Machine Learning for Detecting Virus Infection Hotspots Via Wastewater-Based Epidemiology

The Case of SARS-CoV-2 RNA

Zehnder, Calvin; Béen, Frederic; Vojinovic, Zoran; Savic, Dragan; Torres, Arlex Sanchez; Mark, Ole; Zlatanovic, Ljiljana; Abebe, Yared Abayneh

DOI
10.1029/2023GH000866

Publication date
2023

Document Version
Final published version

Published in
GeoHealth

GeoHealth

RESEARCH ARTICLE
10.1029/2023GH000866

Key Points:

- Using a numerical model of wastewater network and machine learning has the potential for the early detection of viral outbreaks
- The ability to recognize disease hotspots depends on sampling frequency and method
- The use of sewer models can improve the usefulness of wastewater-based epidemiology data

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Y. A. Abebe,
y.abebe@un-ihe.org

Citation:

Zehnder, C., Béen, F., Vojinovic, Z., Savic, D., Torres, A. S., Mark, O., et al. (2023). Machine learning for detecting virus infection hotspots via wastewater-based epidemiology: The case of SARS-CoV-2 RNA. *GeoHealth*, *7*, e2023GH000866. https://doi.org/10.1029/2023GH000866

Received 4 JUL 2023
Accepted 10 SEP 2023

Author Contributions:

Conceptualization: Calvin Zehnder, Frederic Béen, Zoran Vojinovic, Dragan Savic, Arlex Sanchez Torres, Ole Mark
Data curation: Calvin Zehnder, Frederic Béen
Formal analysis: Calvin Zehnder
Investigation: Calvin Zehnder
Methodology: Calvin Zehnder, Frederic Béen, Zoran Vojinovic, Dragan Savic, Arlex Sanchez Torres, Ole Mark, Ljiljana Zlatanovic
Software: Calvin Zehnder, Zoran Vojinovic, Ljiljana Zlatanovic

Machine Learning for Detecting Virus Infection Hotspots Via Wastewater-Based Epidemiology: The Case of SARS-CoV-2 RNA

Calvin Zehnder[1], Frederic Béen[2], Zoran Vojinovic[1,3,4,5], Dragan Savic[2,3,4], Arlex Sanchez Torres[1], Ole Mark[6], Ljiljana Zlatanovic[7,8], and Yared Abayneh Abebe[1,9]

[1]Water Supply, Sanitation and Environmental Engineering Department, IHE Delft Institute for Water Education, Delft, The Netherlands, [2]KWR Water Research Institute, Nieuwegein, The Netherlands, [3]Centre for Water Systems, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK, [4]Faculty of Civil Engineering, University of Belgrade, Belgrade, Serbia, [5]National Cheng Kung University, Tainan, Taiwan, [6]Krüger Veolia, Søborg, Denmark, [7]Sanitary Engineering, Delft University of Technology, Delft, The Netherlands, [8]PWN, Velserbroek, The Netherlands, [9]Department of Hydraulic Engineering, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands

Abstract Wastewater-based epidemiology (WBE) has been proven to be a useful tool in monitoring public health-related issues such as drug use, and disease. By sampling wastewater and applying WBE methods, wastewater-detectable pathogens such as viruses can be cheaply and effectively monitored, tracking people who might be missed or under-represented in traditional disease surveillance. There is a gap in current knowledge in combining hydraulic modeling with WBE. Recent literature has also identified a gap in combining machine learning with WBE for the detection of viral outbreaks. In this study, we loosely coupled a physically-based hydraulic model of pathogen introduction and transport with a machine learning model to track and trace the source of a pathogen within a sewer network and to evaluate its usefulness under various conditions. The methodology developed was applied to a hypothetical sewer network for the rapid detection of disease hotspots of the disease caused by the SARS-CoV-2 virus. Results showed that the machine learning model's ability to recognize hotspots is promising, but requires a high time-resolution of monitoring data and is highly sensitive to the sewer system's physical layout and properties such as flow velocity, the pathogen sampling procedure, and the model's boundary conditions. The methodology proposed and developed in this paper opens new possibilities for WBE, suggesting a rapid back-tracing of human-excreted biomarkers based on only sampling at the outlet or other key points, but would require high-frequency, contaminant-specific sensor systems that are not available currently.

Plain Language Summary Viral outbreaks such as COVID-19 are commonly monitored using traditional surveillance methods (e.g., nose swabs). However, individuals may not follow such testing schemes for different reasons, such as if they do not have symptoms. When people are infected by COVID-19, they excrete SARS-CoV-2, the virus that causes the disease. Hence, analyzing sewage using wastewater-based epidemiology (WBE) methods could help monitor the virus outbreak. Recent literature has identified a gap in combining machine learning and hydraulic models with WBE to detect viral outbreaks. In this study, we combine a hydraulic model that simulates virus transport with a machine learning model that tracks hotspot areas within a sewer network. The methodology developed in this paper was applied to a hypothetical sewer network to detect COVID-19 disease hotspots rapidly. Results showed that the machine learning model has the potential to recognize hotspots but requires high time-resolution data and is dependent on the sewer system's characteristics, such as flow velocity and virus sampling procedure. The developed methodology opens new possibilities for WBE to potentially trace human-excreted viral genetic material based on only sampling at particular points. However, it would require high-frequency sampling sensor systems that are not available currently.

1. Introduction

Health and disease surveillance is one of the cornerstones of public health. High-quality data about disease burdens and epidemiology is required to assess the need for and efficacy of public health interventions such as vaccinations and to monitor the threat of emerging or pandemic diseases (World Health Organization, 2007).

Supervision: Frederic Béen, Zoran Vojinovic, Dragan Savic, Arlex Sanchez Torres, Ole Mark
Visualization: Calvin Zehnder, Zoran Vojinovic
Writing – original draft: Calvin Zehnder, Frederic Béen, Zoran Vojinovic, Dragan Savic, Arlex Sanchez Torres, Ole Mark
Writing – review & editing: Frederic Béen, Zoran Vojinovic, Dragan Savic, Arlex Sanchez Torres, Ole Mark

Much information can be gleaned from the sewage we produce. Certain biomarkers and substances are found in sewage that can tell us information about the population that produced the sewage (e.g., Choi et al., 2018). Wastewater-based epidemiology (WBE) has commonly been used to survey the population for pharmaceutical use as well as exposure to toxic substances such as pesticides, and circulation of pathogens (e.g., Gracia-Lor et al., 2017; Medema, Been, et al., 2020). The advantage of WBE is that it makes it possible to collect information about a population that might be difficult or expensive to collect by other means.

There is currently great interest in using WBE to monitor for the presence of the virus SARS-CoV-2 responsible for the coronavirus disease 2019 (COVID-19). WBE for COVID-19 has the potential to cheaply monitor human-shed viruses in a sewer system's service area as it has been proven that the COVID-19 virus may be detected in wastewater (e.g., Ahmed et al., 2020; Kumar et al., 2020; Medema, Heijnen, et al., 2020; Randazzo et al., 2020; Rimoldi et al., 2020; Vallejo et al., 2022; Wurtzer et al., 2020). This has shown the added value of wastewater surveillance is in the objectivity of the wastewater signal, as not everybody can or wants to get tested, but everybody is using the toilet. An increase in viral load detected in sewage will give an early warning to the community that would not be detected with normal testing until days to a week later (Vallejo et al., 2022; Wurtzer et al., 2020). This is of particular interest considering the situation in which many countries are increasing vaccination coverage. Yet, pockets of not vaccinated and potentially susceptible individuals remain.

To further support the implementation of WBE for disease outbreak monitoring, it is necessary to understand how sewer networks influence the measured signals, and how effective sampling strategies should be developed. In particular, there is interest in testing the potential for using WBE to detect localized outbreaks. Highlighting a hotspot or a zone of re-emergence allows public health interventions to be more targeted. As more and more countries bring COVID-19 under control with vaccines and social distancing, outbreak monitoring is an essential tool for the prevention and eventual elimination of the virus. The ability to rapidly track and trace fecally-shed biomarkers would add valuable information to WBE data.

Sensors have been used to monitor the quality of water resources and their optimal placement is of significant research interest (e.g., Bartos & Kerkez, 2021). The optimization of in-network sensor locations for pollutants has been well-studied. Vonach et al. (2018), Sambito et al. (2020), and Banik et al. (2017) propose methodologies for optimizing the placement of water sensors within a sewer network. Vonach et al. (2018) applied their proposed sensor network to calibrate hydrodynamic models, basing the sensor placement optimization on a hydrodynamic model. Their sensor placement optimization is based on a sewer network model as it is designed to be used in the modeling before any calibration is done for the study area. This methodology illustrates a way to optimize sensor placement for flow measurement, but not for pollutant tracing.

Sambito et al. (2020) and Banik et al. (2017) both proposed methodologies to optimize the placement of sensors within a sewer network for pollutant discharges. Both methodologies are to be able to individualize and identify the source of an illicit discharge, for example, soluble metals from the micro-industry. They rely on online pollutant concentration sensors, which are not currently available for every pollutant or biomarker being studied.

Nourinejad et al. (2021) proposed a methodology for optimizing manhole sampling locations for pinpointing COVID-19 cases within a sewer network for a community that has had no COVID-19 recently but has just had a positive detection at the outlet indicating an emergent case. The method uses wastewater flow volumes to choose sequential sampling points in the network. Each sampling point is chosen to capture 50% of the flow of the remaining network branch with the possibility of containing the COVID-19 source. If the sample is found to be COVID-19 positive, then the source location is upstream of the sampling location, if not, it is downstream. Manholes are tested sequentially until the source location is found. There are several limitations to this method, most importantly that it relies on capturing the positive signal in the wastewater based on sequential rapid sampling. As the virus that causes COVID-19 is shed mostly fecally, there will not necessarily be a constant flow of it through a sewer network. In addition, no currently existing virus-detection technology would facilitate the online automatic sampling method proposed (Mao et al., 2020). Without an automated sensing system, the process of sequentially sampling and testing is time and resource intensive and is cumbersome as a rapid track and trace tool.

A recent comprehensive review article identified that there is a clear gap in using machine learning techniques in combination with WBE to detect viral outbreaks (Abdeldayem et al., 2022). One closer work by Kim et al. (2013) studied the use of outlet-sampled pollutant concentration breakthrough curves for pollutant back-tracing. They

used an artificial neural network (ANN) machine learning tool to identify the source location of a microbial intrusion in a sewer system. The researchers applied the ANN to a model based on a real-life sewer system in Arizona, USA. The network's water quality and quantity were modeled in the sewer modeling software MOUSE. E. coli was introduced into the simulation of four manholes in each sub-basin and the MOUSE model was used to compute breakthrough curves of E. coli at the outlet. Breakthrough curve characteristics were selected to train the ANN to categorize the signal based on its origin. The results were that the ANN could predict the source location correctly 57% of the time. By adding a second sensor location in the network, the accuracy increased to 100%.

While this study had the source contaminant introduced in a known (triangular) load pattern, the introduction of a pathogen into the sewer by a human is much messier. Pathogen loading from one source (one human) may be too insignificant to differentiate. This problem may be overcome by the fact that this proposed project will not focus on the impossible task of individual identification, but rather on sub-catchment identification. Another issue is that Kim et al. (2013) used a sensor that took frequent time-series measurements. This may not be feasible in the real world for real-time polymerase chain reaction (RT-qPCR) testing.

There has not been research on how to apply the methodology developed by Kim et al., 2013 for the purposes of WBE, to rapidly track and trace human-shed biomarkers based on outlet sampling. Additionally, there is currently a knowledge gap on how to optimize wastewater testing locations and methodology for WBE data collection, and how to be able to back-trace biomarkers. There is potential for using a machine learning tool coupled with a numerical sewer model to recognize patterns and aid in biomarker back-tracing, as well as to inform WBE sampling methodologies. This paper provides contribution in this direction.

The goal of this study is to develop a novel methodology/procedure for the rapid track and trace of pathogens in sewer networks, by loosely coupling numerical hydrodynamic models and machine learning. The two modeling types, physically-based and data-driven have been used regularly for understanding the water quality behavior of sewer systems (Jia et al., 2021). In this article, we integrate the models in such a way that the outputs of a physically-based model are used as inputs to a machine learning model. The resulting methodology and its usefulness are then evaluated on a semi-hypothetical test case. The case study area is considered semi-hypothetical as although it is based on a real area in Amsterdam, The Netherlands, it is not based on an existing physical sewer network.

The specific objectives are to develop a numerical model for a test case of a wastewater pipe network; to produce a large number of advection-dispersion model simulations; to train the machine learning model using the results from advection-dispersion model simulations; to test the machine learning model on the test case study under a variety of different sampling methods.

2. Methods

2.1. General Workflow

The overall methodology for this study is illustrated in Figure 1. The methodology began with the generation of realistic water-discharge patterns using the water demand model "SIMDEUM" (Blokker et al., 2010, 2017) for a study area. SIMDEUM generates stochastic water use profiles that can be converted to produce stochastic wastewater discharge profiles (Bailey et al., 2019). The model simulates realistic consumer water use behavior, including the number of toilet flushes. The simulation results are then processed in MATLAB to create the patterns for pathogen loading and wastewater flows for a sewer network model. The MIKE Urban modeling software (https://www.mikepoweredbydhi.com/products/mike-urban) was then used to simulate virus transport. The pollutant (virus) transport simulation data was processed in MATLAB to extract characteristics of the data as input for machine learning to categorize the simulations based on the geographic location of a pathogen hotspot. The methodology was evaluated based on how accurate the machine learning tool was in categorizing the pathogen origins, under a variety of different conditions. The sections below describe in-depth the tools and methods used in this study.

2.2. Numerical Model Case Study Network

A small semi-hypothetical sewer network was created first. The model is based on a real residential block in Amsterdam, consisting of 14 sub-catchments, and four zones (Figure 2). Pipe locations were chosen to run under

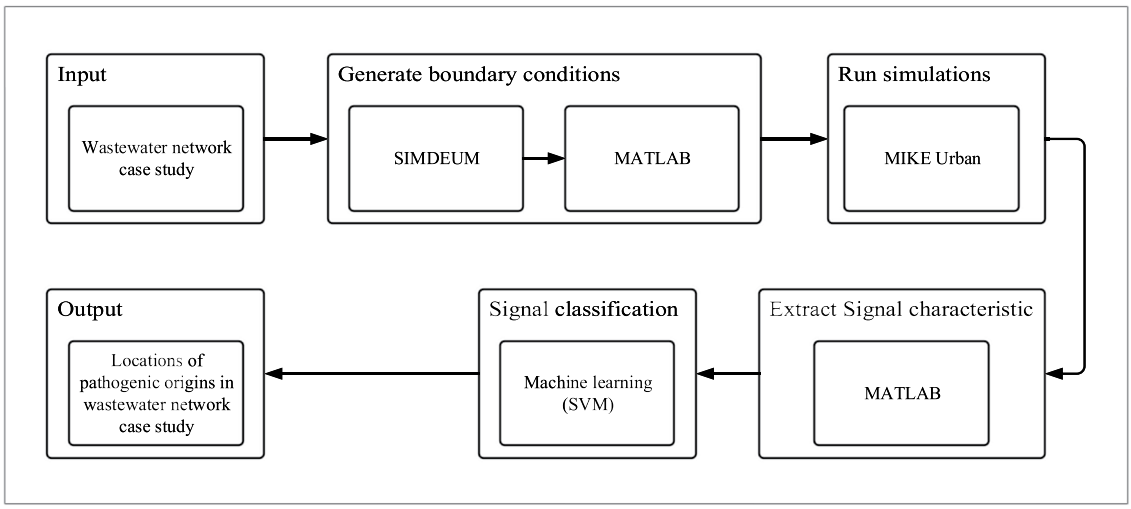

Figure 1. Workflow of methodology showing the main processes starting from input data and boundary conditions generation to hydraulic model simulation, signal classification and virus origins identification.

roads when possible, similar to how Urich et al. (2010) generated realistic virtual sewer networks in both virtual and real city environments.

The network is treated as a separate sewer system; hence we only consider dry weather flow (DWF). In theory, it does not make a difference to the present approach whether the system is separate (only sewage) or combined

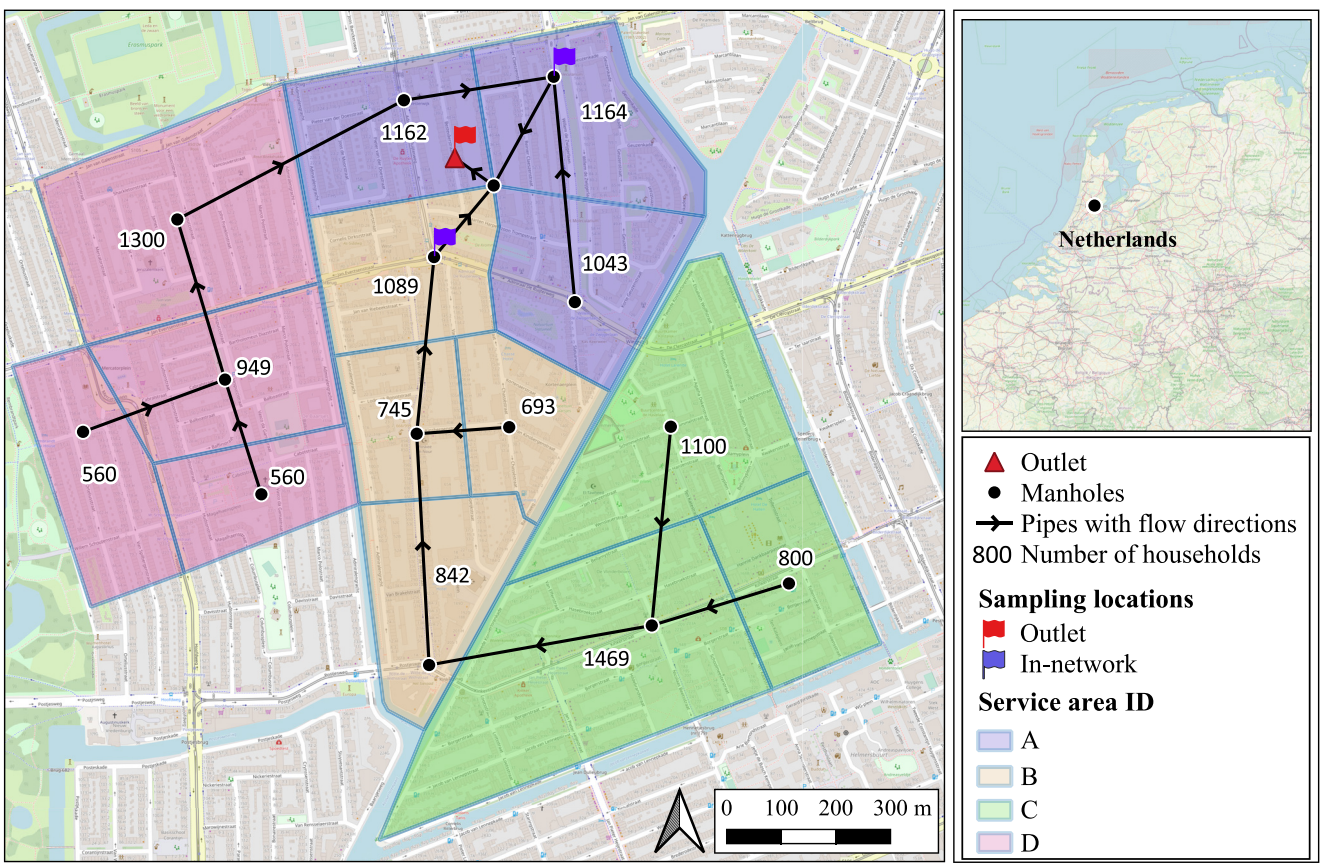

Figure 2. Semi-hypothetical case-study pipe network showing service areas, pipes with flow directions, manholes, sampling locations and an outlet. The service areas are further divided into sub-service areas. The number of households is per sub-service area. The inset map shows the location of the study area within the Netherlands.

(sewage and stormwater). This is because the virus is not thought to accumulate on surfaces to be washed into the sewage in a rain event or accumulate in the network to be flushed out by a rain event. Therefore, the viral "mass" flow for a combined versus separate system should be the same. The concentration, however, would be affected. The exception is that in the event of a combined sewage overflow, some of the virus mass would be lost to the overflow. The pipe and node locations were chosen to balance the precision of the model with its complexity. As the flow generation model SIMDEUM can create wastewater patterns on the individual household member level, it is possible to model every household as a node with its DWF and viral loading pattern. However, for a larger area being modeled, this is unlikely to be necessary.

Each service area is a potential zone for the COVID-19 hotspot. Hotspot creation consists of stochastically determining the number of COVID-19-positive households in one of the four service areas, based on a simulated hotspot prevalence. Each hotspot scenario represents a different service area containing the COVID-19-positive households. Each service area has the same number of households to prevent the machine learning from using the daily total virus levels as a method to determine which area is COVID-positive. If the number of households in each service area was not the same, varying which areas had a certain COVID-19 prevalence would result in the areas that contain more households having a higher number of COVID-19-positive households. This would in turn result in a predictably higher overall viral load, leading the machine learning to use the overall daily load as a feature for categorization, instead of the shape of the signal.

The number of households per service area was determined based on the density of addresses contained within the case study area. The total population was from SIMDEUM simulating 13,476 households, for a result of 22,538 people.

The northern half (areas A and D) has a pipe slope of 0.006, while the rest of the network has a slope of 0.002. This is done to introduce a level of differentiation to the network, to avoid a situation where the network is fully symmetrical, as real-life sewer networks are not normally symmetrical. The pipe routing was created to make the system dendritic from the outlet. The virus transport model developed in MIKE Urban is based on the above-described sewer network characteristics. For the advection-dispersion module, a dispersion factor of 2.00 and a maximum dispersion coefficient of 10.00 were used.

2.3. Simulation Boundary Conditions

2.3.1. Viral Loading

The amount of COVID-19 virus that a person excretes varies greatly between people, and over the course of the illness. An estimated 50% of people excrete the virus via the feces when infected (Parasa et al., 2020).

A loading scheme was derived based on testing in Rotterdam by the Municipal Health Service Organization (GGD Rotterdam-Rijnmond et al., 2021). This results in a log-normal probability distribution function with the following parameters: mu = 9.38, sigma = 0.26, and $n = 9$. A plot of the viral load probability distribution may be seen in Figure S1 in Supporting Information S1. A value sampled from this probability represents the daily viral load for a COVID-19-positive person who is shedding viruses in their feces. For the model, this probability distribution function was sampled every time there is a COVID-19-positive household that is shedding. Every household member is modeled to have the same viral load per flush, while the load level differs between households. The viral loading pattern was then added as a boundary condition to the sewer model.

2.3.2. Dry-Weather Flow Pattern

The SIMDEUM model was used to generate stochastic dry-weather flow patterns, as well as to simulate the loading of the virus by way of toilet flushes. SIMDEUM uses a community profile to generate patterns. The profile used was based on the default, which is accurate in the Netherlands (Blokker et al., 2010). To account for behavior change during the COVID-19 lockdown in the Netherlands, the "home presence" input is modified. From Yerkes et al. (2020), work patterns were adjusted to reflect that only 56% of working women and 34% of working men are in essential professions that require working outside the home during the lockdown period in the Netherlands. To reflect this change, non-crucial workers are assumed to be working from home.

The profile was also altered from the default to reflect people waking up 1 hr later during the lockdown as there is no need to commute while working from home. SIMDEUM has other parameters that may be changed to

reflect the community being simulated, including appliances and household taps. For this study, these are left to the default values as they have been previously found to be accurate (Blokker et al., 2010). The exception is that outdoor taps were removed from the model as they do not contribute to the wastewater. Furthermore, toilet flushing has been modified to change from the water demand profile for flushing to the flushing profile of wastewater. The DWF and toilet use (WC) patterns were exported from SIMDEUM with a time-step of 5 minutes.

2.3.3. Generating COVID-19 Hotspot Disease Prevalence

Simulations were generated with boundary conditions reflecting the potential COVID-19 prevalence in the Netherlands. Data shows that the percentage of infectious individuals has varied from 0% to 1% beginning from February 2020 (RIVM, 2021). This data was normalized using hospitalizations and serological data by the RIVM to account for differences in testing rates, and other factors. It assumes that a person is shedding the virus for a total of 10 days (GGD Rotterdam-Rijnmond et al., 2021). The resultant simulated disease prevalence values for different input hotspot prevalence values can be seen in Figure S2 in Supporting Information S1. A summary of the simulation results created and tested in this study can be found in Table S1 in Supporting Information S1.

2.4. Machine Learning

The machine learning methodology used in the present work is the neural network/support vector machine (SVM) toolbox originally developed by Vojinovic et al. (2003) and Vojinovic and Kecman (2004) and later on enhanced and applied in Vojinovic et al. (2013) and Misra et al. (2018).

The ability of the SVM was checked with a test set of simulated data, stochastically generated with the same boundary conditions as the training data. Additionally, to simulate the way a wastewater stream would be sampled in the physical world, the data from the sewer model was sampled at a specified constant frequency. The effect of pooling sequential samples was also simulated by averaging the concentrations of several consecutive samples. The signal characteristic data was extracted from the sampled and pooled data, not the entire data set.

2.4.1. Training Data

Using the result of the sewer model pollutant transport simulations, pattern characteristics were extracted as training data for the SVM. The dependent variable (SVM output) is the ID of the service area with the disease hotspot (higher COVID-19 prevalence). Through trial and error, it was determined that the best results came from using the ID of the hotspot service area instead of classifying between the north and south sides of the sewer network. Therefore, the SVM classifies between four categories representing the four service areas.

Many different independent variables may be extracted and used in different ways from the simulation results. The necessary processing of data is always done after sampling and pooling data, if applicable. The process for independent variable data creation is shown in Figure 3. This same method is used to extract data sampled at the outlet and the two in-network sampling points. The four signal-independent variables tested in this study were the virus concentration, the fast-Fourier transform (FFT) of the concentration data, the FFT of the viral "mass" flow data, and the times to viral daily accumulation. Ultimately, the FFT of the concentration data was the method chosen based on the analysis presented in Section 3.1.1.

2.4.2. Data Set Size

The data set size was chosen to balance the time required to generate data and the improvements to accuracy that come with increasing the number of simulations used for the training data. Early in the development of this study's methodology, it was found that the larger the training data, the more accurate the SVM will be, with diminishing returns after about 200 simulations per scenario.

Therefore, the data set size chosen was 200 iterations per hotspot location scenario, which means a total of 800 unique simulations were generated per different hotspot prevalence scenarios. However, subsequent testing of the data used in this study found that the size of the training data set can be smaller than 800 simulations with little to no effect on the SVM's accuracy, but with a slight effect on the variance of those results as seen in Table S2 in Supporting Information S1. The test data size used was 40 iterations per scenario, for a total of 160 simulations.

Each loading scenario represents a different service area loaded as the hotspot, with four service areas (A, B, C, and D) there are four different scenarios. This is the same for the training data as for the test data. The only difference is that the number of simulations per scenario generated for the training data is lower.

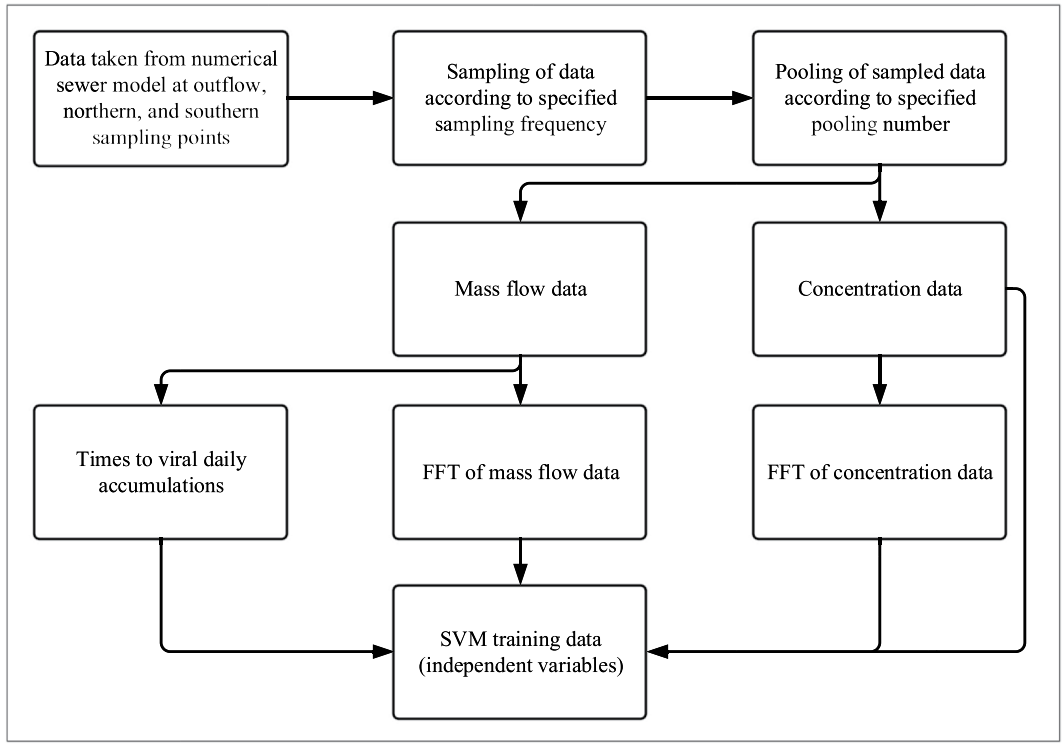

Figure 3. Workflow showing independent variables data generation process for training the SVM model. This is a sub-step of Figure 1.

To test variance, three training and three test data sets were generated with the same boundary conditions, and the size of the data set was varied according to Table S2 in Supporting Information S1. The accuracy of the nine training-test combinations was tested for a constant sampling regime of 360 samples/day. The variance was found to decrease as the data set size increased, but not significantly.

2.5. SARS-CoV-2 Concentrations in Wastewater

Concentrations of SARS-CoV-2 particles in wastewater considered in this study were determined using the methodology established by Medema, Heijnen, et al. (2020). Briefly, 24-hr composite samples were collected at sewer network sampling points to generate viral concentrations in samples used to estimate daily per capita shedding loads as described in Section 2.3.1.

3. Results and Discussion

3.1. Evaluation

After classifying the test data into four categories, the accuracy of the SVM is computed by taking the percentage of correct classifications. The SVM was assessed for its accuracy at the system outlet as well as its accuracy using the in-network sampling locations.

Many of the evaluations presented were done on the data set where the hotspot prevalence was 0.5%, and the rest 0%. This data set was chosen as it balances SVM accuracy results that are high enough such that the effect of various parameter changes can be observed, with a prevalence that is not unrealistic.

Some of the evaluation is done using a "best-case scenario" situation for sampling and pooling data, assuming that there is no limit to the number of samples that can be collected and measured per day. As this is not realistic in real-life, evaluation has also been done to assess how the model handles more realistic data collection methods. For testing SVM accuracy, both its accuracy to differentiate between the four catchments, and the accuracy of the results converted to a binary (north/south) classification are considered.

Table 1
Evaluation of Four Independent Variables for SVM Classification

Characteristics used for SVM	Data dimensionality	SVM accuracy: Outlet: Individual	SVM accuracy: Outlet: Binary from individual
Virus concentration in outlet	360	48%	66%
FFT of virus data	360	68%	71%
FFT of virus concentration data	360	72%	73%
Viral accumulation times	19	29%	55%

Note. The SVM was assessed for its accuracy at the network system outlet.

3.1.1. Independent Variable Selection for SVM

Through trial and error, the best signal characteristic to use as the SVM's independent variables for training was determined to be the FFT of the concentration data. Unless otherwise noted, this method was used for the rest of this study's analysis. The testing was done with a sampling of two minutes/sample and every two samples pooled together, for a total of 360 data points per day. The full evaluation details can be found in Table 1.

SVM categorization based on virus concentration data in the outlet was found to be not as good as other methods (Table 1). This is likely because the data was too unprocessed for the SVM to be able to do reliable categorization. The SVM needs the data to be as clearly and directly helpful as possible, and the unprocessed concentration data does not highlight the crucial differences between signals that allow for the classification of the signals.

The FFT methods had the best resultant SVM classification (Table 1). This supports the theory that the most useful signal characteristics for SVM classification are directly quantified when converting to the frequency-amplitude domain. The amplitude of the signal is an important result of the advection-dispersion, as its dynamics directly affect the way that the peak of the signal is transformed from origin to outlet. The frequency of the signals being processed is overall less important. Therefore, the differences between the advection-dispersion of viral peaks that the FFT quantifies are more important than the overall daily timing of the pattern.

Note that the values that result from the FFT are not important, only the relative values. As the number of data points affects the overall amplitude magnitudes, the number of data points per day must be kept the same for the test and training data.

The time to viral daily accumulations method relies on the timing of the virus mass flow being distinct between the four catchments. For COVID-19, the distribution of the way the virus is loaded into the network (by toilet flushes) is too spread out for the SVM to be able to successfully categorize between the four networks. Additionally, the network is too small to allow the SVM to categorize the signal by the time of arrival, given that the time of loading is not exactly known.

If the pollutant loading occurred at a known time and the network was larger, the time to viral daily accumulations could be a better characteristic to use for the SVM. However, in this case, it was not found to be useful. It is also likely that this method had poor performance due to its low data dimensionality. Even though it was created with the same number of data points as the others, the resultant size of the data is always only 19 points (from a step size of 5% accumulation). This may not be enough to fully depict a useful signal for categorization.

3.1.2. Effect of Boundary Conditions on SVM Classification Accuracy

Figure 4 shows how the prevalence of the simulated hotspot and the number of samples tested per day affect the ability of the SVM to categorize which area (A, B, C, or D) the virus signal originated from. The higher the simulated hotspot disease prevalence is, the easier it is for the SVM to accurately classify the signal (Figure 4). A higher prevalence means that there is more frequent loading of the virus to the system, and therefore more peaks for the SVM to recognize. Adding more virus-shedding events is similar to adding more data points for the SVM with the effect that it has on the SVM result accuracy. The effect of higher shedding concentrations from a different shedding probability distribution is unknown.

The results showed that the boundary conditions strongly affect the accuracy of the SVM classification result. This is due to higher prevalences not only having more differentiation between a hotspot and a non-hotspot zone but that there is more data for the SVM to work with. Results show that the more virus being loaded into the system there is, the more distinct the signals become, as the effect of uncertainty of an individual's fecal virus loading is decreased. As most countries had not achieved a low-enough level of COVID-19 community transmission for the hotspot classification methodology to be applied, low-prevalence hotspot levels are more relevant and should be studied and simulated for maximum added value.

The variation in SVM accuracy was tested by combining three training data sets and three test data sets with the same boundary conditions to make nine unique combinations. The data sets used have a hotspot prevalence

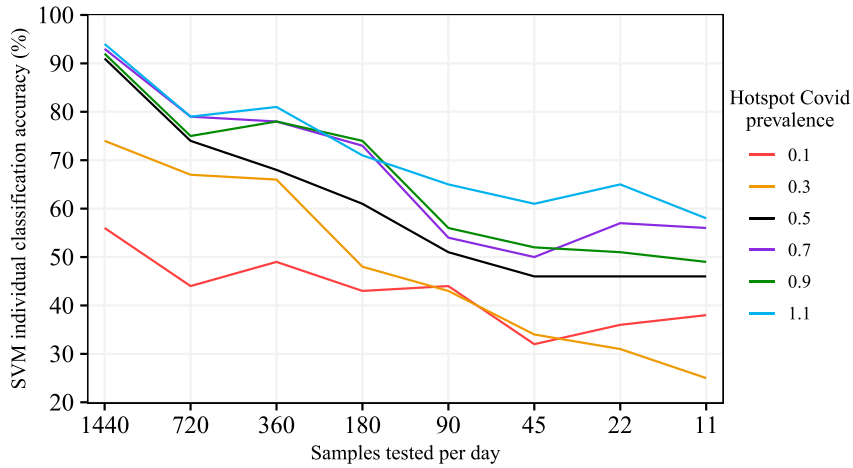

Figure 4. Effect of hotspot prevalence and sampling regime on SVM individual classification accuracy. The sampling regime shows the sampling frequency. For example, 1440 and 90 samples per day mean one sample every minute and 16 min, respectively.

of 0.5%. Table S3 in Supporting Information S1, shows that there was little variation in accuracy between data dimensionalities.

3.1.3. Sampling Locations

The results from sampling at the two in-network locations marked in Figure 2 were compared to the results from sampling at the outlet. Generally, the in-network results follow the same relationship to the hotspot prevalence as seen in Section 3.1.2, but with higher accuracy.

To directly compare the effect of in-network versus outlet sampling, two scenarios are considered. The first is following the methodology described so far in this study and doing an individual classification, based on the outlet. The second is sampling only at the two in-network sampling points, with the SVM result accuracy averaged between the points. For both, the total number of samples tested is kept the same, which means that for a sampling regime of 720 samples tested/day at the outlet, the southern and northern sampling points are sampled at 360 samples tested/day, for a total of 720 samples/day, just like the outlet sampling. Figure S3 in Supporting Information S1, shows a comparison between the two scenarios, for three different simulated hotspot prevalences (0.1%–0.9%), and various amounts of testing per day. Generally, the in-network sampling points have a higher classification accuracy than the outlet, as the large advantage that in-network sampling has over outlet sampling is that it automatically eliminates half of the possible locations for virus hotspots. Each sampling point is only downstream from two possible locations, instead of four like the outlet. This makes the job of the SVM much easier, while still preserving the ability to classify the signal to the individual area.

The disadvantage to the in-network sampling is that in a non-binary loading scenario, where the areas outside the simulated hotspots have a non-zero prevalence, the in-network points can no longer eliminate half of the possible locations. This is because a scenario that would result in a zero signal for a binary loading for an in-network sampling point, would then have a non-zero signal. Therefore, just looking at the signal alone to see if it was zero or non-zero would not be enough to rule out possible hotspot locations.

3.1.4. Effect of Sampling Rate and Pooling on SVM Classification

In this paper, the sensor is considered to work perfectly, able to sample and pool as many samples as needed and to quantify daily any number of samples. However, in a real-life scenario, this is not the case: the sampling is limited in the frequency that an autosampler can handle, the amount of sample pooling that may be done, and most importantly, the number of samples that can be tested and quantified daily. Additionally, the virus quantification methods have an error that could contribute to uncertainty in the signal.

The accuracy of the SVM is highly dependent on the frequency of sampling. This is due to the SVM's reliance on using the dimensions of the viral pulses to categorize the signal. With a sampling frequency that is too low or is pooled too many times, the pulse cannot accurately be captured. This is demonstrated in Figure S4 in Supporting

Information S1, which shows example virus concentration data, sampled at different frequencies. As expected, the less the data is sampled, the more information about the shape of the signal is lost, obscuring the original signal.

If the data transformed by the FFT is under-sampled, to begin with, the FFT method cannot allow for an accurate read on the signal for categorization.

Zooming in on an example concentration data set (Figure S5 in Supporting Information S1), the duration of the signal peaks is observed to be approximately 10–20 min long. The length of the peak is due to the advection-dispersion, and also the way the viral load was modeled.

With SIMDEUM, a 5-min time step was used for pattern generation. Therefore, the original pulse was no shorter than 5 min. Due to the way that MIKE Urban interpolates the pattern data, the signal could result in a duration of 10 min.

The length of the viral peaks affects the sampling rate that is possible for signal categorization. With longer viral peaks, the pattern can be quantified with a lower sampling frequency than with shorter peaks, as it is less likely for the signal's peaks or other important features to fall between samples. Therefore, the time step that the viral loading is modeled with affects the ability for signal recognition and affects the relationship between signal sampling and SVM accuracy.

For the same number of samples tested per day, the effect of changing the pooling number and the sampling rate has little effect. For five different COVID-19 hotspot prevalences, the same data dimensionality was tested (120 samples tested/day), while changing the sample rate, and the number of consecutive pooled samples.

The result of changing the sample rate and pooling number had little to no overall effect on the SVM accuracy (Figure S6 in Supporting Information S1) when controlled for the total number of samples tested daily. This indicates there is no benefit to a high number of pooled samples with high-frequency sampling as compared to a low number of pooled samples with low-frequency sampling, therefore sampling methodology can be based on what is practically possible. If sampling is done with commercially available auto-samplers, the likely maximum sampling speed is 3–5 min between samples (0.33–0.20 samples/min). Depending on the system and method used, there may be additional limits to how many samples can be pooled together.

The ability to sample the wastewater for the pathogen being studied has a large effect on the SVM's results accuracy. With a low number of daily samples being tested, the signal's key advection-dispersion features become lost. Therefore, reducing the number and frequency of samples tested daily restricts the ability of the SVM to recognize and categorize the signal by its location.

3.1.5. Optimizing Sampling

As there is always a limited capacity to test wastewater samples, especially for RT-qPCR testing, the methodology was evaluated for smaller numbers of daily tested samples, based on a data set generated with a 0.5% hotspot prevalence. As the SVM results are highly sensitive to the sampling frequency, it cannot be reduced to decrease the total daily samples while still retaining good SVM accuracy.

To evaluate the potential for optimizing sampling by windowing the sampling period to avoid reducing the sampling frequency while still reducing the total number of samples tested per day, the method of sampling in a particular window was evaluated. The idea is to pick a strategic timeframe in the day to sample to avoid times of the day with few toilet flushes and try to capture as much useful data as possible with as few tested samples as possible.

Limiting the sampling frequency to 3–5 min per sample, and the maximum daily samples tested to 24/day, a subset of the total daily data can be made. The best time of the day is assessed by testing the SVM's accuracy at different periods, for windows of 1.2, 1.6, 2, 2.4, 3.2, and 4 hr. Additional testing parameters are found in Table S4 in Supporting Information S1.

Figure 5 shows the resultant accuracy of the individual classification, for samples centered on different times of the day. It also shows an example diurnal flow pattern for toilet flushing (WC). Figure S7 in Supporting Information S1 shows the resultant accuracy of the binary (north/south) classification. If the two times with the highest SVM accuracies are considered (12 and 10 a.m.), there is a trend toward better accuracy with a smaller window

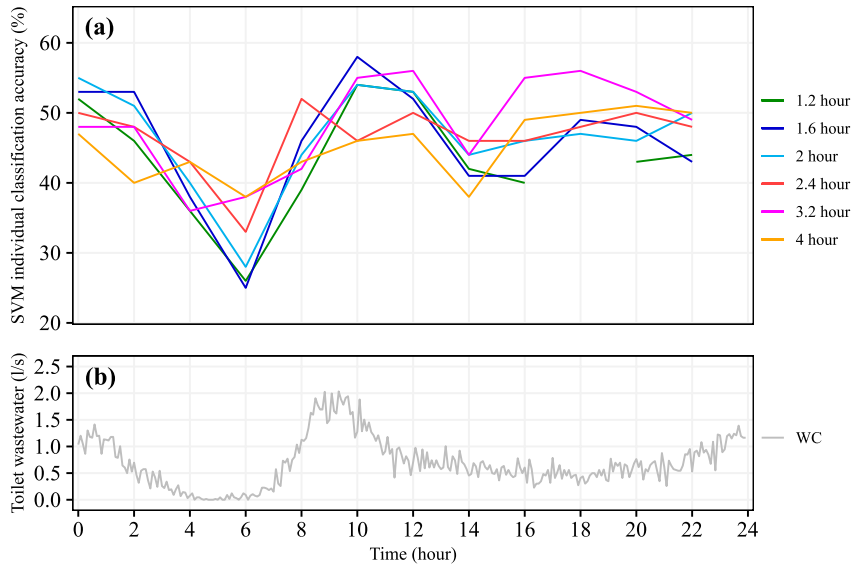

Figure 5. (a) shows SVM individual accuracy versus time of measurement window for different window durations. (b) Shows the diurnal pattern for wastewater flow from toilet flushing.

and higher tested sample frequency (data sets 1.2, 1.6, and 2 hr), as compared to the accuracy for data with a longer window and lower tested sample frequency (data sets 2.4, 3.2, and 4 hr). This indicates that a higher tested sample frequency is more valuable for SVM classification than a longer data collection time.

The data sets with shorter sampling periods also have a higher deviation over the day, suggesting that they are more sensitive to the time of the day of sampling as compared to the samples taken with longer sampling periods.

As a baseline for comparison, different combinations of pooling and sampling rates were tested for 24 hr, with 24 samples tested per day. Results detailed in Table S5 in Supporting Information S1, show that the accuracy for taking samples over the entire day is lower compared to the individual (four) catchment classification in Figure 5.

To test the performance with even smaller data sets, the window of 9–11 a.m. was chosen based on the performance of the previous tests. Results are presented in Table S6 in Supporting Information S1. The results show that the performance with just 12 data points, targeted toward the times of 9–11 a.m. is similar or worse to that of the 24-hr, 24 data point sets.

By sampling at optimized times, the SVM accuracy can be increased as compared with the baseline data with the same number of sampled data points, taken over 24 hr. However, there is a limit to how close to an optimal accuracy the SVM can achieve, with only a small number of data points. When selecting the sampling methodology, the increased accuracy gained by having more samples measured must be weighed against the cost and feasibility of testing large amounts of samples. Given the limited capacity for testing and the lack of an automatic online sensor, optimizing the sampling time is the most promising sampling methodology for the SVM categorization.

The optimized time of day to sample depends on the toilet-use patterns of the service area and the size of the network. In this study, SIMDUEM was used to simulate the toilet-flush patterns. As the optimal time of sampling is dependent on the peak toilet flushing, the model of flush patterns must be validated to determine the peak, for example, by first looking at diurnal patterns of other easier-to-measure biomarkers associated with feces.

For larger sewer networks with longer hydraulic residence times, the peak for toilet flushing may not be as distinct. The patterns for areas close to the outlet and far from the outlet combined may create a diurnal pattern in the outlet with no clear peak in flushing, or a pattern with a less distinct peak. This will decrease the optimization of sampling at peak flushing times.

3.1.6. Effect of Sewer Network's Physical Properties on SVM Accuracy

The physical differences between the four catchments in the studied sewer network are what give the signal a distinctive advection-dispersion and time signature to make the output signal categorizable. Therefore, the

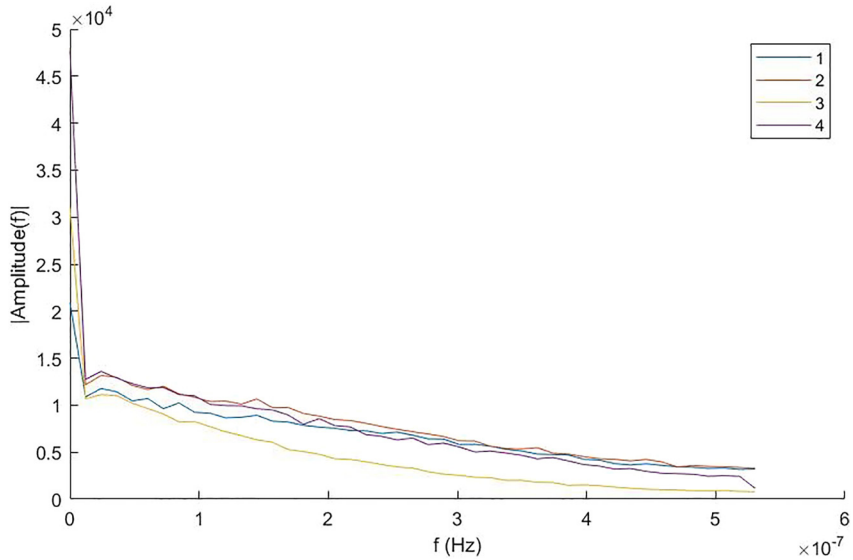

Figure 6. Fast-Fourier transform of concentration data, averaged for each of the four scenarios, 90 samples measured/day, 0.5% hotspot prevalence.

physical properties and layout of the network have an important role in determining the accuracy that the SVM can categorize the output signal. Given an exactly symmetrical network, the SVM would not work, as the same advection-dispersion properties would be happening on either side of the network.

Of the four catchments, catchment "C" (labeled as "3" in the SVM), always had a high accuracy rate as compared with the rest of the catchments. The reason for this lies in its physical attributes of pipe slope, pipe lengths and population density.

Catchment C is in the southern half of the network which has a pipe slope that is lower than the northern half, by design. With lower slopes generally come lower velocities, which is generally true of the two southern catchments. Its distance from the outlet is higher than the center of the two catchments, and this distance combined with its pipe gradients means there is more dispersion than the center two networks, and more dispersion than the northern-most service area (service area D). A more characteristic dispersion signature means that the FFT of the concentration data is on average more distinguishable.

Figure 6 shows the result of averaging the FFT output of viral concentration for 200 iterations for each of the four hotspot scenarios. The *y*-axis has arbitrary units representing the amplitudes of the signals. As can be seen, the FFT of concentration for scenario three, with viral loading coming from service area C, has a lower amplitude over the frequency spectrum.

As this is the training data for the SVM, this means that the SVM can better categorize signals coming from service area C as compared to the other three areas. The result is that the signal from service area C is easily categorized. An example of SVM output showing high accuracy for hotspot zone C can be seen in Figure S8 in Supporting Information S1.

Additionally, the in-network sampling for the southern end of the network was always found to be better than the individual SVM accuracy at the outlet and the northern in-network sampling point. For a range of hotspot prevalences (0.1%–0.9%), and a range of sampling methods, the results for the sampling done in the southern location had a higher accuracy than the northern point (Figure S9 in Supporting Information S1). This is a product of area C's signal being present in the southern sampling point, but not the northern.

As well as the signal being more distinct, the SVM has an easier job differentiating the signal based on the longer length of the viral pulses. With a more dispersed signal, the signal also becomes more spread out, thus allowing it to be better characterized by an infrequent sampling method as compared to a short pulse.

The physical properties and modeling methodology have a large effect on the resultant SVM accuracy as can be seen in service area C. The more different the areas being loaded are, the more diverse the output signals are. Properties that differentiate areas may include the pipe layout, slope, and population density.

The area had four service zones, each with different pipe lengths and configurations. Results showed that the SVM was able to differentiate the area best that had the most dispersion (area C). This was likely due to its lower pipe velocities, distance from the outlet and relatively less dense service in terms of households. However, the way the model was created calls into question the real effect of dispersion on the SVM classification. As flow velocities were low due to pipe over-sizing, the effects of the advection-dispersion are not clear as compared with a more realistic system. It may well be that with higher velocities overall there is less dispersion, even for a larger and longer system. However, it is not the overall level of dispersion that is important for the SVM, rather, the differences between the signal dispersions. Therefore, it is possible for a wastewater system to still have distinctive advection-dispersion patterns.

4. Conclusions

The literature review to date shows a clear gap in combining WBE, machine learning techniques and hydrodynamic models. In this study, we advanced the field of WBE by developing a methodology for rapid track and trace of areas of pathogen introduction in sewer networks. This was accomplished by combining hydrodynamic models and a machine learning tool to categorize pathogen signals measured at the network outlet by their sewer network location.

The results show key takeaway messages from this study:

- The most useful signal characterization method was applying a FFT to the data.
- The larger the disease prevalence difference between the disease hotspot and non-hotspot areas was, the easier it was for the SVM to categorize the signal by location.
- Adding multiple sampling points within the network improves SVM categorization accuracy.
- While the ability of the SVM to accurately categorize signals based on origin is highly dependent on the number of samples tested per day, no discernible relationship between the number of samples pooled or the sampling frequency with SVM categorization accuracy was detected, when controlling for a consistent number of samples tested per day.
- There is potential for optimizing the SVM categorization by increasing the frequency of data collection only during the time of day that the system would expect to have the most toilet flushes.
- The physical properties, such as degree of symmetry, pipe slope, and size of the sewer network, influence the advection-dispersion patterns and thus affect the ability of the SVM to categorize patterns.

The methodology developed in the study shows the theoretical potential for the improvement of existing WBE methods by adding finer resolution spatio-temporal information to the WBE data. This methodology shows potential for enabling rapid back-tracing of biomarkers shed in feces or urine based on only sampling wastewater at the sewer network outlet, or other key points in the network, but would require contaminant-specific high-frequency sensor systems. The ability to determine probable hotspots of a fecally-shed biomarker's origin within a sewer network without testing extensively within the network has never been examined before and would deepen the utility of WBE for disease surveillance if validated in a real-life scenario. However, the SVM model relies on high sampling frequencies, which is larger than what can be currently achieved in practice, where one 24hr composite sample per day per area is the standard approach. It was a theoretical study, that aimed to understand how information about the introduction of contaminants such as SARS-CoV-2 in the sewer and the hydraulics of the sewer network can aid source tracking. For the approach to be feasible in practice, contaminant-specific sensors that allow high-frequency data generation would be needed and these are not available at this time. With the technology that is currently available, it is possible to trace COVID-19 hotspots via the sewer network. It would take four (passive) 24hr composite samples to identify the hotspot accurately in the presented scenarios.

Finally, outside a track and trace methodology, combining hydraulic and pollutant transport models with WBE shows potential for studying the effects of different sampling methodologies and locations. A sewer network being studied using WBE could be modeled and simulated as a method to test the proposed sampling methodology and locations, before using the resources for taking and analyzing real-world samples. Additionally, events such as a combined sewer overflow spill could be simulated to investigate their impact on WBE data collection.

5. Limitations and Future Considerations

Considering that a large volume of data (i.e., characteristics of viral flow patterns) is required to train an ML model, and there is no such amount of measurement data currently available, we used hydrodynamic model results

to train the SVM. However, this can be a limitation as the uncertainty in the hydrodynamic model results may propagate to the final SVM output, which are the hotspots identified. The study was based on a semi-hypothetical sewer network with semi-hypothetical loading schemes, which has certain limitations for application to the real world. The main limitations are the way the hypothetical network was set up and the way the virus loading was simulated. The network model in this study was relatively small compared to the entire study area of the city of Amsterdam. A larger network could mean a more complicated signal categorization, which could decrease the effectiveness of using the SVM for categorization. Additionally, the network was modeled as a separate system: no stormwater, only wastewater. Comparing data from a day with some amount of stormwater flow to a day with only DWF would not work. Care must be taken to ensure that the training data used to create the SVM model accurately reflects the non-hypothetical flow scenario.

To simulate the way that the virus was loaded into the network, the probability distribution function for viral shedding was based on emerging studies. As COVID-19 is a relatively new disease, more research to validate the loading schemes must be done. The way the hotspot simulation was done in this study has limitations in its application to a real-world case study. It is unlikely that in an emergent disease situation, all positive cases would be inside a delineated area with absolutely no negative cases outside of it. The socio-economic correlations with COVID-19 could also become a factor in reality, as different neighbourhoods may have different social characteristics that may affect their shedding or prevalence patterns. It is possible that these differences may overwhelm the differences in the sewer system's physical properties.

The current methodology only considers virus shedding as coming from one delineated residential area. Future work may include people shedding the virus somewhere outside their homes, such as at work, at school, or in a public bathroom. A scenario with small amounts of shedding outside the hotspot location could also be tested as it may have implications on the SVM categorization. The current study only considers one service area loaded as a hotspot at a time. The study could be extended to test the machine learning model's ability to recognize different hotspots by loading multiple service areas as hotspots. The accuracy of the SVM classification results depends on the boundary conditions used in the hydraulic model. To improve the boundary conditions, more research concerning the viral loading dynamics is needed in addition to more data collection on the dynamics of emerging hotspots.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The COVID-19 contagious persons per day data used for generating the COVID-19 hotspot prevalence in the study are available from the Dutch National Institute for Public Health and the Environment (RIVM) at https://data.rivm.nl/covid-19/COVID-19_prevalentie.json with license http://creativecommons.org/publicdomain/mark/1.0/deed.nl.

Acknowledgments
The authors would like to thank Gertjan Medema for his constructive comments and suggestions in the production of this manuscript.

References

Abdeldayem, O. M., Dabbish, A. M., Habashy, M. M., Mostafa, M. K., Elhefnawy, M., Amin, L., et al. (2022). Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: A comprehensive review and outlook. *Science of the Total Environment*, *803*, 149834. https://doi.org/10.1016/j.scitotenv.2021.149834

Ahmed, W., Angel, N., Edson, J., Bibby, K., Bivins, A., O'Brien, J. W., et al. (2020). First confirmed detection of SARS-CoV-2 in untreated wastewater in Australia: A proof of concept for the wastewater surveillance of COVID-19 in the community. *Science of the Total Environment*, *728*, 138764. https://doi.org/10.1016/j.scitotenv.2020.138764

Bailey, O., Arnot, T. C., Blokker, E. J. M., Kapelan, Z., Vreeburg, J., & Hofman, J. (2019). Developing a stochastic sewer model to support sewer design under water conservation measures. *Journal of Hydrology*, *573*, 908–917. https://doi.org/10.1016/j.jhydrol.2019.04.013

Banik, B. K., Di Cristo, C., Leopardi, A., & de Marinis, G. (2017). Illicit intrusion characterization in sewer systems. *Urban Water Journal*, *14*(4), 416–426. https://doi.org/10.1080/1573062X.2016.1176220

Bartos, M., & Kerkez, B. (2021). Observability-based sensor placement improves contaminant tracing in river networks. *Water Resources Research*, *57*(7), e2020WR029551. https://doi.org/10.1029/2020WR029551

Blokker, E. J. M., Agudelo-Vera, C., Moerman, A., van Thienen, P., & Pieterse-Quirijns, I. (2017). Review of applications for SIMDEUM, a stochastic drinking water demand model with a small temporal and spatial scale. *Drinking Water Engineering and Science*, *10*(1), 1–12. https://doi.org/10.5194/dwes-10-1-2017

Blokker, E. J. M., Vreeburg, J., & van Dijk, J. C. (2010). Simulating residential water demand with a stochastic end-use model. *Journal of Water Resources Planning and Management*, *136*(1), 19–26. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000002

Choi, P. M., Tscharke, B., Donner, E., O'Brien, J. W., Grant, S., Kaserzon, S., et al. (2018). Wastewater-based epidemiology biomarkers: Past, present and future. *Trends in Analytical Chemistry*, *105*, 453–469. https://doi.org/10.1016/j.trac.2018.06.004

GGD Rotterdam-Rijnmond, Erasmus MC, & KWR. (2021). Risicogericht Grootschalig Testen: De resultaten van laagdrempelige testmogelijkheden in gemeente Lansingerland en gebied Charlois. Retrieved from https://www.ggdrotterdamrijnmond.nl/nieuws/conclusies-grootschalig-t/20210401-Rapportage-RGT-Lansingerland-en-Charlois-V1.1.pdf

Gracia-Lor, E., Castiglioni, S., Bade, R., Been, F., Castrignanò, E., Covaci, A., et al. (2017). Measuring biomarkers in wastewater as a new source of epidemiological information: Current state and future perspectives. *Environment International*, *99*, 131–150. https://doi.org/10.1016/j.envint.2016.12.016

Jia, Y., Zheng, F., Maier, H. R., Ostfeld, A., Creaco, E., Savic, D., et al. (2021). Water quality modeling in sewer networks: Review and future research directions. *Water Research*, *202*, 117419. https://doi.org/10.1016/j.watres.2021.117419

Kim, M., Choi, C. Y., & Gerba, C. P. (2013). Development and evaluation of a decision-supporting model for identifying the source location of microbial intrusions in real gravity sewer systems. *Water Research*, *47*(13), 4630–4638. https://doi.org/10.1016/j.watres.2013.04.018

Kumar, M., Patel, A. K., Shah, A. V., Raval, J., Rajpara, N., Joshi, M., & Joshi, C. G. (2020). First proof of the capability of wastewater surveillance for COVID-19 in India through detection of genetic material of SARS-CoV-2. *Science of the Total Environment*, *746*, 141326. https://doi.org/10.1016/j.scitotenv.2020.141326

Mao, K., Zhang, H., & Yang, Z. (2020). Can a paper-based device trace COVID-19 sources with wastewater-based epidemiology? *Environmental Science & Technology*, *54*(7), 3733–3735. https://doi.org/10.1021/acs.est.0c01174

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. *Current Opinion in Environmental Science & Health*, *17*, 49–71. https://doi.org/10.1016/j.coesh.2020.09.006

Medema, G., Heijnen, L., Elsinga, G., Italiaander, R., & Brouwer, A. (2020). Presence of SARS-Coronavirus-2 in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in The Netherlands. *Environmental Science and Technology Letters*, *7*(7), 511–516. https://doi.org/10.1021/acs.estlett.0c00357

Misra, A., Vojinovic, Z., Ramakrishnan, B., Luijendijk, A., & Ranasinghe, R. (2018). Shallow water bathymetry mapping using support vector machnie (SRV) technique and multispectral imagery. *International Journal of Remote Sensing*, *39*(13), 4431–4450. https://doi.org/10.1080/01431161.2017.1421796

Nourinejad, M., Berman, O., & Larson, R. C. (2021). Placing sensors in sewer networks: A system to pinpoint new cases of coronavirus. *PLoS One*, *16*(4), e0248893. https://doi.org/10.1371/journal.pone.0248893

Parasa, S., Desai, M., Thoguluva Chandrasekar, V., Patel, H. K., Kennedy, K. F., Roesch, T., et al. (2020). Prevalence of gastrointestinal symptoms and fecal viral shedding in patients with coronavirus disease 2019: A systematic review and meta-analysis. *JAMA Network Open*, *3*(6), e2011335. https://doi.org/10.1001/jamanetworkopen.2020.11335

Randazzo, W., Cuevas-Ferrando, E., Sanjuan, R., Domingo-Calap, P., & Sanchez, G. (2020). Metropolitan wastewater analysis for COVID-19 epidemiological surveillance. *International Journal of Hygiene and Environmental Health*, *230*, 113621. https://doi.org/10.1016/j.ijheh.2020.113621

Rimoldi, S. G., Stefani, F., Gigantiello, A., Polesello, S., Comandatore, F., Mileto, D., et al. (2020). Presence and infectivity of SARS-CoV-2 virus in wastewaters and rivers. *Science of the Total Environment*, *744*, 140911. https://doi.org/10.1016/j.scitotenv.2020.140911

RIVM. (2021). COVID-19 besmettelijke personen per dag (COVID-19 contagious persons per day). [Dataset]. RIVM. Retrieved from https://data.rivm.nl/covid-19/COVID-19_prevalentie.json

Sambito, M., Di Cristo, C., Freni, G., & Leopardi, A. (2020). Optimal water quality sensor positioning in urban drainage systems for illicit intrusion identification. *Journal of Hydroinformatics*, *22*(1), 46–60. https://doi.org/10.2166/hydro.2019.036

Urich, C., Sitzenfrei, R., Möderl, M., & Rauch, W. (2010). An agent-based approach for generating virtual sewer systems. *Water Science and Technology*, *62*(5), 1090–1097. https://doi.org/10.2166/wst.2010.364

Vallejo, J. A., Trigo-Tasende, N., Rumbo-Feal, S., Conde-Perez, K., Lopez-Oriona, A., Barbeito, I., et al. (2022). Modeling the number of people infected with SARS-COV-2 from wastewater viral load in Northwest Spain. *Science of the Total Environment*, *811*, 152334. https://doi.org/10.1016/j.scitotenv.2021.152334

Vojinovic, Z., Abebe, Y. A., Ranasinghe, R., Vacher, A., Martens, P., Mandl, D. J., et al. (2013). A machine learning approach for estimation of shallow water depths from optical satellite images and sonar measurements. *Journal of Hydroinformatics*, *15*(4), 1408–1424. https://doi.org/10.2166/hydro.2013.234

Vojinovic, Z., & Kecman, V. (2004). Contaminant transport modelling with support vector machine model- an alternative to classical advection-dispersion equation. In *Paper presented at 6th international conference on hydroinformatics*. https://doi.org/10.1142/9789812702838_0196

Vojinovic, Z., Kecman, V., & Babovic, V. (2003). Hybrid Approach for modeling wet weather response in wastewater systems. *Journal of Water Resources Planning and Management*, *129*(6), 441–524. https://doi.org/10.1061/(ASCE)0733-9496(2003129:6(511)

Vonach, T., Tscheikner-Gratl, F., Rauch, W., & Kleidorfer, M. (2018). A heuristic method for measurement site selection in sewer systems. *Water*, *10*(2), 122. https://doi.org/10.3390/w10020122

World Health Organization. (2007). *Global framework for immunization monitoring and surveillance: GFIMS (WHO/IVB/07.06)*. World Health Organization. Retrieved from https://apps.who.int/iris/handle/10665/69685

Wurtzer, S., Marechai, V., Mouchel, J. M., Maday, Y., Teyssou, R., Richard, E., et al. (2020). Evaluation of lockdown effect on SARS-CoV-2 dynamics through viral genome quantification in waste water, Greater Paris, France, 5 March to 23 April 2020. *Euro Surveillance*, *25*(50), 2000776. https://doi.org/10.2807/1560-7917.ES.2020.25.50.2000776

Yerkes, M., André, S., Besamusca, J., Remery, C., van der Zwan, R., Kruyen, P., et al. (2020). Werkende ouders in tijden van Corona: Meer maar ook minder genderongelijkheid. Retrieved from https://www.uu.nl/sites/default/files/Policybrief.pdf