Science of the Total Environment 891 (2023) 164519

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Wastewater-based prediction of COVID-19 cases using a random forest algorithm with strain prevalence data: A case study of five municipalities in Latvia

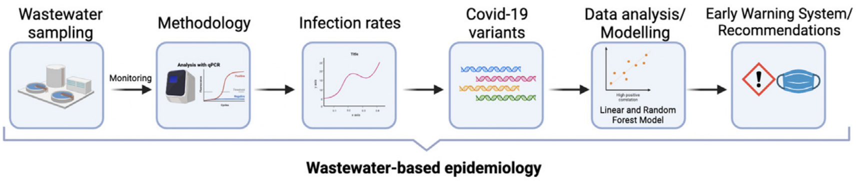
Brigita Dejus [a,*], Pāvels Cacivkins [b], Dita Gudra [c], Sandis Dejus [a], Maija Ustinova [c], Ance Roga [c], Martins Strods [a], Juris Kibilds [d], Guntis Boikmanis [d], Karina Ortlova [d], Laura Krivko [d], Liga Birzniece [c], Edmunds Skinderskis [c], Aivars Berzins [d], Davids Fridmanis [c], Talis Juhna [a]

[a] *Water Research and Environmental Biotechnology Laboratory, Riga Technical University, Kipsala 6A-263, Latvia*
[b] *Exponential Technologies Ltd, Dzerbenes 14, Riga, Latvia*
[c] *Latvian Biomedical Research and Study Centre, Rautsupites 1, Riga, Latvia*
[d] *Institute of Food Safety, Animal Health and Environment BIOR, Lejupes iela 3, Riga, Latvia*

HIGHLIGHTS

- The use of WBE for the SARS-CoV-2 virus and forecast cumulative COVID-19 cases two weeks in advance was investigated.
- RT-qPCR was used to detect the SARS-CoV-2 nucleocapsid 1 (N1), nucleocapsid 2 (N2), and E genes in municipal wastewater.
- The random forest model is more effective in predicting cumulative COVID-19 cases when strain prevalence data are included.
- Strain prevalence has a significant influence on the model predicting COVID-19 outbreaks.

GRAPHICAL ABSTRACT


Wastewater sampling
Methodology
Infection rates
Covid-19 variants
Data analysis/ Modelling
Early Warning System/ Recommendations
Monitoring
Analysis with qPCR
Title
High positive correlation
Linear and Random Forest Model
Wastewater-based epidemiology

ARTICLE INFO

Editor: Kyle Bibby

Keywords:
SARS-CoV-2
Wastewater-based epidemiology
Random forest model
Parameter importance

ABSTRACT

Wastewater-based epidemiology (WBE) is a rapid and cost-effective method that can detect SARS-CoV-2 genomic components in wastewater and can provide an early warning for possible COVID-19 outbreaks up to one or two weeks in advance. However, the quantitative relationship between the intensity of the epidemic and the possible progression of the pandemic is still unclear, necessitating further research. This study investigates the use of WBE to rapidly monitor the SARS-CoV-2 virus from five municipal wastewater treatment plants in Latvia and forecast cumulative COVID-19 cases two weeks in advance. For this purpose, a real-time quantitative PCR approach was used to monitor the SARS-CoV-2 nucleocapsid 1 (N1), nucleocapsid 2 (N2), and E genes in municipal wastewater. The RNA signals in the wastewater were compared to the reported COVID-19 cases, and the strain prevalence data of the SARS-CoV-2 virus were identified by targeted sequencing of receptor binding domain (RBD) and furin cleavage site (FCS) regions employing next-generation sequencing technology. The model methodology for a linear model and a random forest was designed and carried out to ascertain the correlation between the cumulative cases, strain prevalence data, and RNA concentration in the wastewater to predict the COVID-19 outbreak and its scale. Additionally, the factors that impact the model prediction accuracy for COVID-19 were investigated and compared between linear and random forest models. The results of cross-validated model metrics showed that the random forest model is more effective in predicting the

[*] Corresponding author at: Water Research and Environmental Biotechnology Laboratory, Riga Technical University, Kipsala 6A-263, Riga LV – 1048, Latvia.
E-mail address: brigita.dejus@rtu.lv (B. Dejus).

http://dx.doi.org/10.1016/j.scitotenv.2023.164519
Received 1 March 2023; Received in revised form 10 May 2023; Accepted 26 May 2023
Available online 31 May 2023

cumulative COVID-19 cases two weeks in advance when strain prevalence data are included. The results from this research help inform WBE and public health recommendations by providing valuable insights into the impact of environmental exposures on health outcomes.

1. Introduction

By 2023, the COVID-19 pandemic will have persisted for four years, and SARS-CoV-2 will continue to be associated with a significant number of deaths worldwide (Kumar et al., 2022; Westhaus et al., 2021). Consequently, there is an urgent requirement for a surveillance tool that can effectively detect and forecast COVID-19 outbreaks at global and national levels, with a rapid response time, broad coverage, noninvasive characteristics, low installation costs, and are anonymous (Zhu et al., 2021a). Previous research has indicated that SARS-CoV-2 can be detected in human feces and urine throughout the course of the infection, even in asymptomatic cases (Medema et al., 2020). A review by Jones et al. (2020) revealed that SARS-CoV-2 viral shedding can last and is detectable in wastewater samples from 14 to 28 days after infection. Additionally, Ali et al. (2021) and Yanaç et al. (2022) reported that infection with the SARS-CoV-2 virus is associated with persistent shedding of virus RNA in feces in 27 % to 89 % of patients, at densities ranging from 0.8 to 7.5 log10 gene copies per gram. These findings have provided a clear rationale for the use of wastewater-based monitoring to investigate COVID-19 spread in a particular region (Ali et al., 2021; Jones et al., 2020; Wölfel et al., 2020; Yanaç et al., 2022).

The application of the wastewater-based epidemiology (WBE) approach provides a noninvasive and almost instantaneous detection of the SARS-CoV-2 signal in wastewater (Choi et al., 2018). The process involves four essential steps: (1) sampling of wastewater from the selected environment; (2) analysis of wastewater for SARS-CoV-2 genomic components, which includes concentration measurements; (3) data analysis (e.g., normalization of the concentrations of SARS-CoV-2 genomic RNA copies by factors such as the sewage daily flow rates or population size to obtain the daily viral loads); and (4) tracking of the signal and prediction of outbreaks (e.g., tracing the sewer line exhibiting a positive signal back to the contributing communities) (Medema et al., 2020). In 2021, the European Commission published a recommendation for all European Union Member States, including Latvia, to establish wastewater monitoring to track the spread of COVID-19 and its causal agent's variants. All EU Member States have taken prompt action, and currently, approximately 1370 wastewater treatment plants across the EU are under regular surveillance, generating valuable data for WBE (European Commission, 2020; Gudra et al., 2022; Proverbio et al., 2022). To date, the WBE approach has been found to facilitate the predetection of SARS-CoV-2-infection outbreaks by one or two weeks at a city scale (Tiwari et al., 2021). Consequently, the WBE has been proposed as an effective method for epidemiologic monitoring and the creation of a reliable alert or early warning system in the detection of viral RNA content in wastewater samples (Cao and Francis, 2021; Xu et al., 2021). However, the quantitative relationship between the severity of the epidemic and the phase of the pandemic is still unclear (Cheval et al., 2020), which necessitates the use of modeling to resolve these issues and accurately interpret the collected data (Proverbio et al., 2022).

During the past few months, the success of vaccination campaigns has decreased the spread of SARS-CoV-2 infection cases in several countries; however, the risk of the virus or its newer version resurgence may have remained, and we may experience it when the precautionary measures and restrictions are eased (Voigt et al., 2022). Consequently, even if active RT–PCR or antigen testing is discontinued, WBE monitoring should still be conducted, and the observed data should be analyzed (Faria de Moura et al., 2021). Despite its potential, there are still experimental setup, data processing, and modeling procedure-related challenges that should be overcome before the WBE monitoring can be fully applied (Hart and Halden, 2020). For instance, relatively few studies have been conducted to investigate which observed data from wastewater monitoring have the most significant impact on modeling and predicting the COVID-19 outbreak in designing an early warning system (Bibby et al., 2021; le Rutte et al., 2022; Ramos et al., 2021; Zhang et al., 2022; Zhu et al., 2021b). Therefore, the primary objective of this study was to rapidly monitor the SARS-CoV-2 virus in five municipal wastewater treatment plants in Latvia through targeted surveillance via WBE and to determine if it was possible to forecast cumulative COVID-19 cases two weeks in advance using historical data on previous two-week cumulative cases, RNA concentration data, and strain prevalence data.

To achieve this objective, the main tasks of the study were to (i) employ quantitative reverse transcription PCR (RT–qPCR) methodology to monitor SARS-CoV-2 nucleocapsid 1 (N1), nucleocapsid 2 (N2), and E gene concentrations in five Latvian municipal wastewater treatment plants (WWTPs); (ii) compare SARS-CoV-2 RNA signals in municipal wastewater to reported cases; (iii) identify SARS-CoV-2 virus variants; (iv) determine a correlation between dominant virus variants and RNA concentrations in wastewater to predict COVID-19 outbreaks; and (v) investigate and compare the factors that impact the COVID-19 epidemiological model. To the best of our knowledge, this is the first study that employs a generalized cross-validation-based predictive modeling methodology on data from multiple municipalities in Latvia and identifies the most critical parameters for the creation of predictions. The authors of this study believe that the predominant strain variant of the SARS-CoV-2 virus might have a significantly important impact on the model's ability to predict COVID-19 outbreaks, suggesting that these factors should be considered when designing and implementing a national monitoring system for SARS-CoV-2.

2. Material and methods

2.1. Wastewater sampling and monitoring

From July 2021 to September 2022, five relatively different municipalities of Latvia were chosen for consistent monitoring of SARS-CoV-2 presence in wastewater: Liepaja (~74,000 connected inhabitants), Ventspils (~36,000 connected inhabitants), Jurmala (~43,000 connected inhabitants), Jelgava (~53,000 connected inhabitants), and Riga (~635,000 connected inhabitants). A portable autosampler - P6 MINI MAXX (MAXX Mess und Probenahmetechnik GmbH, Germany), was employed to collect 24-h time-dependent daily composite raw wastewater samples (7.2 L) at the inlet of the WWTP. The collected samples (2 L in PET bottle) were immediately stored at 4 °C, transported to the laboratory, and processed within 24 h. Samples from each municipality were collected once or twice per week. To estimate the total volume of wastewater treated at the WWTP during the sample collection time, appropriate water meter reading records were also collected. In total, 525 longitudinal wastewater samples were included in this study: 106 samples from Liepaja, 101 samples from Ventspils, 105 samples from Jurmala, 106 samples from Jelgava, and 107 samples from Riga.

2.2. Sample treatment and RNA extraction

Viral particles were concentrated from wastewater by PEG/NaCl precipitation, and RNA was subsequently extracted using the TRIzol method. In detail, a total of 135 mL of each sample was distributed equally among three 50 mL tubes and centrifuged for 15 min at 4800 ×*g* and 4 °C to pellet the larger particles. The supernatant was poured into a fresh 50 mL tube containing 0.9 g of NaCl (Carl Roth, Germany) and 3.2 g of polyethylene glycol 8000 (Carl Roth, Germany). The tubes were then incubated for

12 h at 4 °C with gentle agitation. Viral particles were deposited on the inside of the tubes by centrifugation for 30 min at 10000 × *g* and 4 °C. After removing the supernatant, the inner walls of the tubes were washed with 1 mL of TRI Reagent™ Solution (Invitrogen, Lithuania), and the rest of the RNA isolation was performed according to the manufacturer's protocol. As the last step, the RNA was dissolved in 70 μL of nuclease-free water (Sigma–Aldrich, United Kingdom).

2.3. PCR analysis

The RNA of SARS-CoV-2 was quantified by RT–qPCR amplification of the nucleocapsid (N) and envelope (E) genes from the viral genome. The E gene was targeted by the E_Sarbeco primers and E_Sarbeco_P1 probe (Corman et al., 2020), and the N gene was targeted by the SARS-CoV-2 N1 + N2 Assay Kit (Qiagen, Germany), which includes N1 and N2 primers and probes from the CDC design (Lu et al., 2020). The total volume of the RT–qPCR was 20 μL, and all measurements were performed in duplicate for each sample. The reaction mixture contained 5 μL of RNA, 10 μL of 2 × Reaction Mix from the SuperScript™ III One-Step RT–PCR System with Platinum™ Taq DNA Polymerase (Invitrogen, USA), 16 nmol of additional $MgSO_4$, 0.5 μL of SuperScript™ III RT/Platinum™ *Taq* Mix, and either of the two primer and probe sets. E_Sarbeco primers and probes were used at final concentrations of 0.4 nM and 0.2 nM, respectively. The N1 + N2 Assay Kit was used at a 20× dilution. The thermal cycling conditions for both E gene and N gene reactions were as described by Corman et al. (2020). Serial dilutions of the synthetic RNA SARS-CoV-2 Positive Run Control (Exact Diagnostics, USA) were used to produce a standard curve.

2.4. Library construction and NGS sequencing

2.4.1. Primer design

Primers were designed specifically for the receptor-binding domain (RBD) and furine cleavage site (FCS) of SARS-CoV-2 spike protein coding domains. For the amplification of the RBD region, the primer pair SCoV2-RBD-2-i5Fw and SCoV2-RBD-2-i7Rs was used, whereas for the amplification of the FCS region, SCoV2-FCS-i5Fw and SCoV2-FCS-i7Rs were used (Table 1).

Synthesis of cDNA and PCR amplification were carried out in one step using a qScript One-Step XLT RT–PCR Kit (Quantabio, USA). For the reaction, 0.5 μL of RNA sample was combined with 12.5 μL of 2x One-step PCR Tough Mix, 1 μL of 25x One-step qScript XLT RT, 1.2 μL of 10 μM primers, and 8.6 μL of nuclease-free water for a total volume of 25 μL. The PCR was performed according to the following thermal cycling parameters: 50 °C for 20 min, 94 °C for 3 min, and 40 cycles of 94 °C for 20 s, 55 °C for 1 min, and 72 °C for 1 min. The quantity and quality of the obtained cDNA were determined by electrophoresis in 1.2 % agarose gel and using the Qubit High Sensitivity dsDNA Assay kit on a Qubit Flex fluorometer (Thermo Fisher Scientific, USA). PCR products were purified using NucleoMag NGS Clean-Up and Size Select kit (Macherey-Nagel, Germany) magnetic beads at a ratio of 1:0.65.

2.4.2. Library construction

During the second PCR stage, Illumina MiSeq i7 and i5 indices were added to the RBD and FCS PCR products using custom-ordered Nextera XT Index (Illumina Inc., USA) primers (Metabion International AG, Germany). For this reaction, 5 ng of the cDNA amplification product was combined with 10 μL of Phusion U Multiplex PCR master mix (Thermo Fisher Scientific), indexed primers, and nuclease-free water for a total reaction volume of 20 μL. The PCR was performed according to the following thermal cycling parameters: 98 °C for 30 s, 35 cycles of 98 °C for 10 s, 67 °C for 15 s, and 72 °C for 15 s, followed by 7 min of 72 °C. PCR products were purified using NucleoMag NGS Clean-Up and Size Select kit magnetic beads at a ratio of 1:0.65. The resulting libraries were quantified using the Qubit High Sensitivity dsDNA assay kit on a Qubit Flex instrument, while the average size in base pairs was assessed using an Agilent High Sensitivity DNA kit on an Agilent 2100 Bioanalyzer (Agilent Technologies, USA).

Table 2
Marker mutations used to calculate the prevalence of SARS-CoV-2 variants in wastewater samples.

SARS-CoV-2 variant	Marker mutations
Delta (B.1.617.2)	L452R, T478K, P681R
Omicron (B.1.1.529)	K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, H655Y, N679K, P681H
Omicron subvariant BA.1	G446S, G496S
Omicron subvariant BA.2[a]	D405N, R408S
Omicron subvariant BA.4/5	L452R, F486V

[a] After the introduction of BA.4/5, the marker mutation for BA.2 was Q493R.

2.4.3. Amplicon sequencing by Illumina MiSeq

Prior to sequencing, all samples were pooled at equal molarities and diluted to 6 pM. They were then paired-end sequenced using a 500-cycle MiSeq Nano Reagent Kit v2 and an Illumina MiSeq instrument (Illumina Inc.). Each run was expected to produce at least 2000 reads per sample. After the sequencing run was completed, the individual sequence reads were filtered using MiSeq Reporter software to remove low-quality sequences.

2.4.4. Sequencing data analysis

Sequence reads were demultiplexed using Illumina's MiSeq Reporter Software and quality filtered using Trimmomatic v.0.39 with the leading and trailing quality of Q20, and the first 28 nt were trimmed as they matched the primer sequences (Bolger et al., 2014). Quality-filtered sequences were then aligned to the SARS-CoV-2 isolate Wuhan-Hu-1 reference sequence (NCBI Reference Sequence ID: NC_045512.2) using bowtie2 v.2.4.5. (Langmead and Salzberg, 2012). The resulting sequence alignment file was converted to a binary alignment (BAM) file, sorted and indexed using samtools v.1.14. (Danecek et al., 2021). Variant calling was performed using LoFreq v.2.1.2. (Wilm et al., 2012) - first indel qualities were inserted into the BAM file, then variants were called using NC_045512.2 reference by disabling all filters and setting minimum coverage to 1. Genetic variant annotations were assigned using the SnpEff toolbox with the NC_045512.2 reference (Cingolani et al., 2012). The resulting variant call format (VCF) files were filtered, and entries with a mapping quality (QUAL) <52 were discarded. Next, VCF files were used to calculate the prevalence of SARS-CoV-2 variants in wastewater based on Pango lineages (O'Toole et al., 2021). Thus, Pango-defined Variants of Concern (VOCs) were used to extract information on marker mutations from the spike regions (Table 2) and the prevalence of SARS-CoV-2 variants was calculated as the mean allelic frequencies of respective marker

Table 1
Summary of primer pairs.

Primer	Region	Sequence	Reference
SCoV2-RBD-2-i5Fw	RBD	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNTGTCTATGCAGATTCATTTK	This study
SCoV2-RBD-2-i7Rs	RBD	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNAGTAGACTTTTTAGGTCCACAA	This study
SCoV2-FCS-i5Fw	FCS	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGNNNNNNCTTCTAACCAGGTTGCTGTT	This study
SCoV2-FCS-i7Rs	FCS	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGNNNNNNGTACAAAAACTGCCATATTG	This study

mutations. Acquired data were visualized using the matplotlib library within the Python environment.

2.5. Data analysis and design of modeling methodology

A generalized cross-validation-based predictive modeling methodology was developed and applied to wastewater monitoring data from all five municipalities of Latvia. For a predictive model of COVID-19 outbreak prediction, observed datasets (normalized by wastewater flow, connected number of inhabitants at the WWTP and standardized to 100,000 inhabitants) from wastewater monitoring were used. In Fig. 1, the conceptual scheme of the modeling methodology for this study is presented.

Datasets from wastewater monitoring were analyzed based on the date of conducted measurements (Fig. 1 A). Additionally, the merged dataset was extended by adding measurements of cumulative COVID-19 cases per 100,000 inhabitants recorded during the previous two weeks acquired from the official statistics database published by the National Centre for Disease Prevention and Control as additional columns (Fig. 1 B). For each cross-validation cycle (total 100), the merged dataset was randomly split into 75 % training and 25 % testing data subsets (Fig. 1 C), and the selected regression modeling algorithm was trained on the training data subset and tested on the testing data subset (Fig. 1 D). Finally, the testing metrics (coefficient of determination R^2 and root mean square error RMSE) were stored for later processing, and several cross-validation cycles were created to approve whether the model could be generalized. When cross-validation cycles were made, the average R^2, RMSE, and their standard deviations were calculated to assess the cross-validated model quality (Fig. 1 E).

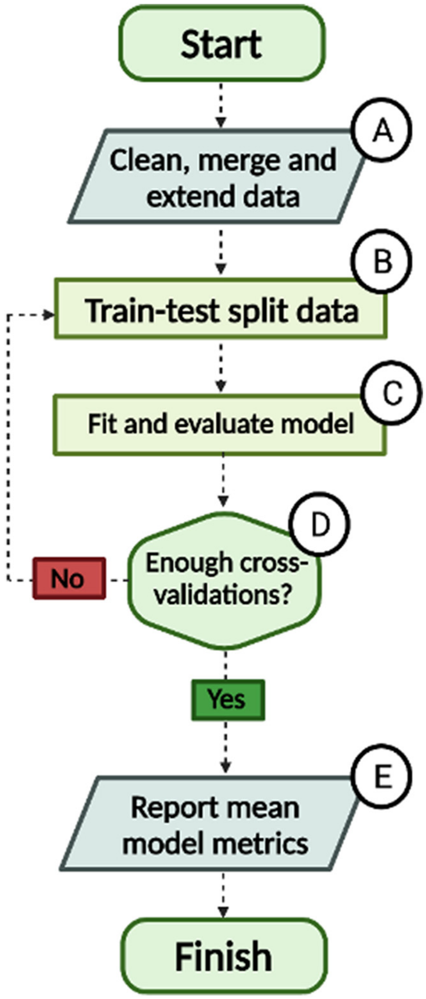

Start
A
Clean, merge and extend data
B
Train-test split data
C
Fit and evaluate model
D
No
Enough cross-validations?
Yes
E
Report mean model metrics
Finish

Fig. 1. Conceptual scheme and description of the modeling methodology (created by Biorender).

3. Results and discussion

3.1. Wastewater monitoring

Latvia was one of the countries that developed and implemented a new approach to WBE during the first stage of the COVID-19 pandemic to obtain additional information about the spread of this viral infectious disease already in 2020 (European Commission, 2020). Based on recommendations of the European Commission, Latvia expanded the previously developed approach as a national monitoring program for the detection of SARS-CoV-2 in wastewater with the purpose of monitoring spreading trends, predicting disease outbreaks, and monitoring the appearance of novel viral variants in a timely manner.

To gain a better understanding of the COVID-19 epidemiological situation in Latvia, we performed both viral RNA quantification and amplification based on viral spike protein gene RBD and FCS region next-generation sequencing. Fig. 2 presents the acquired wastewater monitoring results from five municipalities. Here, the average concentration measurements of SARS-CoV-2 viral RNA in Latvian wastewater samples are overlaid with the reported cumulative incidence of COVID-19 (Fig. 2A) as well as the average SARS-CoV-2 variant prevalence in the Latvian population (Fig. 2B).

The first COVID-19 outbreak was observed in October 2021, with the Delta variant being dominant, while the second outbreak occurred in March 2022, when the Omicron BA.1 and BA.2 variants emerged and outcompeted the former variant. The comparison of these two outbreaks revealed considerable differences between RNA concentrations, which was not detected in the 14-day cumulative incidence per 100,000 inhabitants, indicating that both variants might exhibit different viral shedding rates in feces and urine.

This is consistent with previous studies that have suggested that different variants of the virus can have different impacts on disease severity and transmission (Lin et al., 2021; Parra-Lucares et al., 2022). Moreover, the importance of surveillance strategies for COVID-19 increased in 2021 when the spread of Delta and Omicron variants disrupted Australia's successful public health response to the pandemic (Duckett, 2022). Generally, tracking the overall changes in nucleotide diversity, variant-specific reproduction numbers and emergence of novel mutation constellations in WW allow observation of evolutionary processes, potentially assisting understanding and anticipation of future shifts in circulating virus populations (Amman et al., 2022). For instance, previous studies have found that the Delta variant of SARS-CoV-2 was associated with higher transmission rates compared to other previous variants (Earnest et al., 2022; Meyerowitz and Richterman, 2022).

Overall, the results from monitoring strongly suggest that variants have different transmission and shedding rates, meaning that the effectiveness of containment measures may be different for each variant. In this regard, recent studies have shown that patients infected with the Omicron variant had lower viral shedding rates compared to patients infected with the Delta variant (Prasek et al., 2022, 2023; Puhach et al., 2023). Therefore, the determination of SARS-CoV-2 variants might play an important role in modeling the possible outbreak of the virus Therefore, the determination of SARS-CoV-2 variants might play an important role in modeling the possible outbreak of the virus (Barreiro et al., 2022; Schiøler et al., 2021). Furthermore, the use of data with virus variants might allow a better understanding of how the virus is evolving and how it is spreading, i.e., identifying variants of SARS-CoV-2 can help to predict potential mutations and prepare for their potential consequences (Otto et al., 2021). It can also help to inform decision-makers, as different variants may require different precautionary measures and medical treatments and vaccines (Corrao et al., 2022; Harvey et al., 2021). However, relatively few studies have focused on how to use SARS-CoV-2 variants for modeling the virus outbreak (Hinch et al., 2022; Obermeyer et al., 2022; Rui et al., 2022).

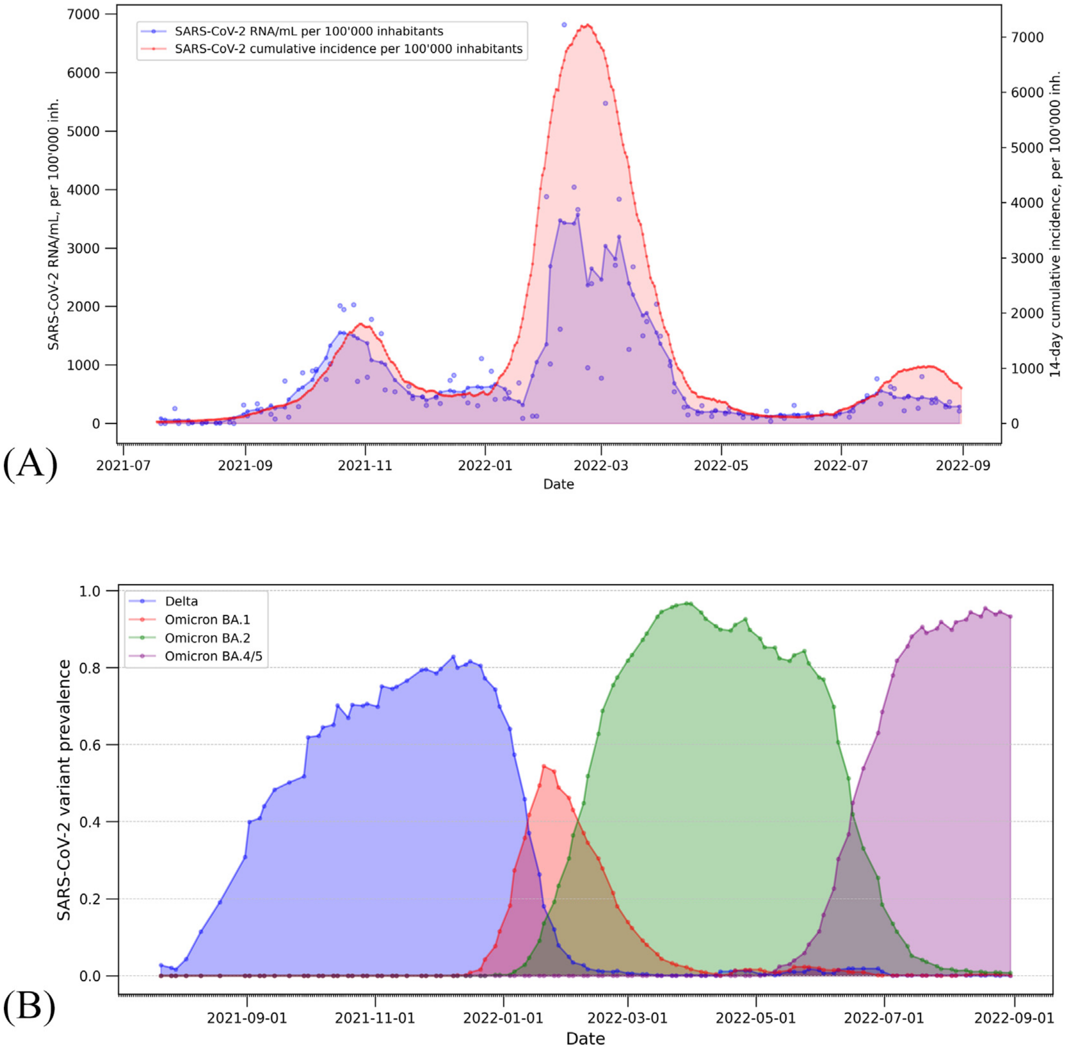


Fig. 2. The epidemiological situation of COVID-19 in Latvia according to wastewater monitoring from July 2021 to September 2022. Section A: Average concentration measurements of SARS-CoV-2 viral RNA in Latvian WWs and its overlap with the reported cumulative incidence of COVID-19. Section B: Average SARS-CoV-2 variant prevalence in the Latvian population as estimated by NGS sequencing of SARS-CoV-2 viral RNA FCS and RBD regions. The measurement values in both sections (A, B) are represented as a five-value centered moving average, excluding the cumulative index.

In this study, to identify the parameters from the national monitoring system that would best inform the model, the design of the model was developed and tested using the data from Liepaja and Jelgava as case studies. Liepaja was selected because it demonstrated a similar tendency in comparison to the overall epidemiological situation of COVID-19 in Latvia, while data from Jelgava were lacking in providing reliable data on average concentration measurements of SARS-CoV-2 viral RNA (Fig. 3). This might be explained by the fact that Jelgava is an industrial city, and municipal wastewater has a relatively high inflow of industrial wastewater compared to other cities in Latvia, which could potentially interfere with PCR, leading to underestimation of viral RNA concentration (Zhang et al., 2022).

The determination of whether it is feasible to create an accurate model with limited data on the average concentration of SARS-CoV-2 viral RNA will therefore also be investigated.

3.2. Design of the modeling methodology

One of the outlined tasks of this study was to investigate the feasibility of predicting cumulative COVID-19 cases two weeks into the future based on the previous two-week cumulative number of COVID-19 cases, RNA concentration data, and SARS-CoV-2 strain prevalence data. Therefore, the model generation strategy depicted in Fig. 4 was developed. In our view, it was optimal because it provided both flexibility and simplicity (Fig. 4) (Jakariya et al., 2022; Rallapalli et al., 2021).

Two regression modeling algorithms, a linear model and a random forest, were selected and compared in this study. Both models were chosen based on the results of similar investigations reported in previous literature, which demonstrated promising outcomes (Koureas et al., 2021; Speiser, 2021; Yuchi et al., 2019). By combining the results of multiple decision trees, random forest models can provide more accurate predictions than linear models (Shanmugasundar et al., 2021). However, the main benefit of using both models for this study is the possibility of gaining a better understanding of the data and making more informed decisions, i.e., each model can provide different insights into the data (Koureas et al., 2021).

A very simple and effective method to assess input parameter importance and test the above assumption is by adding input parameters one by one and testing if the cross-validated metrics such as R^2 score improve (Chicco et al., 2021). Therefore, for testing and validating the model generation strategy, the wastewater monitoring data from Liepāja and Jelgava were individually selected and tested.

3.2.1. Model test for Liepaja

The testing and validation of the modeling methodology to predict future cumulative COVID-19 cases by relying only on prior cumulative COVID-19 cases were performed for Liepaja city. After testing only prior cumulative

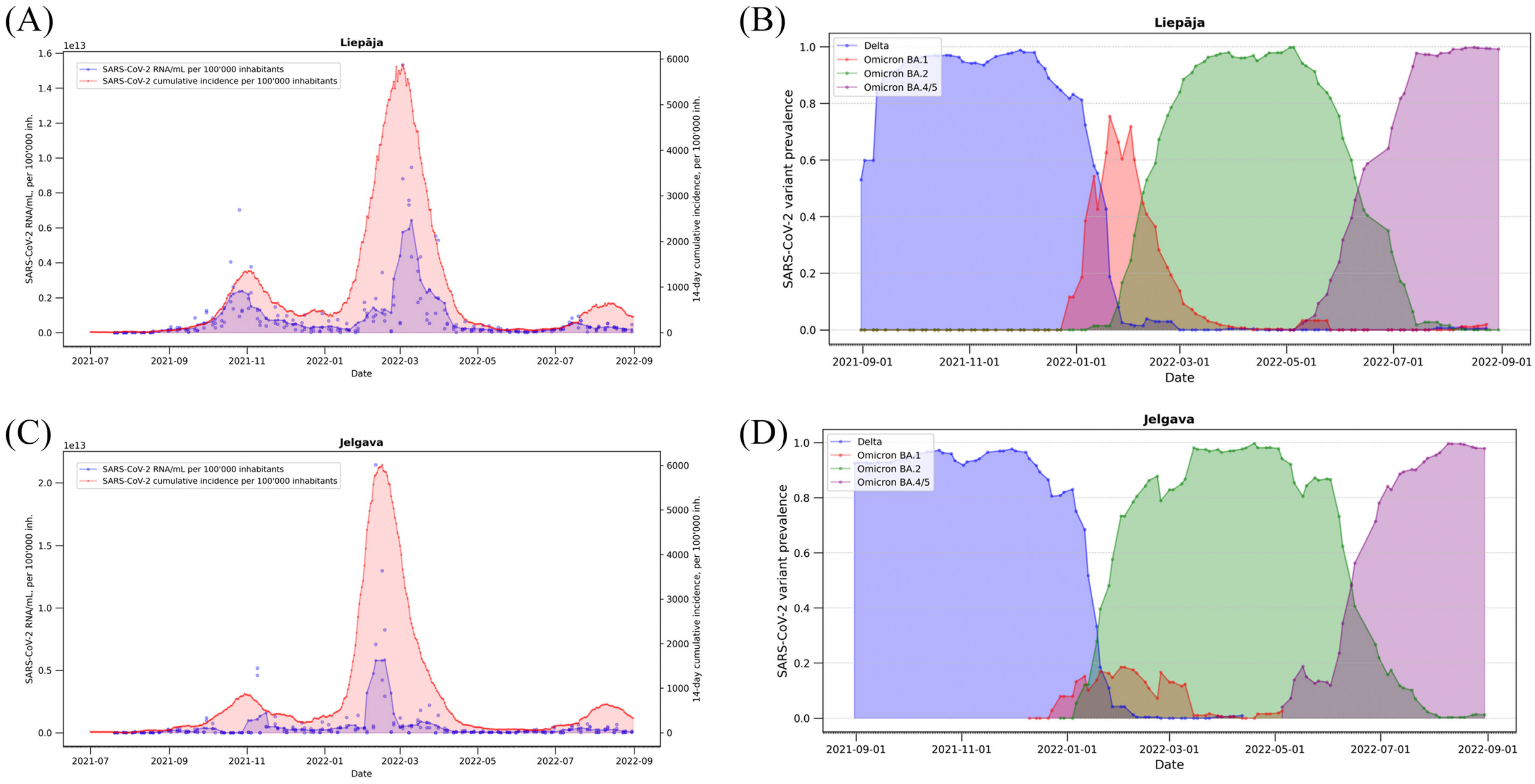


Fig. 3. Spread of detected SARS-CoV-2 lineages in Liepaja and Jelgava municipality wastewater for one year. The prevalence of SARS-CoV-2 lineages is shown as a rolling average of five measurements. Sections A and C: Average concentration measurements of SARS-CoV-2 viral RNA in Liepaja and Jelgava, respectively, WWs and their overlap with the reported cumulative incidence of COVID-19. Sections B and D: average SARS-CoV-2 variant prevalence in Liepaja and Jelgava, respectively, the population as estimated by NGS sequencing of SARS-CoV-2 viral RNA FCS and RBD regions.

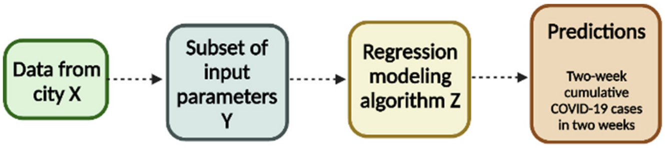

Data from city X
Subset of input parameters Y
Regression modeling algorithm Z
Predictions
Two-week cumulative COVID-19 cases in two weeks

Fig. 4. Simplified schematic diagram of interchangeable components within the developed model generation strategy (created by BioRender).

COVID-19 cases, the RNA concentration input parameter was added, and the test was repeated. Finally, the SARS-CoV-2 strain prevalence input parameter was included. The comparison of the test results is shown in Table 3.

The results demonstrated the improvement of cross-validated model metrics after the inclusion of SARS-CoV-2 strain prevalence data. The random forest model showed relatively better results (CV-R^2 = 0.80, CV-RMSE = 0.54) compared to a linear model (CV-R^2 = 0.75, CV-RMSE = 0.64).

In Fig. 5, the comparison of linear and random forest model cross-validated prediction is presented, where the random forest model demonstrated a relatively more stable prediction for predicted cumulative cases from September 2021 to July 2022 compared to the linear model.

3.2.2. Model test for Jelgava

The model test for Jelgava was performed using the same methodology as for Liepaja. After testing only prior cumulative COVID-19, the RNA concentration input parameter was added and tested. Finally, the SARS-CoV-2 strain prevalence input parameter was included. The comparison of the test results is shown in Table 4.

The results demonstrated that the random forest model showed relatively higher cross-validated scores than the linear model (CV-R^2 = 0.79 with CV-RMSE = 0.54 and CV-R^2 = 0.84 with CV-RMSE = 0.48, respectively). Moreover, as the data on RNA concentration were incomplete, the results showed that the best R^2 and RMSE scores were obtained with the random forest model when the cumulative cases, available data of RNA concentration, and strain prevalence were analyzed by the model compared to the analysis only with the cumulative cases and available data of RNA concentration (CV-R^2 = 0.78 with CV-RMSE = 0.46 and CV-R^2 = 0.84 with CV-RMSE = 0.48, respectively). In Fig. 6, the comparison of linear and random forest model cross-validated prediction is presented, where the random forest model demonstrated a relatively more stable prediction compared to the linear model, similar to Liepaja.

Finally, the model methodology was also applied and tested with the monitoring data from Riga, Jurmala, and Ventspils, where the results demonstrated similar tendencies (see Appendix A, B, and C). Therefore, it can be stated that this modeling methodology can be used to accurately predict COVID-19 outbreaks two weeks in advance for all five municipalities in this study.

Overall, the results from this study demonstrate the effectiveness of the random forest model for predicting cumulative COVID-19 cases. This is especially true when SARS-CoV-2 strain prevalence data are included in the model. These findings suggest that a random forest model is a valuable tool for predicting and managing the spread of COVID-19 in various municipalities. The results from this study are in line with other research on predictive modeling for COVID-19. For instance, a study by Ando et al. (2023) designed a mathematical model for predicting COVID-19 cases. The model was able to successfully predict the cumulative number of newly reported cases within a factor of 2 with a precision of 36–64 %, and the model without recent clinical data was able to successfully predict the number of cases for the following 5 days within a factor of 2 with a precision of 38–66 % (Ando et al., 2023). However, most of the studies do not use data from virus variants to predict the COVID-19 outbreak (Koureas et al., 2021; Speiser, 2021; Yuchi et al., 2019). Thus, Ali AlArjan et al. (2022) concluded that mathematical modeling accuracy might be improved by adding the data of virus variants (AlArjani et al., 2022). Considering this information, the following stage of our study involved the analysis of parameter importance to determine whether the strain prevalence has an influence on the model prediction.

3.3. Analysis of parameter importance

As the highest R^2 and RMSE scores were obtained with the random forest model when the cumulative cases, available data of RNA concentration, and strain prevalence were analyzed by the model, in a further study, this model was used to identify the importance of available input parameters and to test the assumption from this study that SARS-CoV-2 strain prevalence data can improve the quality of predictions. After data analysis, the random forest model reported the Gini ratio (parameter importance) for all used input parameters. The results are presented in Fig. 7.

It was demonstrated that the highest level of parameter importance was achieved by cumulative cases and strain prevalence data (Gini Ratio > 0.05), with only RNA concentration per day following. Consequently, these results have confirmed the hypothesis that strain prevalence has a relatively higher influence compared to other parameters on the model predicting COVID-19 outbreaks.

Finally, the cross-validated model was trained on the natural logarithm of cumulative COVID-19 cases for Latvia due to the distribution of true cumulative COVID-19 cases being positively skewed and thus challenging to model. Therefore, the model can also be used as a tool for an early warning system by predicting the risk level at which infection spreads rapidly (i.e., the logarithm) rather than the true number of cases. For instance, Simkovich et al. (2021) have designed a four-level framework for assessing the risk associated with research activities, enabling risk measures and definitions to be adjusted in response to new evidence that emerges regarding virus transmission.

On the one hand, the results lead to a relatively higher mean absolute error in predicted true case numbers, while the model predicts high natural logarithms of cases. However, the information on risk levels is still useful for decision-makers and epidemiologists since it can be used to predict the level of critical care capacity needed to respond to the outbreak, i.e., the higher the risk level, the more likely patients will require critical care, and therefore, more critical care capacity is needed.

Overall, the use of mathematical modeling with COVID-19 data can increase the knowledge of disease propagation by evaluating prevention measures as well as early and accurate detection of the disease in patients (Mohamadou et al., 2020). In the writing of this paper, the majority of articles found regarding mathematical modeling were related to COVID-19 dynamics (AlArjani et al., 2022; Ramos et al., 2021; Vallejo et al., 2022; Yanaç et al., 2022; Zhu et al., 2022). Modeling can be performed using appropriate datasets to explore the effects of variables such as climate and preventive measures on the spread of COVID-19, as outlined previously. Thus, simulation of the next waves of COVID-19 outbreaks will also be beneficial to enhance surveillance. As countries start to relax social

Table 3
Summary of test results for Liepaja city for 2 different regression modeling algorithms and 3 different subsets of input parameters used for predicting two-week cumulative COVID-19 cases two weeks into the future.

Model	Prior cumulative cases	Prior cumulative cases RNA concentration	Prior cumulative cases RNA concentration Strain prevalence
Linear model	CV-R^2 = 0.65; std = 0.15 CV-RMSE = 0.75; std = 0.13	CV-R^2 = 0.65; std = 0.12 CV-RMSE = 0.63; std = 0.08	CV-R^2 = 0.75; std = 0.10 CV-RMSE = 0.64; std = 0.09
Random forest model	CV-R^2 = 0.70; std = 0.15 CV-RMSE = 0.66; std = 0.12	CV-R^2 = 0.70; std = 0.12 CV-RMSE = 0.51; std = 0.10	CV-R^2 = 0.80; std = 0.10 CV-RMSE = 0.54; std = 0.12

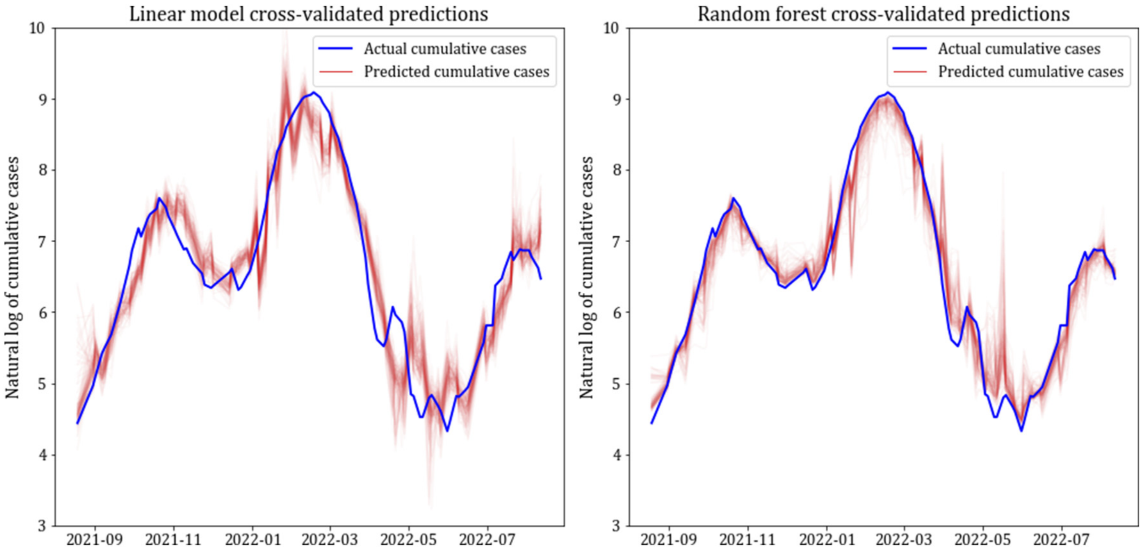


Fig. 5. Comparison of linear and random forest models for Liepaja.

Table 4
Summary of test results for Jelgava city for 2 different regression modeling algorithms and 3 different subsets of input parameters used for predicting two-week cumulative COVID-19 cases two weeks into the future.

Model	Prior cumulative cases	Prior cumulative cases RNA concentration	Prior cumulative cases RNA concentration Strain prevalence
Linear model	CV-R^2 = 0.81; std. = 0.08 CV-RMSE = 0.52; std. = 0.09	CV-R^2 = 0.82; std. = 0.07 CV-RMSE = 0.54; std. = 0.11	CV-R^2 = 0.79; std. = 0.09 CV-RMSE = 0.54; std. = 0.12
Random forest model	CV-R^2 = 0.81; std. = 0.07 CV-RMSE = 0.55; std. = 0.11	CV-R^2 = 0.78; std. = 0.09 CV-RMSE = 0.46; std. = 0.12	CV-R^2 = 0.84; std. = 0.08 CV-RMSE = 0.48; std. = 0.11

restriction measures, it is necessary to conduct a study to estimate potential hotspots for new outbreaks.

4. Conclusions

The presence of different variants of the SARS-CoV-2 virus can have different impacts on the severity and transmission of COVID-19, and this was demonstrated in the data from monitoring the Delta and Omicron variants. Linear models and random forest models can both provide different insights into the data. However, the results of cross-validated model metrics from Liepaja showed that the random forest model was more effective (CV-R^2 = 0.80 and CV-RMSE = 0.54) than the linear model (CV-R^2 = 0.75 and CV-RMSE = 0.64) in predicting the cumulative COVID-19 cases 14 days in advance when strain prevalence data were included. Finally, the results of the risk level can be used to predict the level of critical care capacity needed to respond to the COVID-19 outbreak. Furthermore, the study demonstrated that the highest level of parameter importance was achieved by cumulative cases and strain prevalence data. Overall, this

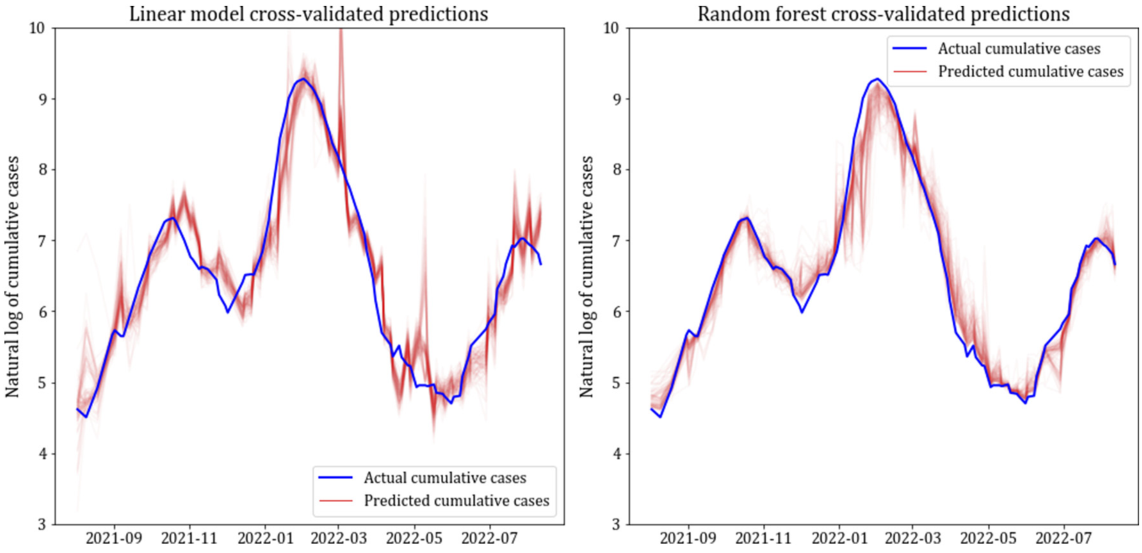


Fig. 6. Comparison of linear and random forest models for Jelgava.

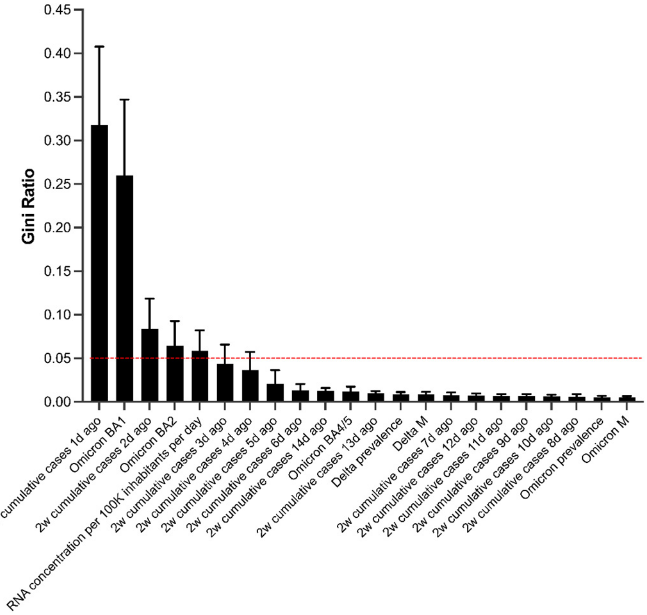


Fig. 7. The random forest model reported the Gini ratio (parameter importance) for all input parameters used.

study has provided valuable insights into the parameters that are important for accurate COVID-19 prediction and can improve preparedness for future outbreaks. Thus, further research is necessary to continue the verification of the modeling methodology with the new variants of COVID-19 and their changing prevalence. Additionally, it is conceivable that this approach could be utilized for WBE's purpose of modeling and constructing early warning systems.

Funding

This work was supported by the project “Platform for the Covid-19 safe work environment” (ID 1.1.1/21/A/011). The project “Platform for the Covid-19 safe work environment” is founded by European Regional Development Fund specific objective 1.1.1 “Improve research and innovation capacity and the ability of Latvian research institutions to attract external funding, by investing in human capital and infrastructure” and co-financed by REACT-EU funding for mitigating the consequences of the pandemic crisis.

Data availability

No data was used for the research described in the article.

Declaration of competing interest

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

Acknowledgment

In addition, we thank the Latvian Water and Wastewater Works Association and employees of the wastewater treatment plant for supporting the sampling and access to wastewater plants.

References

AlArjani, A., Nasseef, M.T., Kamal, S.M., Rao, B.V.S., Mahmud, M., Uddin, M.S., 2022. Application of mathematical modeling in prediction of COVID-19 transmission dynamics. Arab. J. Sci. Eng. 47 (8), 10163–10186. https://doi.org/10.1007/s13369-021-06419-4.

Ali, W., Zhang, H., Wang, Z., Chang, C., Javed, A., Ali, K., Du, W., Niazi, N.K., Mao, K., Yang, Z., 2021. Occurrence of various viruses and recent evidence of SARS-CoV-2 in wastewater systems. J. Hazard. Mater. 414 (February), 125439. https://doi.org/10.1016/j.jhazmat.2021.125439.

Amman, F., Markt, R., Endler, L., Hupfauf, S., Agerer, B., Schedl, A., Richter, L., Zechmeister, M., Bicher, M., Heiler, G., Triska, P., Thornton, M., Penz, T., Senekowitsch, M., Laine, J., Keszei, Z., Klimek, P., Nägele, F., Mayr, M., ... Bergthaler, A., 2022. Viral variant-resolved wastewater surveillance of SARS-CoV-2 at national scale. Nat. Biotechnol. https://doi.org/10.1038/s41587-022-01387-y.

Ando, H., Murakami, M., Ahmed, W., Iwamoto, R., Okabe, S., Kitajima, M., 2023. Wastewater-based prediction of COVID-19 cases using a highly sensitive SARS-CoV-2 RNA detection method combined with mathematical modeling. Environ. Int. 107743. https://doi.org/10.1016/j.envint.2023.107743.

Barreiro, N.L., Govezensky, T., Ventura, C.I., Núñez, M., Bolcatto, P.G., Barrio, R.A., 2022. Modelling the interplay of SARS-CoV-2 variants in the United Kingdom. Sci. Rep. 12 (1). https://doi.org/10.1038/s41598-022-16147-w.

Bibby, K., Bivins, A., Wu, Z., North, D., 2021. Making waves: Plausible lead time for wastewater based epidemiology as an early warning system for COVID-19. Water Research. vol. 202. Elsevier Ltd.. https://doi.org/10.1016/j.watres.2021.117438.

Cao, Y., Francis, R., 2021. On forecasting the community-level COVID-19 cases from the concentration of SARS-CoV-2 in wastewater. Sci. Total Environ. 786. https://doi.org/10.1016/j.scitotenv.2021.147451.

Cheval, S., Adamescu, C.M., Georgiadis, T., Herrnegger, M., Piticar, A., Legates, D.R., 2020. Observed and potential impacts of the covid-19 pandemic on the environment. International Journal of Environmental Research and Public Health (Vol. 17, Issue 11, pp. 1–25). MDPI AG https://doi.org/10.3390/ijerph17114140.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, 1–24. https://doi.org/10.7717/PEERJ-CS.623.

Choi, P.M., Tscharke, 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., Mueller, J.F., 2018. Wastewater-based epidemiology biomarkers: past, present and future. TrAC Trends Anal. Chem. 105, 453–469. https://doi.org/10.1016/j.trac.2018.06.004.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K.W., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M.L., Mulders, D.G.J.C., Haagmans, B.L., van der Veer, B., van den Brink, S., Wijsman, L., Goderski, G., Romette, J.L., Ellis, J., Zambon, M., ... Drosten, C., 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance 25 (3). https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045.

Corrao, G., Franchi, M., Rea, F., Cereda, D., Barone, A., Borriello, C.R., della Valle, P.G., Ercolanoni, M., Fortino, I., Jara, J., Leoni, O., Mazziotta, F., Pierini, E., Preziosi, G., Tirani, M., Galli, M., Bertolaso, G., Pavesi, G., Bortolan, F., 2022. Protective action of natural and induced immunization against the occurrence of delta or alpha variants of SARS-CoV-2 infection: a test-negative case-control study. BMC Med. 20 (1). https://doi.org/10.1186/s12916-022-02262-y.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10 (2). https://doi.org/10.1093/gigascience/giab008.

Duckett, S., 2022. Public health management of the COVID-19 pandemic in Australia: the role of the Morrison government. Int. J. Environ. Res. Public Health 19 (16). https://doi.org/10.3390/ijerph191610400.

Earnest, R., Uddin, R., Matluk, N., Renzette, N., Turbett, S.E., Siddle, K.J., Loreth, C., Adams, G., Tomkins-Tinch, C.H., Petrone, M.E., Rothman, J.E., Breban, M.I., Koch, R.T., Billig, K., Fauver, J.R., Vogels, C.B.F., Bilguvar, K., de Kumar, B., Landry, M.L., ... Grubaugh, N.D., 2022. Comparative transmissibility of SARS-CoV-2 variants Delta and alpha in New England, USA. Cell Rep. Med. 3 (4). https://doi.org/10.1016/j.xcrm.2022.100583.

European Commision, 2020. COMMISSION RECOMMENDATION of 17.3.2021 on a common approach to establish a systematic surveillance of SARS-CoV-2 and its variants in wastewaters in the EU. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:12012E/TXT:en:PDF.

Gudra, D., Dejus, S., Bartkevics, V., Roga, A., Kalnina, I., Strods, M., Rayan, A., Kokina, K., Zajakina, A., Dumpis, U., Ikkere, L.E., Arhipova, I., Berzins, G., Erglis, A., Binde, J., Ansonska, E., Berzins, A., Juhna, T., Fridmanis, D., 2022. Detection of SARS-CoV-2 RNA in wastewater and importance of population size assessment in smaller cities: an exploratory case study from two municipalities in Latvia. Sci. Total Environ. 823, 153775. https://doi.org/10.1016/j.scitotenv.2022.153775.

Hart, O.E., Halden, R.U., 2020. Computational analysis of SARS-CoV-2/COVID-19 surveillance by wastewater-based epidemiology locally and globally: feasibility, economy, opportunities and challenges. Sci. Total Environ. 730. https://doi.org/10.1016/j.scitotenv.2020.138875.

Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J., Robertson, D.L., 2021. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology (Vol. 19, Issue 7, pp. 409–424). Nature Research https://doi.org/10.1038/s41579-021-00573-0.

Hinch, R., Panovska-Griffiths, J., Probert, W.J.M., Ferretti, L., Wymant, C., di Lauro, F., Baya, N., Ghafari, M., Abeler-Dörner, L., Fraser, C., 2022. Estimating SARS-CoV-2 variant fitness and the impact of interventions in England using statistical and geo-spatial agent-based models. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 380 (2233). https://doi.org/10.1098/rsta.2021.0304.

Jakariya, M., Ahmed, F., Islam, M.A., al Marzan, A., Hasan, M.N., Hossain, M., Ahmed, T., Hossain, A., Reza, H.M., Hossen, F., Nahla, T., Rahman, M.M., Bahadur, N.M., Islam, M.T., Didar-ul-Alam, M., Mow, N., Jahan, H., Barceló, D., Bibby, K., Bhattacharya, P., 2022. Wastewater-based epidemiological surveillance to monitor the prevalence of SARS-CoV-2 in developing countries with onsite sanitation facilities. Environ. Pollut. 311. https://doi.org/10.1016/j.envpol.2022.119679.

Jones, D.L., Baluja, M.Q., Graham, D.W., Corbishley, A., McDonald, J.E., Malham, S.K., Hillary, L.S., Connor, T.R., Gaze, W.H., Moura, I.B., Wilcox, M.H., Farkas, K., 2020. Shedding of SARS-CoV-2 in feces and urine and its potential role in person-to-person transmission and the environment-based spread of COVID-19. Sci. Total Environ. 749. https://doi.org/10.1016/j.scitotenv.2020.141364.

Koureas, M., Amoutzias, G.D., Vontas, A., Kyritsi, M., Pinaka, O., Papakonstantinou, A., Dadouli, K., Hatzinikou, M., Koutsolioutsou, A., Mouchtouri, V.A., Speletas, M., Tsiodras, S., Hadjichristodoulou, C., 2021. Wastewater monitoring as a supplementary surveillance tool for capturing SARS-COV-2 community spread. A case study in two Greek municipalities. Environmental Research. vol. 200. Academic Press Inc. https://doi.org/10.1016/j.envres.2021.111749.

Kumar, M., Jiang, G., Kumar Thakur, A., Chatterjee, S., Bhattacharya, T., Mohapatra, S., Chaminda, T., Kumar Tyagi, V., Vithanage, M., Bhattacharya, P., Nghiem, L.D., Sarkar, D., Sonne, C., Mahlknecht, J., 2022. Lead time of early warning by wastewater surveillance for COVID-19: geographical variations and impacting factors. Chemical Engineering Journal. vol. 441. Elsevier B.V. https://doi.org/10.1016/j.cej.2022.135936.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (4), 357–359. https://doi.org/10.1038/nmeth.1923.

Lin, L., Liu, Y., Tang, X., He, D., 2021. The disease severity and clinical outcomes of the SARS-CoV-2 variants of concern. Frontiers in Public Health. vol. 9. Frontiers Media S.A. https://doi.org/10.3389/fpubh.2021.775224.

Lu, X., Wang, L., Sakthivel, S.K., Whitaker, B., Murray, J., Kamili, S., Lynch, B., Malapati, L., Burke, S.A., Harcourt, J., Tamin, A., Thornburg, N.J., Villanueva, J.M., Lindstrom, S., 2020. US CDC real-time reverse transcription PCR panel for detection of severe acute respiratory syndrome Coronavirus 2. Emerg. Infect. Dis. 26 (8), 1654–1665. https://doi.org/10.3201/eid2608.201246.

Medema, G., Been, F., Heijnen, L., Petterson, S., 2020. Implementation of environmental surveillance for SARS-CoV-2 virus to support public health decisions: opportunities and challenges. Curr. Opin. Environ. Sci. Health 17, 49–71. https://doi.org/10.1016/j.coesh.2020.09.006.

Meyerowitz, E.A., Richterman, A., 2022. SARS-CoV-2 transmission and prevention in the era of the Delta variant. Infectious Disease Clinics of North America (Vol. 36, Issue 2, pp. 267–293). W.B. Saunders https://doi.org/10.1016/j.idc.2022.01.007.

Mohamadou, Y., Halidou, A., Kapen, P.T., 2020. A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19. Appl. Intell. 50 (11), 3913–3925. https://doi.org/10.1007/s10489-020-01770-9.

de Moura, Faria, Villela, E., López, R.V.M., Sato, A.P.S., de Oliveira, F.M., Waldman, E.A., van den Bergh, R., Siewe Fodjo, J.N., Colebunders, R., 2021. COVID-19 outbreak in Brazil: adherence to national preventive measures and impact on people's lives, an online survey. BMC Public Health 21 (1). https://doi.org/10.1186/s12889-021-10222-z.

Obermeyer, F., Jankowiak, M., Barkas, N., Schaffner, S. F., Pyle, J. D., Yurkovetskiy, L., Bosso, M., Park, D. J., Babadi, M., Macinnis, B. L., Luban, J., Sabeti, P. C., & Lemieux, J. E. (n.d.). Analysis of 6.4 Million SARS-CoV-2 Genomes Identifies Mutations Associated With Fitness.

Otto, S.P., Day, T., Arino, J., Colijn, C., Dushoff, J., Li, M., Mechai, S., van Domselaar, G., Wu, J., Earn, D.J.D., Ogden, N.H., 2021. The origins and potential future of SARS-CoV-2 variants of concern in the evolving COVID-19 pandemic. Current Biology (Vol. 31, Issue 14, pp. R918–R929). Cell Press https://doi.org/10.1016/j.cub.2021.06.049.

Parra-Lucares, A., Segura, P., Rojas, V., Pumarino, C., Saint-Pierre, G., Toro, L., 2022. Emergence of SARS-CoV-2 variants in the world: how could this happen? Life (Vol. 12, Issue 2). MDPI https://doi.org/10.3390/life12020194

Prasek, S.M., Pepper, I.L., Innes, G.K., Slinski, S., Ruedas, M., Sanchez, A., Brierley, P., Betancourt, W.Q., Stark, E.R., Foster, A.R., Betts-Childress, N.D., Schmitz, B.W., 2022. Population level SARS-CoV-2 fecal shedding rates determined via wastewater-based epidemiology. Sci. Total Environ. 838. https://doi.org/10.1016/j.scitotenv.2022.156535.

Prasek, S.M., Pepper, I.L., Innes, G.K., Slinski, S., Betancourt, W.Q., Foster, A.R., Yaglom, H.D., Porter, W.T., Engelthaler, D.M., Schmitz, B.W., 2023. Variant-specific SARS-CoV-2 shedding rates in wastewater. Sci. Total Environ. 857. https://doi.org/10.1016/j.scitotenv.2022.159165.

Proverbio, D., Kemp, F., Magni, S., Ogorzaly, L., Cauchie, H.M., Gonçalves, J., Skupin, A., Aalto, A., 2022. Model-based assessment of COVID-19 epidemic dynamics by wastewater analysis. Sci. Total Environ. 827. https://doi.org/10.1016/j.scitotenv.2022.154235.

Puhach, O., Meyer, B., Eckerle, I., 2023. SARS-CoV-2 viral load and shedding kinetics. Nature Reviews Microbiology (Vol. 21, Issue 3, pp. 147–161). Nature Research https://doi.org/10.1038/s41579-022-00822-w.

Rallapalli, S., Aggarwal, S., Singh, A.P., 2021. Detecting SARS-CoV-2 RNA prone clusters in a municipal wastewater network using fuzzy-Bayesian optimization model to facilitate wastewater-based epidemiology. Sci. Total Environ. 778. https://doi.org/10.1016/j.scitotenv.2021.146294.

Ramos, A.M., Vela-Pérez, M., Ferrández, M.R., Kubik, A.B., Ivorra, B., 2021. Modeling the impact of SARS-CoV-2 variants and vaccines on the spread of COVID-19. Commun. Nonlinear Sci. Numer. Simul. 102. https://doi.org/10.1016/j.cnsns.2021.105937.

Rui, J., Zheng, J.X., Chen, J., Wei, H., Yu, S., Zhao, Z., Wang, X.Y., Chen, M.X., Xia, S., Zhou, Y., Chen, T., Zhou, X.N., 2022. Optimal control strategies of SARS-CoV-2 omicron supported by invasive and dynamic models. Infect. Dis. Poverty 11 (1). https://doi.org/10.1186/s40249-022-01039-y.

le Rutte, E.A., Shattock, A.J., Chitnis, N., Kelly, S.L., Penny, M.A., 2022. Modelling the impact of omicron and emerging variants on SARS-CoV-2 transmission and public health burden. Commun. Med. 2 (1). https://doi.org/10.1038/s43856-022-00154-z.

Schiøler, H., Knudsen, T., Brøndum, R.F., Stoustrup, J., Bøgsted, M., 2021. Mathematical modelling of SARS-CoV-2 variant outbreaks reveals their probability of extinction. Sci. Rep. 11 (1). https://doi.org/10.1038/s41598-021-04108-8.

Shanmugasundar, G., Vanitha, M., Čep, R., Kumar, V., Kalita, K., Ramachandran, M., 2021. A comparative study of linear, random forest and adaboost regressions for modeling non-traditional machining. Processes 9 (11). https://doi.org/10.3390/pr9112015.

Simkovich, S.M., Thompson, L.M., Clark, M.L., Balakrishnan, K., Bussalleu, A., Checkley, W., Clasen, T., Davila-Roman, V.G., Diaz-Artiga, A., Dusabimana, E., Fuentes, L. de las, Harvey, S., Kirby, M.A., Lovvorn, A., McCollum, E.D., Mollinedo, E.E., Peel, J.L., Quinn, A., Rosa, G., ... Ye, W., 2021. A risk assessment tool for resumption of research activities during the COVID-19 pandemic for field trials in low resource settings. BMC Med. Res. Methodol. 21 (1). https://doi.org/10.1186/s12874-021-01232-x.

Speiser, J.L., 2021. A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data. J. Biomed. Inform. 117. https://doi.org/10.1016/j.jbi.2021.103763.

Tiwari, S.B., Gahlot, P., Tyagi, V.K., Zhang, L., Zhou, Y., Kazmi, A.A., Kumar, M., 2021. Surveillance of Wastewater for Early Epidemic Prediction (SWEEP): environmental and health security perspectives in the post COVID-19 Anthropocene. Environ. Res. 195. https://doi.org/10.1016/j.envres.2021.110831.

Vallejo, J.A., Trigo-Tasende, N., Rumbo-Feal, S., Conde-Pérez, K., López-Oriona, Á., Barbeito, I., Vaamonde, M., Tarrío-Saavedra, J., Reif, R., Ladra, S., Rodiño-Janeiro, B.K., Nasser-Ali, M., Cid, Á., Veiga, M., Acevedo, A., Lamora, C., Bou, G., Cao, R., Poza, M., 2022. Modeling the number of people infected with SARS-COV-2 from wastewater viral load in Northwest Spain. Sci. Total Environ. 811. https://doi.org/10.1016/j.scitotenv.2021.152334.

Voigt, A., Omholt, S., Almaas, E., 2022. Comparing the impact of vaccination strategies on the spread of COVID-19, including a novel household-targeted vaccination strategy. PLoS One 17 (2 February). https://doi.org/10.1371/journal.pone.0263155.

Westhaus, S., Weber, F.A., Schiwy, S., Linnemann, V., Brinkmann, M., Widera, M., Greve, C., Janke, A., Hollert, H., Wintgens, T., Ciesek, S., 2021. Detection of SARS-CoV-2 in raw and treated wastewater in Germany – suitability for COVID-19 surveillance and potential transmission risks. Sci. Total Environ. 751, 141750. https://doi.org/10.1016/j.scitotenv.2020.141750.

Wilm, A., Aw, P.P.K., Bertrand, D., Yeo, G.H.T., Ong, S.H., Wong, C.H., Khor, C.C., Petric, R., Hibberd, M.L., Nagarajan, N., 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 40 (22), 11189–11201. https://doi.org/10.1093/nar/gks918.

Wölfel, R., Corman, V.M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M.A., Niemeyer, D., Jones, T.C., Vollmar, P., Rothe, C., Hoelscher, M., Bleicker, T., Brünink, S., Schneider, J., Ehmann, R., Zwirglmaier, K., Drosten, C., Wendtner, C., 2020. Virological assessment of hospitalized patients with COVID-2019. Nature 581 (7809), 465–469. https://doi.org/10.1038/s41586-020-2196-x.

Xu, X., Zheng, X., Li, S., Lam, N.S., Wang, Y., Chu, D.K.W., Poon, L.L.M., Tun, H.M., Peiris, M., Deng, Y., Leung, G.M., Zhang, T., 2021. The first case study of wastewater-based epidemiology of COVID-19 in Hong Kong. Sci. Total Environ. 790, 148000. https://doi.org/10.1016/j.scitotenv.2021.148000.

Yanaç, K., Adegoke, A., Wang, L., Uyaguari, M., Yuan, Q., 2022. Detection of SARS-CoV-2 RNA throughout wastewater treatment plants and a modeling approach to understand COVID-19 infection dynamics in Winnipeg, Canada. Sci. Total Environ. 825. https://doi.org/10.1016/j.scitotenv.2022.153906.

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B., Naidan, G., Ochir, C., Legtseg, B., Byambaa, T., Barn, P., Henderson, S.B., Janes, C.R., Lanphear, B.P., McCandless, L.C., Takaro, T.K., Venners, S.A., Webster, G.M., Allen, R.W., 2019. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environ. Pollut. 245, 746–753. https://doi.org/10.1016/j.envpol.2018.11.034.

Zhang, D., Duran, S.S.F., Lim, W.Y.S., Tan, C.K.I., Cheong, W.C.D., Suwardi, A., Loh, X.J., 2022. SARS-CoV-2 in wastewater: From detection to evaluation. Materials Today Advances. vol. 13. Elsevier Ltd.. https://doi.org/10.1016/j.mtadv.2022.100211.

Zhu, Y., Oishi, W., Maruo, C., Saito, M., Chen, R., Kitajima, M., Sano, D., 2021a. Early warning of COVID-19 via wastewater-based epidemiology: potential and bottlenecks. Science of the Total Environment. vol. 767. Elsevier B.V. https://doi.org/10.1016/j.scitotenv.2021.145124.

Zhu, Y., Oishi, W., Maruo, C., Saito, M., Chen, R., Kitajima, M., Sano, D., 2021b. Early warning of COVID-19 via wastewater-based epidemiology: potential and bottlenecks. Science of the Total Environment. vol. 767. Elsevier B.V. https://doi.org/10.1016/j.scitotenv.2021.145124.

Zhu, Y., Oishi, W., Maruo, C., Bandara, S., Lin, M., Saito, M., Kitajima, M., Sano, D., 2022. COVID-19 case prediction via wastewater surveillance in a low-prevalence urban community: a modeling approach. J. Water Health 20 (2), 459–470. https://doi.org/10.2166/WH.2022.183.