Original article for **NATURE COMMUNICATIONS**

Highly predictive regression model of active cases of COVID-19 in a population by screening wastewater viral load

Juan A. Vallejo [1]*, Soraya Rumbo-Feal[1]*, Kelly Conde-Pérez[1]*, Ángel López-Oriona[2]*, Javier Tarrío[2#], Rubén Reif[3#], Susana Ladra[4], Bruno K. Rodiño-Janeiro[5], Mohammed Nasser[1], Ángeles Cid[6,3], María C Veiga[3], Antón Acevedo[7], Carlos Lamora[8], Germán Bou[1], Ricardo Cao[2,9#] and Margarita Poza[1,6#]

*Authors contributed equally

[#]Authors for corresponding

[1] Microbiology Research Group, University Hospital Complex (CHUAC) - Institute of Biomedical Research (INIBIC), University of A Coruña (UDC), A Coruña, Spain.

[2] Research Group MODES, Research Center for Information and Communication Technologies (CITIC), University of A Coruña, Spain.

[3] Advanced Scientific Research Center (CICA), University of A Coruña, Spain.

[4] Database Laboratory, Research Center for Information and Communication Technologies (CITIC), University of A Coruña, Spain.

[5] Division of Microbial Ecology, Department for Microbiology and Ecosystem Science, University of Vienna, Austria.

[6] Department of Biology, University of A Coruña, Spain.

[7] Medical Management Department, University Hospital Complex of A Coruña (CHUAC), Spain.

[8] Public wastewater treatment plant company EDAR Bens, S.A., A Coruña, Spain.

[9] Technological Institute for Industrial Mathematics (ITMATI), Universities of A Coruña, Santiago de Compostela and Vigo, Spain

Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

ABSTRACT (150 words)

The quantification of the SARS-CoV-2 load in wastewater has emerged as a useful method to monitor COVID-19 outbreaks in the community. This approach was implemented in the metropolitan area of A Coruña (NW Spain), where wastewater from the treatment plant of Bens was analyzed to track the epidemic's dynamic in a population of 369,098 inhabitants. We developed statistical regression models that allowed us to estimate the number of infected people from the viral load detected in the wastewater with a reliability close to 90%. This is the first wastewater-based epidemiological model that could potentially be adapted to track the evolution of the COVID-19 epidemic anywhere in the world, monitoring both symptomatic and asymptomatic individuals. It can help to understand with a high degree of reliability the true magnitude of the epidemic in a place at any given time and can be used as an effective early warning tool for predicting outbreaks.

RUNNING TITLE: *SARS-CoV-2 reliable surveillance on sewage*

KEYWORDS:

SARS-CoV-2, COVID-19, surveillance, wastewater-based epidemiology, wastewater treatment plant, hospital wastewater, viral load, sewage, outbreaks, imputation, generalized additive models (GAM), kernel smoothing, LOESS, local polynomial regression.

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a novel member of the *Coronaviridae* family and is the pathogen responsible for coronavirus disease 2019 (COVID-19), which has led to a worldwide pandemic. Patients may present with a wide variety of symptoms and the prognosis ranges from mild or moderate disease, to severe disease and death[1]. Importantly, a significant percentage of those infected are asymptomatic, with studies finding that 20% to over 40% of cases show no symptoms[2-4], a condition that helps the silent spread of the disease. SARS-CoV-2 is an enveloped virus with a nucleocapsid made up of single-stranded RNA bound to protein N (Nucleocapsid), surrounded by a lipid membrane that contains structural proteins M (Membrane), E

(Envelop) and S (Spike)[5-8]. The structure of protein S gives the virus its distinctive crown of spikes and is responsible for binding to the angiotensin-converting enzyme 2 (ACE2) receptor, which allows the virus to enter the host cell[9,10]. ACE2 receptors are present in a range of human cell types, with particular abundance in respiratory and gastrointestinal epithelial cells. In fact, an analysis of ACE2 receptor distribution in human tissues found the highest levels of expression in the small intestine[11]. Although respiratory symptoms are the most frequently described in patients with COVID-19, several studies have shown that the gastrointestinal tract can also be affected by SARS-CoV-2. A meta-analysis found that 15% of patients had gastrointestinal symptoms and that around 10% of patients presented gastrointestinal symptoms but not respiratory symptoms[12]. Conversely, SARS-CoV-2 RNA has been found in the faeces of people without gastrointestinal symptoms[13-18]. A systematic literature review found that more than half (53.9%) of those tested for faecal viral RNA were positive, and noted that the virus is excreted in the stool for long periods, in some cases a month or more after the individual has tested negative for their respiratory samples[15,19-24]. The fact that the virus can grow in enterocytes of human small intestine organelles[25,26] and the discovery of infectious virus in faeces highlights the potential for replication in the gastrointestinal epithelium of patients. However, despite several studies suggesting that transmission of the virus could take place through the faecal-oral axis[27,28], there is so far insufficient evidence to confirm this method of contagion[19,29]. Similarly, there has been no evidence of contagion through wastewater, which might reflect SARS-CoV-2's instability in water and its sensitivity to disinfectants[9,29-34] . Viral RNA can, nonetheless, be found in wastewater[33], which has made monitoring of viral RNA load in sewage a promising tool for the epidemiological tracking of the pandemic[35-40].

Wastewater is a dynamic system that can reflect the circulation of microorganisms in the population. Previous studies have evaluated the presence in wastewater of several viruses [41-45]. Processes to monitor SARS-CoV-2 in wastewater were first developed in the Netherlands [35], followed by the USA[46], France[39], Australia[36], Italy[47] and Spain[37,48]. In the Netherlands, no viral RNA was detected 3 weeks before the first case was reported, but genetic material started to appear over time, as the number of cases of COVID-19 increased [35]. A wastewater plant in Massachusetts detected a higher viral RNA load than expected based on the number of confirmed cases at that point, possibly reflecting viral shedding of asymptomatic cases in the community[38]. In Paris, wastewater measurements

over a 7-week period, which included the beginning of the lockdown on March 17th, found that viral RNA loads reflected the number of confirmed COVID-19 cases and decreased as the number of cases went down, roughly following an eight-day delay[39]. Another study in a region of Spain with the lowest prevalence of COVID-19, detected the virus in wastewater before the first COVID-19 cases were reported[37]. A study from Yale University measured the concentration of SARS-CoV-2 RNA in sewage sludge and found that viral RNA concentrations were highest 3 days before peak hospital admissions of COVID-19 cases, and 7 days before peak community COVID-19 cases[40]. Also, Balboa et al,[49] confirmed the presence of SARS-CoV-2 in sludge. These studies show the potential of monitoring SARS-CoV-2 levels in wastewater and sewage sludge to track and even pre-empt outbreaks in the community.

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* modelling approaches. WBE has been extensively used to predict with high accuracy the consumption patterns of numerous substances, such as the use of illicit drugs in different populations or countries. Therefore, the development of epidemiology models based on wastewater analysis has been intensive in the last years. Several examples from the literature show different approaches and strategies to tackle the uncertainty associated with WBE studies. For example, Goulding *et al*.[50] assumed three main sources of uncertainty (fluctuations in flow, uncertainty in analytical determinations, and the actual size of the population served by the wastewater treatment plant) and, using Bayesian statistics, fitted the data to linear regression hierarchical models. Other modelling approaches[51] considered Monte Carlo simulations to deal with uncertainties and with the propagation of errors associated with the parameters that are usually considered in WBE. Again, wastewater inflow variability was highlighted as a prominent source of uncertainty, as well as the stability of the substances in wastewater and their pharmacokinetics. In general, WBE studies show 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 modelling process will overcome such uncertainty, leading to accurate results.

Based on the available data, we have hypothesized that the viral load obtained from wastewater allows modelling that can predict outbreaks with high reliability. In fact, an

earlier study has confirmed the theoretical feasibility of combining WBE approaches with SARS-CoV-2/COVID-19 data [52].

In the present study we set out to monitor the wastewater viral RNA load from a treatment plant in the Northwest of Spain that services a metropolitan area with 369,098 residents, with the aim of developing new statistical regression models to estimate the number of infected people, including symptomatic and asymptomatic persons.

RESULTS

Flow variations during the COVID-19 epidemic

The main objective of the present work was to develop a useful statistical model to determine the entire SARS-CoV-2 infected population, including symptomatic and asymptomatic people, by tracking the viral load present in the wastewater. Since the flow is expected to influence the concentration of SARS-CoV-2 detected in the wastewater treatment plant (WWTP) Bens, a study of this variable was first carried out at the wastewater inlet of the treatment plant before, during and after the lockdown. Hourly mean flow box-plots (Figure 1) at WWTP Bens showed a clear daily trend, with the highest values between 12:00 and 18:00 in period A (before the lockdown). As periods B, C and D pass, a reduction in the level and the variability of the flow was observed. The estimated time for the wastewater to reach the WWTP Bens along the network is between 1.5 h and 3 h, depending on the source in the metropolitan area. Therefore, when interpreting Figure 1, a peak between 12:00 and 18:00 reflects greater human or industrial activity at least 1.5 h beforehand.

In addition, the daily two-minute flow curves performed were exceptionally noisy (Figure 2A), so these curves were smoothed (Figure 2B) to study their real shape. When considering all the daily curves and grouping them into the four time intervals (a-d), clear patterns can be seen in Figure 2C. These patterns were even clearer when plotting central curves (the deepest curves) within every group (Figure 2D).

Estimating COVID-19 active cases in the metropolitan area of A Coruña

To model the viral load, the number of COVID-19 cases needs to be reported or estimated. As explained in the methods section, there were difficulties in determining the number of people infected with SARS-CoV-2. Therefore, mathematical models were developed to estimate the COVID-19 cases, based on data recovered in Dataset S1.
Linear regression models (Figure 3) were successfully used to predict COVID-19 active cases in the A Coruña – Cee health area, where the region of this study is located (Figure S1). This was based on Intensive Care Unit (ICU) patients before April 29th (Figure 3A) and, from this date on, on active cases reported by health authorities in Galicia. Thus, a linear regression trend was fitted to predict the proportion of active cases in the health area of A Coruña – Cee based on the proportion of cumulative cases (Figure 3B). This linear regression fit was finally used to estimate the proportion of active cases in the metropolitan area of A Coruña served by the WWTP Bens, by means of the proportion of cumulative cases in the same area, which was directly obtained from the individual patient data.

Daily variation of viral load in in the metropolitan area of A Coruña

The evolution of the viral load along the day is an important feature for selecting narrower sampling intervals when the viral load was low and difficult to detect. On the other hand, it was important to detect possible differences in daily variation of viral load in the two sampling locations. An analysis of the viral load of the 24-h and 2-h samples collected at the WWTP Bens and CHUAC, as reflected in Dataset S2, was performed. The RT-qPCR results for 24-h samples are included in Dataset S3 and results for 2-h samples are included in Dataset S4. Because of the small sample size, nonparametric LOESS models were used, in order to prevent the possible overfitting of alternatives such as GAM. Figure 4 shows the viral load trends at CHUAC (Figure 4A) and at Bens (Figure 4B) depending on the hour of the day, during four different days. Figure 4A shows the hourly trend at CHUAC, with a maximum around 08:00, whereas the viral load curves at Bens (Figure 4B) attained a minimum around 05:00 and a maximum between 14:00 and 15:00.

Lockdown de-escalation in the metropolitan area of A Coruña

As expected, the mean viral load decreased with time when measured at CHUAC (Figure 5A, late April – mid May) and at WWTP Bens (Figure 5B, mid-April – early June) following an asymptotic type trend (fitted using GAM with cubic regression splines).

The viral load in WWTP Bens was consistent with the number of estimated COVID-19 cases in the metropolitan area of A Coruña, as shown in Figure 6. The number of copies of viral RNA per litre decreased from around 500,000 to less than 1,000, 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.

Wastewater epidemiological models based on viral load for COVID-19 active cases prediction

Based on the correlation analysis (Figure 7), the number of active cases strongly correlate linearly with the logarithm of daily mean viral load at Bens (R=0.923) and with the mean flow (R=-0.362). Nonetheless, there is also a strong inverse linear relationship between active cases and time (R=-0.99), probably because our measurements coincide with the lockdown period. Therefore, fitting a linear model to estimate the active cases as a function of the logarithm of the viral load is reasonable.

Different regression models were used to fit the backcasted number of COVID-19 cases based on the viral load, the flow, and the most relevant atmospheric variables (rainfall, temperature, and humidity). The best results were obtained using GAM models depending on the viral load and the mean flow (Figure 8). The effect of the viral load in the real number of COVID-19 active cases showed a logarithmic shape (Figure 8A), which suggests that the number of COVID-19 active cases can be modelled 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 number of COVID-19 active cases appears to be quadratic (Figure 8B), 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 independent variable that was significant 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. Figure S2 shows the more explicative models for a variable number of predictors using the R^2 maximization criterion, which finds that the only significant predictor was the viral load. In fact, when a multivariate

linear model depending on three predictors (viral load, daily flow, and rainfall) was performed, data showed that the only significant explanatory variable was the viral load (p-value=$1.32 \cdot 10^{-8}$). Table S1 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, 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 (Figure 9A). After removing three outliers, the fit improved slightly (R^2=0.894), as shown in Figure 9B.

The final fitted linear model became: $N = -7079 + 1059 \cdot \log V$

where N denotes the real number of active COVID-19 cases, V is the viral load (number of RNA copies per L) and log stands for the natural logarithm.

For instance, a viral load of V = 150,000 copies per liter would lead to an estimated number of N = 5,543 active cases.

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 active cases in the metropolitan area, and the explanatory variable, the natural logarithm of the viral load. Table S2 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 number of 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% 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.

Figure 10A shows a scatter plot of the number of real active cases in the metropolitan area versus the natural logarithm of the viral load, along with the quadratic LOESS fitted

curve. Figure 10B displays the actual and predicted values of real active cases. The diagonal line was added to compare with the perfect model prediction.

In summary, the quadratic LOESS curve of Figure 10A captured the relationship between the number of active cases and the viral load, avoided overfitting, and showed a good predictive ability (R^2=0.88, RMSPE=478), the best among all the considered models. Besides, the linear model (R^2=0.851, RMSPE=581.94) brings the advantage of simplicity, so both models could be successfully used to predict the number of infected people in a given region based on data about viral load obtained from wastewater.

DISCUSSION

On March 9th 2020, the city of A Coruña, in the region of Galicia, reported the circulation of SARS-CoV-2 for the first time, with data of a COVID-19 outbreak in a civic center that affected 11 people, as well as data on a few more dispersed cases. 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[53]. During these initial days of the COVID-19 epidemic in Galicia, most of the cases were in 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[54]. 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[55], reaching a peak of 1667 active cases. The cases were distributed in an area that covers the municipalities of A Coruña and Cee, as shown by data provided by SERGAS (Galician Health Service) in https://www.datawrapper.de/_/QrkrZ. This area does not coincide with the area that discharges its wastewater into the WWTP Bens, which serves the municipalities of A Coruña, Oleiros, Cambre, Culleredo, and Arteixo, but the figures give an idea of the magnitude of the epidemic at that stage. In this context, an exploratory sampling and analysis was carried out on April 15th, which showed the presence of viral genetic material in the wastewater of the WWTP Bens. From April 19th, the 24-hour composite samples

were continuously analyzed until early June for this study, although surveillance will continue at WWTP Bens until the virus disappears.

The data from wastewater obtained from April 19th onwards 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. Moreover, the seroprevalence studies carried out by the Spanish Centre for Epidemiology showed that cases in A Coruña represented about 1.8 % of the local population. 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 6,644, 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) is estimated to be 1:4, we calculated that reported cases contributing their wastewater into WWTP Bens would be around 1,661, which is close to the maximum number of cases reported in the A Coruña-Cee area (1,667 cases on April 28th). It must 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 Figure 6, where both a decrease in the viral load and in the estimated COVID-19 cases can be observed from mid-April to early June.

An initial study of flows was made to analyze their variability, which could have influenced the concentration of the virus in the wastewater. For instance, it was expected that on rainy days, the viral load detected at the entrance of WWTTP Bens would be less than for dry days, with the same number of COVID-19 cases. Therefore, a study of flow rates was first carried out at the wastewater inlet of the treatment plant before, during and after the lockdown. The daily flow analysis showed that the usage of the sewage network changed over time, especially when comparing the pre-pandemic period with the three phases during the lockdown. The mean flow curves exhibited higher levels before the lockdown (period a) and their levels decreased as time passed (b-d). In addition, the mean flow curves tended to shift to the right when moving from a to b, c, and d periods. This is probably due to the change in habits related to the restrictions to work activity, the confinement in people's homes and the increasing paralysis of the economic activity during the state of alarm in Spain.

The level of the curve at WWTP Bens at the beginning of May was much greater than that corresponding to the 11th May. This is due to the effectiveness of the confinement measures applied in Spain. The 12th May daily curve at CHUAC showed a higher viral load than the one corresponding to the 11th May at WWTP Bens, showing the viral load measured at the hospital tends to be higher than at WWTP Bens, as expected.

In the present work, nonparametric and even simple parametric regression models have been shown to be useful tools to construct prediction models for 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 since, to our knowledge, WBE studies available are still 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[35,37,46-48]**.** The only precedent[52] combines computational analysis and modelling with a theoretical approach in order to identify useful variables and confirm the feasibility and cost-effectiveness of WBE as a prediction tool. Other examples of 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[56] was optimized incorporating environmental data. 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 associated scarcity of epidemiologic information. Considering this, our statistical model has minimized the uncertainty implementing a complete set of hydraulic information from the sewer network of the city of A Coruña, available thanks of a joint effort from different local authorities. This *ad hoc* model can be adapted to other scenarios as long as similar hydraulic information from the area where it will be used can be obtained.

Other 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 rainfall fluctuated little: its median was 0, its mean was 2.88 L/m^2 and its standard deviation was 6.59 L/m^2.

Therefore, 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 percentage variability explained by the model was reasonably high (85.1%) but three outliers were detected in the sample. Excluding these three observations, no further outliers were detected, and the percentage of variability explained by the model increased up to 89.4%. Alternative, more flexible models, such as GAM and LOESS, were also fitted. They produced slightly better results in terms of R^2 and RMSPE. However, the final fit for the quadratic LOESS model was similar to the linear model fit.

As a conclusion, a simple linear model that relates the logarithm of the viral load to the number of COVID-19 active cases gave a good fitting, explaining nearly 90% of the variability of the response. Of course, this is a simple model. Although a quadratic LOESS model improved the prediction error to a certain degree, the final fit was similar. These two models (linear or quadratic LOESS) can be used as useful new epidemiological tools, and complementary to seroprevalence or RT-PCR tests carried out by healthcare institutions.

Our models, as described, are only applicable to the metropolitan area of A Coruña, the region for which these models have been developed. 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.

These are the first highly reliable wastewater-based epidemiological statistical models that could be adapted for use anywhere in the world. The models allow the actual number of infected patients to be determined with around 90% reliability, since it takes into account the entire population, whether symptomatic or asymptomatic. These statistical

models can estimate the true magnitude of the epidemic at a specific location and their cost-effectiveness and sampling speed can help alert health authorities of possible new outbreaks, which will help to protect the local population.

METHODS

Sample Collection

The WWTP of Bens (43° 22' 8.4" N 8° 27' 10.7" W, A Coruña, Spain) serves 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 (Figure S1). 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 of 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 (Dataset S2). In addition, samples were collected at 2-h intervals for 24 h on specific days at the WWTP Bens and at CHUAC (Dataset S2). The 24-h composite samples were collected by automatic samplers taking 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.

Sample processing

Samples of 100 mL were processed immediately after collection at 4 ºC. Firstly, 100 mL samples were centrifuged for 30 min at 4000 *g* and then filtered through 0.22 µm membranes. Samples were then concentrated and dialyzed using Amicon Ultra Filters 30 KDa (Merckmillipore) in 500 µL of a buffer containing 50 mM Tris-HCL, 100 mM NaCl y 8 mM $MgSO_4$. Samples were preserved in RNAlater reagent (Sigma-Aldrich) at -80 °C.

RNA extraction and qRT-PCR assays

RNA was extracted from the concentrates 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 an RNase-free buffer. The quality and quantity of the RNA was checked using a Nanodrop Instrument and an Agilent Bionanalyzer. Samples were kept at -80ºC 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. Reaction mix (15 µL) consisted of: 5 µL 4x RT-PCR Mix containing DNA polymerase, dNTPs, PCR buffer and a VIC internal control; 1 µL of Reaction Mix 1 containing primers and FAM probe for N gene; and 0.2 µL of reverse transcriptase enzyme. The cycling parameters were 50 °C for 20 minutes 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) (Figure S3). To build the straight pattern, the decimal logarithm of SARS-CoV-2 RNA copies/µL ranging from 5 to 500 were plotted against Ct (threshold cycle) values. Calibration was done amplifying the N gene.

Data collection

The present study also includes data gathered from different sources. More specifically, two-minute flow measurements at WWTP Bens for the period January 1st – May 14th were provided by the company EDAR Bens S.A. (Dataset S5). Daily observations at the meteorological station of Coruña-Bens for the period March 1st – May 31st, 2020, including rainfall, temperature, and humidity, were obtained from the Galician Meteorology Agency, MeteoGalicia (Dataset S6). Finally, several series of cumulative and active number of COVID-19 cases in the metropolitan area of A Coruña and from the health area A Coruña – Cee for the period March 1st – May 31st were obtained from the Galician Health Service (SERGAS), the General Directorate of Public Health

(Automous Government of Galicia) and the University Hospital of A Coruña (CHUAC) (Dataset S1).

Exploratory flow study

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 in Bens during the lockdown period has been performed. Two-minute data of volume of pumped water in the entrance of the WWTP in the period January 1st – May 4th 2020 were collected. This period was split into four time intervals corresponding to: (a) regular conditions (January 1st – March 13th), (b) initial alarm state period (March 14th –26th), (c) strict lockdown period (March 27th – April 9th), and (d) lockdown de-escalation (April 10th – May 3rd). All these data are recovered in Dataset S5.

The mean hour flow at the WWTP has been computed and a multivariate exploratory data analysis was performed. Box-plots for flow at every hour have been computed using the data in every time period a-d. They were plotted as a function of the hour of the day.
Exploratory functional data analyses have been also performed. Since the raw flow curves are exceedingly noisy, local polynomial methods[57] have been used to obtain smooth curves. A direct plug-in method[58] has been used to choose the smoothing parameter. The collection of smoothed daily curves has been analyzed and the deepest curve[59] among every time period a-d has been computed.

Backcasting of COVID-19 active cases

Preliminary statistical methods have been devised to backcast the real number of COVID-19 active cases based on reported official figures.

Follow-up times (available only until May 7th) for anonymized individual official COVID-19 cases in Galicia (NW Spain where A Coruña belongs, Figure S1) have been used to count the number of cases by municipality based on patient zip codes. Since the epidemiological discharge time 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 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 predict 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 estimated official number of COVID-19 cases into the real number, the ratio mean of real cases / mean of official cases is estimated using the official figures of cumulative cases. The results of the seroprevalence study carried out by the National Center of Epidemiology in Spain[60] were used to estimate the numbers of actual active cases in Galicia: 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 S1.

Nonparametric setting of viral load overtime

Generalized Additive Models (GAM) using a basis of cubic regression splines[61] have been used to fit the viral load 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 have been removed from the data, corresponding to unexpected and intensive pipeline cleaning episodes (8-hour 70 ºC water cleaning during the Thursday-Friday nights) carried out in April 23rd-24th, April 30th - May 1st and May 7th-8th.

Viral load models

GAM and LOESS[62] nonparametric regression models have been used to explain the viral load (number of RNA copies) as a function of time, to describe the trend of response variable during a day and through the days.

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. The latter ones allowed 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 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 analysis[63]. Namely, the mgcv library[64] was applied to fit GAM models and ggplot2 and GGally[65,66] to perform correlation analysis, obtain graphical output and fit LOESS models, respectively. The caret R package was used to fit and evaluate regression models.

Although some RT-qPCR replicates could not be measured due the limitation of the detection technique (some errors occurred when the number of copies/L was under 10,000), 74% of the assays led to three or more measured replications, which gives a good statistical approach. However, conditional mean imputation[67] was used for unmeasured replications. Thus, unmeasured replications in an assay were replaced by the sample mean of observed measurements in that assay. In the only assay with all (six) unmeasured replications, the number of RNA copies was imputed using the minimum of measured viral load along the whole set of assays.

ACKNOWLEDGEMENTS

This work was funded by Project INV04020 from the University of A Coruña Foundation (FUAC-UDC) and EDAR Bens S.A., A Coruña, Spain, awarded to MP, by Projects

PI15/00860 to GB and PI17/01482 to MP, all within in the National Plan for Scientific Research, Development and Technological Innovation 2013-2016 and funded by the ISCIII - General Subdirection of Assessment and Promotion of the Research-European Regional Development Fund (FEDER) “A way of making Europe”. The study was also funded by project IN607A 2016/22 (GAIN- Xunta de Galicia, Spain) awarded to GB. This work was also supported by Planes Nacionales de I+D+i 2008-2011/2013-2016 and Instituto de Salud Carlos III, Subdirección General de Redes y Centros de Investigación Cooperativa, Ministerio de Economía y Competitividad, Spanish Network for Research in Infectious Diseases (REIPI RD16/0016/006) co-financed by European Regional Development Fund "A way to achieve Europe" and operative program Intelligent Growth 2014-2020. The publication was also supported by the European Virus Archive Global (EVA-GLOBAL) project that has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 871029. SR-F was financially supported by REIPI RD16/0016/006, KC-P by IN607A 2016/22 and the Spanish Association against Cancer (AECC) and JAV by IN607A 2016/22. This research has been also supported by MINECO grant MTM2017-82724-R awarded to RC, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015, awarded to RC, ED431C 2017/58, awarded to SL, and Centro Singular de Investigación de Galicia ED431G 2019/01, awarded to RC and SL, all of them through the ERDF, and ED431C 2017/66, awarded to MCV.

The authors wish to acknowledge the NORMAN ‘Collaboration in the time of Covid19’ European network.

Authors would like to give special thanks to the Board of Directors from EDAR Bens. Also, we would like to thank Fernanda Rodríguez from the Research Support Services (SAI) at the University of A Coruña, Laura Larriba, from SERGAS, who helped in samples and data collection in CHUAC, Francisco Pérez, Javier Fernández Romero and Cristina Rodríguez Freire from Cadaqua for their help in sample collection at WWTP Bens, Andrés Paz-Ares and Xurxo Hervada, from SERGAS, who provided the anonymized patient database and Amalia Jácome, Ana López-Cheda, Rebeca Peláez and Wende Safari, from CITIC at UDC, who processed that database to produce the epidemiological series at the municipality level, and Fiona Veira McTiernan for editing.

REFERENCES

1 Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C. & Di Napoli, R. in *StatPearls* (StatPearls Publishing Copyright © 2020, StatPearls Publishing LLC., 2020).

2 Yang, R., Gui, X. & Xiong, Y. Comparison of Clinical Characteristics of Patients with Asymptomatic vs Symptomatic Coronavirus Disease 2019 in Wuhan, China. *JAMA network open* **3**, e2010182, doi:10.1001/jamanetworkopen.2020.10182 (2020).

3 Day, M. Covid-19: four fifths of cases are asymptomatic, China figures indicate. *BMJ (Clinical research ed.)* **369**, m1375, doi:10.1136/bmj.m1375 (2020).

4 Bi, Q. *et al.* Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. *The Lancet. Infectious diseases*, doi:10.1016/s1473-3099(20)30287-5 (2020).

5 Astuti, I. & Ysrafil. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. *Diabetes & metabolic syndrome* **14**, 407-412, doi:10.1016/j.dsx.2020.04.020 (2020).

6 Zeng, W. *et al.* Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochemical and biophysical research communications* **527**, 618-623, doi:10.1016/j.bbrc.2020.04.136 (2020).

7 Srinivasan, S. *et al.* Structural Genomics of SARS-CoV-2 Indicates Evolutionary Conserved Functional Regions of Viral Proteins. *Viruses* **12**, doi:10.3390/v12040360 (2020).

8 Ou, X. *et al.* Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nature communications* **11**, 1620, doi:10.1038/s41467-020-15562-9 (2020).

9 Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *Journal of virology* **94**, doi:10.1128/jvi.00127-20 (2020).

10 Zemlin, A. E. & Wiese, O. J. Coronavirus disease 2019 (COVID-19) and the renin-angiotensin system: A closer look at angiotensin-converting enzyme 2 (ACE2). *Annals of Clinical Biochemistry* 4563220928361, doi:10.1177/0004563220928361 (2020).

11 Li, M. Y., Li, L., Zhang, Y. & Wang, X. S. Expression of the SARS-CoV-2 cell receptor gene ACE2 in a wide variety of human tissues. *Infectious diseases of poverty* **9**, 45, doi:10.1186/s40249-020-00662-x (2020).

12 Mao, R. *et al.* Manifestations and prognosis of gastrointestinal and liver involvement in patients with COVID-19: a systematic review and meta-analysis. *The lancet. Gastroenterology & hepatology* **5**, 667-678, doi:10.1016/s2468-1253(20)30126-6 (2020).

13 Pan, Y., Zhang, D., Yang, P., Poon, L. L. M. & Wang, Q. Viral load of SARS-CoV-2 in clinical samples. *The Lancet. Infectious diseases* **20**, 411-412, doi:10.1016/s1473-3099(20)30113-4 (2020).

14 Wang, W. *et al.* Detection of SARS-CoV-2 in Different Types of Clinical Specimens. *Jama* **323**, 1843-1844, doi:10.1001/jama.2020.3786 (2020).

15 Chen, Y. *et al.* The presence of SARS-CoV-2 RNA in the feces of COVID-19 patients. *Journal of Medical Virology* **92**, 833-840, doi:10.1002/jmv.25825 (2020).

16 Lescure, F. X. *et al.* Clinical and virological data of the first cases of COVID-19 in Europe: a case series. *The Lancet. Infectious diseases* **20**, 697-706, doi:10.1016/s1473-3099(20)30200-0 (2020).

17 Zhang, J., Wang, S. & Xue, Y. Fecal specimen diagnosis 2019 novel coronavirus-infected pneumonia. *Journal of medical virology* **92**, 680-682, doi:10.1002/jmv.25742 (2020).

18 Zhang, W. *et al.* Molecular and serological investigation of 2019-nCoV infected patients: implication of multiple shedding routes. *Emerging Microbes & Infections* **9**, 386-389, doi:10.1080/22221751.2020.1729071 (2020).

19 Gupta, S., Parker, J., Smits, S., Underwood, J. & Dolwani, S. Persistent viral shedding of SARS-CoV-2 in faeces - a rapid review. *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland* **22**, 611-620, doi:10.1111/codi.15138 (2020).

20 Xing, Y. H. *et al.* Prolonged viral shedding in feces of pediatric patients with coronavirus disease 2019. *Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi* **53**, 473-480, doi:10.1016/j.jmii.2020.03.021 (2020).

21 Zhang, T. *et al.* Detectable SARS-CoV-2 viral RNA in feces of three children during recovery period of COVID-19 pneumonia. *Journal of Medical Virology* **92**, 909-914, doi:10.1002/jmv.25795 (2020).

22 Wölfel, R. *et al.* Virological assessment of hospitalized patients with COVID-2019. *Nature* **581**, 465-469, doi:10.1038/s41586-020-2196-x (2020).

23 Wu, Y. *et al.* Prolonged presence of SARS-CoV-2 viral RNA in faecal samples. *The lancet. Gastroenterology & hepatology* **5**, 434-435, doi:10.1016/s2468-1253(20)30083-2 (2020).

24 Xu, Y. *et al.* Characteristics of pediatric SARS-CoV-2 infection and potential evidence for persistent fecal viral shedding. *Nature Medicine* **26**, 502-505, doi:10.1038/s41591-020-0817-4 (2020).

25 Lamers, M. M. *et al.* SARS-CoV-2 productively infects human gut enterocytes. *Science*, doi:10.1126/science.abc1669 (2020).

26 Zhou, J., Li, C., Liu, X. & Chiu, M. C. Infection of bat and human intestinal organoids by SARS-CoV-2. *Nature Medicine* doi:10.1038/s41591-020-0912-6 (2020).

27 Amirian, E. S. Potential fecal transmission of SARS-CoV-2: Current evidence and implications for public health. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* **95**, 363-370, doi:10.1016/j.ijid.2020.04.057 (2020).

28 Ding, S. & Liang, T. J. Is SARS-CoV-2 Also an Enteric Pathogen With Potential Fecal-Oral Transmission? A COVID-19 Virological and Clinical Review. *Gastroenterology*, doi:10.1053/j.gastro.2020.04.052 (2020).

29 Cha, M. H., Regueiro, M. & Sandhu, D. S. Gastrointestinal and hepatic manifestations of COVID-19: A comprehensive review. *World journal of gastroenterology* **26**, 2323-2332, doi:10.3748/wjg.v26.i19.2323 (2020).

30 Yeo, C., Kaushal, S. & Yeo, D. Enteric involvement of coronaviruses: is faecal-oral transmission of SARS-CoV-2 possible? *The lancet. Gastroenterology & hepatology* **5**, 335-337, doi:10.1016/s2468-1253(20)30048-0 (2020).

31 Gu, J., Han, B. & Wang, J. COVID-19: Gastrointestinal Manifestations and Potential Fecal-Oral Transmission. *Gastroenterology* **158**, 1518-1519, doi:10.1053/j.gastro.2020.02.054 (2020).

32 Hindson, J. COVID-19: faecal-oral transmission? *Nature reviews. Gastroenterology & hepatology* **17**, 259, doi:10.1038/s41575-020-0295-7 (2020).

33 Lodder, W. & de Roda Husman, A. M. SARS-CoV-2 in wastewater: potential health risk, but also data source. *The lancet. Gastroenterology & hepatology* **5**, 533-534, doi:10.1016/s2468-1253(20)30087-x (2020).

34 Chin, A. *et al.* Stability of SARS-CoV-2 in different environmental conditions. *The Lancet Microbe* **1**, e10, doi:https://doi.org/10.1016/S2666-5247(20)30003-3 (2020).

35 Medema, G., Heijnen, L., Elsinga, G., Italiaander, R. & Brouwer, A. Presence of SARS-Coronavirus-2 in sewage. Preprint at https://www.medrxiv.org/content/10.1101/2020.03.29.20045880v1 (2020).

36 Ahmed, W. *et al.* 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. *The Science of the total environment* **728**, 138764, doi:10.1016/j.scitotenv.2020.138764 (2020).

37 Randazzo, W. *et al.* SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. *Water research* **181**, 115942, doi:10.1016/j.watres.2020.115942 (2020).

38 Wu, F. *et al.* SARS-CoV-2 titers in wastewater are higher than expected from clinically confirmed cases. Preprint at https://www.medrxiv.org/content/10.1101/2020.04.05.20051540v1 (2020).

39 Wurtzer, S. *et al.* Evaluation of lockdown impact on SARS-CoV-2 dynamics through viral genome quantification in Paris wastewaters. Preprint at https://www.medrxiv.org/content/10.1101/2020.04.12.20062679v2 (2020).

40 Peccia, J. *et al.* SARS-CoV-2 RNA concentrations in primary municipal sewage sludge as a leading indicator of COVID-19 outbreak dynamics. Preprint at https://www.medrxiv.org/content/10.1101/2020.05.19.20105999v2 (2020).

41 Lizasoain, A. *et al.* 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. *Letters in applied microbiology* **66**, 215-221, doi:10.1111/lam.12839 (2018).

42 Hellmér, M. *et al.* Detection of pathogenic viruses in sewage provided early warnings of hepatitis A virus and norovirus outbreaks. *Applied and environmental microbiology* **80**, 6771-6781, doi:10.1128/aem.01981-14 (2014).

43 Ehlers, M. M., Grabow, W. O. & Pavlov, D. N. Detection of enteroviruses in untreated and treated drinking water supplies in South Africa. *Water research* **39**, 2253-2258, doi:10.1016/j.watres.2005.04.014 (2005).

44 Hovi, T. *et al.* Role of environmental poliovirus surveillance in global polio eradication and beyond. *Epidemiology and infection* **140**, 1-13, doi:10.1017/s095026881000316x (2012).

45 Mancini, P., Bonanno Ferraro, G., Iaconelli, M. & Suffredini, E. Molecular characterization of human Sapovirus in untreated sewage in Italy by amplicon-based Sanger and next-generation sequencing. *Journal of Applied Microbiology* **126**, 324-331, doi:10.1111/jam.14129 (2019).

46 Nemudryi, A. *et al.* Temporal detection and phylogenetic assessment of SARS-CoV-2 in municipal wastewater. Preprint at https://www.medrxiv.org/content/10.1101/2020.04.15.20066746v1 (2020).

47 La Rosa, G. *et al.* First detection of SARS-CoV-2 in untreated wastewaters in Italy. *The Science of the total environment* **736**, 139652, doi:10.1016/j.scitotenv.2020.139652 (2020).

48 Randazzo, W., Cuevas-Ferrando, E., Sanjuan, R., Domingo-Calap, P. & Sanchez, G. Metropolitan Wastewater Analysis for COVID-19 Epidemiological Surveillance. Preprint at https://www.medrxiv.org/content/10.1101/2020.04.23.20076679v2 (2020).

49 Balboa, S. *et al.* The fate of SARS-CoV-2 in wastewater treatment plants points out the sludge line as a suitable spot for incidence monitoring. Preprint at https://www.medrxiv.org/content/10.1101/2020.05.25.20112706v1 (2020).

50 Goulding, N. & Hickman, M. A comparison of trends in wastewater-based data and traditional epidemiological indicators of stimulant consumption in three locations. *Addiction* **115**, 462-472, doi:10.1111/add.14852 (2020).

51 Croft, T. L., Huffines, R. A., Pathak, M. & Subedi, B. 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. *Journal of hazardous materials* **384**, 121306, doi:10.1016/j.jhazmat.2019.121306 (2020).

52 Hart, O. E. & Halden, R. U. Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: Feasibility, economy, opportunities and challenges. *The Science of the total environment* **730**, 138875, doi:10.1016/j.scitotenv.2020.138875 (2020).

53 *O foco do centro cívico da Coruña suma xa 11 casos de coronavirus*, https://www.gciencia.com/saude/centro-civico-coruna-coronavirus/ (2020).

54 *Galicia suma 90 casos de SARS-CoV-2*, https://www.gciencia.com/extra/galicia-incidencia-casos-coronavirus/ (2020).

55 Rey, M. *María José Pereira: "Estamos preparados, pero será esencial a responsabilidade cidadá"*, https://www.gciencia.com/saude/maria-jose-pereira-estamos-preparados-pero-a-responsabilidade-cidada-sera-esencial/ (2020).

56 Brouwer, A. F. *et al.* Epidemiology of the silent polio outbreak in Rahat, Israel, based on modeling of environmental surveillance data. *Proceedings of the National Academy of Sciences of the United States of America* **115**, E10625-e10633, doi:10.1073/pnas.1808798115 (2018).

57 Fan, J. & Gijbels, I. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. (Taylor & Francis, 1996).

58 Ruppert, D., Sheather, S. J., Wand, M. P. & Management, A. G. S. o. *An Effective Bandwidth Selector for Local Least Squares Regression*. (Australian Graduate School of Management, University of New South Wales, 1993).

59 Fraiman, R. & Muniz, G. *Trimmed Means for Functional Data*. (Universidad de San Andrés, 2001).

60 *National Study of SARS-CoV-2 sero-Epidemiology in Spain (ENE-Covid19)*, https://portalcne.isciii.es/enecovid19/ (consulted on June 10th)

61 Hastie, T., Tibshirani, R. & LLC., C. P. *Generalized Additive Models*. (Chapman and Hall, 1990).

62 Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association* **74**, 829-836, doi:10.1080/01621459.1979.10481038 (1979).

63 Team, R. C. *R: A language and environment for statistical computing*, https://www.R-project.org/

64 Wood, S. *Generalized Additive Models: An Introduction with R*. (Taylor & Francis, 2006).

65 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing, 2016).

66 GGally: Extension to 'ggplot2' v. R package version 1.5.0 (2020).

67 Enders, C. K. *Applied missing data analysis*. (Guilford Press, 2010).

AUTHOR CONTRIBUTIONS

M.P., R.C., C.L., J.A.V., J.T. and R.R. conceived and designed the study. J.A.V., S.R-F., M.N. and K. C. performed wastewater analysis, A.L-O. and J.T. performed statistical models and data analysis, S.L. managed and analyzed data, B.K.R-J. assisted in the study design and analysis, A.A. assessed in data collection, A.C. supervised the wastewater analysis, M.C.V. assessed in wastewater sampling, G.B. and M.P. supervised the microbiology team, M.P, J.A.V., R.C., R.R. and J.T. wrote the manuscript. M.P. and R.C. supervised the team and coordinated all tasks.

COMPETING INTERESTS

The authors have no conflicts of interest to declare.

MATERIALS & CORRESPONDENCE

Margarita Poza Domínguez
Email: margarita.poza.dominguez@sergas.es
Address: Microbiology Service, 3ª planta, Edificio Sur, Hospital Universitario, As Xubias 15006, A Coruña, Spain
Ricardo Cao Abad
Email: ricardo.cao@udc.es
Address: Facultade de Informática, Campus de Elviña s/n, 15071 A Coruña, Spain.
Javier Tarrío Saavedra
javier.tarrio@udc.es
Address: Escola Politécnica Superior, Mendizábal, s/n, Campus de Esteiro, 15403 Ferrol (A Coruña), Spain.
Rubén Reif López
ruben.reif@udc.es
Address: Advanced Scientific Research Center (CICA), University of A Coruña, As Carballeiras, s/n, Campus de Elviña 15071 A Coruña, Spain.

DAVAILABILITY

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.

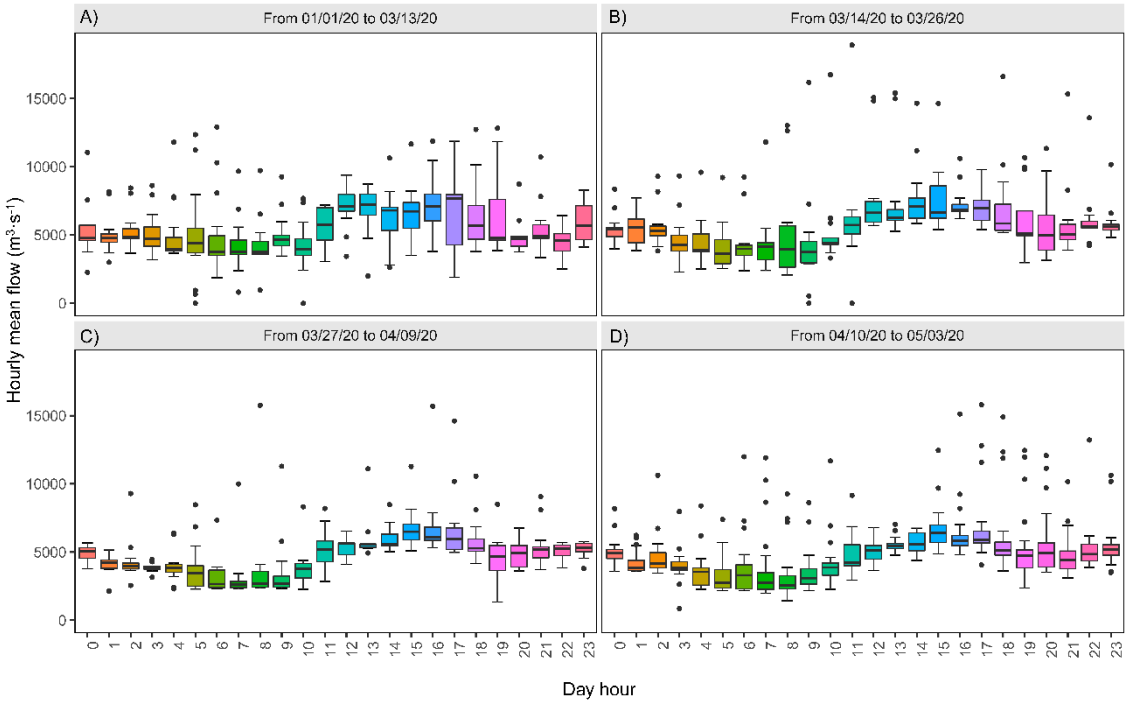

Figure 1. Hourly mean flow box-plots for the 24 h in the day corresponding to the four time intervals, A-D, considered during the lockdown period in the metropolitan area of A Coruña. A) Regular conditions (January 1st – March 13th). B) Initial alarm state period (March 14th –26th). C) Strict lockdown period (March 27th – April 9th). D) Lockdown de-escalation (April 10th – May 3rd).

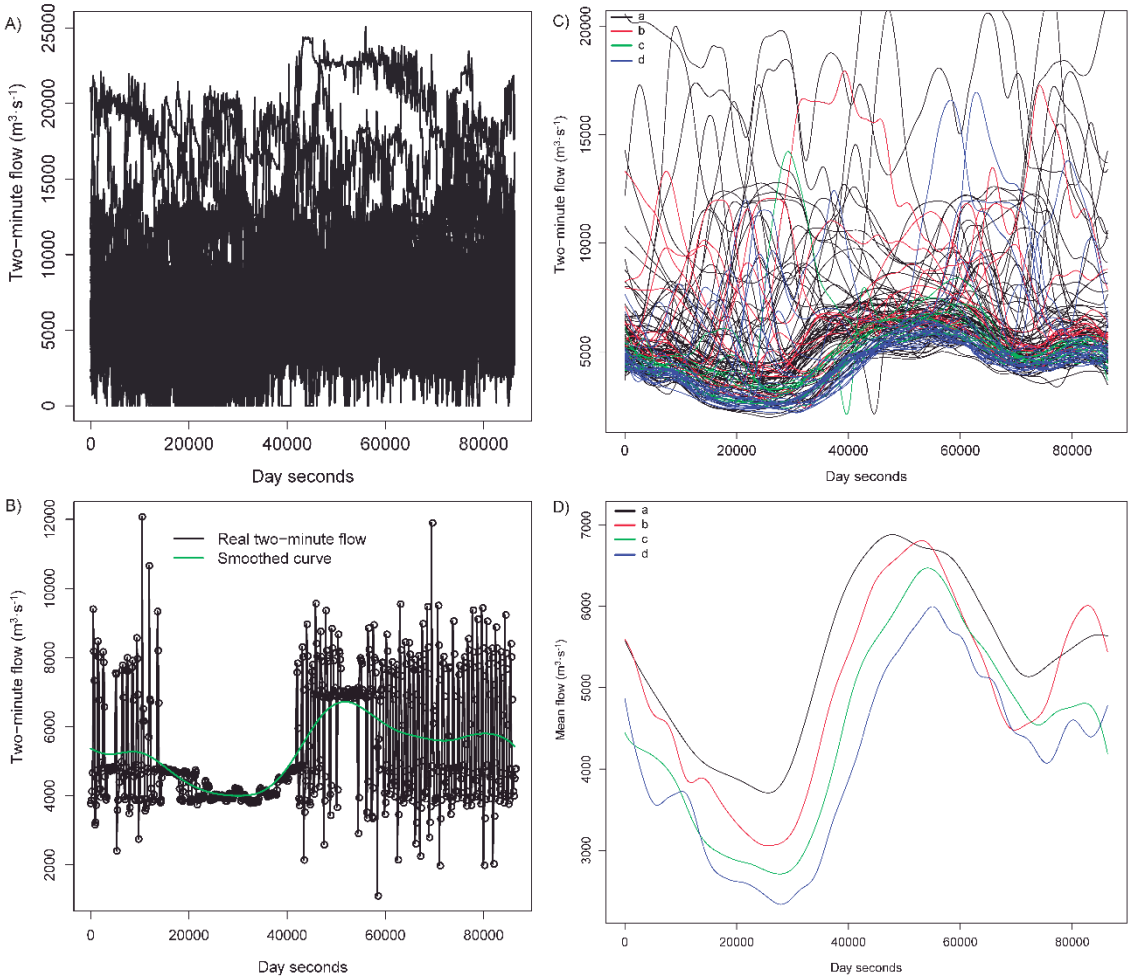

Figure 2. Exploratory flow study during the COVID-19 epidemic in the metropolitan area of A Coruña. A) Set of daily two-minute flow curves for the period January 1st – May 3rd. B) One two-minute flow curve (black) and its smoothed version (green) using a local linear fit. C) Smooth daily two-minute flow curves for the four time intervals considered, a (black), b (red), c (green) and d (blue) during the lockdown period. D) Deepest daily flow smoothed curves for every time interval, a (black), b (red), c (green) and d (blue), during the lockdown period.

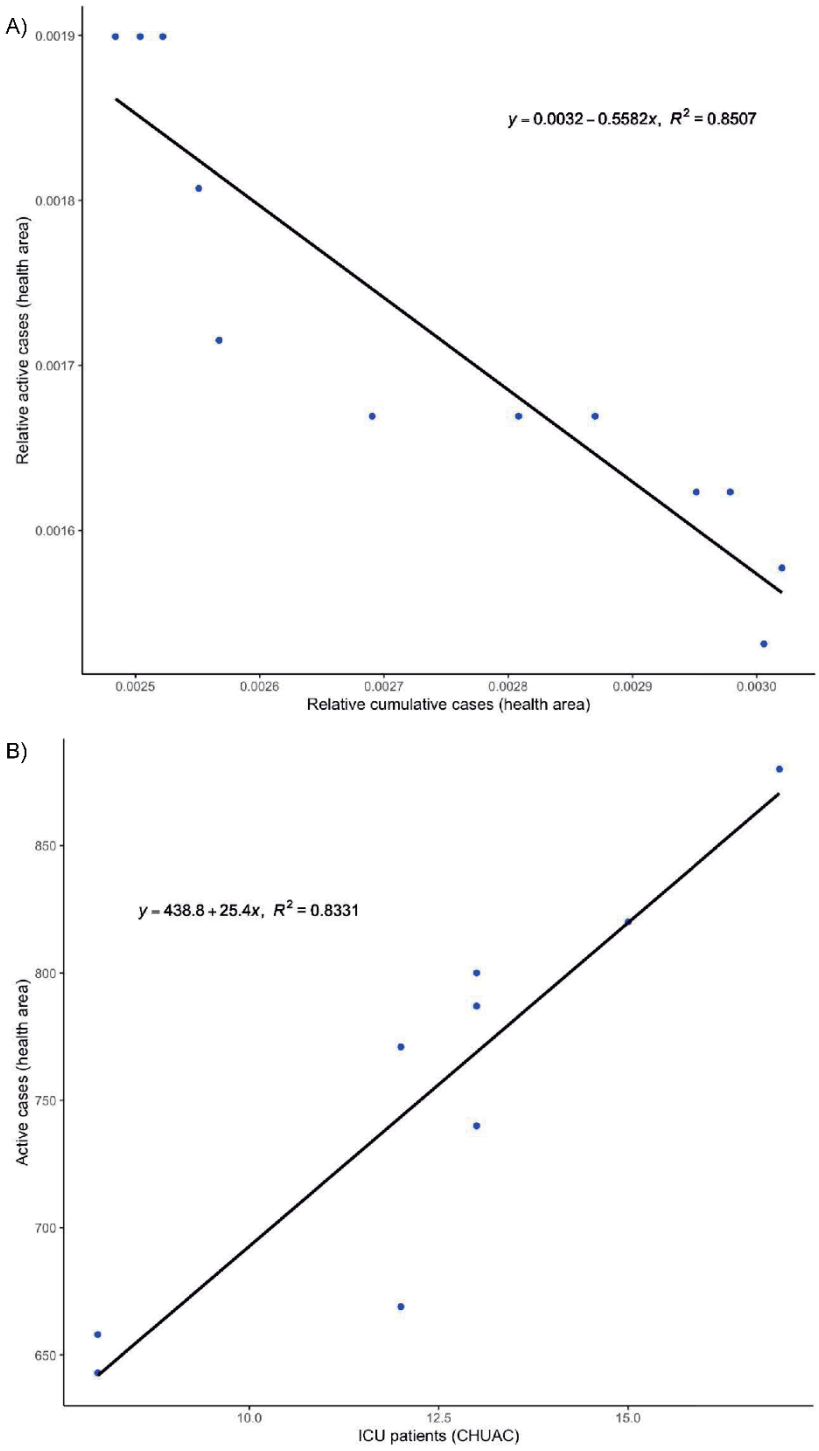

Figure 3. Estimation of the COVID- active cases using simple linear regression models. A) ICU patients versus active cases in A Coruña – Cee health area, and its linear fit for the period April 29th on, when active cases were reported. B) Relative cumulative cases versus relative active cases in the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

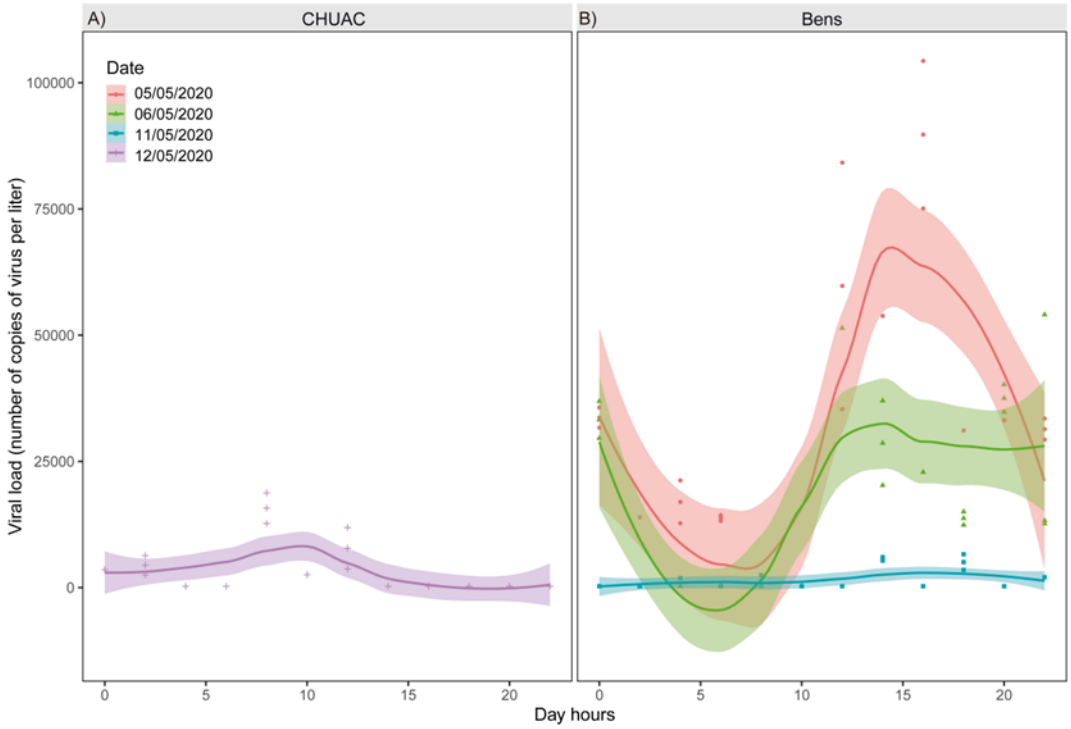

Figure 4. Viral load trend during the day. A) Viral load with respect to the hour of the day in CHUAC during the 05/12/2020 and nonparametric LOESS fitted model (span parameter equal to 0.75) with 95% confidence interval. B) Viral load with respect to the hour of the day in Bens for three different days of May, and nonparametric models (span parameter equal to 0.75) with 95% confidence interval fitted to the data of each day separately.

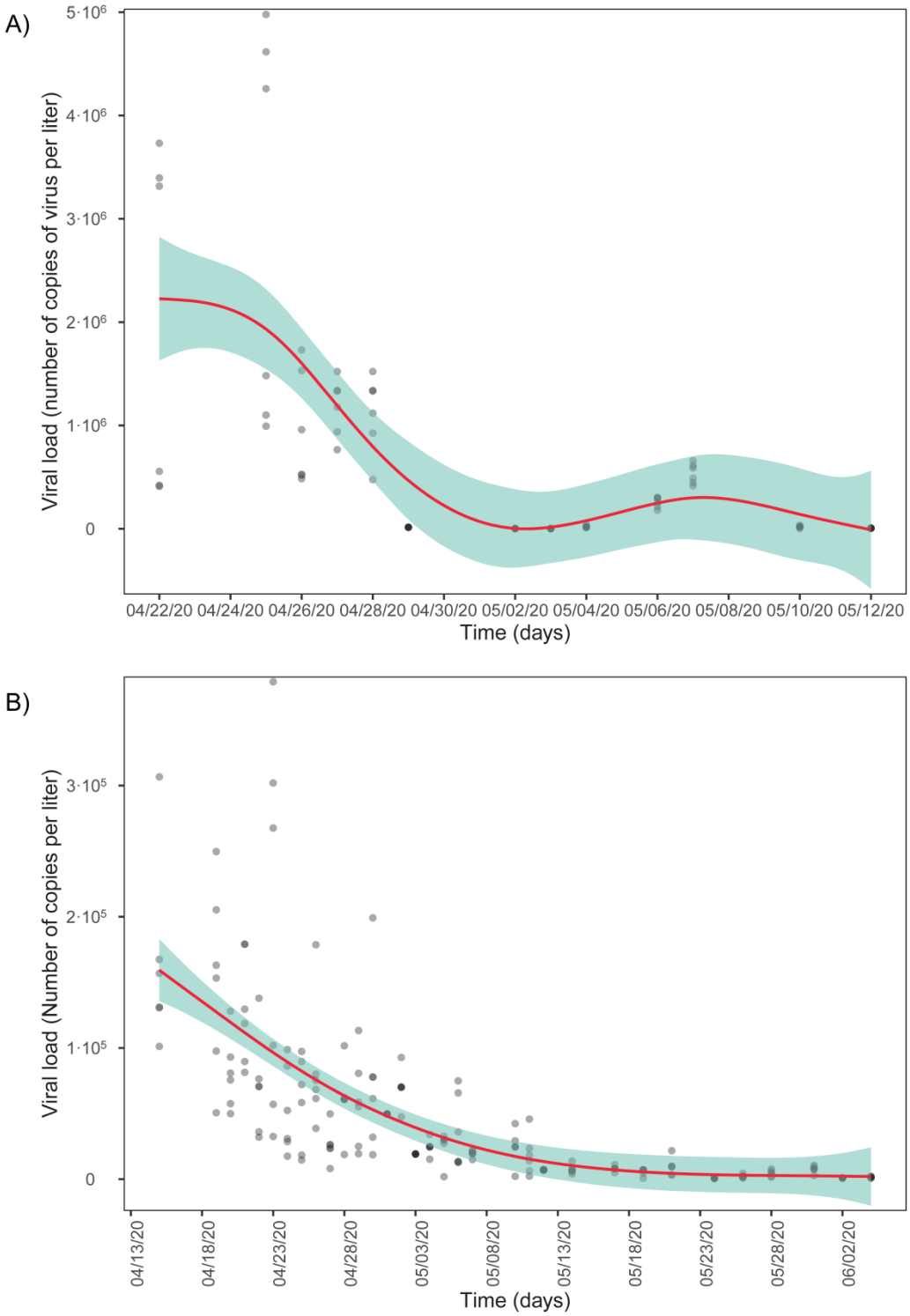

Figure 5. A) Nonparametric estimator (GAM) of the date effect in the viral load at A) CHUAC after removing the data corresponding to intensive cleaning episodes (April 23rd – 24th, April 30th – May 1st, May 8th – 9th) and the outliers in May 5th and B) at WWTP Bens after removing the data corresponding to intensive cleaning episodes (April 23rd – 24th, April 30th – May 1st, May 8th – 9th).

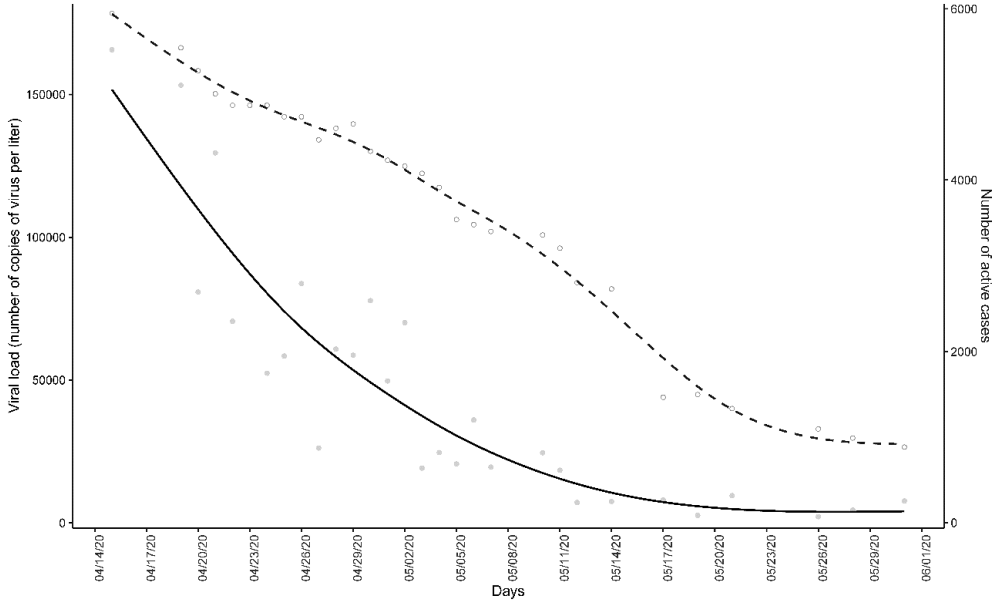

Figure 6. Viral load detected in the influent of the WWTP Bens (solid line) and number of estimated people with COVID-19 (dashed line) in the metropolitan area of A Coruña. Time course quantitative detection of SARS-CoV-2 in A Coruña WWTP Bens wastewater correlates with COVID-19 cases.

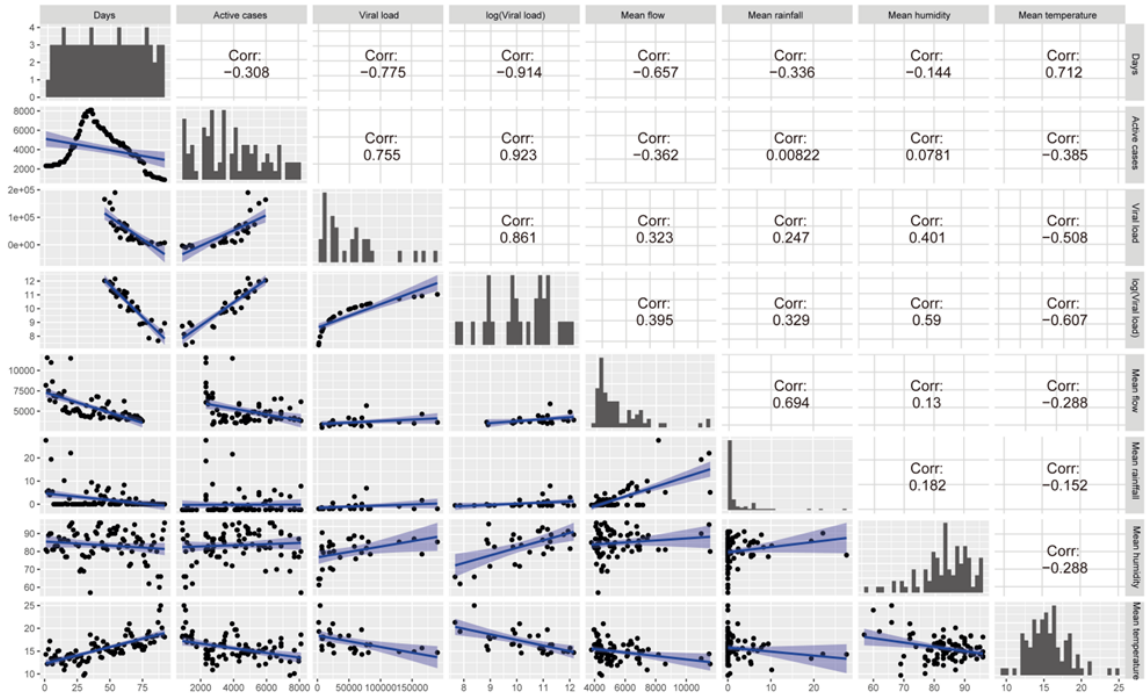

Figure 7. Scatterplot matrix with fitted linear models and linear correlation coefficients of each pair of variables: time (measured in days from the beginning of reported COVID-19 cases), estimated active cases, daily mean viral load measured in Bens, mean flow of sewage water in Bens, rainfall, daily mean humidity and daily mean temperature.

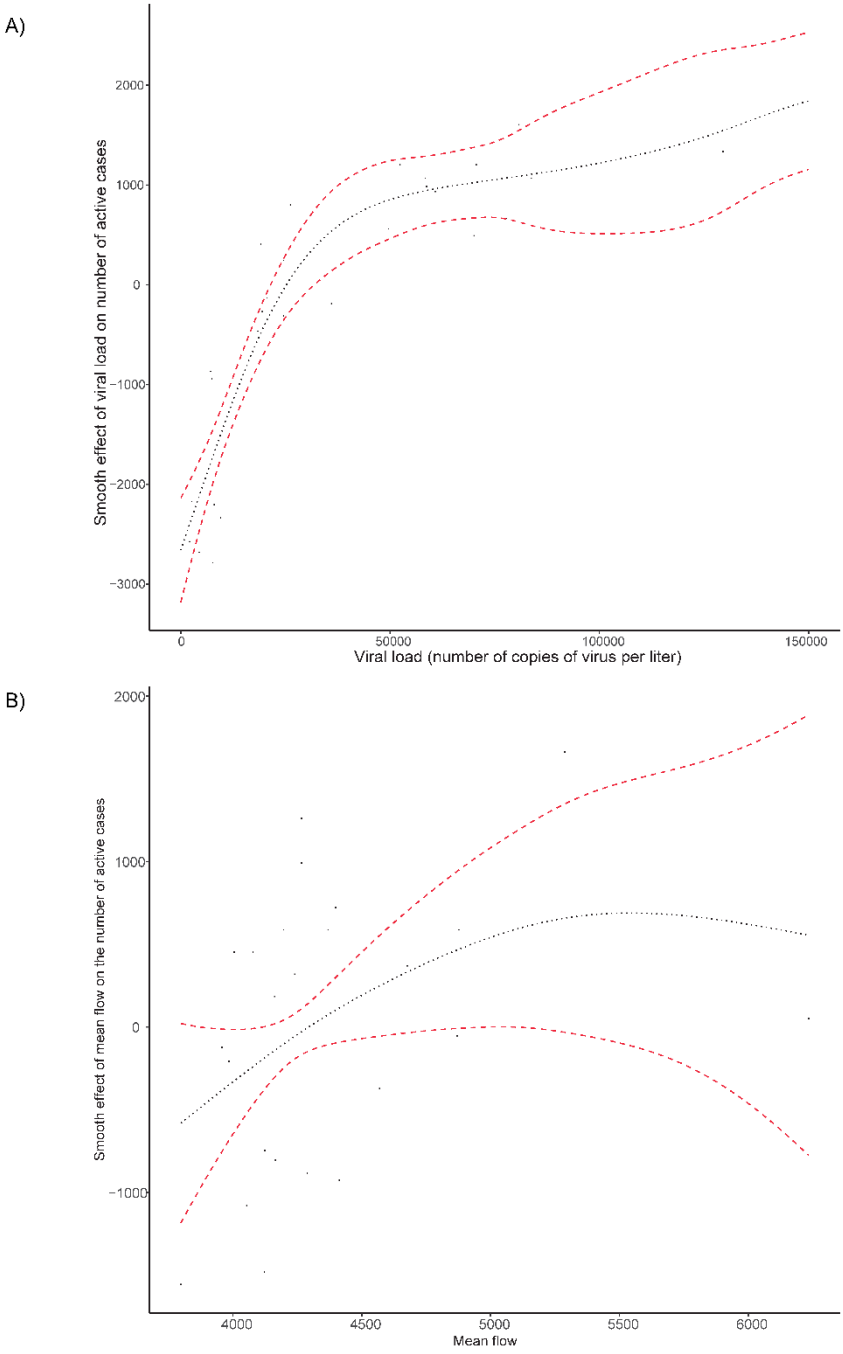

Figure 8. Nonparametric effect and confidence band for the viral load (A) and for the mean flow (B) in the real number of COVID-19 active cases when fitting a GAM model.

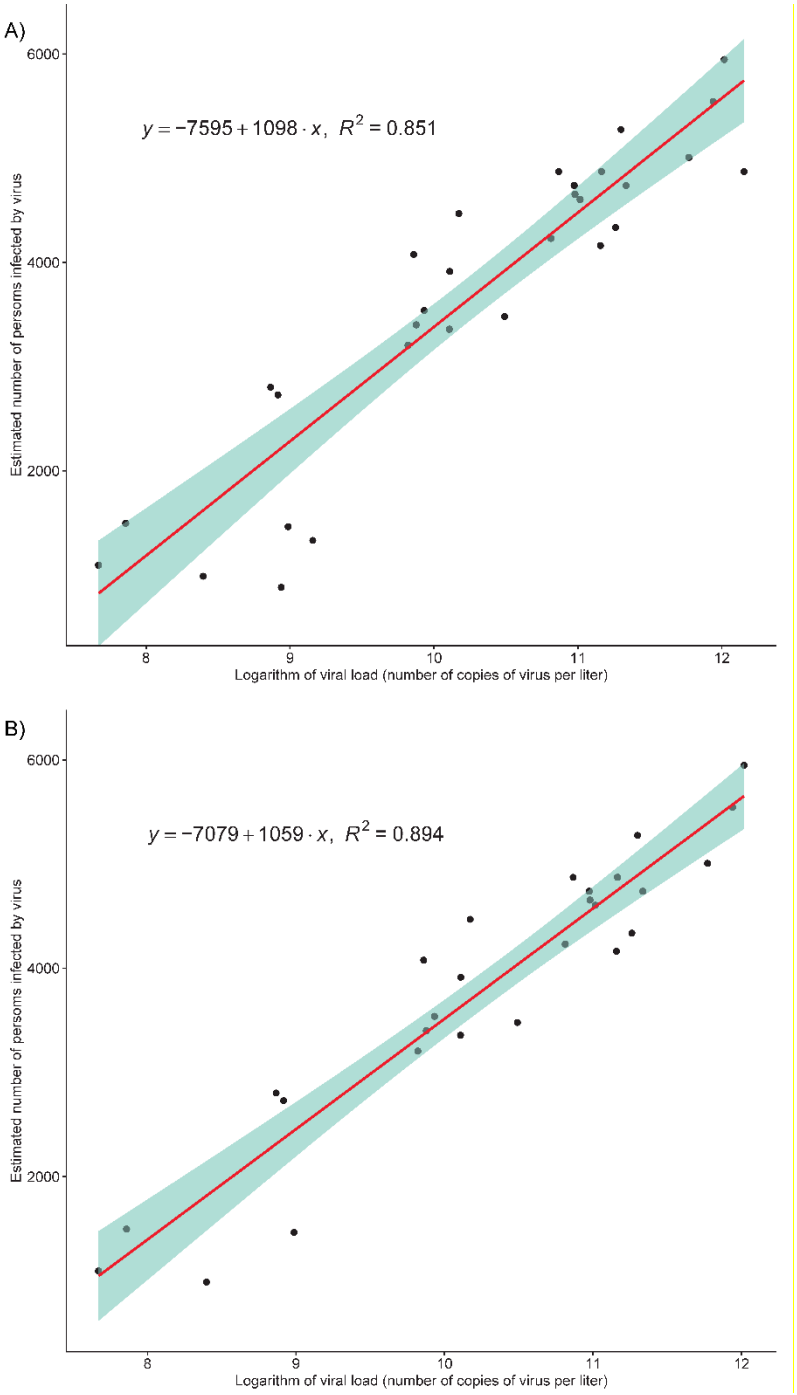

Figure 9. Scatterplot of the logarithm of the viral load measured in WWTP Bens and the estimated number of COVID-19 active cases before (A) and after (B) removing the three outliers detected. The linear fit (red line) and the confidence band (blue shaded area) are also included.

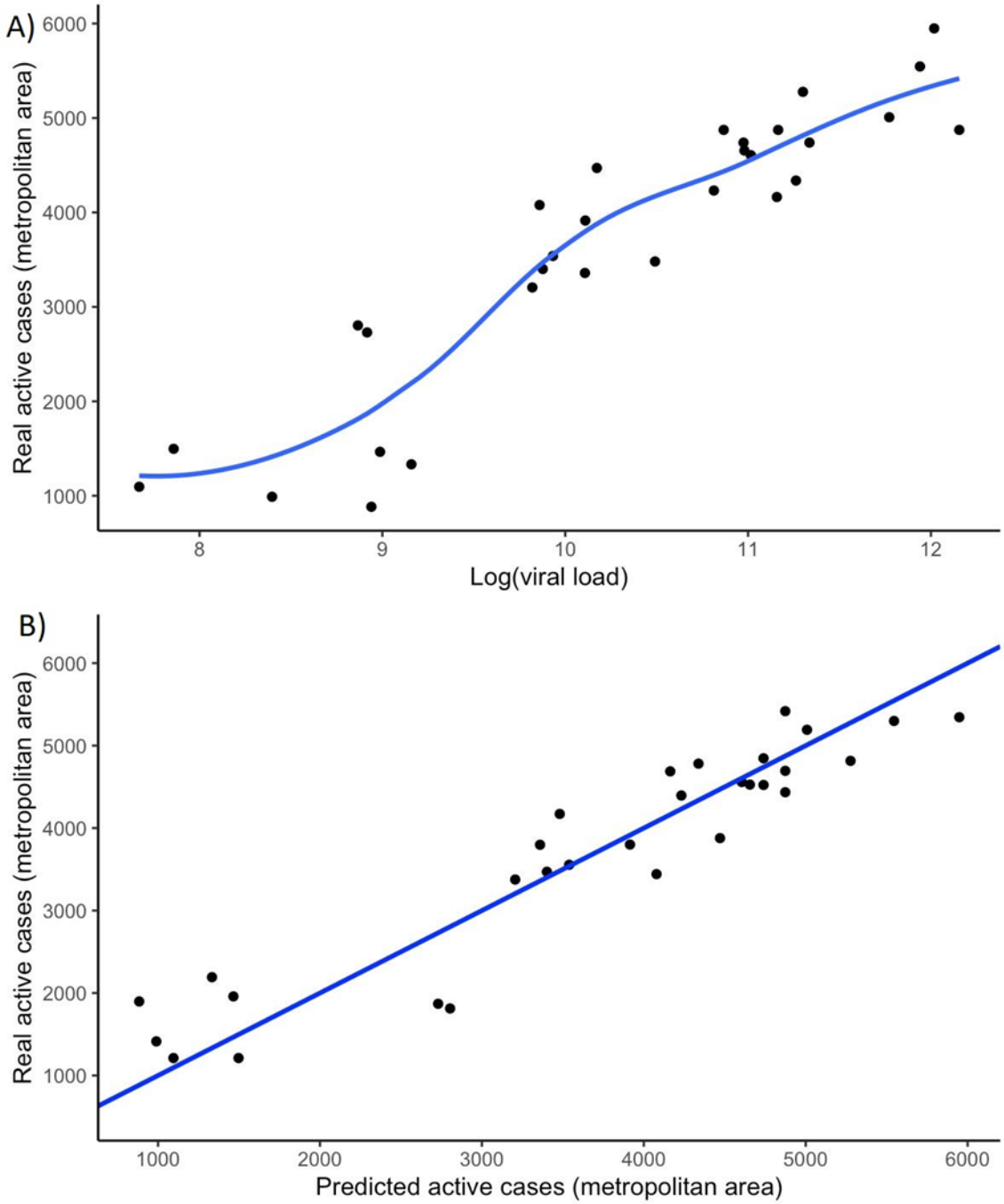

Figure 10. Estimation of the COVID- real active cases using LOESS. A) Real active cases in the metropolitan area vs the natural logarithm of the viral load, along with the quadratic LOESS fit. B) Real active cases in the metropolitan area vs the corresponding predicted values according to the quadratic LOESS fit in A), and the diagonal line.