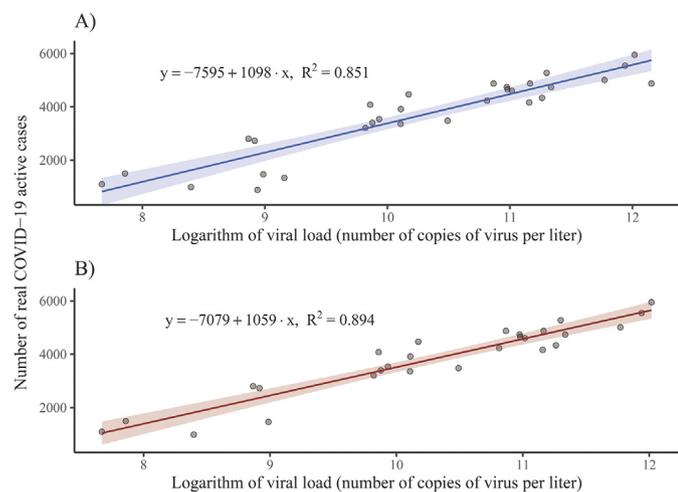
Fig. 7. Estimation of the number of COVID–19 real active cases using a linear regression model. Scatterplot represents the logarithm of the viral load measured in WWTP Bens and the estimated number of COVID–19 real active cases before (A) and after (B) removing the three outliers detected. The linear fit (solid line) and the confidence band (shaded area) are also included.

Table 2
R^2 and root mean squared prediction error (RMSPE) corresponding to different regression models explaining the number of real active cases in the metropolitan area as a function of the natural logarithm of the viral load. RMSPE was obtained using a 6-fold cross validation procedure.

Model	R^2	RMSPE
Linear	0.8515	581.94
GAM	0.8767	508.62
LOESS (linear)	0.8695	487.97
LOESS (quadratic)	0.8833	478.33

in WBE. In our case, before starting to model, the study correlations study between several variables was done, of course including flow. Since this work was done at time of drought, it was found that, contrary to expectations, flow was not a variable that entered the model. The strongest correlation was established between the log of the viral load and the number of active cases (R^2 = 0.93). Of course, our models can be affected in the rainy season. However, as soon as the rainy season came, large distortions in our models were detected caused by the appearance of the Alpha and Delta variants (data not shown). Also, the water temperature remains stable in our geographical area and never raises enough to cause excessive degradation of viral RNA, so this variable does not enter the models either, as expected. It is important to highlight that, as a consequence of the results of the GAM fit, a simple linear model was considered to fit the estimated number of COVID–19 active cases as a function of the logarithm of the viral load. The log-linear relationship between viral concentration and the number of active cases of COVID–19 can be explained by exponential degradation of SARS–CoV–2 RNA during transportation through the sanitary network. As pointed out by Deagle et al. (2006) for DNA samples, if RNA damage occurs according to a random Poisson process, then an exponential decline in the amount of amplifiable product with increasing product size is expected. If this is the case and the logarithm of the amount of amplifiable product is linearly related with the original amount of SARS–CoV–2 RNA material on excretion. The percentage of variability explained by this linear
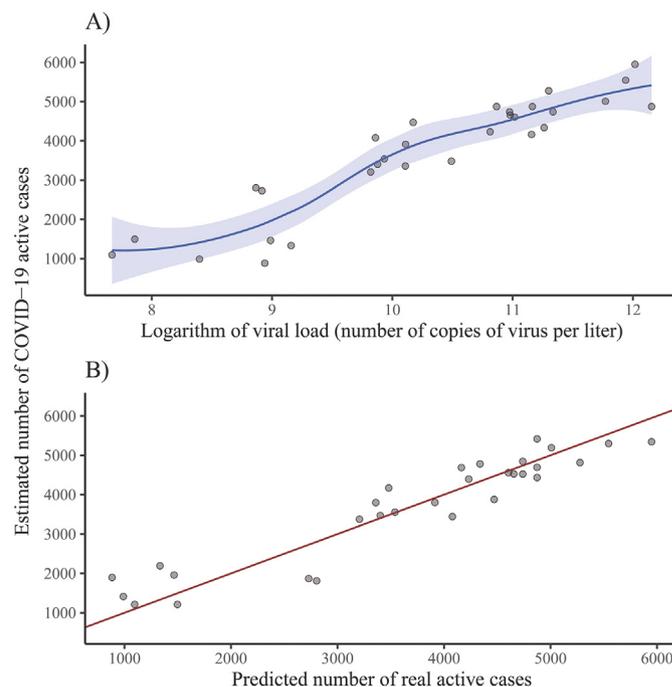
Fig. 8. Estimation of number of real COVID–19 active cases using the quadratic LOESS model. A) Estimated COVID–19 active cases in the metropolitan area vs the natural logarithm of the viral load. B) Estimated number of COVID–19 active cases in the metropolitan area vs the predicted number of COVID–19 real active cases.
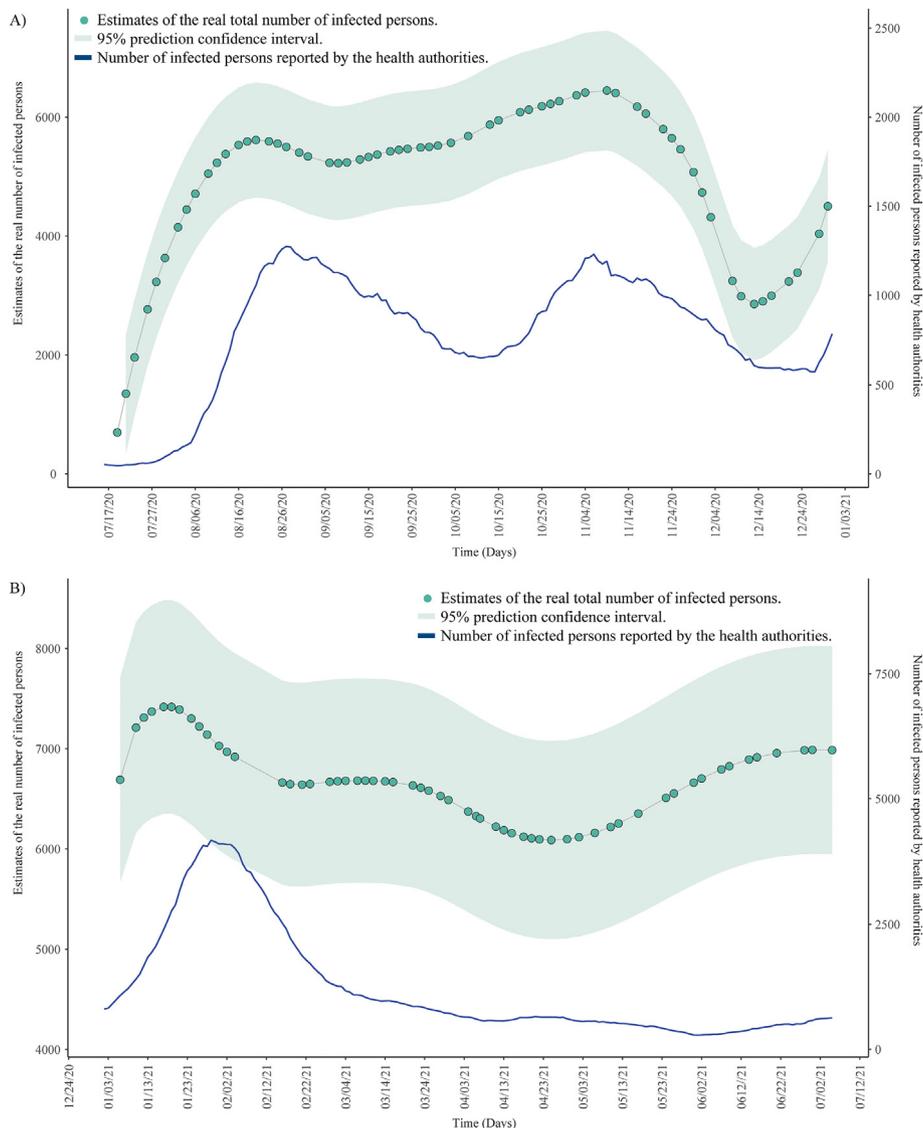
Fig. 9. Estimation of number of real COVID–19 active cases using the linear model (in green), including the prediction confidence intervals. These estimates are compared with the number of cases of infected persons reported by the Galician health services (in blue). A) Trends corresponding to 2020. B) Trends corresponding to 2021.

model was reasonably high (85.1%) increasing up to 89.4% when three outliers detected were excluded. Some outliers were data obtained from samples collected during cleaning episodes, where viral RNA was almost totally degraded, leading to discordant data. Alternative, more flexible models, such as GAM and LOESS, were also fitted. They produced slightly better results in terms of R^2 and RMSPE. The quadratic LOESS model avoided overfitting and showed a good predictive ability (R^2 = 0.88, *RMS PE* = 478), the best among all the considered models. However, similar results were found for the linear model that also brings the advantage of simplicity. Therefore, both models, linear and LOESS quadratic, could be successfully used to estimate the number of infected people in a given region based on viral load data obtained from wastewater.

The proposed models, as described, are only applicable to the metropolitan area of A Coruña, the region for which they have been developed, although can be adapted to any other location. This area has Atlantic weather and it may rain substantially in autumn and winter, which could lead to explanatory variables such as rainfall and/or mean flow becoming significant for those seasons and needing to enter the prediction model. Thus, when applying these models to the same location but in seasons with different climatic behavior, they might need to be reformulated.

In addition, the methodology used to build these statistical models could be used at other locations for epidemiological COVID–19 outbreak detection, or even for other epidemic outbreaks caused by other microorganisms. Of course, in that case a detailed data analysis would have to be carried out as well, since specific features of the sewage network or the climate may affect the model itself.

Overall results showed that the estimations of the real total number of infected persons are significantly higher than the number of reported infected persons. Both curves are very different in terms of scale. However, regarding the shape of both trends, it is very similar, providing relevant information of the evolution of the pandemic in Spain. Moreover, it is important to note that the curve of estimates is slightly out of phase with respect to the number of active cases reported by health authorities. In fact, it is slightly ahead (around two weeks) contributing to early information about the beginning of each outbreak, such as those corresponding to August 2020, November 2020 or January 2021 (data not shown). This information has provided support for decision making to the public authorities, involving decisions such as confinement, mobility and time restrictions. In addition, the model estimates have given an idea of the real progress of the pandemic in our region, complementing the information

on active cases reported by public administrations. This work has proposed a methodology to estimate the real number of infected persons of COVID–19 (symptomatic and asymptomatic) in A Coruña metropolitan area from the viral load measured in its wastewater. It is a comprehensive and multi-disciplinary procedure that includes all the steps from the wastewater sample collection to the estimation of actual infected persons by parametric or nonparametric regression models. The close collaboration between the public wastewater company and an interdisciplinary research team of microbiologists, chemists, engineers, statisticians and computer scientists has resulted in the implementation of an alert system in the metropolitan area of A Coruña. The results have been weekly delivered since July 2020 up to now to public authorities of A Coruña metropolitan area, Galicia and Spain to support their decision making during the pandemic, being one of the pioneer systems in Spain. The usefulness of this work has transcended the administrative sphere to provide a service open to all citizens of the metropolitan area, initially through the media (La Voz de Galicia, 2021; La Opinión de A Coruña, 2021; Bravo, 2021) and then through its own website (COVIDBENS, 2021) with an important social impact. The successive web reports have helped citizens to know about the evolution of the pandemic in our metropolitan area, even anticipating each of the successive outbreaks, also facilitating decision making at personal, family and business level.

5. Study limitations

The first serious limitation to start this work was the beginning of the total confinement of the first outbreak decreed by the national government when the COVIDBENS research team was created on purpose for this work. This limited, of course, interaction between the researchers. We had to ask for special permits to be able to carry out the laboratory work in the hospital and do the sample collection. Then the lack of reagents came; for RNA purification, for PCR, for suitable devices and personal protective material such as masks, gloves, etc. The lack of reagents did not allow us to make replicates at the very first time, which could only be done later. The control of recovery efficient was also impossible due to the unavailability of marker viruses at that moment. Also, the lack of reliable data of COVID–19 cases was an important limitation, therefore statisticians had to create a statistical model to be able to estimate the real number of cases, since the information at that first moment of the pandemic was very biased and incomplete. Moreover, some periodically cleaning events made with hot water in the wastewater collection system of the hospital originated confusing data during the first weeks of the study, since hot water practically eliminates viral RNA. Due to lack of several materials and reagents a great effort in sample collection and processing had to be done and finally the number of samples was enough to develop the statistical models as soon as possible. Variability found in data obtained by PCR when viral load was very low was another important limitation, which has been partially solved by making six replications of each RT-qPCR. Also, ratios fixed for the calculation of symptomatic cases and total cases are another source of uncertainties that should be taken into account. The mean ratio between real COVID–19 cases and official COVID–19 cases was estimated using the Spanish sero–epidemiological survey, which age a mean ratio estimation of 5.29. The mean ratio was rather stable over the time in our study since it ranged between 5.316 (May 11th) and 5.254 (June 1st). However, this mean ratio may be quite different in further periods of the pandemic. The statistical models described in the present study were done with data recovered only during the first epidemic outbreak. This is a limitation but, nevertheless, this fact allowed having early statistical models that, since July 2020 were able to estimate the number of infected people in the population and collaborate with the public health system. Other important limitation is that the models described here did not depend on rainfall, probably because the data were obtained in spring, a season of little rain in NW Spain. So in order for the models to be used in rainy seasons, they must be adapted. Also adaptation has to be done for using this type of models in other locations. However, the authors believe they could be of help to adapt these models to any other place in the world where there is a wastewater sanitation system. Finally, despite the serious limitations, we decided to exploit the statistical models as soon as possible to help in the fight against the pandemic.

6. Conclusions

This work aimed to evaluate the viral load measured in wastewater as a predictor of the real number of COVID–19 active cases. Different regression models were fitted to predict the real number of COVID–19 active cases based on the viral load, flow and the most relevant atmospheric variables. These statistical models have been able to provide estimates, from April 2020 up to now, of the real magnitude of the epidemic at the metropolitan area of A Coruña. These estimates have been weekly reported to the local public authorities. A variable selection method based on the maximization of the adjusted R^2 was used. The only predictor with a significant effect on the real number of infected persons was the viral load, both when GAM and multivariate linear models were fitted. In the GAM model, the effect of the viral load on the real number of active cases had a logarithmic shape. Thus, the viral load was introduced in the multivariate linear model using its log-arithmic transformation. As a result, very explanatory GAM ($R^2 = 0.86$) and linear ($R^2 = 0.851$) models were fitted. Their goodness-of-fit increased when outliers are removed ($R^2 = 0.894$). It is important to stress that this simple model, $Y = -7079 + 1059 \times \ln X$, is valid for this specific case of the metropolitan area of A Coruña. In other locations, with different weather, or in other seasons, the role of the other potential explanatory variables could be more important. In addition, the reliability of the model predictions could change during time due to different causes such as the change of SARS–CoV–2 variants. Moreover, the prediction ability of the linear, the GAM, and the linear and quadratic LOESS models has been evaluated using a 6–fold cross validation through the prediction R^2 and the RMSPE index. The prediction errors of all the models were relatively small, supporting their use for prediction tasks. The lowest prediction error corresponded to the quadratic LOESS model. Finally, we have used the linear model to predict the number of real active cases using viral load. The reasons are its low prediction error, its simplicity, and the fact that it provides a parametric function that defines the relationship between variables. The present methodology, including the fitting of the parametric linear or nonparametric statistical models (GAM, LOESS), can be extended to estimate the real number of infected people at other locations. In addition, it is important to note that their sampling cost–effectiveness and speed can help to early alert health authorities about potential new outbreaks, thereby helping to protect the local population.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.152334.

Data availability

The authors declare that all data supporting the findings of this study are available within the article and Supplementary information files, and also are available from the corresponding authors on reasonable request.

CRediT authorship contribution statement

MP, RC, CL, JAV, JT-S and RR conceived and designed the study. MP, JAV, NT, SR-F, MN and KC-P performed wastewater processing and viral analysis, RC, AL-O, IB, MV and JT-S performed statistical models and data analysis, SL managed and analyzed data, BKR-J assisted in the study design and analysis, AA assessed in data collection, AC supervised the wastewater analysis, MCV assessed in wastewater sampling, GB and MP supervised the microbiology team, MP, JAV, RC, RR and JT-S wrote the manuscript. MP, JT-S and RC supervised the team and coordinated all tasks.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

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., et al., 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, 138764.

Ahmed, W., Simpson, S.L., Bertsch, P.M., Bibby, K., Bivins, A., Blackall, L.L., Bofill-Mas, S., Bosch, A., Brandão, J., Choi, P.M., et al., 2022. Minimizing errors in RT-PCR detection and quantification of SARS-CoV-2 RNA for wastewater surveillance. Sci. Total Environ. 805, 149877. https://doi.org/10.1016/j.scitotenv.2021.149877.

Balboa, S., Mauricio-Iglesias, M., Rodríguez, S., Martínez-Lamas, L., Vasallo, F.J., Regueiro, B., Lema, J.M., 2021. The fate of SARS-COV-2 in wwtps points out the sludge line as a suitable spot for detection of COVID-19. Sci. Total Environ. 772, 145268.

Bandala, E.R., Kruger, B.R., Cesarino, I., Leao, A.L., Wijesiri, B., Goonetilleke, A., 2021. Impacts of covid-19 pandemic on the wastewater pathway into surface water: a review. Sci. Total Environ. 774, 145586.

Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S.A., Zhang, T., et al., 2020. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. Lancet Infect. Dis. 20 (8), 911–919.

Bravo, I., 2021. La depuradora de bens remite todas las semanas a la xunta sus datos sobre el coronavirus en aguas residuales (the Bens wastewater treatment plant sends its data on coronavirus in wastewater to the Xunta –government of Galicia– every week). Cadena SER radio. https://cadenaser.com/emisora/2020/11/17/radio_coruna/1605620618_965641. html;. (Accessed 14 July 2021).

Breidt, F.J., Opsomer, J.D., 2000. Local polynomial regression estimators in survey sampling. Ann. Stat. 1026–1053.

Brouwer, A.F., Eisenberg, J.N., Pomeroy, C.D., Shulman, L.M., Hindiyeh, M., Manor, Y., Grotto, I., Koopman, J.S., Eisenberg, M.C., 2018. Epidemiology of the silent polio outbreak in Rahat, Israel, based on modeling of environmental surveillance data. Proc. Natl. Acad. Sci. 115 (45), E10625–E10633.

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., et al., 2009. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-time PCR Experiments.

Cao, Y., Francis, R., 2021. On forecasting the community-level COVID-19 cases from the concentration of SARS-CoV-2 in wastewater. Sci. Total Environ. 786, 147451.

Cevik, M., Tate, M., Lloyd, O., Maraolo, A.E., Schafers, J., Ho, A., 2021. SARS–CoV–2, SARS–CoV, and MERS–CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. Lancet Microbe 2 (1), 13–22.

Chavarria-Miró, G., Anfruns-Estrada, E., Martínez-Velázquez, A., Vázquez-Portero, M., Guix, S., Paraira, M., Galofré, B., Sánchez, G., Pintó, R.M., Bosch, A., 2021. Time evolution of severe acute respiratory syndrome coronavirus 2 (SARS–CoV–2) in wastewater during the first pandemic wave of covid-19 in the metropolitan area of barcelona, spain. Appl Environ Microbiol 87 (7) e02750–20.

Chen, Y., Chen, L., Deng, Q., Zhang, G., Wu, K., Ni, L., Yang, Y., Liu, B., Wang, W., Wei, C., et al., 2020. The presence of SARS-CoV-2 RNA in the feces of COVID-19 patients. J. Med. Virol. 92 (7), 833–840.

Cleveland, W.S., 1979. Robust locally weighted regression and smoothing scatterplots. J. Am. Stat. Assoc. 74 (368), 829–836.

COVIDBENS, P., 2021. COVIDBENS Project Web Site. https://edarbens.es/covid19/. (Accessed 16 November 2021).

Croft, T.L., Huffines, R.A., Pathak, M., Subedi, B., 2020. Prevalence of illicit and prescribed neuropsychiatric drugs in three communities in Kentucky using wastewater-based epidemiology and Monte Carlo simulation for the estimation of associated uncertainties. J. Hazard. Mater. 384, 121306.

Curtis, K., Keeling, D., Yetka, K., Larson, A., Gonzalez, R., 2020. Wastewater SARS–CoV–2 concentration and loading variability from grab and 24–hour composite samples. MedRxiv https://doi.org/10.1101/2020.07.10.20150607. https://www.medrxiv.org/content/early/2021/07/21/2020.07.10.20150607. (Accessed 16 December 2021).

Day, M., 2020. Covid-19: Four Fifths of Cases are Asymptomatic, China Figures Indicate.

Deagle, B.E., Eveson, J.P., Jarman, S.N., 2006. Quantification of damage in DNA recovered from highly degraded samples–a case study on DNA in faeces. Front. Zool. 3 (1), 1–10.

Ehlers, M.M., Grabow, W., Pavlov, D.N., 2005. Detection of enteroviruses in untreated and treated drinking water supplies in South Africa. Water Res. 39 (11), 2253–2258.

EMCDDA, E.B., 2020. Wastewater Analysis and Drugs: A European Multi-city Study. European Monitoring Center for Drugs and Drug Addiction, pp. 1–14.

Enders, C.K., 2010. Applied Missing Data Analysis. Guilford press.

Fan, J., Gijbels, I., 2018. Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66. Routledge.

GCiencia, E.B., 2020. Galicia acumulates 90 cases of SARS-CoV-2 (Galicia suma 90 casos de SARS-CoV-2). https://www.gciencia.com/extra/galicia-incidencia-casos-coronavirus/. (Accessed 31 March 2021).

GCiencia, E.B., 2020. The outbreak of the civic center of A Coruña accumulates 11 cases of coronavirus (O foco do centro cívico da Coruña suma xa 11 casos de coronavirus). https://www.gciencia.com/saude/centro-civico-coruna-coronavirus/. (Accessed 31 March 2021).

Ginestet, C., 2011. ggplot2: elegant graphics for data analysis. Wiley Online Library.

Goulding, N., Hickman, M., Reid, M., Amundsen, E.J., Baz-Lomba, J.A., O'Brien, J.W., Tscharke, B.J., de Voogt, P., Emke, E., Kuijpers, W., et al., 2020. A comparison of trends in wastewater-based data and traditional epidemiological indicators of stimulant consumption in three locations. Addiction 115 (3), 462–472.

Gupta, S., Parker, J., Smits, S., Underwood, J., Dolwani, S., 2020. Persistent viral shedding of SARS-CoV-2 in faeces–a rapid review. Color. Dis. 22 (6), 611–620.

Hart, O.E., Halden, R.U., 2020. Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: feasibility, economy, opportunities and challenges. Sci. Total Environ. 730, 138875.

Hasan, S.W., Ibrahim, Y., Daou, M., Kannout, H., Jan, N., Lopes, A., Alsafar, H., Yousef, A.F., 2021. Detection and quantification of SARS–CoV–2 RNA in wastewater and treated effluents: surveillance of COVID–19 epidemic in the United Arab Emirates. Sci. Total Environ. 764, 142929.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. 43. CRC press.

Hellmér, M., Paxéus, N., Magnius, L., Enache, L., Arnholm, B., Johansson, A., Bergström, T., Norder, H., 2014. Detection of pathogenic viruses in sewage provided early warnings of hepatitis a virus and norovirus outbreaks. Appl. Environ. Microbiol. 80 (21), 6771–6781.

Hovi, T., Shulman, L., Van Der Avoort, H., Deshpande, J., Roivainen, M., De Gourville, E., 2012. Role of environmental poliovirus surveillance in global polio eradication and beyond. Epidemiol. Infect. 140 (1), 1–13.

Kuhn, M., et al., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26.

Kumar, M., Joshi, M., Shah, A.V., Srivastava, V., Dave, S., 2021. Wastewater surveillance-based city zonation for effective COVID-19 pandemic preparedness powered by early warning: a perspectives of temporal variations in SARS-CoV-2-RNA in Ahmedabad, India. Sci. Total Environ. 792, 148367.

La Opinión de A Coruña, N., 2021. News Dealing With COVIDBENS Project in La Opinion, Media in A Coruña Metropolitan Area. https://www.laopinioncoruna.es/buscador?text=covidbens. (Accessed 14 July 2021).

La Rosa, G., Iaconelli, M., Mancini, P., Ferraro, G.B., Veneri, C., Bonadonna, L., Lucentini, L., Suffredini, E., 2020. First detection of SARS-CoV-2 in untreated wastewaters in Italy. Sci. Total Environ. 736, 139652.

La Voz de Galicia, N., 2021. News Dealing With COVIDBENS Project in La Voz de Galicia, Most Read Newspaper in Galicia. https://www.lavozdegalicia.es/buscador/q/?text=covidbens. (Accessed 14 July 2021).

Lahrich, S., Laghrib, F., Farahi, A., Bakasse, M., Saqrane, S., El Mhammedi, M., 2021. Review on the contamination of wastewater by COVID-19 virus: impact and treatment. Sci. Total Environ. 751, 142325.

Li, X., Kulandaivelu, J., Zhang, S., Shi, J., Sivakumar, M., Mueller, J., Luby, S., Ahmed, W., Coin, L., Jiang, G., 2021. Data-driven estimation of COVID-19 community prevalence through wastewater-based epidemiology. Sci. Total Environ. 789, 147947.

Lizasoain, A., Tort, L., García, M., Gillman, L., Alberti, A., Leite, J., Miagostovich, M., Pou, S.A., Cagiao, A., Razsap, A., et al., 2018. Human enteric viruses in a wastewater treatment plant: evaluation of activated sludge combined with UV disinfection process reveals different removal performances for viruses with different features. Lett. Appl. Microbiol. 66 (3), 215–221.

Lodder, W., de Roda Husman, A.M., 2020. SARS-CoV-2 in wastewater: potential health risk, but also data source. Lancet Gastroenterol. Hepatol. 5 (6), 533–534.

Maity, A., 2017. Nonparametric functional concurrent regression models. Wiley Interdiscip. Rev. Comput. Stat. 9 (2), e1394.

Mancini, P., Bonanno Ferraro, G., Iaconelli, M., Suffredini, E., Valdazo-González, B., Della Libera, S., Divizia, M., La Rosa, G., 2019. Molecular characterization of human sapovirus in untreated sewage in Italy by amplicon-based sanger and next-generation sequencing. J. Appl. Microbiol. 126 (1), 324–331.

McMahan, C.S., Self, S., Rennert, L., Kalbaugh, C., Kriebel, D., Graves, D., Colby, C., Deaver, J.A., Popat, S., Karanfil, T., Freedman, D.L., 2021. COVID–19 wastewater epidemiology: a model to estimate infected populations. Lancet Planet. Health 5 (12), e874–e881. https://doi.org/10.1016/S2542-5196(21)00230-8.

Medema, G., Heijnen, L., Elsinga, G., Italiaander, R., Brouwer, A., 2020. Presence of SARS-Coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in the Netherlands. Environ. Sci. Technol. Lett. 7 (7), 511–516.

METEOGALICIA, 2021. Website of Galician Meteorology Agency. https://www.meteogalicia.gal/observacion/estacionshistorico/historico.action?idEst=14010. (Accessed 16 November 2021).

Miura, F., Kitajima, M., Omori, R., 2021. Duration of SARS–CoV–2 viral shedding in faeces as a parameter for wastewater-based epidemiology: re-analysis of patient data using a shedding dynamics model. Sci. Total Environ. 769, 144549.

Nemudryi, A., Nemudraia, A., Wiegand, T., Surya, K., Buyukyoruk, M., Cicha, C., Vanderwood, K.K., Wilkinson, R., Wiedenheft, B., 2020. Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater. Cell Rep. Med. 1 (6), 100098.

Peccia, J., Zulli, A., Brackney, D.E., Grubaugh, N.D., Kaplan, E.H., Casanovas-Massana, A., Ko, A.I., Malik, A.A., Wang, D., Wang, M., et al., 2020. Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. Nat. Biotechnol. 38 (10), 1164–1167.

Pillay, L., Amoah, I.D., Deepnarain, N., Pillay, K., Awolusi, O.O., Kumari, S., Bux, F., 2021. Monitoring changes in COVID-19 infection using wastewater-based epidemiology: a south african perspective. Sci. Total Environ. 786, 147273.

Pollán, M., Pérez-Gómez, B., Pastor-Barriuso, R., Oteo, J., Hernán, M.A., Pérez-Olmeda, M., Sanmartín, J.L., Fernández-García, A., Cruz, I., de Larrea, N.F., et al., 2020. Prevalence of sars–cov–2 in Spain (ene–covid): a nationwide, population–based seroepidemiological study. Lancet 396 (10250), 535–544.

Polo, D., Quintela-Baluja, M., Corbishley, A., Jones, D.L., Singer, A.C., Graham, D.W., Romalde, J.L., 2020. Making waves: wastewater-based epidemiology for COVID-19–approaches and challenges for surveillance and prediction. Water Res. 186, 116404.

Team, R.Core, 2021. R: A Language and Environment for Statistical Computing Vienna, Austria.

Randazzo, W., Cuevas-Ferrando, E., Sanjuán, R., Domingo-Calap, P., Sánchez, G., 2020a. Metropolitan wastewater analysis for COVID-19 epidemiological surveillance. Int. J. Hyg. Environ. Health 230, 113621.

Randazzo, W., Truchado, P., Cuevas-Ferrando, E., Simón, P., Allende, A., Sánchez, G., 2020b. SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. Water Res. 181, 115942.

Rey, M., 2020. María José Pereira: We are Ready, but Citizen Responsability is Essential. https://www.gciencia.com/saude/maria-jose-pereira-estamos-preparados-pero-a-responsabilidade-cidada-sera-esencial/. (Accessed 31 March 2021).

Saththasivam, J., El-Malah, S.S., Gomez, T.A., Jabbar, K.A., Remanan, R., Krishnankutty, A.K., Ogunbiyi, O., Rasool, K., Ashhab, S., Rashkeev, S., et al., 2021. COVID-19 (SARS–CoV–2) outbreak monitoring using wastewater-based epidemiology in Qatar. Sci. Total Environ. 774, 145608.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., Crowley, J., 2020. GGally: Extension to 'ggplot2', pp. 1–130 R package version 2.0.0.

SERGAS, W., 2021. COVID–19 Reported Cases in a Coruña Area Shown in the Galician Health Service Web site. https://www.datawrapper.de/_/QrkrZ. (Accessed 16 November 2021).

Shokrzadeh, S., Jozani, M.J., Bibeau, E., 2014. Wind turbine power curve modeling using advanced parametric and nonparametric methods. IIEEE Trans. Sustain. Energy 5 (4), 1262–1269.

Ministery, Spanish Health, 2021. Website of the Spanish Ministry of Health, Consumer Affairs and Social Welfare. https://www.mscbs.gob.es/profesionales/saludPublica/ccayes/alertasActual/nCov/situacionActual.htm. (Accessed 16 November 2021).

Vallejo, J.A., Rumbo-Feal, S., Conde-Pérez, K., López-Oriona, Á., Tarrío, J., Reif, R., Ladra, S., Rodiño-Janeiro, B.K., Nasser, M., Cid, Á., et al., 2020. Highly predictive regression model of active cases of covid-19 in a population by screening wastewater viral load. MedRxiv https://doi.org/10.1101/2020.07.02.20144865. https://www.medrxiv.org/content/early/2020/07/06/2020.07.02.20144865. (Accessed 16 December 2021).

Wand, M.P., Jones, M.C., 1994. Kernel Smoothing. CRC press.

Weidhaas, J., Aanderud, Z.T., Roper, D.K., VanDerslice, J., Gaddis, E.B., Ostermiller, J., Hoffman, K., Jamal, R., Heck, P., Zhang, Y., Torgersen, K., Laan, J.V., LaCross, N., 2021. Correlation of SARS-CoV-2 in wastewater with COVID-19 disease burden in sewersheds. Sci. Total Environ. 775, 145790.

Wood, S.N., 2017. Generalized Additive Models: An Introduction With R. CRC press.

Wu, F., Zhang, J., Xiao, A., Gu, X., Lee, W.L., Armas, F., Kauffman, K., Hanage, W., Matus, M., Ghaeli, N., et al., 2020a. SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. Msystems 5 (4).

Wu, Y., Guo, C., Tang, L., Hong, Z., Zhou, J., Dong, X., Yin, H., Xiao, Q., Tang, Y., Qu, X., et al., 2020b. Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. Lancet Gastroenterol. Hepatol. 5 (5), 434–435.

Wurtzer, S., Marechal, V., Mouchel, J.M., Maday, Y., Teyssou, R., Richard, E., Almayrac, J.L., Moulin, L., 2020. Evaluation of lockdown impact on SARS-CoV-2 dynamics through viral genome quantification in Paris wastewaters. MedRxiv https://doi.org/10.1101/2020.04.12.20062679. https://www.medrxiv.org/content/early/2020/05/06/2020.04.12.20062679. (Accessed 16 December 2021).

Xing, Y.H., Ni, W., Wu, Q., Li, W.J., Li, G.J., Wang, W.D., Tong, J.N., Song, X.F., Wong, G.W.K., Xing, Q.S., 2020. Prolonged viral shedding in feces of pediatric patients with coronavirus disease 2019. J. Microbiol. Immunol. Infect. 53 (3), 473–480.

Xu, Y., Li, X., Zhu, B., Liang, H., Fang, C., Gong, Y., Guo, Q., Sun, X., Zhao, D., Shen, J., et al., 2020. Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. Nat. Med. 26 (4), 502–505.

Zhang, T., Cui, X., Zhao, X., Wang, J., Zheng, J., Zheng, G., Guo, W., Cai, C., He, S., Xu, Y., 2020. Detectable SARS-CoV-2 viral RNA in feces of three children during recovery period of COVID-19 pneumonia. J. Med. Virol. 92 (7), 909–914.