Estimating SARS-CoV-2 prevalence from large-scale wastewater surveillance: insights from combined analysis of 44 sites in England

Mario Morvan

UCL

Anna Lo Jacomo

JBC

Celia Souque

JBC

Matthew Wade

JBC

Till Hoffmann

JBC

Koen Pouwels

University of Oxford

Andrew Singer

UK Centre for Ecology and Hydrology <https://orcid.org/0000-0003-4705-6063>

Joshua Bunce

JBC

Andrew Engeli

JBC

Jasmine Grimsley

Department for Health and Social Care

Kathleen O'Reilly (✉ lsh1603970@lshtm.ac.uk)

LSHTM

Leon Danon

Bristol

Article

Keywords: SARS-CoV-2, COVID-19, wastewater, machine learning

Posted Date: August 2nd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-770963/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Version of Record: A version of this preprint was published at Nature Communications on July 25th, 2022. See the published version at <https://doi.org/10.1038/s41467-022-31753-y>.

Estimating SARS-CoV-2 prevalence from large-scale wastewater surveillance: insights from combined analysis of 44 sites in England

Mario Morvan^{1,2§}, Anna Lo Jacomo^{1,3§}, Celia Souque^{1,4}, Matthew J. Wade^{1,5}, Till Hoffmann^{1,6}, Koen Pouwels^{7,8}, Andrew C Singer⁹, Joshua T. Bunce^{1,10}, Andrew Engeli¹, Jasmine Grimsley¹, Kathleen M. O'Reilly^{1,11§}, Leon Danon^{1,3§}

¹ Joint Biosecurity Centre, Department of Health and Social Care, London WC1B 4DA, UK

² Department of Physics and Astronomy, University College London, London WC1E 6BT, UK

³ Department of Engineering Mathematics, Ada Lovelace Building, University Walk, Bristol BS8 1TW, UK

⁴ Department of Zoology, University of Oxford, Oxford, OX1 3SZ, UK

⁵ School of Engineering, Newcastle University, Newcastle-upon-Tyne, NE1 7RU, UK

⁶ Department of Mathematics, Imperial College London, London, SW7 2AZ, UK

⁷ NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England, Oxford, UK

⁸ Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Oxford, UK

⁹ UK Centre for Ecology & Hydrology, Wallingford OX10 8BB, UK

¹⁰ Department for Environment, Food and Rural Affairs, London, SW1P 4DF, UK

¹¹ Faculty of Epidemiology and Population Health and Centre for Mathematical Modelling of Infectious Disease, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK

§ Contributed equally

Abstract

Accurate surveillance of the COVID-19 pandemic can be weakened by under-reporting of cases, particularly due to asymptomatic or pre-symptomatic infections, resulting in bias. Quantification of SARS-CoV-2 RNA in wastewater (WW) can be used to infer infection prevalence, but uncertainty in sensitivity and considerable variability has meant that accurate measurement remains elusive. Data from 44 sewage sites in England, covering 31% of the population, shows that SARS-CoV-2 prevalence is estimated to within 1.1% of estimates from representative prevalence surveys (with 95% confidence). Using machine learning and phenomenological models, differences between sampled sites, particularly the WW flow rate, influence prevalence estimation and require careful interpretation. SARS-CoV-2 signals in WW appear 4-5 days earlier in comparison to clinical testing data but are coincident with prevalence surveys suggesting that WW surveillance can be a leading indicator for asymptomatic viral infections. Wastewater-based epidemiology complements and strengthens traditional surveillance, with significant implications for public health.

Introduction

Estimates of SARS-CoV-2 infection prevalence are essential to understand COVID-19 disease burden and the impact of public health interventions (Franceschi *et al.*, 2020; Lancet, 2020)(Pouwels *et al.*, 2021). The sensitivity of surveillance for COVID-19 varies for several epidemiological, administrative, political and financial reasons, meaning that reported cases are likely to be an underestimate of actual cases (Noushad & Al-Saqqaf, 2021; Richterich, 2020; F. Wu *et al.*, 2020; S. L. Wu *et al.*, 2020) . Further, only a proportion of infections result in symptomatic disease; estimates range from 22 to 98% with an average of 75%, (Alene *et al.*, 2021; Subramanian *et al.*, 2021), so there is a disconnect between infections that result in transmission and reported cases via clinical surveillance. To monitor the trajectory of the COVID-19 pandemic and reduce the impact of bias due to any single source of information, it is essential to have multiple measures of prevalence with well-understood sources of bias and uncertainty.

Based on its proven success applied to other infectious diseases and markers of human health and behaviour (See, for example, Choi *et al.*, 2018; Feng *et al.*, 2018), wastewater (WW) monitoring for SARS-CoV-2 has been established in many countries, including England, since the start of the pandemic (Baraniuk, 2020; Hillary *et al.*, 2021). Early studies indicated that fragments of RNA corresponding to the SARS-CoV-2 viral genome were detectable in WW (Ahmed *et al.*, 2020), and in quantities that some quantitative measure could be established. However, laboratory protocols required refinement to establish a method that provides consistent measures of RNA with enough sensitivity to be used at scale. Quantitative reverse-transcriptase PCR (RT-qPCR) of target genes of the virus genome is now routinely performed on concentrated samples from wastewater.

A challenge inherent to WW monitoring is the potential impact of environmental and biochemical attributes on the detection and quantification of the virus concentration. In England and other countries that utilise combined sewer networks to transport sewage, the wastewater inflowing at the sewage treatment works (STWs) typically comprises a combination of raw sewage, household

effluent (e.g. from washing and cleaning), agricultural run-off, rainwater/snow melt, and trade waste from industry (Boogaerts et al. 2021). The percentage volume of human derived excreta likely to contain virus RNA (i.e., urine, faeces, nasal discharge, sputum, blood) (Tran *et al.*, 2020) in the collected wastewater sample is likely to vary because of additional inflow detailed above, which will dilute and may degrade the target analyte concentration, reducing the sensitivity of lab assays (Wilder *et al.*, 2021). In order to overcome these challenges and infer an estimate of prevalence from WW samples, additional data and statistical models can be used, and validation of model outputs using reliable estimates of prevalence is critical. However, the associated biases in clinical surveillance and the impact of uncertainties associated with environmental monitoring of viruses in sewer networks present significant challenges when considering prevalence estimation using WW measurements.

Back-calculating or estimating the quantity of chemical compounds (e.g., licit and illicit pharmaceuticals) or stressors (e.g., pathogens) by targeting indicative analytes present in WW is a common feature of wastewater-based epidemiology (WBE). For example, WBE has been applied successfully in estimating illegal drug consumption (Feng *et al.*, 2018), the degree of antibiotic resistance in a population (Hutinel *et al.*, 2019), among other applications. While most studies have used WW to track disease trends (i.e. increase/decrease), a number of studies have attempted to directly quantify prevalence from SARS-CoV-2 measurements, along with biological and hydrological parameters (Ahmed *et al.*, 2020; Arora *et al.*, 2020; F. Wu *et al.*, 2020). Broadly, studies using back-calculation for SARS-CoV-2 generally consider that disease prevalence is equal to the load of RNA in the sample, divided by the load of RNA produced by one infected person (Ahmed *et al.*, 2020). The underlying assumptions are that viral RNA is released proportionally to wastewater and perfectly mixed in the sewers, and that there are no significant losses of virus RNA in the network that lead to a decrease in measurement representativeness. Variations of this hypothesis have been suggested to account for additional 'signal loss' factors, for example decay, flow dilution, and temporal

shedding patterns in the population (Ahmed *et al.*, 2020; Arora *et al.*, 2020; Kantor *et al.*, 2021; Li *et al.*, 2021; Wade *et al.*, 2021).

The Covid Infection Survey (CIS) was established across England and Wales by the Office of National Statistics early in the pandemic to assess the prevalence of infection in the community through stratified nasopharyngeal sampling of individuals in households (Pouwels *et al.*, 2021). This survey has been essential to understand the dynamics of SARS-CoV-2 by estimating community positivity rates, and further to estimate these rates at regional and sub-regional scales. The wide availability of WW samples from July 2020 in England and sub-regional positivity estimates from the CIS provides a unique dataset to investigate and validate WW as a reliable estimate of prevalence to support public health actions. In this study, we estimate the prevalence of SARS-CoV-2 infection in the community and establish what additional data and analyses are required to have accurate and robust estimates of prevalence from WW.

Results

We analysed data collected between July 2020 and March 2021 from 44 sewage treatment works (STWs) across England (Fig. 1) covering an estimated 31% of the population. For each site, an average of four samples were collected per week, by either grab (41%) or composite (59%) sampling. Additional meta-data were collected on inorganics and other wastewater characteristics (see Supplementary Table S1 for further details).

Fig. 1 Geographical summary of the data used to estimate SARS-CoV-2 from wastewater. A) Map of Coronavirus Infection Survey (CIS) regions (outlined in blue) and the locations of wastewater (WW) catchments used in this study (in red). B) Regional 7-day rolling averages of CIS prevalence estimates (black) with 95% credible intervals using Bayesian modelling (grey regions), with corresponding predictions of prevalence from WW data only (blue) with 95% confidence interval from bootstrapping (blue vertical lines), and raw SARS-Cov-2 concentrations (yellow, right axis). The WW prevalence estimates are provided at a sub-regional level and combined to produce regional estimates for comparison.

Translation of raw WW data to prevalence estimates are illustrated using a phenomenological model that considers infection prevalence, shedding and stool generation, and the volume of water in the sewage column (see Methods). The assumptions of the model results in a linear relationship between prevalence and RNA concentrations. Sensitivity analysis illustrated that viral concentration in stool is the largest source of uncertainty in this approach (Fig. S2). Using average values of shedding rate from clinical studies (Hoffmann & Alsing, 2021) gives a relatively good fit with

observations from wastewater and CIS data in terms of average magnitude, but with high variability across individual samples (Fig. 2A). However, that variability is commensurate with the uncertainty in the appropriate hydrological and biological values used for the calculation. Comparing these model estimates to data indicates that for more than half of the CIS sub-regions in the study (60%) the model assumptions illustrate a valid relationship with the data (more than 60% of sample points fall within the 50% confidence interval of the model). Sites showing a poorer fit, have either relatively low (28%) or relatively high (12%) concentrations per positivity rate (Fig. 2A and 2B). Low concentrations could be caused by unusually high per capita flow rates (such as groundwater infiltration), or degradation of RNA during transit due to physical or chemical characteristics of the network (such as numerous pumping stations, or consistently different pH). The method of sample collection, together with limited homogenisation of the 'sewage parcel', could also lead to unrepresentative (either low or high) concentrations, or indeed unaccounted sewage discharge could also affect measurement (Hammond et al. 2021; Ort et al., 2010). Including an additional factor to account for degradation might provide a better model assumption for sites showing relatively low concentrations (e.g. Sub-region B in Figure 2A). In some sub-regions, the relationship between concentrations and prevalence is not well explained by the (static) linear model. A possible reason may be interactions between disease dynamics and shedding where infection in a growing epidemic appears to have increased viral load (Hay et al., n.d.); inclusion of this in the model would require a non-linear model.

Fig. 2 Comparison of the outputs from the phenomenological model to CIS prevalence estimates.

A) Example fit between the phenomenological model estimates (green region) and the wastewater and prevalence data from three CIS sub-regions (blue dots), selected to illustrate three cases: sub-region A (good correspondence), sub-region B (concentrations tend to be low) and sub-region C (concentrations tend to be high). Overall, model estimates are in the same order of magnitude and follow the shape of the relationship between concentrations and prevalence using distributions of likely parameter values, but confidence intervals are wide. The combined uncertainty in parameter values exceeds the variability seen in the data. **B)** The percentage of data points within each sub-region that fall within the 50% credible interval of the phenomenological model. **C)** The median concentrations per positivity rate. Only CIS sub-regions that overlap with the original 44 wastewater catchment sites are shown. Sites with a poor fit to the model (yellow in sub-plot B) show either relatively low (dark blue) or relatively high (dark red) concentrations in sub-plot C. Sites with a good fit to the model (dark green) tend to show intermediate concentrations (white).

Combining WW data, site-level and sample-level variables within a statistical modelling framework (exemplified using an extreme gradient boosted regression tree model (XGB), see Methods) to estimate prevalence of SARS-CoV-2 provides reliable metrics across regions and throughout the evolving epidemiology of the COVID-19 pandemic in England. Using this model, SARS-CoV-2 prevalence was tracked within 1.1% (with 95% confidence) from the Coronavirus infection survey (Fig. 1B). When aggregated to regional level, an average Mean Absolute Error (MAE) was obtained, with the East of England performing above average and the North West performing below average

(MAE of 0.24 and 0.39, respectively) (Fig. 2A). We focus the results of the modelling to a regional level here, but have carried out the analysis at a sub-regional level to inform the public health response at these smaller levels of aggregation (Fig. S3).

The XGB model was found to be the best of the candidate models that were developed to interrogate the data and identify what variables, in addition to the raw RNA concentrations, would provide accurate estimates of prevalence. Different candidate models (linear model, a linear model with random effects and XGB), were evaluated and compared using the MAE between predictions and median positivity rates estimates from CIS. The addition of temporally varying data such as ammonia concentration, fraction of samples below the limit of quantification and site-specific data such as population coverage greatly improved the overall fit with MAE (0.25 to 0.10 for the XGB) and average MAE per CIS subregion (0.22 to 0.10) when compared with a model trained on SARS-COV-2 concentrations alone. This addition reduced both intra- and inter-regional variation in performance (Fig. 3). Further collation of additional site-level characteristics through consultation with water companies and characterisation of the catchment area showed an additional reduction in bias in the model residuals distribution against these characteristics (Fig. S3) highlighting the robustness of our final model to wastewater network differences. While the XGB model will be applied to estimate prevalence in England, the relative contributions of each variable and partial dependency plots (Fig. S4) are used to illustrate the direction of their effects and provide guidance for use outside of this application. However, exploration of the site-specific random effects (within the random effects model) illustrated that there was considerable variability in MAE within sites that had yet to be fully accounted for.

Fig. 3 Extreme gradient boosted regression tree (XGB) model performance across regions of England, A) trained using SARS-COV-2 concentration alone, and B) including the full set of time-varying and site-specific parameters. Lower and upper hinges of the box plot corresponds to first and third quartile with middle line corresponding to the median. The overall subregion median score for each model is represented (dotted line).

These WW data were collected in England across a time period where the prevalence of infection has varied considerably as a result of epidemic emergence and suppression through non-pharmaceutical interventions. The statistical modelling presented here illustrates that prevalence estimates are accurate and precise across a wide range of prevalence values (Fig. 4). The prevalence is tracked within 1.1% (with 95% confidence) for the XGB model and is more precise at higher values of prevalence. Comparison between the random effects and XGB model illustrates reduced precision and over estimation of prevalence at lower values of prevalence for the random effects model.

Fig. 4 Conditional predictions colour-mapped as a function of the CIS values in the $\log_{10}$ space for the A) Random Effects model and B) Extreme gradient boosted regression tree (XGB) model.

A lead and lag analysis was performed using the regression models on CIS estimates. Sampling dates for WW were shifted between -10 and 20 days with daily increments, while training a model at each step to predict the CIS positivity rates in outputs, whose dates had been fixed. At each step, the models were evaluated using the bootstrapped MAE, producing a curve of prediction errors as a function of the wastewater lag (Fig. 5). Results show a minimum value of the smoothed error curve between 0 and 2 shifted days, indicating no clear advantage to predict CIS backward or forward in time from WW data. For comparison, this analysis was replicated on Pillar 1&2 data from Test and Trace. In this case the regression outputs were the case rates reported by Test and Trace until May 17th 2021, smoothed with a 7-day centred window to remove weekly periodicity and preserve consistency of reporting dates. In addition, the WW dataset was stripped to contain samples only up to 20 days before May 17th 2021 to ensure the stability of dataset sizes during the analysis. In this case the MAE is minimal between +3 and +5 days shift, suggesting a ≈ 4 days lead of WW monitoring date over reported Test and Trace cases (Fig. 5B).

Fig. 5 Lead and lag analysis performed by observing the changes of performance of the gradient boosting model when shifting the wastewater sampling dates backward and forward in time while keeping the output cases dates fixed, using both A) CIS and B) Test and Trace cases. The shift associated with the minimal error is indicated by the red dotted line. A minimal error reached for a positive number of WW shifted days can be interpreted as a lead from wastewater by as many days. No clear lead of WW over CIS has been observed in this experiment, but a ~4 days lead over T&T has been observed.

Discussion

We have shown that concentrations of SARS-CoV-2 RNA collected from wastewater in 44 sites in England, combined with essential related variables can provide reliable estimates of prevalence of SARS-CoV-2 infection within a population. Site specific characteristics including limit of detection, dilution, network characteristics and other unexplained data, mean that a mechanistic model alone fails to capture the full variability in the data. However, by using locally explicit information and hierarchical models, we can understand these differences and account for them in a data-driven manner. We used our best fitting models to observe a 4-day lead in WW concentration over mainly symptomatic testing through routine surveillance. This lead in can be explained by transmission often occurring prior to symptomatic illness (He *et al.* 2020), which, it would be reasonable to assume, would be reflected in shedding of SARS-CoV-2 RNA in stool, and subsequently detected in WW.

A strength of our analysis comes from using data on prospective surveys of infection prevalence as well as reported cases of COVID-19. The ONS CIS survey has been designed to minimise bias in prevalence estimates by incentivising participation, and accounting for under-representation of specific groups of individuals (Gao et al., 2021; ONS 2021). SARS-CoV-2 is shed in faeces of both symptomatic and asymptomatic individuals and so even with perfect surveillance for cases of COVID-19 the correlation between infection and case reporting will be imperfect due to differences in the probability of being symptomatic across regions. Furthermore, variability in reporting of symptomatic disease to the Test and Trace programme (and alternative surveillance systems specific to other countries) across socio-demographic strata is well documented (Byambasuren et al., 2021; Kendall et al., 2020; Russell et al., 2021), raising questions on whether clinical surveillance is an accurate and reliable estimate of disease burden. Indeed, previous comparisons of WW to disease incidence have provided variable results where under-reporting of cases has been hypothesised (Hillary et al., 2021; Peccia et al., 2020; Randazzo et al., 2020; F. Wu et al., 2020; Xiao et al., 2021).

For these reasons amongst others, we argue that inferred estimates of prevalence from wastewater RNA concentrations, a measure agnostic to the source of virus, are an informative addition to clinical disease surveillance. Furthermore, WW monitoring in England has since been expanded to include additional smaller sites within networks which may enable estimates of prevalence at a smaller geographical scale than described here.

The validation of prevalence estimates derived from wastewater illustrates the critical value of collecting meta-data in addition to raw RNA concentrations alone, and confirms findings from more established WW monitoring systems (O'Reilly *et al.*, 2020). Several of the variables (suspended solids, ammonia concentration, phosphate) indicate organic and inorganic substances and likely approximate dilution of raw sewage with other sources of flow. The relationship is not likely to be linear as these variables also indicate the presence of agricultural runoff and so can also act as confounders. Inclusion of sample pH appears to further improve model prediction, and may reflect degradation of viral fragments at lower pH values that may reduce the sensitivity of rt-PCR, and has been previously observed in WW surveillance for poliovirus (Hamisu *et al.*, 2020). While the composite fraction, limit of quantification, reception delay, percentage of the catchment population, the catchment area and population fraction indicate a linear relationship, their inclusion still improves model prediction. It is interesting to note that the percentage of samples from each WW site that were composite (as opposed to 'grab') had only a moderate effect on prevalence estimation. The use of composite samplers was dependent on their availability during the pandemic, with grab samples used for convenience as opposed to strategic intent. Ideally, this finding should be investigated using a comparative study design to investigate the possible added benefit of estimating SARS-Cov-2. Further work will establish how these indicators should be used in settings with no measure of infection prevalence to improve inference of WW data.

This analysis has illustrated the predictive ability of WW at a time when comparatively few individuals were vaccinated against COVID-19 in England (by 01 March 2021 30% had received at least one dose). As vaccination increases the relationship between infection and faecal shedding may change and the predictive ability of the model will need to be monitored and potentially adapted to account for this. Studies of viral load in vaccinated but infected individuals have illustrated that less virus is shed in the nasopharynx by vaccinated individuals (Levine-Tiefenbrun *et al.*, 2021; McEllistrem *et al.*, 2021; Pritchard *et al.*, 2021), but there are currently no data on shedding of SARS-CoV-2 viral fragments in stool from infected but vaccinated individuals. Moreover, a study of healthcare workers where the alpha variant was dominant reported no difference in viral load by vaccination status (Ioannou *et al.* 2021), and preliminary analysis from England has not identified noticeable differences in WW through to March 2021 (Hoffman *et al.* 2021b).

The use of WW to infer prevalence is reliant on a converging sewer network that samples a sufficient proportion of the population of interest. Remote populations, however, such as islands or rural communities, may be served by septic tank systems, leading to blind-spots in observations, especially if there is a relationship between income and centralised waste removal provision. Consequently, the benefits of wastewater-based estimates are less obvious for low density settings. Additionally, the impact of sewage effluent from hospitals has not been accounted for in this analysis which could result in an over-estimation of prevalence within a population when compared to the CIS (as hospitalised individuals are not included in the sampling). Further work will investigate the impact of hospitals and other potential sources of bias. Our analysis illustrated the added benefit of including additional meta-data within a statistical model to infer prevalence highlighting the importance of site-specific characteristics. Therefore, the use of our inference model outside of the setting presented here should be avoided in the absence of further external validation and local information.

Nonetheless, close monitoring of emerging SARS-CoV-2 variants with changed phenotypic properties (Challen, *et al.*, 2021; Davies, Abbott, *et al.*, 2021; Davies, Jarvis, *et al.*, 2021) will continue to be needed; meta-genomic analysis of wastewater samples provides insight into the genomic diversity of virus in the community. Also, sampling of sites that cover small catchment areas will remain capable of revealing localised spikes in incidence used to detect hotspots and inform local public health authorities (ONS 2021). Finally, the investment that has already been made in WW surveillance systems across England, combined with insights from analysis pipelines such as ours, can be leveraged to other communicable and noncommunicable diseases and human behaviours impacting personal and community health (Hassard *et al.*, 2021; Polo *et al.*, 2020). As the pandemic continues to evolve, and the threat to society becomes less acute, it is likely that surveillance of SARS-CoV-2 will need to become more sustainable, making the most of those investments. These data streams will remain an important feature for public health monitoring, complementing traditional clinical surveillance as the country emerges out of the acute phase of the COVID-19 pandemic.

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., Tschärke, B., Verhagen, R., Smith, W. J. M., Zaugg, J., Dierens, L., Hugenholtz, P., Thomas, K. V., & Mueller, J. F. (2020). First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Science of the Total Environment*, 728. <https://doi.org/10.1016/j.scitotenv.2020.138764>
- Alene M, Yismaw L, Assemie MA, Ketema DB, Mengist B, et al. (2021) Magnitude of asymptomatic COVID-19 cases throughout the course of infection: A systematic review and meta-analysis. *PLOS ONE* 16(3): e0249090. <https://doi.org/10.1371/journal.pone.0249090>
- Arora, S., Nag, A., Sethi, J., Rajvanshi, J., Saxena, S., Shrivastava, S. K., & Gupta, A. B. (2020). Sewage surveillance for the presence of SARS-CoV-2 genome as a useful wastewater based epidemiology (WBE) tracking tool in India. *Water Science and Technology*, 82(12). <https://doi.org/10.2166/wst.2020.540>
- Baraniuk, C. (2020). Sewage monitoring is the UK's next defence against covid-19. *Bmj*, 370.
- Blangiardo M., and Cameletti M. (2015) Spatial and Spatio-temporal Bayesian Models with R - INLA. Wiley and Sons. ISBN: 978-1-118-32655-8
- Boogaerts T., Ahmed F., Choi P.M., Tschärke B., O'Brien J., De Loof H, Gao J., Thai P., Thomas K., Mueller J.F. , Hall W., Covaci A, van Nuijs A.L.N., (2021) Current and future perspectives for wastewater-based epidemiology as a monitoring tool for pharmaceutical use. *Science of The Total Environment* 789, 0048-9697
- Cardoso A., Prigiobbe V., Giulianelli M., Baer E., De Bénédittis J., Coelho S.T. (2006) Assessing the impact of infiltration and exfiltration in sewer systems using performance indicators: case studies of the APUSS project. *Water Practice and Technology* 1 March 2006; 1 (1): wpt2006019.

doi: <https://doi.org/10.2166/wpt.2006.019>

- Challen, R., Brooks-Pollock, E., Read, J. M., Dyson, L., Tsaneva-Atanasova, K., & Danon, L. (2021). Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study. *Bmj*, 372.
- Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, 73, 914–920.
- Choi, P. M., Tschärke, B. J., Donner, E., O’Brien, J. W., Grant, S. C., Kaserzon, S. L., Mackie, R., O’Malley, E., Crosbie, N. D., Thomas, K. V., & others. (2018). Wastewater-based epidemiology biomarkers: past, present and future. *TrAC Trends in Analytical Chemistry*, 105, 453–469.
- Davies, N. G., Abbott, S., Barnard, R. C., Jarvis, C. I., Kucharski, A. J., Munday, J. D., Pearson, C. A. B., Russell, T. W., Tully, D. C., Washburne, A. D., & others. (2021). Estimated transmissibility and impact of SARS-CoV-2 lineage B. 1.1. 7 in England. *Science*, 372(6538).
- Davies, N. G., Jarvis, C. I., Edmunds, W. J., Jewell, N. P., Diaz-Ordaz, K., & Keogh, R. H. (2021). Increased mortality in community-tested cases of SARS-CoV-2 lineage B. 1.1. 7. *Nature*, 593(7858), 270–274.
- Feng, L., Zhang, W., & Li, X. (2018). Monitoring of regional drug abuse through wastewater-based epidemiology—A critical review. *Science China Earth Sciences*, 61(3), 239–255.
- Franceschi, V. B., Santos, A. S., Glaeser, A. B., Paiz, J. C., Caldana, G. D., Machado Lessa, C. L., de Menezes Mayer, A., Küchle, J. G., Gazzola Zen, P. R., Vigo, A., & others. (2020). Population-based prevalence surveys during the Covid-19 pandemic: a systematic review. *Reviews in Medical Virology*, e2200.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

- Gao, Y., Kennedy, L., Simpson, D., Gelman, A. (2021) Improving Multilevel Regression and Poststratification with Structured Priors. *Bayesian Analysis* <https://doi.org/10.1214/20-BA1223>
- Hamisu, A. W., Blake, I. M., Sume, G., Braka, F., Jimoh, A., Dahiru, H., Bonos, M., Dankoli, R., Mamuda Bello, A., Yusuf, K. M., & others. (2020). Characterizing environmental surveillance sites in Nigeria and their sensitivity to detect poliovirus and other enteroviruses. *The Journal of Infectious Diseases*.
- Hammond, P., Suttie, M., Lewis, V.T. et al. Detection of untreated sewage discharges to watercourses using machine learning. *npj Clean Water* 4, 18 (2021). <https://doi-org.ez.lshtm.ac.uk/10.1038/s41545-021-00108-3>
- Hassard, F., Lundy, L., Singer, A. C., Grimsley, J., & Di Cesare, M. (2021). Innovation in wastewater near-source tracking for rapid identification of COVID-19 in schools. *The Lancet Microbe*, 2(1), e4--e5.
- Hay, J. A., Kennedy-Shaffer, L., Kanjilal, S., Lipsitch, M., & Mina, M. J. (n.d.). *Estimating epidemiologic dynamics from single cross-sectional viral load distributions*. *medRxiv [Internet]*. 2020 Oct 13 [cited 2021 Jan 6]; 2020.10. 08.20204222.
- Hillary, L. S., Farkas, K., Maher, K. H., Lucaci, A., Thorpe, J., Distaso, M. A., Gaze, W. H., Paterson, S., Burke, T., Connor, T. R., & others. (2021). Monitoring SARS-CoV-2 in municipal wastewater to evaluate the success of lockdown measures for controlling COVID-19 in the UK. *Water Research*, 200, 117214.
- Hoffmann, T., & Alsing, J. (2021a). Faecal shedding models for SARS-CoV-2 RNA amongst hospitalised patients and implications for wastewater-based epidemiology. *MedRxiv*.
- Hoffmann, T., McIntyre-Nolan, S., Bunce, J. T., Grimsley, J., Hart, A., Lo Jacomo, A., Morvan, M., Robins, K., Wade, M J., Watts, G., Engeli, A., Henderson, G., (2021b) Current environmental monitoring cannot constrain the effect of vaccines on SARS-CoV-2 transmission: Report for

SAGE 08/04/2021

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/979870/S1192_Current_environmental_monitoring_cannot_constrainthe_effectof_vaccines_onSARS-CoV-2_transmission_Report_for_SAGE.pdf

He, X, Lau, EHY, Wu, P, Deng, XL, Wang, J, Hao, XX, Lau, YC, Wong, JSY, Guan, YJ, Tan, XH, Mo, XN, Chen, YQ, Liao, BL, Chen, WL, Hu, FY, Zhang, Q, Zhong, MQ, Wu, YR, Zhao, LZ, Zhang, FC, Cowling, BJ, Li, F, Leung, GM (2020) Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*. 26 (5),DI 10.1038/s41591-020-0869-5.

Hillary, L.s., Farkas, K., Maher, K. H., Lucaci, A., Thorpe, J. Distaso, M.A., Gaze, W.H., Paterson, S., Burke, T., Connor, T.C., McDonald, J. E., Malham, S.K., Jones, D.J. (2021) Monitoring SARS-CoV-2 in municipal wastewater to evaluate the success of lockdown measures for controlling COVID-19 in the UK. *Water Research*, 200(117214). <https://doi.org/10.1016/j.watres.2021.117214>.

Hutinel, M., Huijbers, P. M. C., Fick, J., Åhrén, C., Larsson, D. G. J., & Flach, C.-F. (2019). Population-level surveillance of antibiotic resistance in *Escherichia coli* through sewage analysis. *Eurosurveillance*, 24(37), 1800497.

Ioannou, P., Karakostas, S., Astrinaki, E., Saplamidou, S., Vitsaxaki, E., Hamilos, G., Sourvinos, G & Kofteridis, D. P. (2021): Transmission of SARS-CoV-2 variant B.1.1.7 among vaccinated health care workers, *Infectious Diseases*, DOI: 10.1080/23744235.2021.1945139

Kendall, M., Milsom, L., Abeler-Dörner, L., Wymant, C., Ferretti, L., Briers, M., Holmes, C., Bonsall, D., Abeler, J., & Fraser, C. (2020). Epidemiological changes on the Isle of Wight after the launch of the NHS Test and Trace programme: a preliminary analysis. *The Lancet Digital Health*, 2(12), e658--e666.

Kojima, N. et al. Self-collected oral fluid and nasal swab specimens demonstrate comparable sensitivity to clinician-collected nasopharyngeal swab specimens for the detection of SARS-

CoV-2. (2020) Clin. Infect. Dis. <https://doi.org/10.1093/cid/ciaa1589>.

Lancet, T. (2020). COVID-19: fighting panic with information. *Lancet (London, England)*, 395(10224), 537.

Levine-Tiefenbrun M, Yelin I, Katz R, et al. Initial report of decreased SARS-CoV-2 viral load after inoculation with the BNT162b2 vaccine. *Nat Med*. 2021;27(5):790–792.

McEllistrem MC, Clancy CJ, Buehrle DJ, et al. Single dose of a mRNA SARS-CoV-2 vaccine is associated with lower nasopharyngeal viral load among nursing home residents with asymptomatic COVID-19. *Clin Infect Dis*. 2021;ciab263.

Noushad, M., & Al-Saqqaf, I. S. (2021). COVID-19 case fatality rates can be highly misleading in resource-poor and fragile nations: the case of Yemen. *Clinical Microbiology and Infection*, 27(4), 509–510.

Office for National Statistics (2021) Interactive Map (of COVID-19 cases)
<https://coronavirus.data.gov.uk/details/interactive-map/cases>

Ort, C., Lawrence, M. G., Reungoat, J., & Mueller, J. F. (2010). Sampling for PPCPs in wastewater systems: comparison of different sampling modes and optimization strategies. *Environmental Science & Technology*, 44(16), 6289–6296.

O'Reilly, K. M., Allen D. J. Fine, P., Asghar H. (2020) The challenges of informative wastewater sampling for SARS-CoV-2 must be met: lessons from polio eradication. *Lancet Microbe*
[https://doi.org/10.1016/S2666-5247\(20\)30100-2](https://doi.org/10.1016/S2666-5247(20)30100-2)

Office for National Statistics (ONS) Coronavirus (COVID-19) Infection Survey, UK: 23 July 2021 - latest release
<https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveyspilot/23july2021> (Accessed 25th July 2021).

- 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., Warren, J. L., Weinberger, D. M., Arnold, W., & Omer, S. B. (2020). Measurement of SARS-CoV-2 RNA in wastewater tracks community infection dynamics. *Nature Biotechnology*, 38(10). <https://doi.org/10.1038/s41587-020-0684-z>
- Penn, R., Ward, B. J., Strande, L., & Maurer, M. (2018). Review of synthetic human faeces and faecal sludge for sanitation and wastewater research. *Water Research*, 132, 222–240.
- Pianosi, F., Sarrazin, F., & Wagener, T. (2015). A Matlab toolbox for global sensitivity analysis. *Environmental Modelling & Software*, 70, 80–85.
- Polo, D. D., Quintela-Baluja, M., Corbishley, A., Jones, D. L., Singer, A. C., Graham, D. W., & Romalde, P. J. L. (2020). Making waves: Wastewater-based epidemiology for SARS-CoV-2--Developing robust approaches for surveillance and prediction is harder than it looks. *Water Research*, 116404.
- Pouwels, K. B., House, T., Pritchard, E., Robotham, J. V., Birrell, P. J., Gelman, A., Vihta, K.-D., Bowers, N., Boreham, I., Thomas, H., & others. (2021). Community prevalence of SARS-CoV-2 in England from April to November, 2020: results from the ONS Coronavirus Infection Survey. *The Lancet Public Health*, 6(1), e30--e38.
- Pritchard, E., Matthews, P. C., Stoesser, N., Eyre, D. W., Gethings, O., et al. (2021) Impact of vaccination on new SARS-CoV-2 infections in the United Kingdom. *Nature Medicine* 1546-170X <https://doi.org/10.1038/s41591-021-01410-w>
- Randazzo, W., Truchado, P., Cuevas-Ferrando, E., Simón, P., Allende, A., & Sánchez, G. (2020). SARS-CoV-2 RNA in wastewater anticipated COVID-19 occurrence in a low prevalence area. *Water Research*, 181. <https://doi.org/10.1016/j.watres.2020.115942>
- Richterich, P. (2020). Severe underestimation of COVID-19 case numbers: effect of epidemic growth rate and test restrictions. *MedRxiv*.

- Rose, C., Parker, A., Jefferson, B., & Cartmell, E. (2015). The characterization of feces and urine: A review of the literature to inform advanced treatment technology. *Critical Reviews in Environmental Science and Technology*, 45(17).
<https://doi.org/10.1080/10643389.2014.1000761>
- Russell, T. W., Wu, J. T., Clifford, S., Edmunds, W. J., Kucharski, A. J., Jit, M., & others. (2021). Effect of internationally imported cases on internal spread of COVID-19: a mathematical modelling study. *The Lancet Public Health*, 6(1), e12--e20.
- Subramanian, R., He, Q., & Pascual, M. (2021). Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proceedings of the National Academy of Sciences*, 118(9).
- Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, 158, 1533–1543.
- Tran, H. N., Le, G. T., Nguyen, D. T., Juang, R.-S., Rinklebe, J., Bhatnagar, A., Lima, E. C., Iqbal, H. M. N., Sarmah, A. K., & Chao, H.-P. (2020). SARS-CoV-2 coronavirus in water and wastewater: A critical review about presence and concern. *Environmental Research*, 110265.
- Volz, E., Mishra, S., Chand, M., Barrett, J. C., Johnson, R., Geidelberg, L., Hinsley, W. R., Laydon, D. J., Dabrera, G., O'Toole, Á., & others. (2021). Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England. *Nature*, 593(7858), 266–269.
- Wade, M., Lo Jacomo, A. Armenise, A., Brown, M., Bunce, J., Cameron, G., Fang, Z. et al. (2021) Understanding and Managing Uncertainty and Variability for Wastewater Monitoring beyond the Pandemic: Lessons Learned from the United Kingdom National COVID-19 Surveillance Programmes. Preprint. Earth and Space Science Open Archive. Earth and Space Science Open Archive. World. 26 July 2021. <https://doi.org/10.1002/essoar.10507606.2>.
- Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable

information criterion in singular learning theory. *Journal of Machine Learning Research* 11, 3571–3594.

Wilder, M. L., Middleton, F., Larsen, D. A., Du, Q., Fenty, A., Zeng, T., Insaf, T., Kilaru, P., Collins, M., Kmush, B., & others. (2021). Co-quantification of crAssphage increases confidence in wastewater-based epidemiology for SARS-CoV-2 in low prevalence areas. *Water Research X*, 11, 100100.

Wohland P, Burkitt M, Norman P, Rees P, Boden P and Durham H, ETHPOP Database, ESRC Follow on Fund "Ethnic group population trends". www.ethpop.org. Date accessed 25th July 2021.

Wu, F., Zhang, J., Xiao, A., Gu, X., Lee, W. L., Armas, F., Kauffman, K., Hanage, W., Matus, M., Ghaeli, N., Endo, N., Duvallet, C., Poyet, M., Moniz, K., Washburne, A. D., Erickson, T. B., Chai, P. R., Thompson, J., & Alm, E. J. (2020). SARS-CoV-2 Titers in Wastewater Are Higher than Expected from Clinically Confirmed Cases. *MSystems*, 5(4). <https://doi.org/10.1128/msystems.00614-20>

Wu, S. L., Mertens, A. N., Crider, Y. S., Nguyen, A., Pokpongkiat, N. N., Djajadi, S., Seth, A., Hsiang, M. S., Colford, J. M., Reingold, A., & others. (2020). Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications*, 11(1), 1–10.

Xiao, A., Wu, F., Bushman, M., Zhang, J., Imakaev, M., Chai, P. R., Duvallet, C., Endo, N., Erickson, T. B., Armas, F., & others. (2021). Metrics to relate COVID-19 wastewater data to clinical testing dynamics. *MedRxiv*.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308–324.

Methods

1. Data

a. Wastewater data from 44 sites in England collected by the Environment Agency

The WW and metadata variables used for the analysis are summarised in Table S1 (A correlation matrix between variables is also provided Fig. S1). Untreated influent WW were collected from each sewage treatment works located across England. Samples were either collected as 'grab' samples or from a composite (24hr) sampler, and were transported to the laboratory and stored at 4°C until processed (within 24hrs). Physio-chemical analyses were carried out and extraction of pre-concentration viral nucleic acids. The physio-chemical analyses include quantification of pH, ammonia, orthophosphates and suspended solids, with methods described further in Hilary *et al.* (2021). For quantification of SARS-CoV-2 RNA 1-2 L of each sample was concentrated using tangential flow ultrafiltration, followed by further secondary concentration. The centrifugation pellets were then re-suspended for nucleic acid extraction, and subjected to q(RT-)PCR (shortened to PCR) assays. The PCR assays focus on detection of the N1 and E gene, and here our analyses is on quantification of the N1 gene. Again, full details of the methods are described in Hilary *et al.* (2021) and are summarised here. A serial dilution of DNA standards within the range of 10^5 – 10^0 gc μL^{-1} was used for quantification. For each sample the PCR cycle threshold was translated to the equivalent mean gc per 100ml⁻¹ based on two duplicates of the sample and the standard curve generated from the serial dilution. The limit of detection (LoD - the lowest concentration where all replicates were positive) and limit of quantification (LoQ - the lowest concentration where the coefficient of variance was below 0.25) were determined by running WW samples spiked with low concentrations of SARS-CoV-2. For the N1 gene the LoD was 1.7 gc L⁻¹ and the LoQ was 11.8 gc L⁻¹.

Some values were missing or below limit of detection, they were replaced by their mean values over week number and sites. Remaining missing values are then replaced by the median for each week and for all sites together first, and then by the median for each site and for all dates otherwise.

Finally, a $\log_{10}$ transformation was applied to all the concentration variables and the target variables to reduce the heavy skewness of their distribution.

b. The ONS Coronavirus Infection Survey

The ONS COVID-19 Infection Survey data are used to infer subregional estimates of positivity (Pouwels *et al.*, 2021). CIS is a large household survey with longitudinal follow-up (ISRCTN21086382). The study received ethical approval from the South Central Berkshire B Research Ethics Committee (20/SC/0195). Private households are randomly selected on a continuous basis from address lists and previous surveys to provide a representative sample across the UK. For the current study, only data from England was used. At the first visit, participants were asked for (optional) consent for follow-up visits every week for the next month, then monthly for 12 months from enrolment. At each visit, enrolled household members provided a nose and throat self-swab following instructions from the study worker. These were comparable to or even more sensitive than swabs performed by healthcare workers (Kojima *et al.* 2020).

c. Linkage of WW data to CIS data

CIS subregional data was computed weekly, and interpolated to daily using linear interpolation between adjacent points for each subregion. This allows comparison with the WW data for which 4 values are available every week. WW sites are mapped onto CIS subregions using a lookup table built from Layer Super Output Areas (LSOA) to WW site using knowledge of extent of catchment area (available from WW companies). With the details of the number of unique properties included or not in each LSOA/catchment intersection, the subregional coverage (population covered per WW site) is then inferred. This data is then used to estimate the percentage of the subregion covered by each WW site. From the 44 sites and CIS sub-regions we estimate infection prevalence from a total of 83 locations (Fig. S3).

d. The Test & Trace Pillar 1&2 case rates at LSOA (Layer Super Output Areas) level

Case rates of the number of new people infected per 100,000 individuals from Pillar 1&2 data were available from Public Health England. These data were aggregated to wastewater catchment level using a mapping from LSOA to catchment areas. Furthermore, the case rates were smoothed using a 7-day moving average to remove any artefactual weekly periodicity and therefore provide a better estimate of incidence.

2. Models

a. Phenomenological Model

The phenomenological model considers that prevalence (P , or the proportion of population infected) is related to C (the measured viral concentration in the sample, in units of gene copies per litre), by (equation 1):

$$P = \frac{C \times Q_P}{S \times V}, \quad (1)$$

where Q_P is the wastewater generated by one person per day (in L/day), S is the mean shedding rate, or the concentration of virus in stool of infected people (in gc/ml) and V is the mean volume of stool per person per day (in ml/day). A similar equation is used in Ahmed et al. (2020) using total flow rather than per person flow. We use the same model for all sites. Mean flow is set to 400 L/person/day, based data from 15 sites where flow data is routinely collected. Mean shedding rate is assumed to be 1.9×10^6 from Hoffmann & Alsing (2021). A mean stool volume of 128 g/person/day is assumed based on a review of clinical studies – the majority being UK-based (Rose *et al.*, 2015), and factor of 1.06 ml/g is used to convert from *g* to *ml* (Penn *et al.*, 2018).

The model is used to identify which variables result in a considerable variability in the output (P). Two methods were utilised in the sensitivity analysis; variance based sensitivity analysis (VBSA) and PAWN. The VBSA is a global sensitivity analysis where the variance of the output is decomposed into fractions attributed to the inputs. The PAWN approach considers the entire distribution of the

outcome (using the cumulative distribution function), which can be useful in cases where variance is not an adequate proxy of uncertainty.

b. Spatiotemporal MRP to obtain CIS prevalence estimates

Bayesian multilevel regression and poststratification is an increasingly used statistical technique to obtain representative estimates of prevalence or preferences at the national and smaller regional levels (Pouwels *et al.*, 2021). By using random effects in the multilevel model stable estimates can be obtained for subnational levels from relative small samples or relatively rare outcomes. However, if there is an spatial underlying structure this needs to be captured by the MPR methodology to avoid biased estimates based on a model that assumes independent group-level errors. Gao *et al.* (2021) recently proposed a spatial MRP using a Besag-York-Mollié (BYM2) specification for the regional effect.

Here we extended the spatial MRP approach proposed by Gao *et al.* to a spatio-temporal context by adding a temporal component to the model. For the temporal components we use autoregressive or random walk processes with discrete time indices (weeks) to capture likely temporal effects in the MRP model. The choice of the type of directed conditional distribution for the time effect (random walk or autoregressive) type of space-time interaction (type I-IV, Blangiardo & Cameletti 2015) and inclusion of additional covariates was guided by comparing the Watanabe-Akaike information criterion (WAIC) of the models. A type I space-time interaction, which assumes no spatial and/or temporal structure on the interaction, with first-order autoregressive terms were selected based on the WAIC (Watanabe 2010).

The following covariates and interactions were considered for the MRP: age (2-11, 12-16, 17-24, 25-34, 35-49, 50-69, 70+); sex; ethnicity (white / non-white), CIS area; region (9 regions in England); time and two-way interactions of age and time, ethnicity and time, area and time, region and ethnicity, and region and age. After running the spatiotemporal regression model, post-stratification was used to obtain representative estimates of the outcome prevalence in the target population.

Post-stratification tables were based on the conditional distribution of age and sex by area from ONS. The conditional distribution of ethnicity by these categories were obtained from the ETHPOP database (Wohland *et al.*, 2021). Using the population sizes of each poststratification cell of the target population, MRP adjusts for residual non-representative by post-stratifying by the percentage of each type in the actual overall population (Gao *et al.*, 2021). The outputs of these analyses consist of median estimates of percent positivity rates and associated 95% credible intervals available weekly at CIS subregional aggregation level between August 31st 2020 and February 14th 2021. For application in this study the estimates were up-sampled daily by linear interpolation between weekly estimates for each subregion and joined to the wastewater dataset using mappings from LTLA (Lower Tier Local Authority) to CIS subregions and from LTLA to catchment areas. Note that the resulting joined dataset was at CIS subregional level, where STW's contributions were weighted by their population covered in each overlapping subregion.

C. Modelling the population prevalence from WW and associated meta-data

The prevalence of SARS-CoV-2 at a sub-regional level was estimated using WW and associated meta-data. We implicitly assume that an estimate of prevalence from a specific site is a random sample of prevalence within the sub-region, where estimation of prevalence at the sub-region level is weighted by the size of the catchment population if >1 catchments are present in the sub-region (eg. smaller sites contribute less to the estimate of prevalence). A set of candidate statistical models were used to examine the relationship of RNA concentration to the median of the posterior estimate of the CIS positivity rate. These were: 1) linear regression, 2) linear regression with random effects intercept, 3) linear regression with random effects intercept and slope. Their out-of-sample predictive ability used to determine which model most effectively estimates prevalence.

i. Random effects models

The “random effects” model is a linear model where WW and the meta-data have random effects on the slope and the intercept. The Python *statsmodels* package was used to implement these models

(examples for this model [here](#)). For a Bayesian description of the model with another application case in WBE, see Hoffman *et al.*, (2021a). The linear regression and linear regression with random effects intercept had the poorest performance and are not described further.

ii. *Gradient Boosting Trees Regressor*

This model consists in a linear combination of non-linear predictors (also known as decision trees) trained by gradient descent (Friedman, 2001). Its performance has been shown on many regression examples (Chang *et al.*, 2018; Touzani *et al.*, 2018; Zhang & Haghani, 2015), and it is especially good at combining a large number or variety of input variables in a non-linear way. The implementation chosen here is an “Extreme Gradient Boosting” from python *XGBoost* package, which simply refers to an efficiently optimised Gradient Boosting Trees regressor using second order optimisers.

d. *Model Evaluation*

The models were compared using the mean absolute (MAE) from out-of-sample prediction (to limit over-fitting of the models). The out-of-sample prediction was carried out using bootstrapping: 50 random splits were generated in the available dataset, in each bootstrap 80% of the data were retained for training and the remaining 20% for testing. The MAE was generated for each dataset and the combined MAE were obtained by averaging the test results from all the bootstraps. If $\mathbf{y}_i$, ($i = 1, \dots, n$) are the testing data for bootstrap i , and $\hat{\mathbf{y}}_i$ the predictions from the model, the mean absolute error associated with the i^{th} bootstrap test set of size n is:

$$MAE_i = \frac{1}{n} \sum_{i=1}^n |10^{y_i} - 10^{\hat{y}_i}|$$

The estimated bootstrapped MAE, and associated standard deviation and 95% confidence intervals are then computed as:

$$MAE = \frac{1}{N_{bootstraps}} \sum_{i \in bootstraps} MAE_i$$

$$\sigma_{MAE} = \sqrt{\frac{\sum_{i \in bootstraps} (MAE_i - MAE)^2}{N_{bootstraps}}}$$

$$CI_{95\%} \approx \left[\frac{-2\sigma_{MAE}}{\sqrt{N_{bootstraps}}}, \frac{2\sigma_{MAE}}{\sqrt{N_{bootstraps}}} \right]$$

e. Model performance across STWs' characteristics

In October and November 2020, interviews were conducted with nine water utilities in England to document information on sewer network characteristics that could impact model performance (average daily flow, proportion of pumping in catchment, combined vs foul sewers etc.) The knowledge gathered from the interviews and a related questionnaire provided both qualitative and quantitative data, and transformed into variables used to assess model performance. An average MAE was obtained for each wastewater treatment plant by averaging the scores of CIS subregions included in the catchment weighted by the population covered. These resulting WTP MAE were then plotted across sites characteristics.

Acknowledgements

We are grateful to Professor Rowland Kao, and Dr Chris Jewell for helpful discussions in model development and interpretation, and for members of the “SPI-M” COVID-19 group for helpful feedback as the modelling developed. We are extremely grateful to representatives from the many water companies in England that provided insight into their sewage treatment works, and individuals in the labs responsible for processing of the samples.

Author Contributions

MM, ALJ, CS, MW, JG, KMO and LD conceptualised the analysis and were involved in the development of the approach. MM, AJL and CS analysed the data and generated the results. All authors contributed to writing the first version of the manuscript and have approved the manuscript

for peer-review. The views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the Department of Health and Social Care

Data availability

The full data that support the findings of this study are available on request from the Environmental Monitoring for Health Protection (EMHP) programme (through correspondence with JG). The raw data available on the following website: <https://www.gov.uk/government/publications/monitoring-of-sars-cov-2-rna-in-england-wastewater-monthly-statistics-june-2021/environmental-monitoring-for-health-protection-emhp-wastewater-monitoring-of-sars-cov-2-in-england-june-2021>

This study made use of [Numpy](#), [Pandas](#), [Sklearn](#), [Statsmodels](#) and [XGBoost](#) python open packages.

[[Code availability: the code will be made available open source there]]

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryInformationWWprevalencepaper.docx](#)